ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Αυτοματοποιημένη διαδικασία συλλογής και προεπεξεργασίας δεδομένων για προβλεπτικά μοντέλα: Η περίπτωση του COVID-19

ΒΑΓΓΕΛΑΤΟΣ ΓΕΩΡΓΙΟΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Υπεύθυνος

ΚΑΡΑΝΙΚΑΣ ΧΑΡΑΛΑΜΠΟΣ

ΛΕΚΤΟΡΑΣ

Λαμία, 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Αυτοματοποιημένη διαδικασία συλλογής και προεπεξεργασίας δεδομένων για προβλεπτικά μοντέλα: Η περίπτωση του COVID-19

ΒΑΓΓΕΛΑΤΟΣ ΓΕΩΡΓΙΟΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων
ΚΑΡΑΝΙΚΑΣ ΧΑΡΑΛΑΜΠΟΣ

Λαμία, 2021

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

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

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

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

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

Ημερομηνία: 03/06/2021

Ο – Η Δηλ.
Βαγγελάτος Γεώργιος

(Υπογραφή)

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

Αυτοματοποιημένη διαδικασία συλλογής και προεπεξεργασίας δεδομένων για προβλεπτικά μοντέλα: Η περίπτωση του COVID-19

ΒΑΓΓΕΛΑΤΟΣ ΓΕΩΡΓΙΟΣ

Τριμελής Επιτροπή:

Καρανίκας Χαράλαμπος,Λέκτορας

Βασίλης Πλαγιανάκος, Καθηγητής

Σωτήρης Τασουλής, Επίκουρος Καθηγητής

Περιεχόμενα

ΠΕΡΙΛΗΨΗ

Βασικό αντικείμενο αυτής της εργασίας, όπως μαρτυρά και ο τίτλος της είναι η δημιουργία μιας αξιόπιστης, γρήγορης και εύκολης στη χρήση συνάρτησης συλλογής δεδομένων σχετικών με τον COVID-19. Η επιστημονική κοινότητα, έχοντας στρέψει το ενδιαφέρον της, σχεδόν ολοκληρωτικά, στην αντιμετώπιση και την κατανόηση του COVID-19, διαρκώς προσπαθεί να συγκεντρώσει και να χρησιμοποιήσει δεδομένα από τις αμέτρητε διαθέσιμες πηγές. Τα δεδομένα αυτά ωστόσο είναι διάσπαρτα, μη σχεδιασμένα να συνδυάζονται, σε διαφορετικές χρονικές περιόδους και ο όγκος τους συνεχώς αυξάνεται. Για τον λόγο αυτό δημιουργείται μια συνάρτηση που αυτόματα συλλεγεί, καθαρίζει, ομαδοποιεί και συνδυάζει δεδομένα για έναν μεγάλο αριθμό χωρών. Τα δεδομένα αυτά αφορούν τον COVID-19 αλλά περιπλανώνται και αλλού είδους δεδομένα για κάθε χώρα, που πιθανόν να παρουσιάζουν ερευνητικό ενδιαφέρον πόσον άφορα την κατανόησή της Πορείας του COVID-19. Η συνάρτηση εξάγει δεδομένα, σε μορφή ημερήσιας χρονοσειράς, σε μικρό χρονικό διάστημα, τα οποία είναι απευθείας έτοιμα να χρησιμοποιηθούν για αναλύσεις. Επιπλέον δημιουργήθηκε ένα αναλυτικό σετ κοκόνων και γράφτηκε αναλυτικά η μεθοδολογία για την δημιουργία μοντέλων ARIMA για πρόβλεψη χρονοσειρών. Με βάση τους κανόνες δημιουργικέ ένα μοντέλο ARIMA, που χρησιμοποιεί τα δεδομένα που εξάγει η συνάρτηση, για τη δημιουργία προβλέψεων. Η προσφορά της πτυχιακής αυτής συντελείται από τον κώδικα της συνάρτησης, το αναλυτικό σετ με κανόνες και τη μεθοδολογία δημιουργίας μοντέλων ARIMA καθώς και ο υπόλοιπος κώδικας που περιέχει τα προβλεπτικά μοντέλα και άλλες χρήσιμες συναρτήσεις.

ΚΕΦΑΛΑΙΟ 1
Data Science

Η επιστήμη των δεδομένων είναι ένα διεπιστημονικό πεδίο που χρησιμοποιεί επιστημονικές μεθόδους, διαδικασίες, αλγόριθμους και συστήματα για την εξαγωγή γνώσεων και πληροφοριών από πολλά δομημένα και μη δομημένα δεδομένα. Η επιστήμη δεδομένων σχετίζεται με την εξόρυξη δεδομένων, τη μηχανική μάθηση και τα μεγάλα δεδομένα.

Αποτελεί μια «έννοια για την ενοποίηση στατιστικών, την ανάλυση δεδομένων και τις σχετικές μεθόδους τους», προκειμένου να «κατανοήσουν και να αναλύσουν πραγματικά φαινόμενα» με τα δεδομένα. Χρησιμοποιεί τεχνικές και θεωρίες που προέρχονται από πολλά πεδία στο πλαίσιο των μαθηματικών, των στατιστικών, της επιστήμης των υπολογιστών, της γνώσης τομέα και της επιστήμης της πληροφορίας. Ο νικητής του βραβείου Turing, *Jim Grey*, φαντάστηκε την επιστήμη των δεδομένων ως «*τέταρτο παράδειγμα*» της επιστήμης (εμπειρική, θεωρητική, υπολογιστική και τώρα βάσει δεδομένων) και ισχυρίστηκε ότι «*τα πάντα σχετικά με την επιστήμη αλλάζουν λόγω του αντίκτυπου της πληροφορικής*» και ο κατακλυσμός των δεδομένων.

Επιπτώσεις της επιστήμης δεδομένων

Τα μεγάλα δεδομένα (Big Data) γίνονται πολύ γρήγορα ένα ζωτικό εργαλείο για επιχειρήσεις και εταιρείες όλων των μεγεθών. Η διαθεσιμότητα και η ερμηνεία των μεγάλων δεδομένων έχει αλλάξει τα επιχειρηματικά μοντέλα των παλαιών βιομηχανιών και επέτρεψε τη δημιουργία νέων. Οι επιχειρήσεις που βασίζονται σε δεδομένα αξίζουν 1,2 τρισεκατομμύρια δολάρια συλλογικά το 2020, αύξηση από 333 δισεκατομμύρια δολάρια το 2015. Οι επιστήμονες δεδομένων είναι υπεύθυνοι για την κατανομή Big Data σε χρήσιμες πληροφορίες και τη δημιουργία λογισμικού και αλγορίθμων που βοηθούν τις εταιρείες και τους οργανισμούς να καθορίσουν τις

βέλτιστες λειτουργίες. Καθώς τα Big Data συνεχίζουν να έχουν σημαντικό αντίκτυπο στον κόσμο, η επιστήμη των δεδομένων το κάνει επίσης λόγω της στενής σχέσης μεταξύ των δύο.

Η επιστήμη δεδομένων έχει βρει τις εφαρμογές της σε σχεδόν κάθε κλάδο.

Φροντίδα υγείας

Οι εταιρείες υγειονομικής περίθαλψης χρησιμοποιούν την επιστήμη δεδομένων για να δημιουργήσουν εξελιγμένα ιατρικά εργαλεία για τον εντοπισμό και τη θεραπεία ασθενειών.

Παιχνίδια

Τα βιντεοπαιχνίδια και τα ηλεκτρονικά παιχνίδια δημιουργούνται τώρα με τη βοήθεια της επιστήμης δεδομένων και αυτό έχει οδηγήσει την εμπειρία παιχνιδιού στο επόμενο επίπεδο.

Αναγνώριση εικόνας

Ο εντοπισμός μοτίβων σε εικόνες και ο εντοπισμός αντικειμένων σε μια εικόνα είναι μία από τις πιο δημοφιλείς εφαρμογές επιστήμης δεδομένων.

Συστήματα σύστασης

Το Netflix και το Amazon δίνουν προτάσεις για ταινίες και προϊόντα με βάση αυτό που επιθυμεί ο πελάτης να παρακολουθήσει, να αγοράσει ή να περιηγηθεί στις πλατφόρμες τους.

Logistics

Η επιστήμη δεδομένων χρησιμοποιείται από εταιρείες logistics για τη βελτιστοποίηση των δρομολογίων για τη διασφάλιση της ταχύτερης παράδοσης προϊόντων και την αύξηση της λειτουργικής αποτελεσματικότητας.

Ανίχνευση απάτης

Τα τραπεζικά και χρηματοπιστωτικά ιδρύματα χρησιμοποιούν την επιστήμη των δεδομένων και τους συναφείς αλγόριθμους για τον εντοπισμό δόλιων συναλλαγών.

Επιστήμη δεδομένων υγειονομικής περίθαλψης

Η επιτυχία στη σημερινή βιομηχανία υγειονομικής περίθαλψης με γνώμονα τα δεδομένα θα καθορίζεται όλο και περισσότερο από ηγέτες που κατανοούν την επιστήμη των δεδομένων. Αυτή η γνώση θα είναι κρίσιμη καθώς τα στελέχη δημιουργούν και καθοδηγούν ομάδες προς ένα αρμονικό, καλά σχεδιασμένο όραμα για τη βελτίωση της υγειονομικής περίθαλψης που αξιοποιεί πλήρως τις δυνατότητες των δεδομένων.

Έως και το 30 τοις εκατό των παγκόσμιων αποθηκευμένων δεδομένων προέρχεται από τη βιομηχανία υγειονομικής περίθαλψης. Υπάρχει σημαντική ευκαιρία για βελτίωση της υγειονομικής περίθαλψης σε αυτήν την προσωρινή μνήμη πληροφοριών, συμπεριλαμβανομένης της εξοικονόμησης κόστους περίπου 300 δισεκατομμυρίων. Ωστόσο, η βιομηχανία μπορεί να καλωσορίσει αυτές τις προοπτικές μόνο εάν τα συστήματα υγείας αξιοποιούν πλήρως τα δεδομένα για να προσδιορίσουν τομείς βελτίωσης και να προωθήσουν τεκμηριωμένη φροντίδα. Ακόμη και με αυτό το τεράστιο δυναμικό δεδομένων, η υγειονομική περίθαλψη βασίζεται πολύ συχνά σε ξεπερασμένη τεχνολογία. Για παράδειγμα, έως και 75 τοις εκατό της ιατρικής επικοινωνίας εξακολουθεί να συμβαίνει μέσω φαξ (σε μια εποχή όπου οι εταιρείες αυτοκινήτων χρησιμοποιούν την επιστήμη δεδομένων για να προσθέσουν δυνατότητες πλοήγησης στα αυτοκίνητα).

Η τεχνολογία έχει καθορίσει τις ευκαιρίες, αλλά, για να πραγματοποιήσει κέρδη στην ψηφιακή εποχή, οι ηγέτες της υγειονομικής περίθαλψης πρέπει να κατανοήσουν

την επιστήμη δεδομένων και τον επείγοντα χαρακτήρα επένδυσης σε πόρους επιστήμης δεδομένων (τεχνολογία και άνθρωποι).

Η καρδιά της επιστήμη δεδομένων υγειονομικής περίθαλψης: Μοντέλα μηχανικής εκμάθησης που αποδίδουν βαθύτερες γνώσεις

Η καρδιά της επιστήμης δεδομένων είναι μοντέλα μηχανικής μάθησης, τα οποία είναι βασικά στατιστικά μοντέλα που μπορούν να χρησιμοποιηθούν για την εξαγωγή μοτίβων από δεδομένα. Η επιστήμη δεδομένων και η μηχανική εκμάθηση μπορούν επίσης να θεωρηθούν ότι χρησιμοποιούν τη δύναμη του σύγχρονου υπολογιστή για να αξιοποιήσουν τη στατιστική. Ορισμένα μοντέλα μηχανικής εκμάθησης, όπως η κανονικοποιημένη παλινδρόμηση και τα δέντρα αποφάσεων, προσφέρονται για να αντλήσουν πληροφορίες και να εξηγήσουν μοτίβα στα δεδομένα (π.χ. ποιοι γιατροί χρησιμοποιούν υπερβολικά δαπανηρό υλικό). Άλλα μοντέλα μηχανικής μάθησης, όπως τυχαία δάση και νευρωνικά δίκτυα (βαθιά μάθηση), χρησιμοποιούνται κυρίως για την πρόβλεψη.

Η Επιστήμη Δεδομένων Υγείας είναι το κλειδί για ταχύτερη διάγνωση, καλύτερη θεραπεία

Η υγειονομική περίθαλψη βασίστηκε εδώ και πολύ καιρό σε ανάλυση δεδομένων για να κατανοήσει ζητήματα που σχετίζονται με την υγεία και να βρει αποτελεσματικές θεραπείες. Για παράδειγμα, οι ερευνητές έχουν χρησιμοποιήσει διπλές τυφλές ελεγχόμενες με εικονικό φάρμακο μελέτες ως το θεμέλιο φαρμάκων που βασίζονται σε στοιχεία. Τέτοιες μελέτες δημιουργούν δεδομένα σχετικά με τη θεραπεία υπό αξιολόγηση και αναλύουν αυτά τα δεδομένα για να προσδιορίσουν εάν η θεραπεία είναι αποτελεσματική, καθώς και κατανοούν τις παρενέργειές της. Ως μέθοδος δημιουργίας δεδομένων και πληροφοριών, αυτή η διαδικασία μελέτης λειτουργεί με πνεύμα παρόμοιο με την επιστήμη των δεδομένων, αλλά είναι δαπανηρότερο και πιο χρονοβόρο.

Σήμερα, η υγειονομική περίθαλψη χρειάζεται δεδομένα για τη βελτιστοποίηση των αποτελεσμάτων των ασθενών με τεκμηριωμένες πρακτικές περισσότερο από ποτέ. Αυτές οι πληροφορίες περιμένουν να ανακαλυφθούν σε δεδομένα που έχουν

ήδη συλλεχθεί. Με την επιστήμη των δεδομένων, η βιομηχανία μπορεί να βρει αποτελεσματικούς, οικονομικά αποδοτικούς τρόπους για να αξιοποιήσει τεράστιες ποσότητες υπαρχόντων δεδομένων υγειονομικής περίθαλψης - για να μεγιστοποιήσει τις δυνατότητές της να μετατρέψει την υγειονομική περίθαλψη σε ταχύτερη, ακριβέστερη διάγνωση και πιο αποτελεσματική θεραπεία χαμηλού κινδύνου.

Η χρήση της επιστήμης δεδομένων στην υγειονομική περίθαλψη

Ανάλυση ιατρικών εικόνων

Η επιστήμη δεδομένων παίζει σημαντικό ρόλο με την εφαρμογή της στην ιατρική απεικόνιση. Με τη χρήση δημοφιλών τεχνικών απεικόνισης, μπορεί κανείς να ανακαλύψει αμέτρητες μεθόδους για να βρει τη διαφορά στην τροποποίηση, τη διαφορά των εικόνων και την ανάλυση. Οι τεχνικές απεικόνισης περιλαμβάνουν ακτινογραφία, υπολογιστική τομογραφία, μαγνητική τομογραφία (MRI) και μαστογραφία. Πολλά νέα εργαλεία αναπτύσσονται για να παρέχουν τα πιο ακριβή μέσα για την εξαγωγή δεδομένων από τις εικόνες με αποτελεσματική ποιότητα. Με τη βοήθεια της επιστήμης των δεδομένων, η ερμηνεία της εικόνας έγινε απρόσκοπτη. Οι αλγόριθμοι Deep Learning παρέχουν σε βάθος διαγνώσεις με την πιο ικανή ερμηνεία.

Ανακάλυψη φαρμάκων

Για μια περίοδο 12 ετών, η διαδικασία ανακάλυψης και δημιουργίας φαρμάκων κοστίζει περίπου 2,6 δισεκατομμύρια δολάρια, και ένας μόνο τύπος περνά μέσα από ένα εκατομμύριο διαδικασίες δοκιμής έως ότου εγκριθεί. Στις περισσότερες περιπτώσεις, ακόμη και αφού επενδυθεί τόσος πολύς χρόνος, προσπάθεια και χρήματα, ο τύπος απορρίπτεται.

Ωστόσο, με τη χρήση της επιστήμης των δεδομένων, η διαδικασία συντομεύεται και γίνεται πολύ πιο αποτελεσματική. Η μηχανική μάθηση προσθέτει βήματα για την αρχική εξέταση κάθε συστατικού και προβλέπει τα ποσοστά επιτυχίας μέσω διαφόρων βιολογικών παραγόντων. Οι αλγόριθμοι που εμπλέκονται έχουν την ικανότητα να προβλέπουν την απόκριση και την αντίδραση μιας συγκεκριμένης ένωσης με το σώμα. Αντί να επιλέξει εργαστηριακά πειράματα, η χρήση της

τεχνολογίας εφαρμόζει προσομοιώσεις και μαθηματικά μοντέλα για την ανάλυση της καταλληλότερης ένωσης.

Γενετική και γονιδιωματική

Με τις εξελίξεις στη γενετική και τη γονιδιωματική, έχει εφαρμοστεί η έννοια της εξατομικευμένης θεραπείας. Ο στόχος είναι να αναλυθεί η επίδραση του DNA και η απόκριση διαφόρων φαρμάκων στην υγεία ενός ατόμου σχετικά με τις βιολογικές του συνδέσεις. Για να εξασφαλιστεί μια εις βάθος ανάλυση, η επιστήμη δεδομένων ενθαρρύνει την ενσωμάτωση διαφόρων δεδομένων με γονιδιωματικά δεδομένα. Εξασφαλίζει αποτελεσματική έρευνα για τις ασθένειες.

Επιπλέον, το Deep Genomics συμβάλλει σημαντικά στην πρόβλεψη των μοριακών επιδράσεων των διαφόρων γενετικών σε ερμηνείες του DNA. Με αυτόν τον τρόπο, οι ερευνητές μπορούν να προβλέψουν πώς οι γενετικές παραλλαγές επηρεάζουν τον γενετικό κώδικα.

Προγνωστική ιατρική

Η εφαρμογή της επιστήμης δεδομένων βοηθά στη διεξαγωγή προγνωστικής ανάλυσης. Οι ερευνητές συλλέγουν τα δεδομένα των ασθενών, βρίσκουν τους συσχετισμούς, αναλύουν τις κλινικές σημειώσεις, συσχετίζουν τα συμπτώματα, τις συνήθειες, τα γνωστά προηγούμενα και στη συνέχεια προβαίνουν σε προβλέψεις. Επιπλέον, βιοϊατρικοί παράγοντες που περιλαμβάνουν κλινικές μεταβλητές και η δομή του γονιδιώματος χρησιμοποιούνται επίσης για την εκτέλεση της πρόβλεψης και για τον προσδιορισμό της εξέλιξης ορισμένων ασθενειών., βοηθώντας στη βελτίωση της ποιότητας ζωής και διασφαλίζοντας αποτελεσματική πρόβλεψη ασθενειών. Η μείωση του κινδύνου και η αποφυγή των αρνητικών αποτελεσμάτων είναι μερικά από τα τελικά προϊόντα που δημιουργούνται από την εφαρμογή της επιστήμης δεδομένων.

Ανάλυση δεδομένων

Η *ανάλυση δεδομένων* είναι μια διαδικασία επιθεώρησης, καθαρισμού, μετατροπής και μοντελοποίησης δεδομένων με στόχο την εύρεση χρήσιμων

πληροφοριών, την εξαγωγή συμπερασμάτων και την υποστήριξη της λήψης αποφάσεων. Η ανάλυση δεδομένων έχει πολλαπλές πτυχές και προσεγγίσεις, που περιλαμβάνουν διαφορετικές τεχνικές με μια ποικιλία ονομάτων και χρησιμοποιείται σε διαφορετικούς τομείς επιχειρήσεων, επιστημών και κοινωνικών επιστημών. Στον σημερινό επιχειρηματικό κόσμο, η ανάλυση δεδομένων παίζει ρόλο στη λήψη αποφάσεων πιο επιστημονικών και βοηθά τις επιχειρήσεις να λειτουργούν πιο αποτελεσματικά.

Η *εξόρυξη δεδομένων* είναι μια συγκεκριμένη τεχνική ανάλυσης δεδομένων που εστιάζει στη στατιστική μοντελοποίηση και στην ανακάλυψη γνώσεων για προγνωστικούς και όχι καθαρά περιγραφικούς σκοπούς, ενώ η επιχειρηματική ευφυΐα καλύπτει την ανάλυση δεδομένων που βασίζεται σε μεγάλο βαθμό στη συγκέντρωση, εστιάζοντας κυρίως στις επιχειρηματικές πληροφορίες.

Η διαδικασία της ανάλυσης δεδομένων

Ο όρος «ανάλυση», αναφέρεται στον διαχωρισμό ενός συνόλου στα ξεχωριστά του συστατικά για ατομική εξέταση. Η ανάλυση δεδομένων, είναι μια διαδικασία για την απόκτηση ανεπεξέργαστων δεδομένων, και στη συνέχεια τη μετατροπή τους σε πληροφορίες χρήσιμες για τη λήψη αποφάσεων από τους χρήστες. Τα δεδομένα, συλλέγονται και αναλύονται για να απαντήσουν σε ερωτήσεις, να δοκιμάσουν υποθέσεις ή να διαψεύσουν θεωρίες. Υπάρχουν διάφορες φάσεις που μπορούν να διακριθούν. Οι φάσεις είναι επαναληπτικές, καθώς η ανατροφοδότηση από μεταγενέστερες φάσεις μπορεί να οδηγήσει σε επιπλέον εργασία σε προηγούμενες φάσεις.

Απαιτήσεις δεδομένων

Τα δεδομένα είναι απαραίτητα ως είσοδοι στην ανάλυση, η οποία καθορίζεται με βάση τις απαιτήσεις εκείνων που κατευθύνουν την ανάλυση ή των πελατών (που θα χρησιμοποιήσουν το τελικό προϊόν της ανάλυσης). Ο γενικός τύπος οντότητας στον οποίο θα συλλέγονται τα δεδομένα αναφέρεται ως πειραματική μονάδα (π.χ. άτομο ή πληθυσμός ανθρώπων). Μπορούν να προσδιοριστούν και να ληφθούν συγκεκριμένες μεταβλητές σχετικά με έναν πληθυσμό (π.χ. ηλικία και εισόδημα). Τα

δεδομένα μπορεί να είναι αριθμητικά ή κατηγορηματικά (δηλαδή, μια ετικέτα κειμένου για αριθμούς).

Συλλογή δεδομένων

Τα δεδομένα συλλέγονται από διάφορες πηγές. Οι απαιτήσεις μπορούν να κοινοποιούνται από αναλυτές στους θεματοφύλακες των δεδομένων, όπως το προσωπικό της Πληροφορικής εντός ενός οργανισμού. Τα δεδομένα μπορούν επίσης να συλλεχθούν από αισθητήρες στο περιβάλλον, όπως κάμερες κυκλοφορίας, δορυφόρους, συσκευές εγγραφής κ.λπ. Μπορεί επίσης να ληφθούν μέσω συνεντεύξεων, λήψεων από διαδικτυακές πηγές ή ανάγνωσης εγγράφων.

Επεξεργασία δεδομένων

Τα δεδομένα, όταν αποκτηθούν αρχικά, πρέπει να υποβληθούν σε επεξεργασία ή να οργανωθούν για ανάλυση. Για παράδειγμα, αυτά μπορεί να περιλαμβάνουν τοποθέτηση δεδομένων σε σειρές και στήλες σε μορφή πίνακα (γνωστά ως δομημένα δεδομένα) για περαιτέρω ανάλυση, συχνά μέσω της χρήσης λογιστικού φύλλου ή στατιστικού λογισμικού.

Καθαρισμός δεδομένων

Μετά την επεξεργασία και την οργάνωση, τα δεδομένα ενδέχεται να είναι ελλιπή, να περιέχουν διπλότυπα ή να περιέχουν σφάλματα. Η ανάγκη για καθαρισμό δεδομένων, θα προκύψει από προβλήματα στον τρόπο εισαγωγής και αποθήκευσης των δεδομένων.

Ο καθαρισμός δεδομένων είναι η διαδικασία πρόληψης και διόρθωσης αυτών των σφαλμάτων. Οι συνηθισμένες εργασίες περιλαμβάνουν την αντιστοίχιση εγγραφών, τον προσδιορισμό της ακρίβειας των δεδομένων, τη συνολική ποιότητα των υπαρχόντων δεδομένων, τη διπλή αναπαραγωγή και την τμηματοποίηση στηλών. Τέτοια προβλήματα δεδομένων μπορούν επίσης να εντοπιστούν μέσω μιας ποικιλίας αναλυτικών τεχνικών.

Ασυνήθιστα ποσά, πάνω ή κάτω από προκαθορισμένα όρια, μπορούν επίσης να αναθεωρηθούν. Υπάρχουν διάφοροι τύποι καθαρισμού δεδομένων, που εξαρτώνται

από τον τύπο δεδομένων στο σύνολο. Αυτό θα μπορούσε να είναι αριθμοί τηλεφώνου, διευθύνσεις email, εργοδότες ή άλλες τιμές. Μπορούν να χρησιμοποιηθούν ποσοτικές μέθοδοι δεδομένων για ανίχνευση ασυνήθιστων τιμών, για να απαλλαγούμε από δεδομένα που φαίνεται να έχουν μεγαλύτερη πιθανότητα εσφαλμένης εισαγωγής. Οι ορθογραφικοί έλεγχοι δεδομένων κειμένου, μπορούν να χρησιμοποιηθούν για τη μείωση του όγκου των λανθασμένα δακτυλογραφημένων λέξεων, ωστόσο, είναι πιο δύσκολο να διαπιστωθεί εάν οι ίδιες οι λέξεις είναι σωστές.

Διερευνητική ανάλυση δεδομένων

Μόλις καθαριστούν τα σύνολα δεδομένων, μπορεί στη συνέχεια να αναλυθούν. Οι αναλυτές μπορούν να εφαρμόσουν μια ποικιλία τεχνικών, που αναφέρονται ως διερευνητική ανάλυση δεδομένων, για να αρχίσουν να κατανοούν τα μηνύματα που περιέχονται στα ληφθέντα δεδομένα.

Η διαδικασία εξερεύνησης δεδομένων μπορεί να έχει ως αποτέλεσμα πρόσθετο καθαρισμό δεδομένων ή πρόσθετα αιτήματα για δεδομένα. Περιγραφικά στατιστικά στοιχεία, όπως ο μέσος όρος, μπορούν να δημιουργηθούν για να βοηθήσουν στην κατανόηση των δεδομένων. Η οπτικοποίηση δεδομένων είναι επίσης μια τεχνική που χρησιμοποιείται, στην οποία ο αναλυτής είναι σε θέση να εξετάσει τα δεδομένα σε γραφική μορφή προκειμένου να λάβει πρόσθετες πληροφορίες σχετικά με τα μηνύματα εντός των δεδομένων.

Μοντελοποίηση και αλγόριθμοι

Μαθηματικοί τύποι ή μοντέλα (γνωστοί ως αλγόριθμοι), μπορούν να εφαρμοστούν στα δεδομένα προκειμένου να προσδιοριστούν οι σχέσεις μεταξύ των μεταβλητών. για παράδειγμα, χρησιμοποιώντας συσχέτιση ή αιτιώδη συνάφεια. Σε γενικές γραμμές, τα μοντέλα μπορούν να αναπτυχθούν για την αξιολόγηση μιας συγκεκριμένης μεταβλητής με βάση άλλες μεταβλητές που περιέχονται στο σύνολο δεδομένων, με κάποιο υπολειπόμενο σφάλμα ανάλογα με την ακρίβεια του εφαρμοζόμενου μοντέλου. Τα συμπεραστικά στατιστικά στοιχεία, περιλαμβάνουν τη χρήση τεχνικών που μετρούν τις σχέσεις μεταξύ συγκεκριμένων μεταβλητών. Για παράδειγμα, η ανάλυση παλινδρόμησης μπορεί να χρησιμοποιηθεί για να

μοντελοποιήσει εάν μια αλλαγή στην ανεξάρτητη μεταβλητή X, παρέχει μια εξήγηση για τη διακύμανση της εξαρτώμενης μεταβλητής Y.

Προϊόν δεδομένων

Ένα προϊόν δεδομένων, είναι μια εφαρμογή υπολογιστή που λαμβάνει εισόδους δεδομένων και παράγει εξόδους, τροφοδοτώντας τα πίσω στο περιβάλλον. Μπορεί να βασίζεται σε μοντέλο ή αλγόριθμο.

Πηγές δεδομένων για ανάλυση δεδομένων

Τα δεδομένα που πρόκειται να αναλυθούν πρέπει να συλλέγονται από διαφορετικές έγκυρες πηγές. Ο κύριος στόχος της συλλογής δεδομένων είναι η συλλογή δεδομένων πλούσιων σε πληροφορίες. Η συλλογή ξεκινά με την υποβολή ορισμένων ερωτήσεων, όπως τι είδους δεδομένα πρέπει να συλλεχθούν και ποια είναι η πηγή συλλογής.

Τα περισσότερα από τα δεδομένα που συλλέγονται είναι δύο τύπων γνωστά ως «ποιοτικά δεδομένα» που είναι μια ομάδα μη αριθμητικών δεδομένων όπως λέξεις, οι προτάσεις εστιάζουν κυρίως στη συμπεριφορά και τις ενέργειες της ομάδας και ένας άλλος είναι «ποσοτικά δεδομένα» που είναι αριθμητικά και μπορούν να υπολογιστούν χρησιμοποιώντας διαφορετικά επιστημονικά εργαλεία και δεδομένα δειγματοληψίας.

Τα δεδομένα που είναι ακατέργαστα, πρωτότυπα και εξάγονται απευθείας από τις επίσημες πηγές είναι γνωστά ως πρωτεύοντα δεδομένα. Αυτός ο τύπος δεδομένων συλλέγεται απευθείας εκτελώντας τεχνικές όπως ερωτηματολόγια, συνεντεύξεις και έρευνες.

Λίγες μέθοδοι συλλογής πρωτογενών δεδομένων:

Μέσω συνεντεύξεων

Τα δεδομένα που συλλέγονται κατά τη διάρκεια αυτής της διαδικασίας είναι μέσω της συνέντευξης του κοινού-στόχου. Ορισμένες βασικές ερωτήσεις σχετικά με την επιχείρηση ή το προϊόν τίθενται και σημειώνονται με τη μορφή σημειώσεων, ήχου ή βίντεο και αυτά τα δεδομένα αποθηκεύονται για επεξεργασία.

Μέθοδος έρευνας:

Η μέθοδος της έρευνας είναι η διαδικασία της έρευνας όπου υποβάλλεται μια λίστα με σχετικές ερωτήσεις και οι απαντήσεις σημειώνονται με τη μορφή κειμένου, ήχου ή βίντεο. Η μέθοδος έρευνας μπορεί να ληφθεί τόσο σε διαδικτυακή όσο και σε κατάσταση εκτός σύνδεσης, όπως μέσω φόρμας ιστότοπου και email. Στη συνέχεια, οι απαντήσεις της έρευνας αποθηκεύονται για την ανάλυση δεδομένων. Παραδείγματα είναι διαδικτυακές έρευνες ή έρευνες μέσω δημοσκοπήσεων κοινωνικών μέσων.

Μέθοδος παρατήρησης

Η μέθοδος παρατήρησης είναι μια μέθοδος συλλογής δεδομένων στην οποία ο ερευνητής παρατηρεί προσεκτικά τη συμπεριφορά και τις πρακτικές του κοινού-στόχου χρησιμοποιώντας κάποιο εργαλείο συλλογής δεδομένων και αποθηκεύει τα παρατηρούμενα δεδομένα με τη μορφή κειμένου, ήχου, βίντεο ή οποιωνδήποτε πρώτων μορφών. Σε αυτήν τη μέθοδο, τα δεδομένα συλλέγονται απευθείας δημοσιεύοντας μερικές ερωτήσεις στους συμμετέχοντες.

Πειραματική μέθοδος

Η πειραματική μέθοδος είναι η διαδικασία συλλογής δεδομένων μέσω της εκτέλεσης πειραμάτων, έρευνας και διερεύνησης.

ΚΕΦΑΛΑΙΟ 2

Εισαγωγή στις χρονοσειρές

Μια χρονοσειρά είναι μια ακολουθία ή σειρά αριθμητικών σημείων δεδομένων που καθορίζονται σε συγκεκριμένη χρονολογική σειρά. Στις περισσότερες περιπτώσεις, μια χρονοσειρά είναι μια ακολουθία που λαμβάνεται σε καθορισμένα χρονικά διαστήματα. Αυτό μας επιτρέπει να προβλέψουμε με ακρίβεια ανάλογα με τις ανάγκες της έρευνας.

Οι χρονοσειρές χρησιμοποιούν γραφήματα γραμμών για να δείξουν εποχιακά μοτίβα, τάσεις και σχέση με εξωτερικούς παράγοντες. Οι χρονοσειρές χρησιμοποιούνται στις περισσότερες περιπτώσεις στην πραγματική ζωή, για αναφορές καιρού, πρόβλεψη σεισμών, αστρονομία, μαθηματική οικονομία και σε μεγάλο βαθμό σε οποιοδήποτε πεδίο εφαρμοσμένης επιστήμης και μηχανικής. Προσφέρει βαθύτερες γνώσεις στον τομέα εργασίας και η πρόβλεψη βοηθά στην αύξηση της αποδοτικότητας της παραγωγής.

Στα Μαθηματικά, μια χρονική σειρά είναι μια σειρά σημείων δεδομένων που έχουν ευρετηριαστεί (ή παρατίθενται ή γράφονται) με χρονική σειρά. Συνήθως, μια χρονοσειρά είναι μια ακολουθία που λαμβάνεται σε διαδοχικά ισόποσα χρονικά σημεία. Έτσι είναι μια ακολουθία δεδομένων διακριτού χρόνου.

Οι χρονοσειρές σχεδιάζονται πολύ συχνά μέσω διαγραμμάτων εκτέλεσης (ένα χρονικό διάγραμμα γραμμών). Οι χρονοσειρές χρησιμοποιούνται σε στατιστικά στοιχεία, επεξεργασία σήματος, αναγνώριση προτύπων, οικονομετρία, μαθηματική οικονομία, πρόγνωση καιρού, πρόβλεψη σεισμού, ηλεκτροεγκεφαλογραφία, μηχανική ελέγχου, αστρονομία, μηχανική επικοινωνιών και σε μεγάλο βαθμό σε οποιονδήποτε τομέα εφαρμοσμένης επιστήμης και μηχανικής που περιλαμβάνει χρονικές μετρήσεις.

Η ανάλυση χρονοσειρών περιλαμβάνει μεθόδους για την ανάλυση δεδομένων χρονοσειρών με σκοπό την εξαγωγή σημαντικών στατιστικών και άλλων χαρακτηριστικών των δεδομένων.

Η πρόβλεψη χρονοσειρών είναι η χρήση ενός μοντέλου για την πρόβλεψη μελλοντικών τιμών βάσει των τιμών που παρατηρήθηκαν προηγουμένως. Ενώ η

ανάλυση παλινδρόμησης χρησιμοποιείται συχνά με τέτοιο τρόπο ώστε να ελέγχει τις σχέσεις μεταξύ μιας ακόμη διαφορετικής χρονοσειράς, αυτός ο τύπος ανάλυσης δεν ονομάζεται συνήθως «ανάλυση χρονοσειρών», η οποία αναφέρεται συγκεκριμένα στις σχέσεις μεταξύ διαφορετικών χρονικών σημείων σε μία μόνο σειρά .

Η ανάλυση διακοπών χρονοσειρών χρησιμοποιείται για τον εντοπισμό αλλαγών στην εξέλιξη μιας χρονοσειράς από πριν έως μετά από κάποια παρέμβαση που μπορεί να επηρεάσει την υποκείμενη μεταβλητή.

Τα δεδομένα χρονοσειρών έχουν φυσική χρονική σειρά. Αυτό καθιστά την ανάλυση χρονοσειρών διαφορετική από τις μελέτες διατομής, στις οποίες δεν υπάρχει φυσική σειρά των παρατηρήσεων (π.χ. εξήγηση των μισθών των ανθρώπων με αναφορά στα αντίστοιχα επίπεδα εκπαίδευσης τους, όπου τα δεδομένα των ατόμων θα μπορούσαν να εισαχθούν με οποιαδήποτε σειρά).

Η ανάλυση χρονοσειρών διαφέρει επίσης από την ανάλυση χωρικών δεδομένων όπου οι παρατηρήσεις σχετίζονται συνήθως με γεωγραφικές τοποθεσίες (π.χ. λογιστικές τιμές των κατοικιών από την τοποθεσία καθώς και τα εγγενή χαρακτηριστικά των κατοικιών). Ένα στοχαστικό μοντέλο για μια χρονοσειρά θα αντικατοπτρίζει γενικά το γεγονός ότι οι παρατηρήσεις κοντά στο χρόνο θα σχετίζονται στενότερα από τις παρατηρήσεις που απέχουν περισσότερο.

Επιπλέον, τα μοντέλα χρονοσειρών συχνά χρησιμοποιούν τη φυσική μονόδρομη μεταβολή του χρόνου, έτσι ώστε οι τιμές για μια δεδομένη περίοδο να εκφράζονται ως προερχόμενες κατά κάποιο τρόπο από τις προηγούμενες τιμές και όχι από τις μελλοντικές τιμές.

Η ανάλυση χρονοσειρών μπορεί να εφαρμοστεί σε πραγματικά, συνεχή, διακριτά αριθμητικά ή διακριτά συμβολικά δεδομένα (δηλαδή ακολουθίες χαρακτήρων, όπως γράμματα και λέξεις στην αγγλική γλώσσα).

Η χρονοσειρά είναι ένας τύπος δεδομένων πίνακα. Τα δεδομένα πίνακα είναι η γενική κλάση, ένα πολυδιάστατο σύνολο δεδομένων, ενώ ένα σύνολο δεδομένων χρονοσειρών είναι ένα μονοδιάστατο πλαίσιο (όπως είναι ένα σύνολο δεδομένων διατομής).

Ένα σύνολο δεδομένων μπορεί να εμφανίζει χαρακτηριστικά τόσο των δεδομένων πίνακα όσο και των δεδομένων χρονοσειρών. Ένας τρόπος για να ξεχωριστούν είναι να παρατηρηθεί τι κάνει μια εγγραφή δεδομένων μοναδική από τις

άλλες εγγραφές. Εάν η απάντηση είναι το πεδίο δεδομένων χρόνου, τότε πρόκειται για υποψήφιο συνόλου δεδομένων χρονοσειρών. Εάν ο καθορισμός μιας μοναδικής εγγραφής απαιτεί ένα πεδίο δεδομένων χρόνου και ένα πρόσθετο αναγνωριστικό που δεν σχετίζεται με το χρόνο (αναγνωριστικό μαθητή, σύμβολο μετοχής, κωδικός χώρας), τότε είναι υποψήφιος για δεδομένα πίνακα. Εάν η διαφοροποίηση βρίσκεται στο μη αναγνωριστικό χρόνου, τότε το σύνολο δεδομένων είναι υποψήφιο σύνολο δεδομένων διατομής.

Πρόβλεψη χρονοσειρών

Η πρόβλεψη χρονοσειρών είναι μια μέθοδος χρήσης ενός μοντέλου για την πρόβλεψη μελλοντικών τιμών βάσει των τιμών των χρονοσειρών που έχουν παρατηρηθεί προηγουμένως. Οι χρονοσειρές είναι ένα σημαντικό μέρος της μηχανικής μάθησης. Καταγράφει ένα εποχιακό μοτίβο ή τάση στα δεδομένα των χρονοσειρών που παρατηρούνται και τα χρησιμοποιεί για μελλοντικές προβλέψεις. Η πρόβλεψη περιλαμβάνει τη λήψη μοντέλων πλούσιων σε ιστορικά δεδομένα και τη χρήση τους για την πρόβλεψη μελλοντικών παρατηρήσεων.

Ένα από τα πιο χαρακτηριστικά της πρόβλεψης χρονοσειρών είναι ότι δεν προβλέπει ακριβώς το μέλλον, απλώς μας δίνει μια υπολογισμένη εκτίμηση του τι έχει ήδη συμβεί για να μας δώσει μια ιδέα για το τι θα μπορούσε να συμβεί.

Γενικές μέθοδοι πρόβλεψης

Οι μέθοδοι ανάλυσης χρονοσειρών μπορούν να χωριστούν σε δύο κατηγορίες: μεθόδους πεδίου συχνότητας και μέθοδοι πεδίου χρόνου. Τα πρώτα περιλαμβάνουν φασματική ανάλυση και ανάλυση κύματος. Τα τελευταία περιλαμβάνουν ανάλυση αυτόματης συσχέτισης και διασταυρούμενης συσχέτισης. Στο χρονικό πεδίο, η συσχέτιση και η ανάλυση μπορούν να γίνουν με φίλτρο, χρησιμοποιώντας κλιμακωτή συσχέτιση, ελαχιστοποιώντας έτσι την ανάγκη λειτουργίας στο πεδίο συχνότητας.

Επιπλέον, οι τεχνικές ανάλυσης χρονοσειρών μπορούν να χωριστούν σε παραμετρικές και μη παραμετρικές μεθόδους. Οι παραμετρικές προσεγγίσεις

υποθέτουν ότι η υποκείμενη στατική στοχαστική διαδικασία έχει μια συγκεκριμένη δομή η οποία μπορεί να περιγραφεί χρησιμοποιώντας έναν μικρό αριθμό παραμέτρων (για παράδειγμα, χρησιμοποιώντας ένα μοντέλο αυτοσυσχέτισης ή κινούμενου μέσου όρου).

Σε αυτές τις προσεγγίσεις, ο στόχος είναι να εκτιμηθούν οι παράμετροι του μοντέλου που περιγράφει τη στοχαστική διαδικασία. Αντιθέτως, οι μη παραμετρικές προσεγγίσεις εκτιμούν ρητά τη συνδιακύμανση ή το φάσμα της διαδικασίας χωρίς να υποθέσουμε ότι η διαδικασία έχει κάποια συγκεκριμένη δομή.

Οι μέθοδοι ανάλυσης χρονοσειρών μπορούν επίσης να χωριστούν σε γραμμικές και μη γραμμικές, και μίας μεταβλητής και πολλαπλών μεταβλητών.

Η ποιοτική πρόβλεψη χρησιμοποιείται γενικά όταν τα ιστορικά δεδομένα δεν είναι διαθέσιμα και θεωρείται εξαιρετικά αντικειμενικά και κρίσιμα.

Η ποσοτική πρόβλεψη είναι όταν έχουμε μεγάλες ποσότητες δεδομένων από το παρελθόν και θεωρείται εξαιρετικά αποτελεσματική εφόσον δεν υπάρχουν ισχυροί εξωτερικοί παράγοντες στο παιχνίδι.

Η ικανότητα ενός μοντέλου πρόβλεψης χρονοσειρών καθορίζεται από την αποτελεσματικότητά του στην πρόβλεψη του μέλλοντος. Αυτό συχνά κοστίζει το να μπορούμε να εξηγήσουμε γιατί έγινε μια συγκεκριμένη πρόβλεψη, διαστήματα εμπιστοσύνης ή και ακόμη την κατανόηση των υποκειμενικών παραγόντων πίσω από το πρόβλημα.

Εφαρμογή προγνώσεων χρονοσειρών

Η χρήση μοντέλων χρονοσειρών είναι διπλή:

- Απόκτηση κατανόησης των υποκείμενων δυνάμεων και της δομής που παρήγαγε τα δεδομένα
- Ταίριασμα ενός μοντέλου και πρόβλεψη.

Η ανάλυση χρονοσειρών έχει πολλούς διαφορετικούς στόχους, ανάλογα με το πεδίο εφαρμογής. Αυτά περιλαμβάνουν την πρόβλεψη μελλοντικών τιμών της σειράς, την εξαγωγή ενός σήματος κρυμμένου σε θορυβώδη δεδομένα, την ανακάλυψη του

μηχανισμού με τον οποίο δημιουργούνται τα δεδομένα, την προσομοίωση ανεξάρτητων συνειδητοποιήσεων της σειράς για να δει πώς μπορεί να συμπεριφέρεται στο μέλλον (και ως εκ τούτου, για παράδειγμα, εκτίμηση της πιθανότητας ακραίων γεγονότων όπως πλημμύρες) και εξαλείφοντας την εποχιακή συνιστώσα από σύνολα δεδομένων, προκειμένου να αποκαλυφθεί πιο ξεκάθαρα η υποκείμενη τάση.

Για όλες αυτές τις εφαρμογές, η ανάλυση χρονοσειρών συνήθως ξεκινά με μια προσπάθεια εύρεσης ενός μαθηματικού μοντέλου που παρέχει μια καλή αναπαράσταση των παρατηρούμενων δεδομένων.

Αν και σπάνια συμβαίνει ότι υπάρχει ένα πραγματικό μαθηματικό μοντέλο που να βασίζεται σε εμπειρικά δεδομένα, έχουν αναπτυχθεί συστηματικές διαδικασίες για την επιλογή ενός καλού μοντέλου, σύμφωνα με σαφώς καθορισμένα κριτήρια. Μόλις επιτευχθεί αυτό, το επιλεγμένο μοντέλο μπορεί να χρησιμοποιηθεί για την αντιμετώπιση των ερωτήσεων που τέθηκαν στην προηγούμενη παράγραφο.

Υπάρχει σχεδόν μια ατελείωτη εφαρμογή προβλημάτων πρόβλεψης χρονοσειρών.

Ακολουθούν μερικά από τα παραδείγματα από μια σειρά βιομηχανιών που κάνουν τις έννοιες της ανάλυσης χρονολογικών σειρών και των προβλέψεων πιο ισχυρές.

- Προβλέποντας την απόδοση του ρυζιού σε τόνους από το κράτος κάθε χρόνο.
- Η πρόβλεψη εάν ένα ίχνος EEG σε δευτερόλεπτα δείχνει ότι ένας ασθενής έχει καρδιακή προσβολή ή όχι.
- Πρόβλεψη της τιμής κλεισίματος των αποθεμάτων κάθε μέρα.
- Πρόβλεψη του ποσοστού γέννησης ή θανάτου σε όλα τα νοσοκομεία μιας πόλης κάθε χρόνο.
- Πρόβλεψη πωλήσεων προϊόντων σε μονάδες που πωλούνται καθημερινά.
- Πρόβλεψη του αριθμού των επιβατών που κάνουν κράτηση εισιτηρίων πτήσης κάθε μέρα.
- Πρόβλεψη ανεργίας για ένα κράτος κάθε τρίμηνο
- Πρόβλεψη του μεγέθους του πληθυσμού της τίγρης σε μια κατάσταση κάθε σεζόν αναπαραγωγής.

Στο πλαίσιο της στατιστικής, της οικονομετρίας, της ποσοτικής χρηματοδότησης, της σεισμολογίας, της μετεωρολογίας και της γεωφυσικής, ο πρωταρχικός στόχος της ανάλυσης χρονοσειρών είναι η πρόβλεψη. Στο πλαίσιο της επεξεργασίας σήματος, της μηχανικής ελέγχου και της μηχανικής επικοινωνίας χρησιμοποιείται για την ανίχνευση σήματος. Άλλες εφαρμογές είναι η εξόρυξη δεδομένων, η αναγνώριση προτύπων και η μηχανική εκμάθηση, όπου η ανάλυση χρονοσειρών μπορεί να χρησιμοποιηθεί για ομαδοποίηση, ταξινόμηση, ερώτημα βάσει περιεχομένου, ανίχνευση ανωμαλιών καθώς και πρόβλεψη.

Ένας απλός τρόπος για να εξεταστεί εάν μια κανονική χρονολογική σειρά είναι χειροκίνητα με ένα γράφημα γραμμών.

Άλλες τεχνικές περιλαμβάνουν:

- Ανάλυση αυτοσυσχέτισης για εξέταση της σειριακής εξάρτησης
- Φασματική ανάλυση για την εξέταση της κυκλικής συμπεριφοράς που δεν χρειάζεται να σχετίζεται με την εποχικότητα. Για παράδειγμα, η δραστηριότητα στον ήλιο κυμαίνεται σε κύκλους 11 ετών. Άλλα κοινά παραδείγματα περιλαμβάνουν ουράνια φαινόμενα, καιρικές συνθήκες, νευρική δραστηριότητα, τιμές εμπορευμάτων και οικονομική δραστηριότητα.
- Διαχωρισμός σε στοιχεία που αντιπροσωπεύουν τάση, εποχικότητα, αργή και γρήγορη διακύμανση και κυκλική ανωμαλία.

Η προσαρμογή καμπύλης είναι η διαδικασία κατασκευής μιας καμπύλης, ή μαθηματικής συνάρτησης, η οποία ταιριάζει καλύτερα σε μια σειρά σημείων δεδομένων, πιθανώς υπόκεινται σε περιορισμούς. Η προσαρμογή καμπύλης μπορεί να περιλαμβάνει είτε παρεμβολή, όπου απαιτείται ακριβής προσαρμογή στα δεδομένα, είτε εξομάλυνση, στην οποία κατασκευάζεται μια "ομαλή" λειτουργία που ταιριάζει περίπου στα δεδομένα.

Ένα σχετικό θέμα είναι η ανάλυση παλινδρόμησης, η οποία επικεντρώνεται περισσότερο σε ζητήματα στατιστικών συμπερασμάτων, όπως πόση αβεβαιότητα υπάρχει σε μια καμπύλη που ταιριάζει σε δεδομένα που παρατηρούνται με τυχαία σφάλματα. Οι προσαρμοσμένες καμπύλες μπορούν να χρησιμοποιηθούν ως βοήθημα για την οπτικοποίηση δεδομένων, για να συναγάγουν τιμές μιας συνάρτησης όπου δεν υπάρχουν διαθέσιμα δεδομένα και για να συνοψίσουν τις σχέσεις μεταξύ δύο ή περισσότερων μεταβλητών.

Η παρέκταση αναφέρεται στη χρήση μιας προσαρμοσμένης καμπύλης πέρα από το εύρος των παρατηρούμενων δεδομένων και υπόκειται σε βαθμό αβεβαιότητας, καθώς μπορεί να αντικατοπτρίζει τη μέθοδο που χρησιμοποιείται για την κατασκευή της καμπύλης όσο αντανακλά τα παρατηρούμενα δεδομένα.

Η κατασκευή οικονομικών χρονοσειρών περιλαμβάνει την εκτίμηση ορισμένων συνιστωσών για ορισμένες ημερομηνίες με παρεμβολή μεταξύ τιμών ("σημεία αναφοράς") για προηγούμενες και μεταγενέστερες ημερομηνίες. Η παρεμβολή είναι εκτίμηση μιας άγνωστης ποσότητας μεταξύ δύο γνωστών ποσοτήτων (ιστορικά δεδομένα) ή εξαγωγή συμπερασμάτων σχετικά με τις ελλείπουσες πληροφορίες από τις διαθέσιμες πληροφορίες ("ανάγνωση μεταξύ των γραμμών"). Η παρεμβολή είναι χρήσιμη όταν τα δεδομένα που περιβάλλουν τα δεδομένα που λείπουν είναι διαθέσιμα και είναι γνωστή η τάση, η εποχικότητα και οι μακροπρόθεσμοι κύκλοι της. Αυτό γίνεται συχνά χρησιμοποιώντας μια σχετική σειρά γνωστή για όλες τις σχετικές ημερομηνίες. Εναλλακτικά, χρησιμοποιείται πολυωνυμική παρεμβολή όταν οι πολυωνυμικές λειτουργίες κατά τεμάχιο ταιριάζουν σε χρονικά διαστήματα έτσι ώστε να ταιριάζουν ομαλά μεταξύ τους.

Ένα διαφορετικό πρόβλημα που σχετίζεται στενά με την παρεμβολή είναι η προσέγγιση μιας πολύπλοκης συνάρτησης με μια απλή λειτουργία (ονομάζεται επίσης παλινδρόμηση). Η κύρια διαφορά μεταξύ παλινδρόμησης και παρεμβολής είναι ότι η πολυωνυμική παλινδρόμηση δίνει ένα μόνο πολυώνυμο που μοντελοποιεί ολόκληρο το σύνολο δεδομένων. Η ενδιάμεση παρεμβολή, ωστόσο, αποδίδει μια συνεχή συνάρτηση που αποτελείται από πολλά πολυώνυμα για να μοντελοποιήσει το σύνολο δεδομένων.

Η παρέκταση είναι η διαδικασία εκτίμησης, πέρα από το αρχικό εύρος παρατήρησης, της τιμής μιας μεταβλητής με βάση τη σχέση της με μια άλλη μεταβλητή. Είναι παρόμοιο με την παρεμβολή, η οποία παράγει εκτιμήσεις μεταξύ γνωστών παρατηρήσεων, αλλά η παρέκταση υπόκειται σε μεγαλύτερη αβεβαιότητα και υψηλότερο κίνδυνο παραγωγής αποτελεσμάτων χωρίς νόημα.

Μοντέλα

Τα μοντέλα για δεδομένα χρονοσειρών μπορούν να έχουν πολλές μορφές και να αντιπροσωπεύουν διαφορετικές στοχαστικές διαδικασίες. Κατά τη μοντελοποίηση διακυμάνσεων στο επίπεδο μιας διαδικασίας, τρεις ευρείες κατηγορίες πρακτικής σημασίας είναι τα μοντέλα αυτό-παλινδρόμησης (AR), τα ενσωματωμένα (I) μοντέλα και τα μοντέλα κινητού μέσου όρου (MA). Αυτές οι τρεις κατηγορίες εξαρτώνται γραμμικά από προηγούμενα σημεία δεδομένων.

Οι συνδυασμοί αυτών των ιδεών παράγουν μοντέλα αυτοπαλινδρόμισης κινούμενου μέσου όρου (ARMA) και αυτοπαλινδρόμισης ενσωματωμένου κινούμενου μέσου (ARIMA). Το μοντέλο αυτόματης αύξησης κλασματικά ενσωματωμένου κινούμενου μέσου όρου (ARFIMA) γενικεύει τα τρία πρώτα. Επεκτάσεις αυτών των κλάσεων για την αντιμετώπιση δεδομένων είναι διαθέσιμες υπό τον τίτλο των μοντέλων χρονομετρικών σειρών πολλαπλών παραλλαγών και μερικές φορές τα προηγούμενα ακρωνύμια επεκτείνονται συμπεριλαμβάνοντας ένα αρχικό "V" για "διάνυσμα", όπως στο VAR για αυτό-παλινδρόμηση φορέα.

Ένα επιπλέον σύνολο επεκτάσεων αυτών των μοντέλων είναι διαθέσιμο για χρήση όπου οι παρατηρούμενες χρονοσειρές οδηγούνται από κάποιες χρονοβόρες σειρές "εξαναγκασμού" (οι οποίες ενδέχεται να μην έχουν αιτιώδη επίδραση στις παρατηρούμενες σειρές): η διάκριση από την περίπτωση πολλαπλών παραλλαγών είναι ότι η σειρά εξαναγκασμού μπορεί να είναι ντετερμινιστική ή υπό τον έλεγχο του πειραματιστή. Για αυτά τα μοντέλα, τα ακρωνύμια επεκτείνονται με ένα τελικό "X" για "εξωγενές".

Η μη γραμμική εξάρτηση του επιπέδου μιας σειράς από προηγούμενα σημεία δεδομένων παρουσιάζει ενδιαφέρον, εν μέρει λόγω της δυνατότητας παραγωγής μιας χαοτικής χρονοσειράς. Ωστόσο, το πιο σημαντικό, οι εμπειρικές έρευνες μπορούν να δείξουν το πλεονέκτημα της χρήσης προβλέψεων που προέρχονται από μη γραμμικά μοντέλα, έναντι εκείνων από γραμμικά μοντέλα, όπως για παράδειγμα σε μη γραμμικά εξωγενή μοντέλα.

Μεταξύ άλλων τύπων μη γραμμικών μοντέλων χρονοσειρών, υπάρχουν μοντέλα που αντιπροσωπεύουν τις μεταβολές της διακύμανσης με την πάροδο του χρόνου (ετεροσκεδαστικότητα). Αυτά τα μοντέλα αντιπροσωπεύουν αυτό-

παλινδρόμηση και ετεροσκεδιστικότητα υπό όρους (ARCH) και η συλλογή περιλαμβάνει μια μεγάλη ποικιλία αναπαράστασης (GARCH, TARCH, EGARCH, FIGARCH, CGARCH κ.λπ.). Εδώ οι αλλαγές στη μεταβλητότητα σχετίζονται με, ή προβλέπονται από, πρόσφατες προηγούμενες τιμές της σειράς που παρατηρήθηκε. Αυτό έρχεται σε αντίθεση με άλλες πιθανές αναπαραστάσεις τοπικής μεταβλητότητας, όπου η μεταβλητότητα μπορεί να μοντελοποιηθεί ως καθοδηγούμενη από μια ξεχωριστή διαδικασία που ποικίλλει από το χρόνο, όπως σε ένα διπλά στοχαστικό μοντέλο.

Σε πρόσφατες εργασίες για αναλύσεις χωρίς μοντέλα, οι μέθοδοι που βασίζονται σε μετασχηματισμό κυματοειδών (για παράδειγμα τοπικά σταθερά κύματα και νευρωνικά δίκτυα αποσυντιθέμενων κυματοειδών) έχουν κερδίσει εύνοια. Οι τεχνικές Multiscale (συχνά αναφέρονται ως multiresolution) αποσυνθέτουν μια δεδομένη χρονική σειρά, προσπαθώντας να απεικονίσουν την εξάρτηση από το χρόνο σε πολλές κλίμακες. Αξίζει να σημειωθούν επίσης οι τεχνικές πολλαπλών κλασικών αλλαγών Markov (MSMF) για μοντελοποίηση της εξέλιξης μεταβλητότητας.

Ένα κρυφό μοντέλο Markov (HMM) είναι ένα στατιστικό μοντέλο Markov στο οποίο το σύστημα που μοντελοποιείται θεωρείται μια διαδικασία Markov με μη παρατηρημένες (κρυφές) καταστάσεις. Ένα HMM μπορεί να θεωρηθεί ως το απλούστερο δυναμικό δίκτυο Bayesian. Τα μοντέλα HMM χρησιμοποιούνται ευρέως στην αναγνώριση ομιλίας, για τη μετάφραση μιας σειράς προφορικών λέξεων σε κείμενο.

Συστατικά των χρονοσειρών

Η ανάλυση χρονοσειρών παρέχει μια πληθώρα τεχνικών για την καλύτερη κατανόηση ενός συνόλου δεδομένων. Ίσως το πιο χρήσιμο από αυτά είναι ο διαχωρισμός των χρονοσειρών σε 4 μέρη:

- ***Επίπεδο:*** Η βασική τιμή για τη σειρά αν ήταν ευθεία.
- ***Τάση:*** Η γραμμική αύξηση ή μείωση της συμπεριφοράς της σειράς με την πάροδο του χρόνου.
- ***Εποχικότητα:*** Τα επαναλαμβανόμενα πρότυπα ή κύκλοι συμπεριφοράς με την πάροδο του χρόνου.

- ***Θόρυβος:*** Η μεταβλητότητα στις παρατηρήσεις που δεν μπορεί να εξηγηθεί από το μοντέλο.

Οι σειρές όλων των εποχών έχουν γενικά επίπεδο, θόρυβο, ενώ η τάση και η εποχικότητα είναι προαιρετικές.

Τα κύρια χαρακτηριστικά πολλών χρονοσειρών είναι οι τάσεις και οι εποχιακές παραλλαγές. Ένα άλλο χαρακτηριστικό των περισσότερων χρονοσειρών είναι ότι οι παρατηρήσεις που κλείνουν μαζί στο χρόνο τείνουν να συσχετίζονται

Αυτά τα στοιχεία συνδυάζονται με κάποιο τρόπο για να παρέχουν τις παρατηρούμενες χρονοσειρές. Για παράδειγμα, μπορούν να προστεθούν μαζί για να σχηματίσουν ένα μοντέλο όπως:

Y = επίπεδα + τάσεις + εποχικότητα + θόρυβος

Αυτά τα στοιχεία είναι ο πιο αποτελεσματικός τρόπος για να πραγματοποιηθούν προβλέψεις σχετικά με μελλοντικές τιμές, αλλά μπορεί να μην λειτουργούν πάντα. Αυτό εξαρτάται από την ποσότητα των δεδομένων που έχουμε για το παρελθόν.

Ανάλυση τάσης

Ο έλεγχος δεδομένων για επαναλαμβανόμενη συμπεριφορά στη γραφική αναπαράστασή του είναι γνωστός ως ανάλυση τάσης. Όσο η τάση αυξάνεται συνεχώς ή μειώνεται, αυτό το μέρος της ανάλυσης δεδομένων δεν είναι γενικά πολύ δύσκολο. Εάν τα δεδομένα των χρονοσειρών περιέχουν κάποιο είδος σημαντικού σφάλματος, τότε το πρώτο βήμα στη διαδικασία αναγνώρισης τάσεων είναι ομαλή.

Η εξομάλυνση περιλαμβάνει πάντα κάποια μορφή τοπικού μέσου όρου δεδομένων, έτσι ώστε τα συστατικά των μεμονωμένων παρατηρήσεων να ακυρώνονται μεταξύ τους. Η πιο ευρέως χρησιμοποιούμενη τεχνική είναι ο μέσος όρος κίνησης που αντικαθιστά κάθε στοιχείο της σειράς με έναν απλό ή σταθμισμένο μέσο όρο των γύρω στοιχείων. Οι διάμεσοι χρησιμοποιούνται ως επί το πλείστον αντί για μέσα.

Το κύριο πλεονέκτημα του μέσου όρου σε σύγκριση με την εξομάλυνση του κινούμενου μέσου όρου είναι ότι τα αποτελέσματά του είναι λιγότερο προκατειλημμένα από τα ακραία επίπεδα στο παράθυρο εξομάλυνσης. Το κύριο

μειονέκτημα της μέσης εξομάλυνσης είναι ότι απουσία καθαρών ακραίων ακρών μπορεί να παράγει περισσότερες διαταραγμένες καμπύλες από τον μέσο όρο κίνησης.

Στις άλλες λιγότερο συχνές περιπτώσεις, όταν το σφάλμα μέτρησης είναι μεγάλο, μπορεί να χρησιμοποιηθούν τεχνικές εξομάλυνσης με ελάχιστα τετράγωνα από απόσταση ή αρνητικές εκθετικά σταθμισμένες τεχνικές εξομάλυνσης. Αυτές οι μέθοδοι συνήθως τείνουν να αγνοούν τα ακραία σημεία και να δίνουν μια ομαλή καμπύλη προσαρμογής.

Εάν υπάρχει ένα σαφές μονοτονικό μη γραμμικό στοιχείο, τα δεδομένα πρέπει πρώτα να μετατραπούν για να αφαιρεθεί η μη γραμμικότητα. Συνήθως, η συνάρτηση log, εκθετική ή πολυώνυμη χρησιμοποιείται για να επιτευχθεί αυτό.

Ανάλυση εποχικότητας

Η εποχικότητα είναι η επανάληψη των δεδομένων σε μια συγκεκριμένη χρονική περίοδο. Για παράδειγμα, κάθε χρόνο παρατηρούμε ότι οι άνθρωποι τείνουν να πηγαίνουν διακοπές κατά την περίοδο Δεκεμβρίου - Ιανουαρίου, αυτό είναι εποχικότητα. Είναι ένα άλλο πιο σημαντικό χαρακτηριστικό της ανάλυσης χρονοσειρών. Γενικά μετριέται με αυτοσυσχέτιση μετά την αφαίρεση της τάσης από τα δεδομένα.

Μοντέλο αυτο-παλινδρόμησης (AR)

Το AR είναι ένα μοντέλο χρονοσειρών που χρησιμοποιεί παρατηρήσεις από προηγούμενα βήματα χρόνου ως εισαγωγή σε μια εξίσωση παλινδρόμησης για να προβλέψει την τιμή στο επόμενο βήμα. Ένα μοντέλο παλινδρόμησης όπως η γραμμική παλινδρόμηση έχει τη μορφή:

$$yhat = b0 + (b1 * X1)$$

Αυτή η τεχνική μπορεί να χρησιμοποιηθεί σε χρονοσειρές όπου οι μεταβλητές εισόδου λαμβάνονται ως παρατηρήσεις σε προηγούμενα χρονικά βήματα, που ονομάζονται μεταβλητές υστέρησης. Αυτό θα μοιάζει με:

$$Xt+1 = b0 + (b1 * Xt) + (b2 * Xt-1)$$

Δεδομένου ότι το μοντέλο παλινδρόμησης χρησιμοποιεί δεδομένα από την ίδια μεταβλητή εισόδου σε προηγούμενα χρονικά βήματα, αναφέρεται ως αυτό-παλινδρόμησης.

Μοντέλο κινούμενου μέσου (MA)

Τα υπολειπόμενα σφάλματα από τις προβλέψεις σε μια χρονική σειρά παρέχουν μια άλλη πηγή πληροφοριών που μπορούν να μοντελοποιηθούν. Τα υπολειπόμενα σφάλματα σχηματίζουν χρονοσειρές. Ένα μοντέλο αυτό-παλινδρόμησης αυτής της δομής μπορεί να χρησιμοποιηθεί για την πρόβλεψη του σφάλματος πρόβλεψης, το οποίο με τη σειρά του μπορεί να χρησιμοποιηθεί για τη διόρθωση προβλέψεων.

Η δομή στο υπολειπόμενο σφάλμα μπορεί να αποτελείται από τάση, προκατάληψη και εποχικότητα που μπορούν να μοντελοποιηθούν άμεσα. Κάποιος μπορεί να δημιουργήσει ένα μοντέλο των υπολειπόμενων χρονοσειρών σφάλματος και να προβλέψει το αναμενόμενο σφάλμα του μοντέλου. Το προβλεπόμενο σφάλμα μπορεί στη συνέχεια να αφαιρεθεί από την πρόβλεψη του μοντέλου και με τη σειρά του να προσφέρει μια επιπλέον αύξηση στην απόδοση.

Μια αυτόματη επέκταση του υπολειπόμενου σφάλματος είναι το Moving Average Model.

Ενσωματωμένος Κινούμενος Μέσος όρος Αυτο-παλινδρόμησης (ARIMA)

Ο ενσωματωμένος κινούμενος μέσος όρος αυτο-παλινδρόμησης ή το ARIMA είναι ένα πολύ σημαντικό μέρος των στατιστικών, της οικονομετρίας και ιδίως της ανάλυσης χρονοσειρών.

Το ARIMA είναι μια τεχνική πρόβλεψης που μας δίνει μελλοντικές τιμές βασισμένες εξ ολοκλήρου στην αδράνεια του.

Τα μοντέλα Autoregressive Integrated Moving Average (ARIMA) περιλαμβάνουν ένα σαφές στατιστικό μοντέλο για το ασύμμετρο συστατικό μιας χρονοσειράς που επιτρέπει μη μηδενικές αυτοσυσχέτισεις στο ακανόνιστο στοιχείο

Τα μοντέλα ARIMA ορίζονται για σταθερές χρονοσειρές. Επομένως, επιλέγοντας μια μη στατική χρονοσειρά, θα πρέπει πρώτα να «διαφοροποηθεί» μέχρι να φτάσει σε στατική χρονοσειρά.

Ένα μοντέλο ARIMA μπορεί να δημιουργηθεί χρησιμοποιώντας τη βιβλιοθήκη statsmodels ως εξής:

1. Ορισμός του μοντέλου χρησιμοποιώντας το ARIMA και περνώντας στις παραμέτρους p, d και q.
2. Το μοντέλο προετοιμάζεται στα δεδομένα εκπαίδευσης.
3. Οι προβλέψεις μπορούν να γίνουν χρησιμοποιώντας τη λειτουργία πρόβλεψης και καθορίζοντας τον δείκτη του χρόνου ή των ωρών που θα προβλεφθούν.

ACF και PACF

Δύναται να υπολογιστεί η συσχέτιση για παρατηρήσεις χρονοσειρών με παρατηρήσεις από προηγούμενα χρονικά βήματα, που ονομάζονται καθυστερήσεις Δεδομένου ότι η συσχέτιση των παρατηρήσεων χρονοσειρών υπολογίζεται με τιμές της ίδιας σειράς σε προηγούμενες εποχές, αυτό ονομάζεται σειριακή συσχέτιση ή αυτοσυσχέτιση.

Μια γραφική παράσταση της αυτοσυσχέτισης ενός συνόλου δεδομένων μιας χρονοσειράς κατά υστέρηση ονομάζεται AutoCorrelation Function ή το ακρωνύμιο ACF. Αυτή η γραφική παράσταση καλείται μερικές φορές ένας συσχετισμός ή μια αυτοσυσχέτιση.

Μια μερική αυτοσυσχέτιση ή PACF είναι μια σύνοψη της σχέσης μεταξύ μιας παρατήρησης σε μια χρονολογική σειρά με παρατηρήσεις σε προηγούμενα χρονικά βήματα με τις σχέσεις μεταξύ των παρατηρήσεων που έχουν αφαιρεθεί.

COVID-19

Στα τέλη του 2019, εντοπίστηκε μια μόλυνση άγνωστης προέλευσης με εκδηλώσεις αναπνευστικών ασθενειών καθώς άρχισε να εξαπλώνεται γρήγορα στην επαρχία Hubei της Κίνας, ιδίως στην μεγαλύτερη πόλη της, Wuhan. Λίγο αργότερα, ο Παγκόσμιος Οργανισμός Υγείας χαρακτήρισε την ασθένεια κοροναϊό 2019 (COVID-19). Μέχρι τις 30 Ιανουαρίου 2020, η ταχεία εξάπλωση αυτής της νέας ασθένειας σε όλο τον κόσμο έγινε εμφανής, με αποτέλεσμα ο ΠΟΥ να κηρύξει το COVID-19 ως

έκτακτη ανησυχία για τη δημόσια υγεία και μέχρι τον Μάρτιο, το ξέσπασμα χαρακτηρίστηκε «πανδημία».

Από τις αρχές Ιουνίου 2020, ο συνολικός αριθμός κρουσμάτων COVID-19 σε ολόκληρο τον κόσμο έφτασε τα 7,3 εκατομμύρια και οι συνολικοί θάνατοι έφτασαν τους 414.000. Μόνο στις ΗΠΑ, οι περιπτώσεις είχαν φτάσει σχεδόν τα 2 εκατομμύρια και ο αριθμός των θανάτων έφτασε τους 113.000. Αν και πολλές τοποθεσίες, ιδίως εκείνες οι χώρες και τα κράτη που επλήγησαν περισσότερο από την ασθένεια, έχουν δει δραματικές μειώσεις στον αριθμό των νέων περιπτώσεων και θανάτων, υπάρχουν πολλές ακόμη τοποθεσίες στις οποίες οι τάσεις στις λοιμώξεις και τους θανάτους είναι ασαφείς ή, χειρότερα, αυξάνονται.

Ιδιαίτερη ανησυχία για την επανεμφάνιση μολύνσεων COVID-19 αρχιζει να εμφανίζεται, ενόψει της χαλάρωσης των κανόνων από τις τοπικές κυβερνήσεις. Κανόνων που αποσκοπούν στην πρόληψη της εξάπλωσης της νόσου όπως η καραντίνα, το κλείσιμο των επιχειρήσεων και η κοινωνική απομάκρυνση.

Το πρόβλημα είναι ότι «οι τάσεις της καθημερινής επίπτωσης και των θανάτων του COVID-19 εξακολουθούν να είναι ελάχιστα κατανοητές». Ειδικότερα, οι πηγές δεδομένων παρέχουν ελάχιστη στατιστική ανάλυση των τάσεων. Καθώς οι χώρες αρχίζουν να ανοίγουν ξανά μετά τους περιορισμούς του COVID-19, είναι κρίσιμο για τις κυβερνήσεις, τις επιχειρήσεις και το ευρύ κοινό να βρεθούν μέσα για την αποκωδικοποίηση των αλλαγών σε κρούσματα και θανάτους για τον προσδιορισμό των σχετικών τάσεων στα δεδομένα.

Εξίσου σημαντικό είναι ότι η κριτική ανάλυση των δεδομένων είναι απαραίτητη για να καθοριστεί εάν οι περιορισμοί πρέπει να εφαρμοστούν εκ νέου υπό το φως μιας ουσιαστικής τάσης στην επανεμφάνιση. Ως εκ τούτου, οι ερευνητές ζητούν στατιστική ανάλυση των τάσεων της νόσου βάσει δεδομένων, ώστε να παρέχουν περισσότερο πλαίσιο και εγκυρότητα για σημαντικές πολιτικές αποφάσεις, όπως το άνοιγμα ή ο περιορισμός των οικονομικών δραστηριοτήτων, η παραγγελία και η διοργάνωση ιατρικού εξοπλισμού και οι περιορισμοί ταξιδιού. Η ταχεία ανταπόκριση στις τάσεις των δεδομένων έχει πιστωθεί για τη μείωση της εξάπλωσης του COVID-19 σε χώρες όπως η Κίνα, η Νότια Κορέα και η Σιγκαπούρη, υποδεικνύοντας περαιτέρω τη σημασία της ποιότητας των δεδομένων και την ανάλυση αυτών των δεδομένων.

Αναδυόμενος, λοιπόν, από το Γουχάν της Κίνας, στα τέλη του 2019, ο COVID-19 είναι η τρίτη επιδημία κορωναϊού που εμφανίστηκε τις τελευταίες δύο δεκαετίες. Υποτίθεται ότι ο ιός προήλθε από ζωονοσογόνο μετάδοση από ένα ζώο που πωλείται σε μια «υγρή» αγορά στο Wuhan. Αυτό το μεταφορικό ζώο πιθανότατα πήρε τον ιό από το guano νυχτερίδας κατά την αναζήτηση τροφής κοντά σε τοπικές σπηλιές.

Έγινε γρήγορα εμφανές ότι ο Γουχάν αντιμετώπιζε μια παράξενη, νέα πνευμονική ασθένεια, καθώς οι περιπτώσεις άρχισαν να εμφανίζονται στα τοπικά νοσοκομεία. Τα συμπτώματα περιλάμβαναν πυρετό, αδιαθεσία και βήχα. Προσδιορίστηκε αμέσως μετά την ανακάλυψη ότι τα κύρια μέσα μετάδοσης COVID-19 ήταν μέσω της αποβολής και της εισπνοής αναπνευστικών σταγονιδίων. Ιδιαίτερη ανησυχία ήταν η φαινομενικά γρήγορη μετάδοση μεταξύ του τοπικού πληθυσμού.

Οι προκαταρκτικές εκτιμήσεις των αριθμών αναπαραγωγής ιών κυμάνθηκαν από 1,4 έως 3,8 σε σύγκριση με το H1N1 (1,25) και το σοβαρό οξύ αναπνευστικό σύνδρομο (2.2-3.6). Οι εκτιμήσεις των ποσοστών θνησιμότητας μεταξύ των ασθενών με COVID-19 ποικίλλουν ανάλογα με την τοποθεσία του θύματος, που κυμαίνεται από 15,3% στη Γαλλία έως 1,3% στη Ρωσία. Ο δείκτης θνησιμότητας στις ΗΠΑ από τον Ιούνιο του 2020 ήταν 5,7%.Μέρος των τοπικών διαφορών στη θνησιμότητα πιθανότατα συνδέεται με μεγάλο αριθμό ατόμων που έχουν δημογραφικά χαρακτηριστικά που σχετίζονται με υψηλότερη συχνότητα θανάτων όπως είναι η ηλικία άνω των 80 ετών, η ύπαρξη συννοσηρότητας (όπως ο διαβήτης) και εκείνων με μειωμένη ανοσία.

Ερευνητές και κυβερνήσεις από όλο τον κόσμο άρχισαν αμέσως να παρακολουθούν τις τάσεις στα δεδομένα COVID-19 για να εξακριβώσουν την ταχύτητα και το εύρος της εξάπλωσης του ιού. Περιπλέκοντας αυτές τις προσπάθειες, οι τάσεις στις περιπτώσεις COVID-19 και οι θάνατοι ήταν σε συνεχή ροή από τότε που η ασθένεια εντοπίστηκε και παρακολουθήθηκε ενεργά. Ενώ η δραστηριότητα της νόσου αυξάνεται γρήγορα σε μία τοποθεσία, μπορεί να αυξηθεί αργά μόνο σε άλλες, ενώ σε ορισμένες περιοχές, η ελάχιστη εξάπλωση.

Επιπλέον, καθώς μια περιοχή ή χώρα μπορεί να παρουσιάσει αύξηση της συχνότητας εμφάνισης, άλλες ταυτόχρονα φαίνεται να έχουν ξεπεράσει τον μέγιστο αριθμό περιπτώσεων και θανάτων, ενώ οι λοιμώξεις φαίνεται να μειώνονται. Για παράδειγμα, επιβεβαιωμένα κρούσματα στην Κίνα έφθασαν στα μέσα Φεβρουαρίου

2020, οι περιπτώσεις στην Ιταλία δεν έφτασαν τις μέγιστες τιμές μέχρι περίπου τις 21 Μαρτίου, οι περιπτώσεις (από τις αρχές Ιουνίου 2020) στη Βραζιλία εξακολουθούν να αυξάνονται και στις ΗΠΑ, οι περιπτώσεις έφτασαν σε υψηλό σημείο προς τα τέλη Απριλίου 2020. Ωστόσο, φαίνεται να υπάρχει πιθανότητα επανεμφάνισης σε μεμονωμένες καταστάσεις. Γενικά, οι παγκόσμιες περιπτώσεις και οι θάνατοι συνέχισαν να αυξάνονται με ανησυχητικό ρυθμό.

Λόγω της ταχείας εξάπλωσης και της θανατηφόρας φύσης του COVID-19, οι χώρες έχουν εφαρμόσει μια ποικιλία επιπέδων προληπτικών μηχανισμών και καθοδήγησης σε διάφορες ταχύτητες ή στάδια, για να περιορίσουν την εξάπλωση του ιού. Το γρήγορο κλείδωμα της Κίνας πιστώθηκε με τον περιορισμό της ιογενούς εξάπλωσης σε μεγαλύτερο βαθμό από ό, τι είχε παρατηρηθεί.

Επιπλέον, τα μέτρα κοινωνικής απόστασης και καραντίνας φαίνεται ότι είχαν σημαντικές θετικές επιπτώσεις στη μετάδοση στην Ιταλία. Επιπλέον, παρόμοιες πολιτικές στις Η.Π.Α. έδειξαν αρχικά «ισοπέδωση της καμπύλης» της μόλυνσης στις περισσότερες πολιτείες. Τέτοιες προσπάθειες πιστώνεται επίσης ότι τελικά διακόπτουν την ταχεία επιδημία στην περιοχή της Νέας Υόρκης.

Παρόλο που υπήρξε σημαντική πρόβλεψη και έρευνα, υπάρχουν λίγα διαθέσιμα δεδομένα για την ενεργό παρακολούθηση των τρεχουσών τάσεων δεδομένων και τόσο τη στατιστική όσο και την πρακτική τους σημασία. Προς το παρόν, οι βραχυπρόθεσμες τάσεις περιγράφονται συνήθως από αλλαγές στους αριθμούς και τα ποσοστά σε περιπτώσεις που, όταν χρησιμοποιούνται μεμονωμένα άλλες μετρήσεις, μπορεί να είναι παραπλανητικές.

Συγκεκριμένα, φαίνεται να υπάρχει ισχυρή σχέση μεταξύ του αριθμού των περιπτώσεων που αναφέρθηκαν και της ημέρας της εβδομάδας, πιθανότατα λόγω του τρόπου συλλογής και διάδοσης των δοκιμών και άλλων δεδομένων. Έτσι, ένα άλμα 1 ή 2 ημερών σε περιπτώσεις πρέπει να ληφθεί στο πλαίσιο αυτών των προτύπων. Σε απάντηση στις ανησυχίες σχετικά με τη χρήση αυτών των απλών μέτρων, τα μέσα ενημέρωσης και οι ερευνητές επέλεξαν τη χρήση διαφόρων μέτρησης κινούμενου μέσου όρου, όπως οι απλοί κινούμενοι / κυλιόμενοι μέσοι όροι, οι οποίοι περιλαμβάνουν περιόδους 3, 5 ή 7 ημερών. Παρόλο που αυτά τα βελτιωμένα μέσα οπτικοποίησης των τάσεων δεδομένων είναι χρήσιμα, δεν παρέχουν το μέγεθος και τη σημασία των παραλλαγών. Η ποιότητα των δεδομένων τάσεων έχει αναγνωριστεί ως

κρίσιμη για τη λήψη αποφάσεων τόσο για τους περιορισμούς όσο και για την άρση αυτών, με την επακόλουθη παρακολούθηση τέτοιων τάσεων να θεωρείται κρίσιμη για την αποφυγή της ταχείας επαναμόλυνσης του κοινού.

Δυστυχώς, τα μέσα ενημέρωσης και ακόμη και ορισμένες κυβερνητικές υπηρεσίες, αναφέρουν τάσεις που βασίζονται σε καθημερινές αλλαγές σε περιπτώσεις και θανάτους, οι οποίες δεν είναι μόνο παραπλανητικές, αλλά δυνητικά λανθασμένα, ονομάζοντας προσωρινές διακυμάνσεις ως νέες τάσεις. Αυτό δημιουργεί σύγχυση μεταξύ του κοινού και άλλων ενδιαφερομένων, οδηγώντας σε αλλαγές στη συμπεριφορά που μπορεί να θέσουν σε κίνδυνο την υγεία τους (π.χ. διακοπή της χρήσης μάσκας, έξοδο πιο συχνά και κοινωνικοποίηση σε μεγάλες ομάδες).

Σημασία των κοινωνικών δικτύων

Κάθε χώρα λαμβάνει προληπτικά μέτρα για να πολεμήσει κατά της πανδημίας COVID-19. Η κοινωνική αποστασιοποίηση ή η παραμονή στο σπίτι έγινε η πιο ευρέως χρησιμοποιούμενη οδηγία στον κόσμο. Η κοινωνική απομάκρυνση αναγκάζει τους ανθρώπους να μείνουν στο σπίτι και ως αποτέλεσμα, επηρεάζει τη δημόσια εκδήλωση, τις επιχειρήσεις, την εκπαίδευση, και σχεδόν κάθε άλλη δραστηριότητα που σχετίζεται με την ανθρώπινη ζωή. Οι άνθρωποι χάνουν επίσης τις δουλειές τους και μολύνονται από COVID-19 και έτσι, το άγχος αυξάνεται στο προσωπικό και στο επίπεδο της κοινότητας. Μελέτες οικονομικής συμπεριφοράς δείχνουν ότι τα συναισθήματα (χαρά, θυμός, ανησυχία, αηδία, φόβος κ.λπ.) μπορούν βαθιά επηρεάζουν την ατομική συμπεριφορά και τη λήψη αποφάσεων.

Τα κοινωνικά δίκτυα έχουν την κρυφή δυνατότητα να αποκαλύψουν πολύτιμες γνώσεις για τα ανθρώπινα συναισθήματα σε προσωπικό και κοινοτικό επίπεδο. Η παρακολούθηση των tweets θα μπορούσε να είναι πολύτιμη ιδιαίτερα κατά τη διάρκεια και μετά την πανδημία COVID-19 καθώς η κατάσταση και η αντίδραση των ανθρώπων αλλάζουν και οι δύο κάθε στιγμή κατά τη διάρκεια αυτής της απρόβλεπτης περιόδου. Έτσι, η ανάλυση των δεδομένων του twitter μπορεί να διαδραματίσει καθοριστικό ρόλο για την κατανόηση της συμπεριφοράς και της απόκρισης των ανθρώπων κατά τη διάρκεια της πανδημίας COVID-19.

Πρόσφατες έρευνες έδειξαν ότι τα δεδομένα του twitter και η ανάλυση ανθρώπινων συναισθημάτων μπορεί να είναι χρήσιμα για την πρόβλεψη εγκλημάτων, του χρηματιστηρίου, εκλογών, καταστροφών, διαχείριση και άλλα. Επομένως, είναι πρωταρχικής σημασίας η ανάλυση των δεδομένων κοινωνικών μέσων για την κατανόηση της ανθρώπινης συμπεριφοράς και αντίδραση στη συνεχιζόμενη πανδημία.

Το Twitter έχει αποδειχθεί χρήσιμο για διάφορες εργασίες όπως δίκτυο επικοινωνίας έκτακτης ανάγκης, παρακολούθηση δημόσιων συναισθημάτων, εντοπίστε ανωμαλίες και να δωθεί έγκαιρη προειδοποίηση, κ.λπ. Το Twitter χρησιμοποιείται ως πηγή δεδομένων για την παρακολούθηση της δημόσιας αντίδρασης και υγεία κατά τη διάρκεια καταστροφών (π.χ. τυφώνες, πλημμύρες, σεισμοί, τρομοκρατικοί βομβαρδισμοί, που σχετίζονται με τη δημόσια υγεία διάδοση παραπληροφόρησης και άλλοι), και εκδηλώσεις ασθενειών.

Οι ερευνητές προσπαθούν να ενσωματώσουν διάφορες ιδέες που περιλαμβάνουν τη χρήση του δεδομένα twitter με διάφορους τρόπους. Οι Catherine et al. χρησιμοποίησαν δεδομένα από το twitter για εξερεύνηση και απεικόνιση πέντε διαφορετικών μεθόδων ανάλυσης με θέμα βασικοί όροι και χαρακτηριστικά, διάδοση πληροφοριών και διάδοση και συμπεριφορά δικτύου κατά τη διάρκεια του COVID-19. Οι συγγραφείς χρησιμοποιούν την αντιστοίχιση προτύπων και τη μοντελοποίηση θεμάτων χρησιμοποιώντας το Latent Dirichlet Allocation (LDA) για να επιλέξουν είκοσι διαφορετικά θέματα σχετικά με τη διάδοση περιπτώσεων COVID-19, εργαζομένων στον τομέα της υγειονομικής περίθαλψης και ατομικού προστατευτικού εξοπλισμού (ΜΑΠ).

Χρησιμοποιώντας διάφορες αναλύσεις, οι συγγραφείς μπόρεσαν να ανιχνεύσουν τάσεις θεμάτων υψηλού επιπέδου, σχέση με τις ξαφνικές ειδήσεις ή με τις αιχμές του θέματος και τελικά οι συγγραφείς προσπάθησαν να καταλάβουν πώς τα θέματα εξελίσσονται με το χρόνο.

Πολλοί ερευνητές από διαφορετικές χώρες προσπαθούν να κάνουν συλλογή και κοινή χρήση συνόλων δεδομένων twitter στο COVID-19. Αυτά τα κοινόχρηστα σύνολα δεδομένων περιέχουν tweets με μια ποικιλία λέξεων-κλειδιών από διαφορετικές χώρες. Ενώ κάποια αποθετήρια δεδομένων μοιράζονται μόνο τα αναγνωριστικά tweet, άλλα μοιράζονται επίσης κάποια μορφή προεπεξεργασμένων δεδομένων.

ΚΕΦΑΛΑΙΟ 3

Μια πληθώρα από τεχνολογίες χρησιμοποιήθηκαν για την συγγραφή του κώδικα. Οι τεχνολογίες επιλέχθηκαν με βάση την ευελιξία, τη χρησιμότητα και το επίπεδο χρήσης τους από την επιστημονική κοινότητα.

Jupyter Lab

Το JupyterLab είναι ένα διαδραστικό περιβάλλον ανάπτυξης διαδικτύου για υπολογιστές, κώδικα και δεδομένα. Είναι ευέλικτο επιτρέποντας την διαμόρφωση και τακτοποίηση του περιβάλλοντος εργασίας χρήστη με αποτέλεσμα να υποστηρίζει ένα ευρύ φάσμα ροών εργασίας στην επιστήμη δεδομένων, την επιστημονική πληροφορική και τη μηχανική μάθηση. Επιπλέον, το JupyterLab είναι επεκτάσιμο και αρθρωτό υποστηρίζοντας plugin που προσθέτουν νέα στοιχεία και ενσωματώνονται με υπάρχοντα.

Το JupyterLab είναι το νεότερο περιβάλλον εργασίας χρήστη για το Project Jupyter. Προσφέρει τα δομικά στοιχεία του κλασικού Jupyter Notebook (σημειωματάριο, τερματικό, πρόγραμμα επεξεργασίας κειμένου, πρόγραμμα περιήγησης αρχείων κ.λπ.) σε μια ευέλικτη διεπαφή χρήστη.

Το όνομα του Project Jupyter είναι μια αναφορά στις τρεις βασικές γλώσσες προγραμματισμού που υποστηρίζονται από τον Jupyter, οι οποίες είναι οι Julia, Python και R, και επίσης ένα αφιέρωμα στα σημειωματάρια του Galileo που καταγράφουν την ανακάλυψη των φεγγαριών του Δία. Το Project Jupyter ανέπτυξε και υποστήριξε τα διαδραστικά υπολογιστικά προϊόντα Jupyter Notebook, JupyterHub και JupyterLab.

Σαν γλώσσα προγραμματισμού κατά την διάρκεια της πτυχιακής χρησιμοποιήθηκε η γλώσσα Python.

Python

Η γλώσσα Python ως εργαλείο για ανάλυση δεδομένων

Υπάρχει μια σειρά από εξέχουσες γλώσσες προγραμματισμού για ανάλυση δεδομένων. Οι C, C ++, R, Java, Javascript και Python είναι μερικές από αυτές. Κάθε μία απο αυτές προσφέρει μοναδικές δυνατότητες, επιλογές και εργαλεία που ταιριάζουν στις διαφορετικές απαιτήσεις ανάλογα με τις ανάγκες. Μερικά είναι καλύτερα από άλλα για συγκεκριμένες ανάγκες.

Υπάρχουν δύο κύριοι παράγοντες που καθιστούν την Python μια ευρέως χρησιμοποιούμενη γλώσσα προγραμματισμού στον επιστημονικό τομέα, το εκπληκτικό οικοσύστημα και ένας μεγάλος αριθμός πακέτων χαρακτηριστικών προσανατολισμένων στα δεδομένα που μπορούν να επιταχύνουν και να απλοποιήσουν την επεξεργασία δεδομένων, διευκολύνοντας έτσι την εξοικονόμηση χρόνου.

Εκτός από αυτό, η Python χρησιμοποιείται αρχικά για την πραγματοποίηση της ανάλυσης δεδομένων. Είναι μεταξύ αυτών των γλωσσών που αναπτύσσονται σε συνεχή βάση. Με αυτόν τον τρόπο, η Python ονομάζεται η κορυφαία γλώσσα με υψηλές δυνατότητες στον τομέα της επιστήμης δεδομένων περισσότερο από άλλες γλώσσες προγραμματισμού.

Η Python είναι μια διαλειτουργική, μέγιστα ερμηνευμένη γλώσσα που έχει πολλά πλεονεκτήματα. Η αντικειμενοστραφής γλώσσα προγραμματισμού χρησιμοποιείται συνήθως για τον εξορθολογισμό μεγάλων σύνθετων συνόλων δεδομένων. Επιπλέον, έχοντας μια δυναμική σημασιολογία συν αμετρημένες ικανότητες για ταχεία ανάπτυξη εφαρμογών, η Python χρησιμοποιείται σε μεγάλο βαθμό στο script.

Ένα άλλο πλεονέκτημα της Python είναι η υψηλή αναγνωσιμότητα που βοηθά τους μηχανικούς να εξοικονομήσουν χρόνο πληκτρολογώντας λιγότερες γραμμές κώδικα για την ολοκλήρωση των εργασιών. Όντας γρήγορη, η Python είναι η καλύτερη επιλογή για την ανάλυση δεδομένων. Αυτό οφείλεται στη μεγάλη υποστήριξη καθώς και τη διαθεσιμότητα ενός πλήθους βιβλιοθηκών ανοιχτού κώδικα για διαφορετικούς σκοπούς.

Επομένως, δεν προκαλεί έκπληξη καθόλου ότι υποστηρίζεται σαν η προτιμώμενη γλώσσα προγραμματισμού για την επιστήμη των δεδομένων. Υπάρχει ένα εύρος μοναδικών δυνατοτήτων που καθιστούν την Python νούμερο ένα επιλογή για ανάλυση δεδομένων.

Ένα άλλο δυνατό χαρακτηριστικό της γλώσσας είναι η υπερβολική ευελιξία που κάνει την Python να ζητείται ιδιαίτερα μεταξύ επιστημόνων και αναλυτών δεδομένων. Λόγω αυτού, είναι δυνατή η κατασκευή μοντέλων δεδομένων, η συστηματοποίηση συνόλων δεδομένων, η δημιουργία αλγορίθμων που υποστηρίζονται από ML, οι υπηρεσίες ιστού και η εφαρμογές εξόρυξης δεδομένων για την ολοκλήρωση διαφορετικών εργασιών σε σύντομο χρονικό διάστημα. Ένα τέτοιο πλεονέκτημα καθιστά την Python μια ιδανική λύση που χρειάζεται η βιομηχανία της επιστήμης δεδομένων.

Συγκρίνοντας με άλλες γλώσσες όπως R, Go και Rust, το Python είναι πολύ πιο γρήγορη και επεκτάσιμη. Επομένως, η Python είναι καλή για διαφορετικές χρήσεις σε διάφορους τομείς που μπορούν να λύσουν ένα ευρύ φάσμα προβλημάτων. Γι 'αυτό πολλές εταιρείες έχουν μεταναστεύσει στην Python. Επιπλέον, αυτή η γλώσσα είναι ιδανική για όλα τα είδη RAD.

Είναι ένα πολύ γνωστό γεγονός ότι οι οπτικές πληροφορίες είναι πολύ πιο εύκολα κατανοητές, λειτουργικές και εύκολο να ανακληθούν απο τη μνήμη. Υπάρχει ένα πακέτο διαφορετικών επιλογών οπτικοποίησης που καθιστά την Python ένα απαραίτητο εργαλείο όχι μόνο για την ανάλυση δεδομένων αλλά και για όλη την επιστήμη των δεδομένων. Δίνεται η δυνατότητα να γίνουν τα δεδομένα πιο προσβάσιμα και ευκολότερα στη χρήση μέσω της δημιουργίας διαφόρων γραφημάτων και γραφικών, καθώς και διαδραστικών διαδρομών έτοιμων για τον ιστό.

Τέλος, η Python είναι μια από τις πιο εύκολες στη μάθηση γλώσσες, πολύ απλή στη χρήση, και με ένα εξαιρετικό πακέτο χαρακτηριστικών. Αν και η Python είναι μια γλώσσα ανοιχτού κώδικα, παραμένει καλά υποστηριζόμενη από μια τεράστια κοινότητα. Όλα αυτά κάνουν την Python ιδανική για αρχάριους στον προγραμματισμό. Επιπλέον, η Python είναι επεκτάσιμη και αρκετά ευέλικτη ώστε να εφαρμόζεται σε διαφορετικά πεδία και για διάφορους σκοπούς. Χάρη στο πακέτο επιλογών γραφικών μαζί με εργαλεία οπτικοποίησης που κάνουν τα δεδομένα πιο

προσβάσιμα, η Python ονομάζεται ως η πιο προτιμώμενη γλώσσα μεταξύ των αναλυτών δεδομένων και των επιστημόνων δεδομένων.

Matplotlib

Η Matplotlib είναι μια βιβλιοθήκη σχεδίασης για τη γλώσσα προγραμματισμού Python και την αριθμητική επέκταση μαθηματικών NumPy. Παρέχει ένα αντικειμενοστραφές API για την ενσωμάτωση γραφημάτων σε εφαρμογές χρησιμοποιώντας γενικά εργαλεία GUI όπως Tkinter, wxPython, Qt ή GTK. Υπάρχει επίσης μια διαδικαστική διεπαφή "pylab" (όπως το OpenGL), σχεδιασμένο να μοιάζει πολύ με αυτό του MATLAB, αν και η χρήση του αποθαρρύνεται. Η SciPy χρησιμοποιεί το Matplotlib.

Το Matplotlib γράφτηκε αρχικά από τον John D. Hunter. Έκτοτε, έχει μια ενεργή κοινότητα ανάπτυξης και διανέμεται με άδεια τύπου BSD.

Το Matplotlib 2.0.x υποστηρίζει τις εκδόσεις Python 2.7 έως 3.6. Η υποστήριξη Python 3 ξεκίνησε με το Matplotlib 1.2. Το Matplotlib 1.4 είναι η τελευταία έκδοση που υποστηρίζει το Python 2.6. Ο Matplotlib δεσμεύτηκε να μην υποστηρίξει την Python 2 μετά το 2020 υπογράφοντας τη δήλωση Python 3.

Το Matplotlib αποστέλλεται με πολλές πρόσθετες εργαλειοθήκες, συμπεριλαμβανομένης της τρισδιάστατης σχεδίασης με το mplot3d, των βοηθών αξόνων στο axes_grid1 και των βοηθών άξονα στο axisartist.

Ένας μεγάλος αριθμός πακέτων τρίτων επεκτείνεται και βασίζεται στη λειτουργικότητα του Matplotlib, συμπεριλαμβανομένων αρκετών διεπαφών σχεδίασης υψηλότερου επιπέδου (seaborn, HoloViews, ggplot, ...) και μια εργαλειοθήκη προβολής και χαρτογράφησης (Cartopy).

Pandas

Το pandas είναι μια βιβλιοθήκη λογισμικού γραμμένη για τη γλώσσα προγραμματισμού Python για χειρισμό και ανάλυση δεδομένων. Συγκεκριμένα, προσφέρει δομές δεδομένων και λειτουργίες για χειρισμό αριθμητικών πινάκων και χρονοσειρών. Είναι ελεύθερο λογισμικό που κυκλοφορεί με την άδεια BSD. Το

όνομα προέρχεται από τον όρο "panel data", έναν όρο οικονομετρίας για σύνολα δεδομένων που περιλαμβάνει παρατηρήσεις σε πολλαπλές χρονικές περιόδους για τα ίδια άτομα. Το Pandas χρησιμοποιείται κυρίως για την ανάλυση δεδομένων. Το Pandas επιτρέπει την εισαγωγή δεδομένων από διάφορες μορφές αρχείων, όπως τιμές διαχωρισμένες με κόμμα, JSON, SQL, Microsoft Excel. Το Pandas επιτρέπει διάφορες λειτουργίες χειραγώγησης δεδομένων, όπως συγχώνευση, αναδιαμόρφωση, επιλογή, καθώς και καθαρισμό δεδομένων και λειτουργίες ανταλλαγής δεδομένων.

Χαρακτηριστικά βιβλιοθήκης

- Αντικείμενο DataFrame για χειρισμό δεδομένων με ενσωματωμένη ευρετηρίαση.
- Εργαλεία για ανάγνωση και εγγραφή δεδομένων μεταξύ δομών δεδομένων στη μνήμη και διαφορετικών μορφών αρχείων.
- Ευθυγράμμιση δεδομένων και ολοκληρωμένος χειρισμός δεδομένων που λείπουν.
- Αναδιαμόρφωση και περιστροφή συνόλων δεδομένων.
- Τεμαχισμός βάσει ετικετών, φανταχτερή ευρετηρίαση και υποσύνολο μεγάλων συνόλων δεδομένων.
- Εισαγωγή και διαγραφή στήλης δομής δεδομένων.
- Ομαδοποίηση ανά συνθήκη επιτρέποντας λειτουργίες split-apply-combine σε σύνολα δεδομένων.
- Συγχώνευση και ένωση δεδομένων.
- Ιεραρχικός άξονας ευρετηρίασης για εργασία με δεδομένα υψηλής διάστασης σε δομή δεδομένων χαμηλότερης διάστασης.
- Λειτουργικότητα χρονοσειρών: Δημιουργία εύρους ημερομηνιών [6] και μετατροπή συχνότητας, στατιστικά κινούμενου παραθύρου, γραμμικές παλινδρομήσεις κινούμενου παραθύρου, αλλαγή ημερομηνίας και καθυστέρηση.
- Παρέχει φιλτράρισμα δεδομένων.

Η βιβλιοθήκη είναι εξαιρετικά βελτιστοποιημένη για απόδοση, με κρίσιμες διαδρομές κώδικα γραμμένες σε Cython ή C

Υπάρχουν πολλά οφέλη από τη βιβλιοθήκη Python Pandas, Συγκεκριμένα, αυτά είναι τα βασικά πλεονεκτήματα της χρήσης της βιβλιοθήκης Pandas:

1. Αναπαράσταση δεδομένων

Το Pandas παρέχει εξαιρετικά βελτιωμένες μορφές αναπαράστασης δεδομένων. Αυτό βοηθά στην καλύτερη ανάλυση και κατανόηση των δεδομένων. Η απλούστερη αναπαράσταση δεδομένων διευκολύνει καλύτερα αποτελέσματα για έργα επιστήμης δεδομένων.

2. Λιγότερο γράψιμο και περισσότερη δουλειά

Είναι ένα από τα καλύτερα πλεονεκτήματα του Pandas. Αυτό που θα είχε πάρει πολλές γραμμές στο Python χωρίς καμία βιβλιοθήκη υποστήριξης, μπορεί απλά να επιτευχθεί μέσω 1-2 γραμμών με τη χρήση του Pandas. Έτσι, η χρήση του Pandas βοηθά στη συντόμευση της διαδικασίας χειρισμού δεδομένων. Με το χρόνο που εξοικονομείται, μπορούμε να επικεντρωθούμε περισσότερο στους αλγόριθμους ανάλυσης δεδομένων.

3. Ένα εκτεταμένο σύνολο χαρακτηριστικών

Το Pandas παρέχει ένα τεράστιο σύνολο σημαντικών εντολών και λειτουργιών που χρησιμοποιούνται για την εύκολη ανάλυση των δεδομένων σας. Μπορεί να χρησιμοποιηθεί για να εκτελεστούν διάφορες εργασίες, όπως φιλτράρισμα των δεδομένων σύμφωνα με συγκεκριμένες συνθήκες ή τμηματοποίηση και διαχωρισμός των δεδομένων σύμφωνα με τις προτιμήσεις κ.λπ.

4. Διαχειρίζεται αποτελεσματικά μεγάλα δεδομένα

Ο Wes McKinney, ο δημιουργός του Pandas, έκανε τη βιβλιοθήκη python να χειρίζεται κυρίως μεγάλα σύνολα δεδομένων αποτελεσματικά. Το Pandas βοηθά να εξοικονομηθεί πολύς χρόνος εισάγοντας μεγάλες ποσότητες δεδομένων πολύ γρήγορα.

5. Κάνει τα δεδομένα ευέλικτα και προσαρμόσιμα

Το Pandas παρέχει ένα τεράστιο σύνολο χαρακτηριστικών για εφαρμογή στα δεδομένα, ώστε να είναι δυνατή η προσαρμογή, η επεξεργασία και η περιστροφή. Αυτό βοηθά να αξιοποιούνται στο έπακρο τα δεδομένα.

6. Φτιαγμένο για Python

Η Python έχει γίνει μια από τις πιο περιζήτητες γλώσσες προγραμματισμού στον κόσμο, με τον εκτεταμένο όγκο χαρακτηριστικών και το τεράστιο ποσό παραγωγικότητας που παρέχει. Επομένως, η δυνατότητα κωδικοποίησης του Pandas στην Python, δίνει τη δυνατότητα να αξιοποιηθεί η δύναμη των διαφόρων λειτουργιών και βιβλιοθηκών προσφέρει μόνο η Python. Μερικές από αυτές τις βιβλιοθήκες είναι NumPy, SciPy, MatPlotLib κ.λπ.

Numpy

Το NumPy είναι το θεμελιώδες πακέτο για την επιστημονική πληροφορική στην Python. Πρόκειται για μια βιβλιοθήκη Python που παρέχει ένα πολυδιάστατο αντικείμενο συστοιχίας, διάφορα παράγωγα αντικείμενα (όπως συγκαλυμμένες συστοιχίες και πίνακες) και μια σειρά από ρουτίνες για γρήγορες λειτουργίες σε συστοιχίες, συμπεριλαμβανομένων μαθηματικών, λογικών, χειρισμού σχήματος, ταξινόμησης, επιλογής, I / O , διακριτοί μετασχηματισμοί Fourier, βασική γραμμική άλγεβρα, βασικές στατιστικές λειτουργίες, τυχαία προσομοίωση και πολλά άλλα.

Στον πυρήνα του πακέτου NumPy, βρίσκεται το αντικείμενο ndarray. Αυτό περιλαμβάνει ενδιάμεσες συστοιχίες ομοιογενών τύπων δεδομένων, με πολλές λειτουργίες να εκτελούνται σε μεταγλωττισμένο κώδικα για απόδοση. Υπάρχουν πολλές σημαντικές διαφορές μεταξύ των συστοιχιών NumPy και των τυπικών ακολουθιών Python:

Οι πίνακες NumPy έχουν σταθερό μέγεθος κατά τη δημιουργία, σε αντίθεση με τις λίστες Python (οι οποίες μπορούν να αναπτυχθούν δυναμικά). Η αλλαγή του μεγέθους ενός ndarray θα δημιουργήσει έναν νέο πίνακα και θα διαγράψει το πρωτότυπο.

Τα στοιχεία σε μια συστοιχία NumPy πρέπει όλα να είναι του ίδιου τύπου δεδομένων και έτσι θα έχουν το ίδιο μέγεθος στη μνήμη. Η εξαίρεση: υπαρχει δυνατότητα ύπαρξης πινάκων αντικειμένων (Python, συμπεριλαμβανομένων των NumPy), επιτρέποντας έτσι συστοιχίες στοιχείων διαφόρων μεγεθών.

Οι συστοιχίες NumPy διευκολύνουν προηγμένους μαθηματικούς τύπους και άλλους τύπους λειτουργιών σε μεγάλο αριθμό δεδομένων. Συνήθως, τέτοιες λειτουργίες εκτελούνται πιο αποτελεσματικά και με λιγότερο κώδικα από ό, τι είναι δυνατό χρησιμοποιώντας τις ενσωματωμένες ακολουθίες της Python.

Μια αυξανόμενη πληθώρα επιστημονικών και μαθηματικών πακέτων με βάση το Python χρησιμοποιούν πίνακες NumPy. Αν και αυτά συνήθως υποστηρίζουν είσοδο ακολουθίας Python, μετατρέπουν αυτήν την είσοδο σε πίνακες NumPy πριν από την επεξεργασία και συχνά εξάγουν πίνακες NumPy. Με άλλα λόγια, προκειμένου να χρησιμοποιηθεί αποτελεσματικά μεγάλο μέρος (ίσως ακόμη και το μεγαλύτερο μέρος) του σημερινού επιστημονικού / μαθηματικού λογισμικού που βασίζεται στην Python, η γνώση του πως χρησιμοποιούνται οι ενσωματωμένοι τύποι ακολουθιών της Python είναι ανεπαρκές και πρέπει επίσης να υπάρχει γνώση για των συστοιχιών NumPy.

Statsmodels

Το statsmodels είναι ένα πακέτο της Python που παρέχει classes και λειτουργίες για την εκτίμηση πολλών διαφορετικών στατιστικών μοντέλων, καθώς και για τη διεξαγωγή στατιστικών δοκιμών και την εξερεύνηση στατιστικών δεδομένων. Μια εκτεταμένη λίστα στατιστικών αποτελεσμάτων είναι διαθέσιμη για κάθε εκτιμητή. Τα αποτελέσματα ελέγχονται με βάση τα υπάρχοντα στατιστικά πακέτα για να διασφαλιστεί ότι είναι σωστά. Το πακέτο κυκλοφορεί υπό την άδεια Open Source Modified BSD (3-clause).

Το Statsmodels είναι μέρος της επιστημονικής στοίβας της Python που προσανατολίζεται στην ανάλυση δεδομένων, στην επιστήμη δεδομένων και στη στατιστική. Το Statsmodels είναι χτισμένο πάνω από τις αριθμητικές βιβλιοθήκες NumPy και SciPy, ενσωματώνεται με Pandas για διαχείριση δεδομένων και χρησιμοποιεί το Patsy για μια διεπαφή τύπου R. Οι γραφικές συναρτήσεις βασίζονται

στη βιβλιοθήκη Matplotlib. Το Statsmodels παρέχει το στατιστικό υπόβαθρο για άλλες βιβλιοθήκες Python.

Πηγές δεδομένων

Meteostat

Η Meteostat είναι μια ανοιχτή πλατφόρμα που παρέχει δωρεάν πρόσβαση σε στατιστικές για τον καιρό και το κλίμα για μη εμπορικούς σκοπούς. Το Meteostat παρέχει πολλαπλές διεπαφές για την ανάκτηση δεδομένων καιρού και κλίματος. Επιλέγεται οποιαδήποτε από τις ακόλουθες επιλογές:

- ***Βιβλιοθήκη Python:*** Ανάλυση ιστορικών δεδομένων καιρού για χιλιάδες μετεωρολογικούς σταθμούς
- ***Μαζικά δεδομένα:*** Λήψη δεδομένων μεμονωμένων καιρικών σταθμών
- ***JSON API:*** Απλή και γρήγορη πρόσβαση των δεδομένων Meteostat σε μορφή JSON

Υπάρχουν πολλές κυβερνητικές διεπαφές που παρέχουν ανοιχτή πρόσβαση στα δεδομένα καιρού που διατίθενται από τα εθνικά μετεωρολογικά γραφεία. Τα δεδομένα που παρέχονται από οργανισμούς όπως οι NOAA, DWD και Environment Canada είναι ένας πολύτιμος πόρος για την επιστήμη, την εκπαίδευση, τις επιχειρήσεις και κάθε άτομο που αναζητά δεδομένα καιρού και κλίματος.

Ωστόσο, όλες αυτές οι διεπαφές χρησιμοποιούν διαφορετικές μορφές δεδομένων και διαδικασίες για την πρόσβαση στις πληροφορίες. Απαιτεί τεράστια προσπάθεια και συντήρηση για την ενημέρωση προσωπικής βάση δεδομένων για τις καιρικές και κλιματικές στατιστικές. Και για αυτό είναι το Meteostat. Αντί για την διατήρηση προσωπικής βάση δεδομένων, ρουτίνες εισαγωγής και μέτρα διασφάλισης ποιότητας, δίνει την δυνατότητα να ξεκινήσει η ανάπτυξη εφαρμογών βάσει δεδομένων καιρού και κλίματος σε λίγα λεπτά.

Η Meteostat τροφοδοτεί από ένα μικρό έργο επιστήμης δεδομένων έως τα μεγαλύτερα πανεπιστήμια και επιχειρήσεις στον κόσμο. Σε αντίθεση με τις περισσότερες άλλες εφαρμογές και API που σχετίζονται με τον καιρό, η Meteostat επικεντρώνεται σε ιστορικά δεδομένα καιρού και κλίματος που μετρήθηκαν στην

περιοχή από μετεωρολογικούς σταθμούς σε όλο τον κόσμο. Δίνεται η δυνατότητα ανάκτησης ακατέργαστων παρατηρήσεων μεμονωμένων μετεωρολογικών σταθμών που δεν έχουν παρεμβολή ή η χρήση δεδομένα σημείου απόκτηση δεδομένων ανά γεωγραφική τοποθεσία.

Η βιβλιοθήκη Meteostat Python παρέχει μια απλή διεπαφή προγραμματισμού για πρόσβαση σε δεδομένα ανοιχτού καιρού και κλίματος. Οι ιστορικές παρατηρήσεις και στατιστικά στοιχεία λαμβάνονται από τη μαζική διεπαφή δεδομένων της Meteostat και αποτελούνται από δεδομένα που παρέχονται από διαφορετικές δημόσιες διεπαφές, οι περισσότερες από τις οποίες είναι κυβερνητικές. Μεταξύ των πηγών δεδομένων είναι οι εθνικές μετεωρολογικές υπηρεσίες όπως η Εθνική Ωκεάνια και η Ατμοσφαιρική Διοίκηση (NOAA) και η εθνική μετεωρολογική υπηρεσία της Γερμανίας (DWD).

Σε αντίθεση με άλλες μετεωρολογικές διεπαφές δεδομένων, η Meteostat δεν χρησιμοποιεί καθολικό μοντέλο δεδομένων. Αντ 'αυτού, η Meteostat παρέχει παρατηρήσεις καιρού και μακροπρόθεσμες στατιστικές για το κλίμα για μεμονωμένους μετεωρολογικούς σταθμούς. Είναι κατανοητό ότι κανείς δεν γνωρίζει τα αναγνωριστικά κάθε μετεωρολογικού σταθμού. Επομένως, η Meteostat παρέχει την Station class, μια απλή διεπαφή για την αναζήτηση μετεωρολογικών σταθμών χρησιμοποιώντας διάφορα φίλτρα. Κάθε μετεωρολογικός σταθμός αντιπροσωπεύεται από μια σειρά Pandas DataFrame που παρέχει meta-data σχετικά με το σταθμό.

Αποθετήριο δεδομένων COVID-19 από το Κέντρο Επιστήμης και Μηχανικής Συστημάτων (CSSE) στο Πανεπιστήμιο Johns Hopkins (JHU)

Το Repository αυτό είναι το αποθετήριο δεδομένων για το Visual Dashboard Novel Coronavirus του 2019, το οποίο διαχειρίζεται το Πανεπιστήμιο Johns Hopkins University for Systems Science and Engineering (JHU CSSE). Επίσης, υποστηρίζεται από την ομάδα ESRI Living Atlas και το εργαστήριο εφαρμοσμένης φυσικής Johns Hopkins University (JHU APL).

Συλλογή δεδομένων COVID-19 από το Our World in Data

Το πλήρες σύνολο δεδομένων COVID-19 είναι μια συλλογή των δεδομένων COVID-19 που διατηρούνται από το Our World in Data. Ενημερώνεται καθημερινά και περιλαμβάνει δεδομένα σχετικά με επιβεβαιωμένα περιστατικά, θανάτους, νοσηλεία, εξετάσεις και εμβολιασμούς, καθώς και άλλες μεταβλητές δυνητικού ενδιαφέροντος.

Τα δεδομένα και οι πηγές δεδομένων απο τις οποίες συντελείται απο τα παρακάτω:

- ***Επιβεβαιωμένες περιπτώσεις και θάνατοι:*** τα δεδομένα προέρχονται από το αποθετήριο δεδομένων COVID-19 από το Κέντρο Επιστήμης και Μηχανικής Συστημάτων (CSSE) στο Πανεπιστήμιο Johns Hopkins (JHU). Το σύνολο δεδομένων περιστατικών & θανάτων ενημερώνεται καθημερινά.
- ***Εισαγωγή σε μονάδες νοσηλείας και μονάδας εντατικής θεραπείας (ICU):*** τα δεδομένα προέρχονται από το Ευρωπαϊκό Κέντρο Πρόληψης και Ελέγχου Νόσων (ECDC) για επιλεγμένο αριθμό ευρωπαϊκών χωρών οι οποίες είναι η κυβέρνηση του Ηνωμένου Βασιλείου, το Υπουργείο Υγείας και Ανθρωπίνων Υπηρεσιών για τις Ηνωμένες Πολιτείες και το COVID-19 Tracker για τον Καναδά. Δυστυχώς, δεν παρέχονται δεδομένα για νοσηλεία σε άλλες χώρες καθώς προς το παρόν δεν υπάρχει παγκόσμια, συγκεντρωτική βάση δεδομένων για τη νοσηλεία COVID-19.
- ***Τέστ για COVID-19:*** αυτά τα δεδομένα συλλέγονται από την ομάδα του Our World in Data από επίσημες αναφορές. Το σύνολο δεδομένων δοκιμών ενημερώνεται περίπου δύο φορές την εβδομάδα.
- ***Εμβολιασμοί κατά του COVID-19:*** αυτά τα δεδομένα συλλέγονται από την ομάδα του Our World in Data από επίσημες αναφορές.
- ***Άλλες μεταβλητές:*** αυτά τα δεδομένα συλλέγονται από ποικίλες πηγές (Ηνωμένα Έθνη, Παγκόσμια Τράπεζα, Παγκόσμια Βαρύτητα της Ασθένειας, Σχολή Κυβέρνησης Blavatnik κ.λπ.).

Το πλήρες σύνολο δεδομένων COVID-19 διατίθεται σε μορφές CSV, XLSX και JSON και περιλαμβάνει όλα τα ιστορικά δεδομένα σχετικά με την πανδημία έως την ημερομηνία δημοσίευσης.

Panacea Lab - Georgia State University Covid-19 Twitter chatter dataset for scientific use

Λόγω της παγκόσμιας πανδημίας COVID-19, δημιουργήθηκε το σύνολο δεδομένων για tweets που αποκτήθηκαν από το Twitter Stream που σχετίζονται με τον COVID-19. Από την πρώτη κυκλοφορία, έχουμε συγκεντρωθεί επιπλέον δεδομένα από νέους συνεργάτες, επιτρέποντας σε αυτόν το αποθετήριο να αυξηθεί στο τρέχον μέγεθός του. Η συλλογή ειδικών δεδομένων ξεκίνησε από τις 11 Μαρτίου, αποδίδοντας πάνω από 3,3 εκατομμύρια tweets την ημέρα.

Τα δεδομένα που συλλέγονται από τη ροή καταγράφουν όλες τις γλώσσες, αλλά η υψηλότερη επικράτηση είναι: Αγγλικά, Ισπανικά και Γαλλικά. Συγκεντρώνονται όλα τα tweets και retweets στο αρχείο full_dataset.tsv (990.198.297 μοναδικά tweets) και μια καθαρή έκδοση χωρίς retweets στο full_dataset-clean.tsv (252.342.227 μοναδικά tweets). Για εργασίες NLP και όχι μόνο, παρέχονται οι κορυφαίοι 1000 συχνότεροι όροι στο frequent_terms.csv, οι κορυφαίοι 1000 bigrams στο frequent_bigrams.csv και τα κορυφαία 1000 trigrams στο frequent_trigrams.csv. Ορισμένα γενικά στατιστικά στοιχεία ανά ημέρα περιλαμβάνονται και για τα δύο σύνολα δεδομένων στα αρχεία statistics-full_dataset.tsv και stats-full_dataset-clean.tsv.

ΚΕΦΑΛΑΙΟ 3

Η συνάρτηση createCovidDataFrame()

Τα δεδομένα για τον covid-19 αποτελούν ένα τεράστιο μέρος της ενασχόλησης της επιστημονικής κοινότητας. Τα δεδομένα αυτά ωστόσο στις περισσότερες περιπτώσεις είναι διαθέσιμα από διαφορετικές πήγες, με διαφορετικό format και ακόμη και διαφορετικές μονάδες μέτρησης για διαφορετικά χρονικά διαστήματα. Η συγκέντρωση λοιπών όλων των πληροφοριών σχετικών, ή/και πιθανώς σχετικών, με τον covid-19 και η ενοποίηση τους σε ένα σετ δεδομένων με τυποποιημένα ονόματα, μονάδες μέτρησης και προσδιορισμένα ανά προσδιορισμένα χρονικά περιθώρια, τέθηκε σαν προτεραιότητα κατά την διάρκεια της πτυχιακής εργασίας.

Ένας επιπλέον στόχος της συνάρτησης αυτής ήταν να διαθέτει όσο τον δυνατόν μεγαλύτερο βαθμό αυτοματοποίησης. Στην μορφή στην οποία βρίσκεται κατά την διάρκεια της συγγραφής, η συνάρτηση εάν κληθεί, συλλεγεί, προετοιμάζει και ενοποιεί δεδομένα μέχρι και μια ημέρα πριν από την ημερομηνία που έγινε η κλήση της.

Είναι επιπλέον σημαντικό να σημειωθεί το βάρος που δίνεται στην επεξεργασία των δεδομένων με σκοπό το σετ δεδομένων που εξάγει σαν αποτέλεσμα η συνάρτηση να είναι άμεσα αξιοποιήσιμο για οποιαδήποτε ανάλυση δεδομένων. Κατά την διάρκεια της εκτέλεσης ο αλγόριθμος φροντίζει για οποιαδήποτε ασυνέχεια στα δεδομένα, καθώς και για την ομοιομορφία των χρονικών στιγμών μεταξύ των εγγράφων του σετ δεδομένων.

Η συνάρτηση δέχεται μόνο ένα όρισμα, το οποίο είναι η χωρά για την οποία θα εξαχθούν δεδομένα. Κατά την χρονική στιγμή της συγγραφής η συνάρτηση υποστηρίζει συνολικά 121 χώρες, οι οποίες φαίνονται στον παρακάτω πίνακα.

'Albania'	'Andorra'	'Argentina'	'Austria'	'Bahrain'
'Bangladesh'	'Belarus'	'Belgium'	'Bhutan'	'Bolivia'
'Bosnia and Herzegovina'	'Brazil'	'Bulgaria'	'Chile'	'Colombia'
'Costa Rica'	"Cote d'Ivoire"	'Croatia'	'Cuba'	'Cyprus'
'Denmark'	'Dominica'	'Dominican Republic'	'Ecuador'	'El Salvador'
'Estonia'	'Ethiopia'	'Fiji'	'Finland'	'France'
'Gambia'	'Germany'	'Ghana'	'Greece'	'Guatemala'
'Hungary'	'Iceland'	'India'	'Indonesia'	'Iran'
'Iraq'	'Ireland'	'Israel'	'Italy'	'Jamaica'
'Japan'	'Jordan'	'Kazakhstan'	'Kenya'	'Kuwait'
'Latvia'	'Libya'	'Lithuania'	'Luxembourg'	'Madagascar'
'Malawi'	'Malaysia'	'Malta'	'Mauritania'	'Mexico'
'Mongolia'	'Morocco'	'Mozambique'	'Namibia'	'Nepal'
'Netherlands'	'New Zealand'	'Niger'	'Nigeria'	'Norway'
'Oman'	'Pakistan'	'Panama'	'Paraguay'	'Peru'
'Philippines'	'Poland'	'Portugal'	'Qatar'	'Romania'
'Russia'	'Rwanda'	'Saudi Arabia'	'Senegal'	'Serbia'
'Singapore'	'Slovenia'	'South Africa'	'South Sudan'	'Spain'
'Sri Lanka'	'Sudan'	'Sweden'	'Switzerland'	'Thailand'
'Togo'	'Trinidad and Tobago'	'Tunisia'	'Turkey'	'Uganda'
'Ukraine'	'United Arab Emirates'	'United Kingdom'	'Uruguay'	'Vietnam'
'Zambia'	'Zimbabwe'	'Australia'	'Canada'	'Democratic Republic of Congo'
'China'	'United States'			

Πίνακας 1: Διαθέσιμες Χώρες για τη συνάρτηση createCovidDataFrame()

Το σετ δεδομένων το οποίο εξάγει η συνάρτηση, μετά το πέρας της εκτέλεσης της, αποτελείται από καθημερινά δεδομένα, με την πρώτη ημερομηνία να ξεκινάει στις 01/01/2020 και την τελευταία να είναι η προηγουμένη ημέρα από εκείνη που κλήθηκε η συνάρτηση. Στις τελευταίες ημερομηνίες σε μερικές από τις μεταβλητές από τις οποίες αποτελούνται οι χρονοσειρές, πραγματοποιείται ένα interpolation καθώς δεν έχουν νέα δεδομένα οι πηγές τους. Έτσι η τιμή των μεταβλητών αυτών παραμένει σταθερή για αυτές τις 2 ή/και 3 ημέρες μέχρι να γίνουν διαθέσιμα τα δεδομένα αυτά έως εκείνες τις ημερομηνίες όπως οι υπόλοιπες μεταβλητές.

Συνολικά η συνάρτηση παράγει ένα σετ δεδομένων που αποτελεί μια χρονοσειρά με χρόνο δειγματοληψίας ανά μια ημέρα και 88 παραμέτρους/μεταβλητές για κάθε ημέρα. Οι 88 αυτές μεταβλητές συνδυάζονται και σχηματίζουν τις παρακάτω κατηγορίες:

- Καθημερινά δεδομένα καιρού
- Κυβερνητικά μέτρα και δεδομένα αυστηρότητας μέτρων
- Δεδομένα παγκόσμιας χρήσης όρων για τον COVID-19 στο Twitter
- Στατιστικά δεδομένα υγείας χώρας
- Στατιστικά δεδομένα χώρας
- Καθημερινά και αθροιστικά δεδομένα COVID-19
- Δεδομένα τεστ COVID-19
- Δεδομένα εμβολιασμού
- Καθημερινές εισαγωγές νοσοκομείων και μονάδων εντατικής θεραπείας (ορισμένες χώρες)

Ονομαστικά οι μεταβλητές που καθορίζουν τις διαστάσεις της χρονοσειράς αναφέρονται στον παρακάτω πίνακα.

'total_vaccinations'	'people_vaccinated'	'people_fully_vaccinated'	'daily_vaccinations_raw'
'daily_vaccinations'	'total_vaccinations_per_hundred'	'people_vaccinated_per_hundred'	'people_fully_vaccinated_per_hundred'
'daily_vaccinations_per_million'	'Daily_Tests'	'Cumulative_Tests_Total'	'Cumulative_Total_Tests_Per_Thousand'
'Daily_Change_Cumulative_Total_Tests_Per_Thousand'	'7_Day_Smoothed_Test_Daily_Change'	'7_Day_Smoothed_Test_Daily_Change_Per_Thousand'	'Test_Short_Term_Positive_Rate'
'Short_Term_Tests_Per_Case'	'positive_percent'	'14_Day_Week_Average_Temp'	'14_Day_Week_Average_Wind'
	'Cumulative	'Cumulative Deaths'	'Cumulative

'14_Day_Week_Average_Humidity'	Confirmed'		Recovered'
'Daily Confirmed'	'Daily Deaths'	'Daily Recovered'	'Avg Temp'
'Avg Humidity'	'Avg Wind Speed'	'Daily_Test_Positivity_ewm_03'	'Daily_Test_Positivity_ewm_05'
'Daily_Test_Positivity_ewm_07'	'reproduction_rate'	'icu_patients'	'icu_patients_per_million'
'hosp_patients'	'hosp_patients_per_million'	'weekly_icu_admissions'	'weekly_icu_admissions_per_million'
'weekly_hosp_admissions'	'weekly_hosp_admissions_per_million'	'stringency_index'	'population'
'population_density'	'median_age'	'aged_65_older'	'aged_70_older'
'gdp_per_capita'	'extreme_poverty'	'cardiovasc_death_rate'	'diabetes_prevalence'
'female_smokers'	'male_smokers'	'handwashing_facilities'	'hospital_beds_per_thousand'
'life_expectancy'	'human_development_index'	'school_closing'	'workplace_closing'
'cancel_public_events'	'restrictions_on_gatherings'	'close_public_transport'	'stay_at_home_requirements'
'movementrestrictions'	'internationaltravel'	'containment_health_index'	'income_support'
'government_response_index'	'testing_policy'	'facial_coverings'	'stringency_index_2'
'coronavirus'	'covid'	'covid19'	'lockdown'
'cases'	'pandemic'	'mask'	'deaths'
'quarantine'	'virus'	'14_Day_Week_Average_coronavirus'	'14_Day_Week_Average_lockdown'
'14_Day_Week_Average_cases'	'14_Day_Week_Average_mask'	'14_Day_Week_Average_pandemic'	'14_Day_Week_Average_deaths'

Πίνακας 2: Μεταβλητές Εξόδου της συνάρτησης createCovidDataFrame() (Ονομαστικά)

Παρακάτω είναι διαθέσιμη μια σύντομη περιγραφή για κάθε μια από τις μεταβλητές. Είναι σημαντικό να σημειωθεί ότι όλες οι μεταβλητές διαθέτουν καθημερινά δεδομένα.

ΜΕΤΑΒΛΗΤΗ	ΠΕΡΙΓΡΑΦΗ
'total_vaccinations'	Ο συνολικός αριθμός των χορηγηθεισών δόσεων. Αυτό υπολογίζεται ως εφάπαξ δόση και ενδέχεται να μην ισούται με τον συνολικό αριθμό των ατόμων που

	εμβολιάστηκαν, ανάλογα με το συγκεκριμένο δοσολογικό σχήμα (π.χ. τα άτομα λαμβάνουν πολλαπλές δόσεις). Εάν ένα άτομο λάβει μια δόση του εμβολίου, αυτή η μέτρηση αυξάνεται κατά 1. Εάν λάβει μια δεύτερη δόση, αυξάνεται ξανά κατά 1.
'people_vaccinated'	Ο συνολικός αριθμός ατόμων που έλαβαν τουλάχιστον μία δόση εμβολίου. Εάν ένα άτομο λάβει την πρώτη δόση ενός εμβολίου 2 δόσεων, αυτή η μέτρηση αυξάνεται κατά 1. Εάν λάβει τη δεύτερη δόση, η μέτρηση παραμένει η ίδια.
'people_fully_vaccinated'	Ο συνολικός αριθμός των ατόμων που έλαβαν όλες τις δόσεις που προβλέπονται από το πρωτόκολλο εμβολιασμού. Εάν ένα άτομο λάβει την πρώτη δόση ενός εμβολίου 2 δόσεων, αυτή η μέτρηση παραμένει η ίδια. Εάν λάβουν τη δεύτερη δόση, η μέτρηση αυξάνεται κατά 1.
'daily_vaccinations_raw'	Η καθημερινή μεταβολή στον συνολικό αριθμό των χορηγηθεισών δόσεων Υπολογίζεται μόνο για διαδοχικές ημέρες. Αυτό είναι ένα πρωτογενές μέτρο που παρέχεται για έλεγχο δεδομένων και διαφάνεια
'daily_vaccinations'	Νέες δόσεις χορηγούμενες ανά ημέρα (εξομάλυνση 7 ημερών). Για χώρες που δεν αναφέρουν δεδομένα σε καθημερινή βάση, υποθέτουμε ότι οι δόσεις άλλαξαν εξίσου σε καθημερινή βάση σε οποιεσδήποτε περιόδους κατά τις οποίες δεν αναφέρθηκαν δεδομένα. Αυτό παράγει μια πλήρη σειρά ημερήσιων αριθμών, η οποία στη συνέχεια υπολογίζεται κατά μέσο όρο σε ένα κυλιόμενο παράθυρο 7 ημερών.
'total_vaccinations_per_hundred'	'total_vaccinations' ανά 100 άτομα στο συνολικό πληθυσμό της χώρας.
'people_vaccinated_per_hundred'	“people_vaccinated” ανά 100 άτομα στο συνολικό πληθυσμό της χώρας.
'people_fully_vaccinated_per_hundred'	“people_fully_vaccinated” ανά 100 άτομα στο συνολικό πληθυσμό της χώρας.
'daily_vaccinations_per_million'	“'daily_vaccinations ” ανά 1,000,000 άτομα στο συνολικό πληθυσμό της χώρας.
'Daily_Tests'	Νέα τέστ για COVID-19 (υπολογίζονται μόνο για διαδοχικές ημέρες)
'Cumulative_Tests_Total'	Συνολικός αριθμός τέστ για COVID-19
'Cumulative_Total_Tests_Per_Thousand'	'Cumulative_Tests_Total' ανά 100 άτομα στο συνολικό πληθυσμό της χώρας.
'Daily_Change_Cumulative_Total_Tests_Per_Thousand'	'Cumulative_Tests_Total' ανά 1000 άτομα στο συνολικό πληθυσμό της χώρας.

'7_Day_Smoothed_Test_Daily_Change'	Νέα test για COVID-19 (εξομάλυνση 7 ημερών). Για χώρες που δεν αναφέρουν δεδομένα δοκιμών σε καθημερινή βάση, θεωρείται ότι οι δοκιμές άλλαξαν εξίσου σε καθημερινή βάση για οποιεσδήποτε περιόδους στις οποίες δεν αναφέρθηκαν δεδομένα. Αυτό παράγει μια πλήρη σειρά ημερήσιων αριθμών, η οποία στη συνέχεια υπολογίζεται κατά μέσο όρο σε ένα κυλιόμενο παράθυρο 7 ημερών
'7_Day_Smoothed_Test_Daily_Change_Per_Thousand'	'7_Day_Smoothed_Test_Daily_Change' ανά 1000 άτομα στο συνολικό πληθυσμό της χώρας.
'Test_Short_Term_Positive_Rate'	Το μερίδιο των τέστ COVID-19 που είναι θετικά, δίνεται ως κυλιόμενο μέσο όρο 7 ημερών (αυτό είναι το αντίστροφο των δοκιμών_per_case)
'Short_Term_Tests_Per_Case'	Τέστ που διεξήχθησαν ανά νέα επιβεβαιωμένη περίπτωση COVID-19, που δίδονται ως κυλιόμενος μέσος όρος 7 ημερών (αυτός είναι ο αντίστροφος του θετικού ρυθμού)
'positive_percent'	Το ποσοστό των τέστ που βγήκαν θετικά εκείνη την ημέρα
'14_Day_Week_Average_Temp'	Μέση θερμοκρασία 7 ημερών δύο εβδομάδων πίσω. Ημέρα -14 έως -7 από την τρέχουσα ημέρα
'14_Day_Week_Average_Wind'	Μέση ταχύτητα ανέμου 7 ημερών δύο εβδομάδων πίσω. Ημέρα -14 έως -7 από την τρέχουσα ημέρα
'14_Day_Week_Average_Humidity'	Μέση υγρασία 7 ημερών δύο εβδομάδων πίσω. Ημέρα -14 έως -7 από την τρέχουσα ημέρα
'Cumulative Confirmed'	Σύνολο επιβεβαιωμένων περιπτώσεων COVID-19
'Cumulative Deaths'	Σύνολο επιβεβαιωμένων θανάτων απο COVID-19
'Cumulative Recovered'	Σύνολο επιβεβαιωμένων ασθενών που θεραπεύτηκαν απο COVID-19
'Daily Confirmed'	Νέες επιβεβαιωμένες περιπτώσεις COVID-19
'Daily Deaths'	Νέοι θάνατοι που αποδίδονται στο COVID-19
'Daily Recovered'	Νέοι ασθενείς που θεραπεύτηκαν απο COVID-19
'Avg Temp'	Μέση θερμοκρασία μέρας
'Avg Humidity'	Μέση υγρασία ημέρας
'Avg Wind Speed'	Μέση ταχύτητα ανέμου σε km/h
'Daily_Test_Positivity_ewm_03'	Το ποσοστό των τέστ που βγήκαν θετικά εκείνη την ημέρα με συντελεστή εξομάλυνσης ίσο με 0.3
'Daily_Test_Positivity_ewm_05'	Το ποσοστό των τέστ που βγήκαν θετικά εκείνη την ημέρα με συντελεστή εξομάλυνσης ίσο με 0.5

'Daily_Test_Positivity_ewm_07'	Το ποσοστό των τέστ που βγήκαν θετικά εκείνη την ημέρα με συντελεστή εξομάλυνσης ίσο με 0.7
'reproduction_rate'	Εκτίμηση του πραγματικού ρυθμού αναπαραγωγής (R) του COVID-19
'icu_patients'	Αριθμός ασθενών με COVID-19 σε μονάδες εντατικής θεραπείας (ΜΕΘ) μια δεδομένη ημέρα
'icu_patients_per_million'	Αριθμός ασθενών με COVID-19 σε μονάδες εντατικής θεραπείας (ΜΕΘ) μια δεδομένη ημέρα ανά 1.000.000 άτομα
'hosp_patients'	Αριθμός ασθενών με COVID-19 σε νοσοκομείο μια δεδομένη ημέρα
'hosp_patients_per_million'	Αριθμός ασθενών με COVID-19 σε νοσοκομείο μια δεδομένη ημέρα ανά 1.000.000 άτομα
'weekly_icu_admissions'	Αριθμός ασθενών με COVID-19 που εισήχθησαν πρόσφατα σε μονάδες εντατικής θεραπείας (ΜΕΘ) σε μια δεδομένη εβδομάδα
'weekly_icu_admissions_per_million'	Αριθμός ασθενών με COVID-19 που εισήχθησαν πρόσφατα σε μονάδες εντατικής θεραπείας (ΜΕΘ) σε μια δεδομένη εβδομάδα ανά 1.000.000 άτομα
'weekly_hosp_admissions'	Αριθμός ασθενών με COVID-19 που εισήχθησαν πρόσφατα σε νοσοκομεία σε μια δεδομένη εβδομάδα
'weekly_hosp_admissions_per_million'	Αριθμός ασθενών COVID-19 που εισήχθησαν πρόσφατα σε νοσοκομεία σε μια δεδομένη εβδομάδα ανά 1.000.000 άτομα
'stringency_index'	Δείκτης αυστηρότητας απόκρισης της κυβέρνησης: σύνθετο μέτρο που βασίζεται σε 9 δείκτες απόκρισης, συμπεριλαμβανομένων κλεισίματος σχολείου, κλεισίματος στο χώρο εργασίας και απαγορεύσεων ταξιδιού, επαναπροσδιορίστηκε σε τιμή από 0 έως 100 (100 = αυστηρότερη απόκριση)
'population'	Πληθυσμός το 2020
'population_density'	Αριθμός ατόμων διαιρούμενο ανά έκταση, μετρημένο σε τετραγωνικά χιλιόμετρα, το πιο πρόσφατο διαθέσιμο έτος
'median_age'	Μέση ηλικία του πληθυσμού, πρόβλεψη του ΟΗΕ για το 2020
'aged_65_older'	Μερίδιο του πληθυσμού που είναι 65 ετών και άνω, το πιο πρόσφατο διαθέσιμο έτος
'aged_70_older'	Μερίδιο του πληθυσμού που είναι 70 ετών και άνω, το πιο πρόσφατο διαθέσιμο έτος
'gdp_per_capita'	καθάριστο εγχώριο προϊόν με ισοτιμία αγοραστικής δύναμης (σταθερά διεθνή δολάρια 2011), το πιο πρόσφατο διαθέσιμο έτος

'extreme_poverty'	Μερίδιο του πληθυσμού που ζει σε ακραία φτώχεια, το πιο πρόσφατο έτος διαθέσιμο από το 2010
'cardiovasc_death_rate'	Ποσοστό θανάτου από καρδιαγγειακές παθήσεις το 2017 (ετήσιος αριθμός θανάτων ανά 100.000 άτομα)
'diabetes_prevalence'	Επικράτηση του διαβήτη (% του πληθυσμού ηλικίας 20 έως 79 ετών) το 2017
'female_smokers'	Ποσοστό γυναικών που καπνίζουν, το πιο πρόσφατο διαθέσιμο έτος
'male_smokers'	Ποσοστό ανδρών που καπνίζουν, το πιο πρόσφατο διαθέσιμο έτος
'handwashing_facilities'	Μερίδιο του πληθυσμού με βασικές εγκαταστάσεις πλυσίματος χεριών στις εγκαταστάσεις, το πιο πρόσφατο έτος διαθέσιμο
'hospital_beds_per_thousand'	Νοσοκομειακά κρεβάτια ανά 1.000 άτομα, το πιο πρόσφατο έτος διαθέσιμο από το 2010
'life_expectancy'	Προσδόκιμο ζωής κατά τη γέννηση το 2019
'human_development_index'	Ένας σύνθετος δείκτης που μετρά το μέσο επίτευγμα σε τρεις βασικές διαστάσεις της ανθρώπινης ανάπτυξης - μια μακρά και υγιή ζωή, γνώση και ένα αξιοπρεπές βιοτικό επίπεδο. Τιμές για το 2019
'school_closing'	Καταγραφή κλεισίματος σχολείων και πανεπιστημίων 0 - κανένα μέτρο 1 - συνιστούμε να κλείσετε ή να ανοίξετε όλα τα σχολεία με αλλαγές που έχουν ως αποτέλεσμα σημαντικές διαφορές σε σύγκριση με τις επιχειρήσεις που δεν είναι Covid-19 2 - Απαιτείται κλείσιμο (μόνο ορισμένα επίπεδα ή κατηγορίες, π.χ. μόνο γυμνάσιο ή απλά δημόσια σχολεία) 3 - απαιτείται κλείσιμο όλων των επιπέδων Κενό - χωρίς δεδομένα
'workplace_closing'	Καταγράψτε το κλείσιμο των χώρων εργασίας 0 - κανένα μέτρο 1 - προτείνουμε κλείσιμο (ή συνιστούμε εργασία από το σπίτι) 2 - Απαιτείται κλείσιμο (ή εργασία από το σπίτι) για ορισμένους τομείς ή κατηγορίες εργαζομένων 3 - Απαιτείται κλείσιμο (ή εργασία από το σπίτι) για όλους αλλά όχι απαραίτητους χώρους εργασίας (π.χ. παντοπωλεία, γιατροί) Κενό - χωρίς δεδομένα
'cancel_public_events'	Εγγραφή ακύρωσης δημόσιων εκδηλώσεων

	0 - κανένα μέτρο 1 - συνιστάται η ακύρωση 2 - απαιτείται ακύρωση Κενό - χωρίς δεδομένα
'restrictions_on_gatherings'	Καταγραγή των ορίων των συγκεντρώσεων 0 - χωρίς περιορισμούς 1 - περιορισμοί σε πολύ μεγάλες συγκεντρώσεις (το όριο είναι πάνω από 1000 άτομα) 2 - περιορισμοί στις συγκεντρώσεις μεταξύ 101-1000 ατόμων 3 - περιορισμοί στις συγκεντρώσεις μεταξύ 11-100 ατόμων 4 - περιορισμοί σε συγκεντρώσεις 10 ατόμων ή λιγότερο Κενό - χωρίς δεδομένα
'close_public_transport'	Αρχείο κλεισίματος των δημόσιων συγκοινωνιών 0 - κανένα μέτρο 1 - συνιστούμε να κλείσετε (ή να μειώσετε σημαντικά τον όγκο / διαδρομή / διαθέσιμα μέσα μεταφοράς) 2 - Απαιτείται κλείσιμο (ή απαγόρευση χρήσης από τους περισσότερους πολίτες) Κενό - χωρίς
'stay_at_home_requirements'	Καταγραφή οδηγιών για περιοριορισμό στο σπίτι 0 - κανένα μέτρο 1 - συνιστούμε να μην φύγετε από το σπίτι 2 - να μην αφήσετε το σπίτι με εξαιρέσεις για καθημερινή άσκηση, ψώνια και "απαραίτητα" ταξίδια 3 - Απαιτείται να μην φύγετε από το σπίτι με ελάχιστες εξαιρέσεις (π.χ. επιτρέπεται να φύγει μία φορά την εβδομάδα ή μόνο ένα άτομο μπορεί να φύγει κάθε φορά, κ.λπ.) Κενό - χωρίς δεδομένα
'movementrestrictions'	Καταγραφή περιορισμών στην εσωτερική κίνηση μεταξύ πόλεων / περιοχών 0 - κανένα μέτρο 1 - συνιστάται να μην ταξιδεύετε μεταξύ περιοχών / πόλεων 2 - ισχύουν περιορισμοί εσωτερικής κίνησης Κενό - χωρίς δεδομένα
'internationaltravel'	Αρχείο περιορισμών στα διεθνή ταξίδια 0 - χωρίς περιορισμούς 1 - προβολή αφίξεων 2 - αφίξεις καραντίνας από ορισμένες ή όλες τις περιοχές 3 - απαγόρευση των αφίξεων από ορισμένες περιοχές 4 - απαγόρευση για όλες τις περιοχές ή

	συνολικό κλείσιμο των συνόρων Κενό - χωρίς δεδομένα
'containment_health_index'	
'income_support'	Αρχεία σχετικά με το εάν η κυβέρνηση παρέχει άμεσες πληρωμές σε μετρητά σε άτομα που χάνουν τη δουλειά τους ή δεν μπορούν να εργαστούν. Σημείωση: περιλαμβάνει πληρωμές σε εταιρείες μόνο εάν συνδέονται ρητά με μισθοδοσία / μισθούς 0 - καμία υποστήριξη εισοδήματος 1 - η κυβέρνηση αντικαθιστά λιγότερο από το 50% του χαμένου μισθού (ή εάν ένα κατ 'αποκοπή ποσό, είναι μικρότερο από το 50% του μεσαίου μισθού) 2 - η κυβέρνηση αντικαθιστά το 50% ή περισσότερο του χαμένου μισθού (ή εάν ένα κατ 'αποκοπή ποσό, είναι μεγαλύτερο από το 50% του μεσαίου μισθού) Κενό - χωρίς δεδομένα
'government_response_index'	Συνολικός δείκτης απόκρισης της κυβέρνησης (όλοι οι δείκτες)
'testing_policy'	Αρχείο της κυβερνητικής πολιτικής σχετικά με το ποιος έχει πρόσβαση σε test Σημείωση: αυτό καταγράφει πολιτικές σχετικά με τον έλεγχο για τρέχουσα λοίμωξη (test PCR) και όχι για test ανοσίας (test αντισωμάτων) 0 - καμία πολιτική test 1 - μόνο όσοι και οι δύο (α) έχουν συμπτώματα ΚΑΙ (β) πληρούν συγκεκριμένα κριτήρια (π.χ. βασικοί εργαζόμενοι, που εισήχθησαν στο νοσοκομείο, ήρθαν σε επαφή με μια γνωστή υπόθεση, που επέστρεψαν από το εξωτερικό) 2 - test σε οποιονδήποτε εμφανίζει συμπτώματα Covid-19 3 - ανοιχτά δημόσια test (π.χ. test διαθέσιμα σε ασυμπτωματικά άτομα) Κενό - χωρίς δεδομένα
'facial_coverings'	Αρχείο πολιτικών σχετικά με τη χρήση καλυμμάτων προσώπου εκτός σπιτιού 0 - Χωρίς πολιτική 1 - Συνιστάται 2 - Απαιτείται σε ορισμένους κοινόχρηστους / δημόσιους χώρους έξω από το σπίτι με άλλα άτομα που είναι παρόντα ή σε ορισμένες καταστάσεις όταν δεν είναι δυνατή η κοινωνική απόσταση 3 - Απαιτείται σε όλους τους κοινόχρηστους / δημόσιους χώρους έξω από το σπίτι με άλλα άτομα που είναι παρόντα ή σε όλες τις καταστάσεις όταν δεν είναι δυνατή η κοινωνική απόσταση

	4 - Απαιτείται πάντα έξω από το σπίτι ανεξάρτητα από την τοποθεσία ή την παρουσία άλλων ατόμων
'stringency_index_2'	Δείκτης αυστηρότητας (όλοι οι δείκτες σχετικοί με μέτρα που έλαβε η κυβέρνηση κατα του COVID-19)
'coronavirus'	Συνολικός αριθμός αναφορών της λέξης «'coronavirus'» σε tweets
'covid'	Συνολικός αριθμός αναφορών της λέξης «'covid'» σε tweets
'covid19'	Συνολικός αριθμός αναφορών της λέξης «'covid19'» σε tweets
'lockdown'	Συνολικός αριθμός αναφορών της λέξης «'lockdown'» σε tweets
'cases'	Συνολικός αριθμός αναφορών της λέξης «'cases'» σε tweets
'pandemic'	Συνολικός αριθμός αναφορών της λέξης «'pandemic'» σε tweets
'mask'	Συνολικός αριθμός αναφορών της λέξης «'mask'» σε tweets
'deaths'	Συνολικός αριθμός αναφορών της λέξης «'deaths'» σε tweets
'quarantine'	Συνολικός αριθμός αναφορών της λέξης «'quarantine'» σε tweets
'virus'	Συνολικός αριθμός αναφορών της λέξης «'virus'» σε tweets
'14_Day_Week_Average_coronavirus'	Μέσος αριθμός αναφορών της λέξης «coronavirus» σε tweets, 7 ημερών δύο εβδομάδων πίσω. Ημέρα -14 έως -7 από την τρέχουσα ημέρα στην οποία αναφέρεται η εγγραφή
'14_Day_Week_Average_lockdown'	Μέσος αριθμός αναφορών της λέξης «lockdown» σε tweets, 7 ημερών δύο εβδομάδων πίσω. Ημέρα -14 έως -7 από την τρέχουσα ημέρα στην οποία αναφέρεται η εγγραφή
'14_Day_Week_Average_cases'	Μέσος αριθμός αναφορών της λέξης «cases» σε tweets, 7 ημερών δύο εβδομάδων πίσω. Ημέρα -14 έως -7 από την τρέχουσα ημέρα στην οποία αναφέρεται η εγγραφή
'14_Day_Week_Average_mask'	Μέσος αριθμός αναφορών της λέξης «mask'» σε tweets, 7 ημερών δύο εβδομάδων πίσω. Ημέρα -14 έως -7 από την τρέχουσα ημέρα στην οποία αναφέρεται η εγγραφή
'14_Day_Week_Average_pandemic'	Μέσος αριθμός αναφορών της λέξης «pandemic'» σε tweets, 7 ημερών δύο εβδομάδων πίσω. Ημέρα -14 έως -7 από την τρέχουσα ημέρα στην οποία αναφέρεται η εγγραφή
'14_Day_Week_Average_deaths'	Μέσος αριθμός αναφορών της λέξης

	«deaths'» σε tweets, 7 ημερών δύο εβδομάδων πίσω. Ημέρα -14 έως -7 από την τρέχουσα ημέρα στην οποία αναφέρεται η εγγραφή

Πίνακας 3: Μεταβλητές εξόδου της συνάρτησης createCovidDataFrame() με περιγραφή

Μεθοδολογία επιλογής Παραμέτρων για μοντέλα ARIMA

Τα δεδομένα που παράγονται από τη συνάρτηση **createCovidDataFrame()** είναι σε μεγάλο βαθμό έτοιμα να χρησιμοποιηθούν σε αναλύσεις δεδομένων και σε προβλεπτικά μοντέλα. Αναπτύχθηκε σαν παράδειγμα ένα μοντέλο ARIMA για την πρόβλεψη της μεταβλητής «hosp_patients».

Στη στατιστική και την οικονομετρία, και ιδίως στην ανάλυση χρονοσειρών, ένα μοντέλο αυτοσυσχέτισης ενσωματωμένου κινούμενου μέσου όρου (ARIMA) είναι μια γενίκευση ενός μοντέλου αυτοσυσχέτισης κινούμενου μέσου όρου (ARMA). Και τα δύο αυτά μοντέλα είναι προσαρμοσμένα σε δεδομένα χρονοσειρών είτε για καλύτερη κατανόηση των δεδομένων είτε για πρόβλεψη μελλοντικών σημείων στη σειρά (πρόβλεψη).

Τα μη εποχικά μοντέλα ARIMA δηλώνονται γενικά ARIMA (p, d, q) όπου οι παράμετροι p, d και q είναι μη αρνητικοί ακέραιοι αριθμοί, p είναι η σειρά (αριθμός χρονικών υστέρησης) του αυτο-αναδρομικού μοντέλου, d είναι ο βαθμός διαφοροποίηση (ο αριθμός των φορών που αφαιρέθηκαν οι τιμές στο παρελθόν) και q είναι ο βαθμός του μοντέλου κινούμενου μέσου όρου.

Κατασκευάζεται ένα μοντέλο γραμμικής παλινδρόμησης που περιλαμβάνει τον καθορισμένο αριθμό και τον τύπο των όρων και τα δεδομένα προετοιμάζονται με βαθμό διαφοροποίησης προκειμένου να το κάνουν στάσιμο, δηλαδή να αφαιρούν τάσεις και εποχιακές δομές που επηρεάζουν αρνητικά το μοντέλο παλινδρόμησης.

Μια τιμή 0 μπορεί να χρησιμοποιηθεί για μια παράμετρο, η οποία δείχνει να μην χρησιμοποιείται αυτό το στοιχείο του μοντέλου. Με αυτόν τον τρόπο, το μοντέλο ARIMA μπορεί να διαμορφωθεί ώστε να εκτελεί τη λειτουργία ενός μοντέλου ARMA, ακόμη και ενός απλού μοντέλου AR, I ή MA.

Η υιοθέτηση ενός μοντέλου ARIMA για μια χρονική σειρά προϋποθέτει ότι η υποκείμενη διαδικασία που δημιούργησε τις παρατηρήσεις είναι μια διαδικασία

ARIMA. Αυτό μπορεί να φαίνεται προφανές, αλλά βοηθά να παρακινήσει την ανάγκη επιβεβαίωσης των υποθέσεων του μοντέλου στις πρώτες παρατηρήσεις και στα υπολειπόμενα σφάλματα των προβλέψεων από το μοντέλο.

Το πρώτο βήμα για τη δημιουργία ενός μοντέλου ARIMA είναι η εύρεση του αριθμού d, δηλαδή τον αριθμό χρονικής υστέρησης έτσι ώστε η χρονοσειρά να γίνει στάσιμη. Τα μοντέλα ARIMA είναι, θεωρητικά, η πιο γενική κατηγορία μοντέλων για την πρόβλεψη μιας χρονοσειράς που μπορεί να γίνει «στάσιμη» με διαφοροποίηση (αν είναι απαραίτητο), ίσως σε συνδυασμό με μη γραμμικούς μετασχηματισμούς όπως καταγραφή ή ξεφούσκωμα (εάν είναι απαραίτητο). Μια τυχαία μεταβλητή που είναι μια χρονοσειρά είναι στάσιμη εάν οι στατιστικές της ιδιότητες είναι όλες σταθερές με την πάροδο του χρόνου. Μια στατική σειρά δεν έχει τάση, οι παραλλαγές γύρω από το μέσο όρο έχουν σταθερό πλάτος και κουνάει με συνεπή τρόπο, δηλαδή, τα βραχυπρόθεσμα τυχαία χρονοδιαγράμματα της μοιάζουν πάντα τα ίδια με στατιστική έννοια.

Η τελευταία συνθήκη σημαίνει ότι οι αυτοσυσχέτισμοί της (συσχετίσεις με τις δικές του προηγούμενες αποκλίσεις από το μέσο όρο) παραμένουν σταθερές με την πάροδο του χρόνου ή ισοδύναμα, ότι το φάσμα ισχύος του παραμένει σταθερό με την πάροδο του χρόνου. Μια τυχαία μεταβλητή αυτής της φόρμας μπορεί να θεωρηθεί (ως συνήθως) ως συνδυασμός σήματος και θορύβου και το σήμα (εάν είναι εμφανές) θα μπορούσε να είναι ένα πρότυπο γρήγορης ή αργής μέσης αναστροφής ή ημιτονοειδούς ταλάντωσης ή ταχείας εναλλαγής στο σήμα , και θα μπορούσε επίσης να έχει εποχιακό στοιχείο. Ένα μοντέλο ARIMA μπορεί να θεωρηθεί ως «φίλτρο» που προσπαθεί να διαχωρίσει το σήμα από το θόρυβο και το σήμα στη συνέχεια παρεκτείνεται στο μέλλον για να λάβει προβλέψεις.

Η εξίσωση πρόβλεψης ARIMA για μια σταθερή χρονική σειρά είναι μια γραμμική (δηλαδή, τύπος παλινδρόμησης) στην οποία οι προγνωστικοί παράγοντες αποτελούνται από καθυστερήσεις της εξαρτημένης μεταβλητής ή / και καθυστερήσεις των σφαλμάτων πρόβλεψης. Αυτό είναι:

Προβλεπόμενη τιμή του Y = μια σταθερή και / ή ένα σταθμισμένο άθροισμα μίας ή περισσοτέρων πρόσφατων τιμών του Y και / ή ενός σταθμισμένου αθροίσματος μίας ή περισσότερων πρόσφατων τιμών των σφαλμάτων.

Εάν οι προγνωστικοί παράγοντες αποτελούνται μόνο από υστερούμενες τιμές του Y, είναι ένα καθαρό μοντέλο αυτόματης επέμβασης (“self-regressed”), το οποίο είναι μόνο μια ειδική περίπτωση ενός μοντέλου παλινδρόμησης και το οποίο θα μπορούσε να είναι εξοπλισμένο με τυπικό λογισμικό παλινδρόμησης. Για παράδειγμα, ένα μοντέλο αυτόματης επέμβασης πρώτης τάξης (“AR (1)”) για το Y είναι ένα απλό μοντέλο παλινδρόμησης στο οποίο η ανεξάρτητη μεταβλητή έχει μόνο Y καθυστερήσει κατά μία περίοδο (LAG (Y, 1) στα Στατιστικά ή Y_LAG1 στο RegressIt) .

Εάν ορισμένοι από τους προγνωστικούς παράγοντες είναι καθυστερήσεις των σφαλμάτων, ένα μοντέλο ARIMA ΔΕΝ είναι ένα μοντέλο γραμμικής παλινδρόμησης, επειδή δεν υπάρχει τρόπος να προσδιοριστεί το "σφάλμα της τελευταίας περιόδου" ως ανεξάρτητη μεταβλητή: τα σφάλματα πρέπει να υπολογίζονται σε μια περίοδο-προς-βάση περιόδου όταν το μοντέλο είναι προσαρμοσμένο στα δεδομένα.

Από τεχνική άποψη, το πρόβλημα με τη χρήση σφαλμάτων με καθυστέρηση ως προγνωστικά είναι ότι οι προβλέψεις του μοντέλου δεν είναι γραμμικές συναρτήσεις των συντελεστών, παρόλο που είναι γραμμικές συναρτήσεις των προηγούμενων δεδομένων. Έτσι, οι συντελεστές στα μοντέλα ARIMA που περιλαμβάνουν καθυστερημένα σφάλματα πρέπει να εκτιμώνται με μη γραμμικές μεθόδους βελτιστοποίησης ("αναρρίχηση σε λόφο") και όχι απλώς με την επίλυση ενός συστήματος εξισώσεων.

Η εξίσωση πρόβλεψης κατασκευάζεται ως εξής. Αρχικά, ας υποδηλώσουμε τη d-ωστή διαφορά του Y, που σημαίνει:

Εάν d = 0: yt = Yt

Εάν d = 1: yt = Yt - Yt-1

Αν d = 2: yt = (Yt - Yt-1) - (Yt-1 - Yt-2) = Yt - 2Yt-1 + Yt-2

Σημειώνεται ότι η δεύτερη διαφορά του Y (η περίπτωση d = 2) δεν είναι η διαφορά από 2 περιόδους πριν. Αντίθετα, είναι η πρώτη-διαφορά-της-πρώτης διαφοράς, η οποία είναι το διακριτό ανάλογο ενός δεύτερου παραγώγου, δηλαδή η τοπική επιτάχυνση της σειράς και όχι η τοπική της τάση.

Όσον αφορά το y, η γενική εξίσωση πρόβλεψης είναι:

$$\hat{y}t = \mu + \phi 1\ yt\text{-}1 + \ldots + \phi p\ yt\text{-}p - \theta 1 et\text{-}1 - \ldots - \theta q et\text{-}q$$

Εδώ ορίζονται οι παράμετροι κινούμενου μέσου όρου (θ) έτσι ώστε τα σημάδια τους να είναι αρνητικά στην εξίσωση, σύμφωνα με τη σύμβαση που εισήγαγαν οι Box και Jenkins.

Κανονικά, η σωστή ποσότητα διαφοροποίησης είναι η χαμηλότερη σειρά διαφοροποίησης που αποδίδει μια χρονοσειρά που κυμαίνεται γύρω από μια καλά καθορισμένη μέση τιμή και της οποίας η γραφική παράσταση λειτουργίας συσχέτισης (ACF) αποσυντίθεται αρκετά γρήγορα στο μηδέν, είτε από πάνω είτε από κάτω. Εάν η σειρά εξακολουθεί να παρουσιάζει μακροπρόθεσμη τάση, ή αλλιώς δεν έχει την τάση να επιστρέψει στη μέση τιμή της, ή εάν οι αυτοσυσχέτισμοί της είναι θετικοί σε μεγάλο αριθμό καθυστερήσεων (π.χ. 10 ή περισσότερα), τότε χρειάζεται υψηλότερο σειρά διαφοροποίησης. Θα το ορίσουμε ως τον «πρώτο κανόνα αναγνώρισης των μοντέλων ARIMA»:

Κανόνας 1: Εάν η σειρά έχει θετικές αυτο-συσχετίσεις με μεγάλο αριθμό καθυστερήσεων, τότε πιθανότατα χρειάζεται υψηλότερη σειρά διαφορών.

Η διαφοροποίηση τείνει να προκαλέσει αρνητική συσχέτιση: εάν η σειρά εμφανίζει αρχικά ισχυρή θετική αυτοσυσχέτιση, τότε μια μη εποχική διαφορά θα μειώσει την αυτοσυσχέτιση και ίσως ακόμη και να οδηγήσει την αυτοσυσχέτιση lag-1 σε αρνητική τιμή. Εάν εφαρμοστεί μια δεύτερη μη εποχιακή διαφορά (η οποία είναι περιστασιακά απαραίτητη), η αυτοσυσχέτιση lag-1 θα προχωρήσει ακόμη περισσότερο στην αρνητική κατεύθυνση.

Εάν η αυτοσυσχέτιση lag-1 είναι μηδέν ή ακόμη και αρνητική, τότε η σειρά δεν χρειάζεται περαιτέρω διαφοροποίηση. Ένα από τα πιο συνηθισμένα σφάλματα στη μοντελοποίηση ARIMA είναι η "υπερδιαφορά" της σειράς και καταλήγοντας να προσθεθούν επιπλέον όροι AR ή MA για να αναιρεθεί η ζημιά. Εάν η αυτοσυσχέτιση lag-1 είναι πιο αρνητική από -0,5 (και θεωρητικά μια αρνητική αυτοσυσχέτιση lag-1 δεν πρέπει ποτέ να είναι μεγαλύτερη από 0,5 σε μέγεθος), αυτό μπορεί να σημαίνει ότι η σειρά έχει υπερ-διαφοροποιηθεί. Η πλοκή χρονοσειρών μιας σειράς με υπερβολική διαφοροποίηση μπορεί να φαίνεται αρκετά τυχαία με την πρώτη ματιά, αλλά αν ελεγχθεί πιο κοντά θα εντοπιστεί ένα μοτίβο υπερβολικών αλλαγών από τη μία παρατήρηση στην άλλη.

Κανόνας 2: Εάν η αυτοσυσχέτιση lag-1 είναι μηδενική ή αρνητική ή οι αυτοσυσχέτισεις είναι όλες μικρές και χωρίς μοτίβο, τότε η σειρά δεν χρειάζεται

υψηλότερη σειρά διαφορών. Εάν η αυτοσυσχέτιση lag-1 είναι -0,5 ή περισσότερο αρνητική, η σειρά μπορεί να είναι υπερβολική.

Ένα συνηθισμένο "σφάλμα" στη μοντελοποίηση ARIMA είναι να εφαρμοστεί μια επιπλέον σειρά διαφοροποίησης, επειδή το τρέχον σχέδιο αυτοσυσχέτισης δεν δείχνει μεγάλο μοτίβο. Αν όχι, αυτό είναι καλό, όχι κακό! Ένα άλλο σύμπτωμα πιθανής υπερβολικής διαφοροποίησης είναι η αύξηση της τυπικής απόκλισης, παρά η μείωση, όταν αυξάνεται η σειρά των διαφορών. Αυτό γίνεται ο τρίτος κανόνας:

Κανόνας 3: Η βέλτιστη σειρά διαφοροποίησης είναι συχνά η σειρά διαφοροποίησης στην οποία η τυπική απόκλιση είναι η χαμηλότερη.

Οι δύο πρώτοι κανόνες δεν καθορίζουν πάντα σαφώς το «σωστό» βαθμό διαφοροποίησης. Η «ήπια υποδιαφοροποίηση» μπορεί να αντισταθμιστεί με την προσθήκη όρων AR στο μοντέλο, ενώ η «ήπια υπερβολική διαφοροποίηση» μπορεί να αντισταθμιστεί με την προσθήκη όρων MA.

Κανόνας 4: Ένα μοντέλο χωρίς βαθμό διαφοροποίησης προϋποθέτει ότι η αρχική σειρά είναι στάσιμη (μέση επαναφορά). Ένα μοντέλο με μια σειρά διαφοροποίησης προϋποθέτει ότι η αρχική σειρά έχει σταθερή μέση τάση (π.χ. τυχαίο περπάτημα ή μοντέλο τύπου SES, με ή χωρίς ανάπτυξη). Ένα μοντέλο με δύο βαθμούς συνολικής απόκλισης προϋποθέτει ότι η αρχική σειρά έχει μια τάση που διαφέρει από το χρόνο (π.χ. μια τυχαία τάση ή ένα μοντέλο τύπου LES).

Ένα άλλο ζήτημα για τον προσδιορισμό της σειράς διαφοροποίησης είναι ο ρόλος που διαδραματίζει ο σταθερός όρος στο μοντέλο - εάν περιλαμβάνεται κάποιος. Η παρουσία μιας σταθεράς επιτρέπει έναν μη μηδενικό μέσο όρο στη σειρά εάν δεν πραγματοποιείται διαφορά, επιτρέπει μια μη μηδενική μέση τάση στη σειρά εάν χρησιμοποιείται μία σειρά διαφορών και επιτρέπει μη μηδενικό μέσο όρο trend-in-the-trend (δηλ. καμπυλότητα) εάν υπάρχουν δύο τάξεις διαφοροποίησης. Γενικά δεν υποθέτουμε ότι υπάρχουν τάσεις σε τάσεις, επομένως η σταθερά αφαιρείται συνήθως από μοντέλα με δύο τάξεις διαφοροποίησης. Σε ένα μοντέλο με μια σειρά διαφοροποίησης, η σταθερά μπορεί ή δεν μπορεί να συμπεριληφθεί, ανάλογα με το αν κάνουμε ή δεν θέλουμε να επιτρέψουμε μια μέση τάση. Ως εκ τούτου:

Κανόνας 5: Ένα μοντέλο χωρίς βαθμούς διαφοροποίησης περιλαμβάνει συνήθως έναν σταθερό όρο (που επιτρέπει τη μη μηδενική μέση τιμή). Ένα μοντέλο με δύο βαθμούς ολικής απόκλισης συνήθως δεν περιλαμβάνει έναν σταθερό όρο. Σε

ένα μοντέλο με μια σειρά συνολικών διαφορών, θα πρέπει να συμπεριλαμβάνεται ένας σταθερός όρος εάν η σειρά έχει μη μηδενική μέση τάση.

Για την ανάγκη αυτή δημιουργηθηκε η συνάρτηση **test_stationarity(timeseries, window = 5, cutoff = 0.01).** Η συνάρτηση αυτή δέχεται σαν όρισμα μια χρονοσειρά, ένα «παράθυρο» και ένα όριο. Το «παράθυρο» αποτελεί τον αριθμό των παρατηρήσεων από τις οποίες βγαίνει ο μέσος και η μέση απόκλιση. Από την άλλη μεριά το όριο που δέχεται είναι το όριο εμπιστοσύνης με το οποίο κρίνει εάν η σειρά είναι τελικά στάσιμη. Για παράδειγμα το όριο ίσο με 0.01 σημαίνει ότι πρέπει να είναι 99% ο αλγόριθμος ότι η χρονοσειρά είναι στάσιμη για να καταλήξει σε αυτό το συμπέρασμα.

Αφού μια χρονοσειρά τελειώσει με την διαφοροποίηση, το επόμενο βήμα για την προσαρμογή ενός μοντέλου ARIMA είναι να προσδιοριστεί εάν απαιτούνται όροι AR ή MA για τη διόρθωση τυχόν αυτοσυσχέτισης που παραμένει στη διαφορετική σειρά. Φυσικά, με λογισμικό όπως το Statgraphics, δύναται απλώς να δοκιμαστούν διαφορετικοί συνδυασμοί όρων και να εντοπιστεί τι λειτουργεί καλύτερα.

Υπάρχει όμως ένας πιο συστηματικός τρόπος για να γίνει αυτό. Κοιτάζοντας τις γραφικές παραστάσεις της αυτοσυσχέτισης (ACF) και της μερικής αυτοσυσχέτισης (PACF) των διαφορετικών σειρών, μπορεί να προσδιοριστεί προσωρινά τους αριθμούς των όρων AR και / ή MA που χρειάζονται. Το ACF είναι απλώς ένα γράφημα ράβδων των συντελεστών συσχέτισης μεταξύ μιας χρονοσειράς και των καθυστερήσεών της. Το διάγραμμα PACF είναι ένα διάγραμμα των συντελεστών μερικής συσχέτισης μεταξύ της σειράς και των καθυστερήσεών της.

Γενικά, η «μερική» συσχέτιση μεταξύ δύο μεταβλητών είναι το ποσό συσχέτισης μεταξύ τους, το οποίο δεν εξηγείται από τους αμοιβαίους συσχετισμούς τους με ένα καθορισμένο σύνολο άλλων μεταβλητών. Για παράδειγμα, εάν υποχωρούμε μια μεταβλητή Y σε άλλες μεταβλητές X1, X2 και X3, η μερική συσχέτιση μεταξύ Y και X3 είναι το ποσό συσχέτισης μεταξύ Y και X3 που δεν εξηγείται από τους κοινούς συσχετισμούς τους με X1 και X2. Αυτή η μερική συσχέτιση μπορεί να υπολογιστεί ως η τετραγωνική ρίζα της μείωσης της διακύμανσης που επιτυγχάνεται με την προσθήκη X3 στην παλινδρόμηση του Y στα X1 και X2.

Μερική αυτοσυσχέτιση είναι το ποσό συσχέτισης μεταξύ μιας μεταβλητής και μιας υστέρησης από μόνη της που δεν εξηγείται από συσχετίσεις καθυστερημένων καθυστερήσεων. Η αυτοσυσχέτιση μιας χρονοσειράς Y στο lag 1 είναι ο συντελεστής συσχέτισης μεταξύ Yt και Yt-1, ο οποίος πιθανώς είναι επίσης ο συσχετισμός μεταξύ Yt-1 και Yt-2. Αλλά αν το Yt συσχετίζεται με το Yt-1 και το Yt-1 συσχετίζεται εξίσου με το Yt-2, τότε θα πρέπει επίσης να περιμένουμε να βρούμε συσχέτιση μεταξύ Yt και Yt-2. Στην πραγματικότητα, το ποσό συσχέτισης που πρέπει να περιμένουμε στο lag 2 είναι ακριβώς το τετράγωνο της συσχέτισης lag-1. Έτσι, η συσχέτιση στην υστέρηση 1 «διαδίδεται» στην υστέρηση 2 και πιθανώς σε υστέρες υψηλότερης τάξης. Η μερική αυτοσυσχέτιση στο lag 2 είναι επομένως η διαφορά μεταξύ της πραγματικής συσχέτισης στο lag 2 και της αναμενόμενης συσχέτισης λόγω της διάδοσης της συσχέτισης στο lag 1.

Η μερική αυτοσυσχέτιση σε όλες τις καθυστερήσεις μπορεί να υπολογιστεί προσαρμόζοντας μια σειρά διαδοχικών μοντέλων με αυξανόμενο αριθμό καθυστερήσεων. Συγκεκριμένα, η μερική αυτοσυσχέτιση στο lag k είναι ίση με τον εκτιμώμενο συντελεστή AR (k) σε ένα αυτοεκτελεστικό μοντέλο με όρους k - δηλαδή, ένα μοντέλο πολλαπλής παλινδρόμησης στο οποίο το Y παλινδρομεί σε LAG (Y, 1), LAG (Y , 2) κ.λπ., έως LAG (Y, k). Έτσι, με απλή επιθεώρηση του PACF, μπορεί να προσδιοριστεί πόσους όρους AR πρέπει να χρησιμοποιηθεί για να εξηγηθεί το μοτίβο αυτοσυσχέτισης σε μια χρονολογική σειρά: εάν η μερική αυτοσυσχέτιση είναι σημαντική στο lag k και δεν είναι σημαντική σε καθυστερήσεις υψηλότερης τάξης - δηλαδή, εάν το PACF «κόψει» στο lag k - τότε αυτό υποδηλώνει ότι θα πρέπει να δοκιμαστεί η εφαρμογή ενός μοντέλου αυτόματης εξόδου του βαθμού k.

Εάν το PACF εμφανίζει απότομη αποκοπή ενώ το ACF αποσυντίθεται πιο αργά (δηλαδή, έχει σημαντικές αιχμές σε υψηλότερες καθυστερήσεις), λέμε ότι η σταθεροποιημένη σειρά εμφανίζει μια "υπογραφή AR", που σημαίνει ότι το μοτίβο αυτοσυσχέτισης μπορεί να εξηγηθεί πιο εύκολα προσθέτοντας AR όροι παρά προσθέτοντας όρους MA. Πιθανότατα θα διαπιστωθεί ότι μια υπογραφή AR συσχετίζεται συνήθως με θετική αυτοσυσχέτιση στην υστέρηση 1 - δηλαδή, τείνει να προκύπτει σε σειρές που είναι ελαφρώς υποδιαιρούμενες.

Ο λόγος για αυτό είναι ότι ένας όρος AR μπορεί να λειτουργήσει σαν «μερική διαφορά» στην εξίσωση πρόβλεψης. Για παράδειγμα, σε ένα μοντέλο AR (1), ο όρος

AR ενεργεί σαν μια πρώτη διαφορά αν ο συντελεστής αυτοανάπτυξης είναι ίσος με 1, δεν κάνει τίποτα εάν ο συντελεστής αυτοανάπτυξης είναι μηδέν και ενεργεί σαν μερική διαφορά εάν ο συντελεστής είναι μεταξύ 0 και 1. Έτσι, εάν η σειρά είναι ελαφρώς υποδιαιρούμενη - δηλ Εάν το μη στατικό μοτίβο θετικής αυτοσυσχέτισης δεν έχει εξαλειφθεί πλήρως, θα "ζητήσει" μια μερική διαφορά εμφανίζοντας μια υπογραφή AR. Ως εκ τούτου, προκύπτει ο ακόλουθος κανόνας για τον καθορισμό του πότε θα προστείθενται όροι AR:

Κανόνας 6: Εάν το PACF της διαφορετικής σειράς εμφανίζει απότομη αποκοπή ή / και η αυτοσυσχέτιση lag-1 είναι θετική - δηλαδή, εάν η σειρά εμφανίζεται ελαφρώς "υποδιαιρούμενη" - τότε δύναται να προστεθεί ένας όρος AR στο μοντέλο. Η καθυστέρηση στην οποία διακόπτεται το PACF είναι ο υποδεικνυόμενος αριθμός όρων AR.

Κατ 'αρχήν, οποιοδήποτε μοτίβο αυτοσυσχέτισης μπορεί να αφαιρεθεί από μια σταθεροποιημένη σειρά με την προσθήκη αρκετών όρων αυτόματης αύξησης (καθυστερήσεις της σταθεροποιημένης σειράς) στην εξίσωση πρόβλεψης και το PACF σας λέει πόσους τέτοιους όρους είναι πιθανό να χρειάζονται. Ωστόσο, αυτός δεν είναι πάντα ο απλούστερος τρόπος για να εξηγηθεί ένα δεδομένο μοτίβο αυτοσυσχέτισης: μερικές φορές είναι πιο αποτελεσματικό να προστεθούν όροι MA (καθυστερήσεις των σφαλμάτων πρόβλεψης).

Η λειτουργία αυτοσυσχέτισης (ACF) παίζει τον ίδιο ρόλο για τους όρους MA που παίζει το PACF για τους όρους AR - δηλαδή, το ACF λέει πόσους όρους MA είναι πιθανό να χρειαστούν για να αφαιρεθεί η εναπομένουσα αυτοσυσχέτιση από τη διαφορετική σειρά. Εάν η αυτοσυσχέτιση είναι σημαντική στο lag k αλλά όχι σε υψηλότερες καθυστερήσεις - δηλαδή, εάν το ACF "διακόπτει" στο lag k, αυτό δείχνει ότι ακριβώς οι όροι k MA πρέπει να χρησιμοποιούνται στην εξίσωση πρόβλεψης. Στην τελευταία περίπτωση, λέμε ότι η σταθεροποιημένη σειρά εμφανίζει μια "υπογραφή MA", που σημαίνει ότι το μοτίβο αυτοσυσχέτισης μπορεί να εξηγηθεί πιο εύκολα με την προσθήκη όρων MA παρά με την προσθήκη όρων AR.

Μια υπογραφή MA συσχετίζεται συνήθως με αρνητική αυτοσυσχέτιση στην υστέρηση 1 - δηλαδή, τείνει να προκύψει σε σειρές που είναι ελαφρώς υπερβολικές. Ο λόγος για αυτό είναι ότι ένας όρος MA μπορεί να "ακυρώσει μερικώς" μια σειρά διαφοροποίησης στην εξίσωση πρόβλεψης. Για να ελεγχθεί αυτό, αρκεί να ληφθεί

υπόψιν ότι ένα μοντέλο ARIMA (0,1,1) χωρίς σταθερά είναι ισοδύναμο με ένα μοντέλο Simple Exponential Smoothing. Η εξίσωση πρόβλεψης για αυτό το μοντέλο είναι:

$$\hat{Y}t = \mu + Yt\text{-}1 - \theta 1et\text{-}1$$

όπου ο συντελεστής MA (1) θ1 αντιστοιχεί στην ποσότητα 1- α στο μοντέλο SES. Εάν το θ1 είναι ίσο με 1, αυτό αντιστοιχεί σε ένα μοντέλο SES με α = 0, το οποίο είναι απλώς ένα σταθερό μοντέλο επειδή η πρόβλεψη δεν ενημερώνεται ποτέ. Αυτό σημαίνει ότι όταν το θ1 είναι ίσο με 1, στην πραγματικότητα ακυρώνει τη λειτουργία διαφοροποίησης που επιτρέπει συνήθως στην πρόβλεψη SES να αγκυρωθεί ξανά στην τελευταία παρατήρηση. Από την άλλη πλευρά, εάν ο συντελεστής κινούμενου μέσου όρου είναι 0, αυτό το μοντέλο μειώνεται σε ένα μοντέλο τυχαίας πορείας - δηλαδή, αφήνει μόνη της τη λειτουργία διαφοροποίησης. Έτσι, εάν το θ1 είναι κάτι μεγαλύτερο από 0, είναι σαν να ακυρώνουμε εν μέρει μια σειρά διαφοροποίησης. Εάν η σειρά είναι ήδη ελαφρώς υπερβολική - δηλαδή, εάν έχει εισαχθεί αρνητική αυτοσυσχέτιση - τότε θα "ζητήσει" μια διαφορά να ακυρωθεί εν μέρει εμφανίζοντας μια υπογραφή MA. Εξ ου και ο ακόλουθος πρόσθετος κανόνας:

Κανόνας 7: Εάν το ACF της διαφορετικής σειράς εμφανίζει απότομη αποκοπή ή / και η αυτοσυσχέτιση lag-1 είναι αρνητική - δηλαδή, εάν η σειρά εμφανίζεται ελαφρώς "υπερβολικά διαφοροποιημένη" - τότε δύναται να προστεθεί ένας όρος MA στο μοντέλο. Η καθυστέρηση στην οποία διακόπτεται το ACF είναι ο υποδεικνυόμενος αριθμός όρων MA.

Έτσι, για παράδειγμα, ας θεωρηθεί ότι το "σωστό" μοντέλο για μια χρονική σειρά είναι ένα μοντέλο ARIMA (0,1,1), αλλά αντ 'αυτού χρησιμοποιείται ένα μοντέλο ARIMA (1,1,2), δηλαδή συμπεριλαμβάνεται ένας επιπλέον όρος AR και ένας επιπλέον όρος MA. Τότε οι πρόσθετοι όροι μπορεί να καταλήξουν να φαίνονται σημαντικοί στο μοντέλο, αλλά εσωτερικά μπορεί απλώς να λειτουργούν ο ένας εναντίον του άλλου. Οι προκύπτουσες εκτιμήσεις παραμέτρων μπορεί να είναι διφορούμενες και η διαδικασία εκτίμησης παραμέτρων μπορεί να πάρει πάρα πολλές (π.χ. περισσότερες από 10) επαναλήψεις για σύγκλιση. Ως εκ τούτου:

Κανόνας 8: Είναι πιθανό ένας όρος AR και ένας όρος MA να ακυρώσουν τα αποτελέσματα του άλλου, οπότε αν ένα μικτό μοντέλο AR-MA φαίνεται να ταιριάζει

στα δεδομένα, δοκιμάστε επίσης ένα μοντέλο με έναν μικρότερο όρο AR και έναν λιγότερο όρο MA ιδιαίτερα εάν οι εκτιμήσεις των παραμέτρων στο αρχικό μοντέλο απαιτούν περισσότερες από 10 επαναλήψεις για σύγκλιση.

Για αυτόν τον λόγο, τα μοντέλα ARIMA δεν μπορούν να προσδιοριστούν με προσέγγιση "απο πίσω προς τα εμπρός" που περιλαμβάνει τόσο τους όρους AR όσο και τους MA. Με άλλα λόγια, δεν μπορούν να ξεκινήσουν να συμπεριλαμβάνονται διάφοροι όροι κάθε είδους και στη συνέχεια να πεταχτούν αυτοί των οποίων οι εκτιμώμενοι συντελεστές δεν είναι σημαντικοί. Αντ 'αυτού, συνήθως ακολουθείτε μια "εμπρός σταδιακή" προσέγγιση, προσθέτοντας όρους του ενός ή του άλλου, όπως υποδεικνύεται από την εμφάνιση των γραφημάτων ACF και PACF.

Εάν μια σειρά είναι υπερβολικά υπό- ή υπερ διαφοροποιημένη, δηλαδή εάν πρέπει να προστεθεί ή να ακυρωθεί μια ολόκληρη σειρά διαφοροποιήσεων, αυτό σηματοδοτείται συχνά από μια "μονάδα ρίζας" στους εκτιμώμενους συντελεστές AR ή MA του μοντέλου. Ένα μοντέλο AR (1) λέγεται ότι έχει ρίζα μονάδας εάν ο εκτιμώμενος συντελεστής AR (1) είναι σχεδόν ακριβώς ίσος με το 1. (Με το "ακριβώς ίσο" εννοώ ότι δεν είναι σημαντικά διαφορετικός από, όσον αφορά το τυπικό σφάλμα του ίδιου του συντελεστή .) Όταν συμβεί αυτό, σημαίνει ότι ο όρος AR (1) μιμείται ακριβώς μια πρώτη διαφορά, οπότε θα πρέπει να αφαιρεθεί ο όρος AR (1) και να προστεθεί μια σειρά διαφοροποίησης. Σε ένα μοντέλο AR υψηλότερης τάξης, υπάρχει μια ρίζα μονάδας στο τμήμα AR του μοντέλου εάν το άθροισμα οι συντελεστές AR είναι ακριβώς ίσοι με 1. Σε αυτήν την περίπτωση θα πρέπει να μειωθεί η σειρά του όρου AR κατά 1 και να προστεθεί μια σειρά διαφοροποίησης. Μια χρονική σειρά με ρίζα μονάδας στους συντελεστές AR είναι μη στατική - δηλαδή, χρειάζεται υψηλότερη σειρά διαφοροποίησης.

Κανόνας 9: Εάν υπάρχει ρίζα μονάδας στο τμήμα AR του μοντέλου, δηλαδή εάν το άθροισμα των συντελεστών AR είναι σχεδόν ακριβώς 1, θα πρέπει να μειωθεί ο αριθμός των όρων AR κατά έναν και να αυξηθεί η σειρά των διαφορών κατά ένας.

Ομοίως, ένα μοντέλο MA (1) λέγεται ότι έχει ρίζα μονάδας εάν ο εκτιμώμενος συντελεστής MA (1) είναι ακριβώς ίσος με 1. Όταν συμβεί αυτό, σημαίνει ότι ο όρος MA (1) ακυρώνει ακριβώς μια πρώτη διαφορά, στο Σε αυτήν την περίπτωση, θα πρέπει να καταργηθεί ο όρος MA (1) και επίσης να μειωθεί η σειρά των διαφορών

κατά έναν βαθμό. Σε ένα μοντέλο MA υψηλότερης τάξης, υπάρχει μια ρίζα μονάδας εάν το άθροισμα των συντελεστών MA είναι ακριβώς ίσο με 1.

Κανόνας 10: Εάν υπάρχει μια ρίζα μονάδας στο μέρος MA του μοντέλου - δηλαδή, εάν το άθροισμα των συντελεστών MA είναι σχεδόν ακριβώς 1 - θα πρέπει να μειωθεί ο αριθμός των όρων MA κατά έναν και να μειωθεί η σειρά των διαφορών κατά έναν βαθμό.

Για παράδειγμα, εάν ταιριαστεί σε ένα γραμμικό εκθετικό μοντέλο εξομάλυνσης (ένα μοντέλο ARIMA (0,2,2)) όταν ένα απλό εκθετικό μοντέλο εξομάλυνσης (ένα μοντέλο ARIMA (0,1,1)) θα ήταν επαρκές, μπορεί να βρεθεί οτι το άθροισμα των δύο συντελεστών MA είναι σχεδόν ίσο με το 1. Μειώνοντας τη σειρά MA και τη σειρά των διαφορών από το ένα, αποκτάται το καταλληλότερο μοντέλο SES. Ένα μοντέλο πρόβλεψης με ρίζα μονάδας στους εκτιμώμενους συντελεστές MA λέγεται ότι είναι μη αναστρέψιμο, πράγμα που σημαίνει ότι τα υπολείμματα του μοντέλου δεν μπορούν να θεωρηθούν ως εκτιμήσεις του «αληθινού» τυχαίου θορύβου που δημιούργησε τις χρονοσειρές.

Ένα άλλο σύμπτωμα μιας ρίζας μονάδας είναι ότι οι προβλέψεις του μοντέλου μπορεί να "εκραγούν" ή να συμπεριφερθούν αλλιώς παράξενα. Εάν το χρονοδιάγραμμα των μακροπρόθεσμων προβλέψεων του μοντέλου φαίνεται περίεργο, θα πρέπει να ελεγθούν οι εκτιμώμενοι συντελεστές του μοντέλου για την παρουσία μιας ρίζας μονάδας.

Κανόνας 11: Εάν οι μακροπρόθεσμες προβλέψεις εμφανίζονται ακανόνιστες ή ασταθείς, ενδέχεται να υπάρχει ρίζα μονάδας στους συντελεστές AR ή MA.

Μεθοδολογία επιλογής Παραμέτρων για μοντέλα SARIMA

Ένα μοντέλο SARIMA αποτελείται από δύο μέρη:

- ένα μη εποχιακό μέρος
- ένα εποχιακό μέρος.

Το SARIMA σημειώνεται ARIMA (p, d, q) x (P, D, Q).

Ένα μοντέλο SARIMA αποτελείται από τέσσερα κομμάτια,

- ***Εποχιακότητα (S):*** Λογαριασμοί για εποχικότητα που συμβαίνουν σε μια καθορισμένη χρονική περίοδο.

- ***Autoregressive (AR):*** Λογαριασμός για τυχόν μακροπρόθεσμες τάσεις στα δεδομένα παλινδρομώντας μελλοντικές τιμές σε προηγούμενες τιμές.
- ***Ενσωματωμένο (I):*** Εξασφαλίζει σταθερότητα στα δεδομένα, που αποτελεί παραδοχή του μοντέλου SARIMA.
- ***Κινητός μέσος όρος (MA):*** Λογαριασμός για τυχόν ξαφνικά σοκ στα δεδομένα υποχωρώντας μελλοντικές τιμές σε προηγούμενα σφάλματα.

Για να διακρίνουμε ποια κομμάτια θα συμπεριλάβουμε, είναι χρήσιμο να δούμε την εποχική αποσύνθεση των δεδομένων. Ένα μεγάλο πράγμα για το οποίο μπορούμε να χρησιμοποιήσουμε την ανάλυση χρονοσειρών είναι να την αποσυνθέσουμε. Για παράδειγμα, στο παρακάτω γράφημα χρονοσειρών, μπορεί να παρατηρηθεί ότι έχει ανοδική τάση, αλλά έχει και κάποια εποχικότητα σε αυτό. Θα ήταν χρήσιμο αν μπορούσε να γίνει αποσυνθέση αυτής της χρονικής σειράς στα συστατικά της.

Η εποχιακή αποσύνθεση επιτρέπει να διασπαστούν (ή να «αποσυνθεθούν») δεδομένα χρονοσειρών σε εποχιακά στοιχεία, τάσεις και υπολείμματα. Αναλύοντας αυτά τα στοιχεία, είμαστε σε θέση να εντοπίσουμε ορισμένα κομμάτια του μοντέλου SARIMA που θα συμπεριλάβουμε. Δύναται να το πραγματοποιηθεί χρησιμοποιώντας μια βιβλιοθήκη που ονομάζεται statsmodels. Συγκεκριμένα χρησιμοποιώντας μια συνάρτηση που ονομάζεται seasonal_decompose. Αυτή η συνάρτηση αποσυντίθεται έπειτα σε τρία μέρη:

Τάση: Αυτή είναι η γενική κατεύθυνση του πώς η συγκεκριμένη στήλη κάνει υπερωρίες. Μπορούμε να έχουμε ανοδικές τάσεις, οριζόντιες ή σταθερές τάσεις και πτωτικές τάσεις. Μια τάση υπάρχει όταν υπάρχει μακροπρόθεσμη αύξηση ή μείωση στα δεδομένα. Δεν χρειάζεται να είναι γραμμικό. Μερικές φορές θα αναφερόμαστε σε μια τάση ως «αλλαγή κατεύθυνσης», όταν μπορεί να μεταβαίνει από μια αυξανόμενη τάση σε μια φθίνουσα τάση.

Εποχικότητα: Αυτό δείχνει πώς η στήλη-στόχος μας διαφέρει ανάλογα με το χρόνο. Σαν παράδειγμα μπορεί να ληφθεί η γη της Disney κατά τη διάρκεια των θερινών διακοπών έναντι ας πούμε τον Φεβρουάριο ή στα τέλη Οκτωβρίου. Προφανώς οι καλοκαιρινοί μήνες είναι πολύ πιο πολυσύχναστοι. Ένα εποχιακό μοτίβο εμφανίζεται όταν μια χρονοσειρά επηρεάζεται από εποχιακούς παράγοντες, όπως η ώρα του έτους ή η ημέρα της εβδομάδας. Η εποχικότητα είναι πάντα σταθερή και γνωστή.

Σφάλμα ή θόρυβος: Ως επιστήμονες δεδομένων που γνωρίζουμε στον πραγματικό κόσμο, υπάρχουν πάντα ανεξήγητες και απροσδόκητες παραλλαγές που αποδίδουμε σε θόρυβο ή λάθη.

Πολλές χρονολογικές σειρές περιλαμβάνουν τάση, κύκλους και εποχικότητα. Κατά την επιλογή μιας μεθόδου πρόβλεψης, θα πρέπει πρώτα να προσδιορίσουμε τα μοτίβα χρονοσειρών στα δεδομένα και, στη συνέχεια, να επιλέξουμε μια μέθοδο που μπορεί να καταγράψει σωστά τα μοτίβα.

Στην ανάλυση και την πρόβλεψη χρονοσειρών, συνήθως θεωρείται ότι τα δεδομένα είναι ένας συνδυασμός τάσης, εποχικότητας και θορύβου και θα μπορούσαω να σχηματίσουω ένα μοντέλο πρόβλεψης, αποτυπώνοντας τα καλύτερα από αυτά τα στοιχεία. Συνήθως, υπάρχουν δύο μοντέλα αποσύνθεσης για χρονοσειρές: πρόσθετο και πολλαπλασιαστικό. Πιστεύεται ότι το πρόσθετο μοντέλο είναι χρήσιμο όταν η εποχιακή διακύμανση είναι σχετικά σταθερή με την πάροδο του χρόνου, ενώ το πολλαπλασιαστικό μοντέλο είναι χρήσιμο όταν η εποχιακή διακύμανση αυξάνεται με την πάροδο του χρόνου.

Τα προβλήματα του πραγματικού κόσμου είναι ακατάστατα και ο θόρυβος, όπως πχ η τάση δεν είναι μονότονη και το πραγματικό μοντέλο θα μπορούσε να έχει τόσο πρόσθετα όσο και πολλαπλασιαστικά στοιχεία. Ωστόσο, αυτά τα μοντέλα αποσύνθεσης παρέχουν έναν δομημένο και απλό τρόπο ανάλυσης και πρόβλεψης των δεδομένων.

Ως εκ τούτου, ο εντοπισμός της εποχικότητας σε μια χρονική σειρά θα μπορούσε να βοηθήσει να δημιουργηθεί ένα καλύτερο μοντέλο. Αυτό μπορεί να συμβεί με τους ακόλουθους τρόπους:

- ***Καθαρισμός δεδομένων:*** η αφαίρεση της εποχιακής συνιστώσας θα σας δώσει μια σαφέστερη σχέση μεταξύ εισόδου και εξόδου.
- ***Ερμηνευσιμότητα:*** παρέχει περισσότερες πληροφορίες για χρονοσειρές.

Το εποχιακό μέρος ενός μοντέλου ARIMA έχει την ίδια δομή με το μη εποχιακό μέρος: μπορεί να έχει έναν παράγοντα AR, έναν παράγοντα MA και / ή μια σειρά διαφορών. Στο εποχιακό μέρος του μοντέλου, όλοι αυτοί οι παράγοντες λειτουργούν σε πολλαπλάσια του lag s (ο αριθμός των περιόδων σε μια σεζόν).

Ένα εποχιακό μοντέλο ARIMA ταξινομείται ως μοντέλο ARIMA (p, d, q) x (P, D, Q), όπου P = αριθμός εποχιακών όρων αυτοεπιθετικής (SAR), D = αριθμός εποχιακών διαφορών, Q = αριθμός εποχιακών Όροι κινούμενου μέσου όρου (SMA) Κατά τον προσδιορισμό ενός εποχιακού μοντέλου, το πρώτο βήμα είναι να προσδιοριστεί αν απαιτείται εποχική διαφορά, επιπλέον ή ίσως αντί μιας μη εποχιακής διαφοράς. Θα πρέπει να ελεγχθούν τα χρονικά διαγράμματα και τα γραφήματα ACF και PACF για όλους τους πιθανούς συνδυασμούς 0 ή 1 μη εποχιακής διαφοράς και 0 ή 1 εποχιακής διαφοράς.

Προσοχή: Μην χρησιμοποιηθούν ΠΟΤΕ περισσότερες από ΜΙΑ εποχιακές διαφορές, ούτε περισσότερες από ΔΥΟ συνολικές διαφορές (εποχιακές και μη εποχιακές συνδυασμένες). Εάν το εποχιακό μοτίβο είναι ισχυρό και σταθερό με την πάροδο του χρόνου (π.χ. υψηλό το καλοκαίρι και χαμηλό το χειμώνα ή το αντίστροφο), τότε πιθανότατα θα πρέπει να χρησιμοποιηθεί μια εποχιακή διαφορά ανεξάρτητα από το αν χρησιμοποιείται μια μη εποχιακή διαφορά, καθώς αυτό θα αποτρέψει το εποχιακό μοτίβο να «πεθάνει» στις μακροπρόθεσμες προβλέψεις. Ας προσθέθεί αυτό στη λίστα κανόνων για τον προσδιορισμό μοντέλων.

Κανόνας 12: Εάν η σειρά έχει ένα ισχυρό και συνεπές εποχιακό μοτίβο, τότε θα πρέπει να χρησιμοποιηθεί μια σειρά εποχιακών διαφορών - αλλά ποτέ να μην χρησιμοποιηθούν περισσότερες από μία βαθμίδες εποχιακών διαφορών ή περισσότερες από 2 βαθμίδες συνολικής διαφοράς (εποχιακά + μη εποχιακά).

Η υπογραφή της καθαρής συμπεριφοράς SAR ή της καθαρής SMA είναι παρόμοια με την υπογραφή της καθαρής συμπεριφοράς AR ή της καθαρής MA, εκτός από το ότι το μοτίβο εμφανίζεται σε πολλαπλάσια υστέρησης στο ACF και PACF. Για παράδειγμα, μια καθαρή διαδικασία SAR (1) έχει αιχμές στο ACF σε υστέρηση s, 2s, 3s, κλπ., Ενώ το PACF διακόπτεται μετά την καθυστέρηση. Αντίθετα, μια καθαρή διαδικασία SMA (1) έχει αιχμές στο PACF σε υστέρηση s, 2s, 3s, κ.λπ., ενώ η ACF διακόπτεται μετά την καθυστέρηση. Μια υπογραφή SAR εμφανίζεται συνήθως όταν η αυτοσυσχέτιση κατά την εποχική περίοδο είναι θετική, ενώ η υπογραφή SMA εμφανίζεται συνήθως όταν η εποχική αυτοσυσχέτιση είναι αρνητική, επομένως:

Κανόνας 13: Εάν η αυτοσυσχέτιση κατά την εποχική περίοδο είναι θετική, δύναται να προσθεθεί ένας όρος SAR στο μοντέλο. Εάν η αυτοσυσχέτιση κατά την εποχική περίοδο είναι αρνητική, δύναται να προσθεθεί ένας όρος SMA στο μοντέλο.

Δεν συνίσταται ο συνδυασμός όρων SAR και SMA στο ίδιο μοντέλο και η χρήση περισσότερων από έναν από τους δύο τύπους.

Συνήθως αρκεί ένας όρος SAR (1) ή SMA (1). Σπάνια θα βρεθεί μια γνήσια διαδικασία SAR (2) ή SMA (2), και ακόμη πιο σπάνια υπάρχουν αρκετά δεδομένα για να εκτιμηθούν 2 ή περισσότεροι εποχιακοί συντελεστές χωρίς ο αλγόριθμος εκτίμησης να μπει σε ένα «βρόχο ανάδρασης».

Αν και ένα εποχιακό μοντέλο ARIMA φαίνεται να έχει μόνο λίγες παραμέτρους, πρέπει να ληφθεί υπόψιν ότι το back-forecasting απαιτεί την εκτίμηση μιας έμμεσης παραμέτρου αξίας μίας ή δύο εποχών για την αρχικοποίησή του. Επομένως, θα πρέπει να υπάρχου διαθέσιμες τουλάχιστον 4 ή 5 σεζόν δεδομένων για να ταιριάζει σε ένα εποχιακό μοντέλο ARIMA.

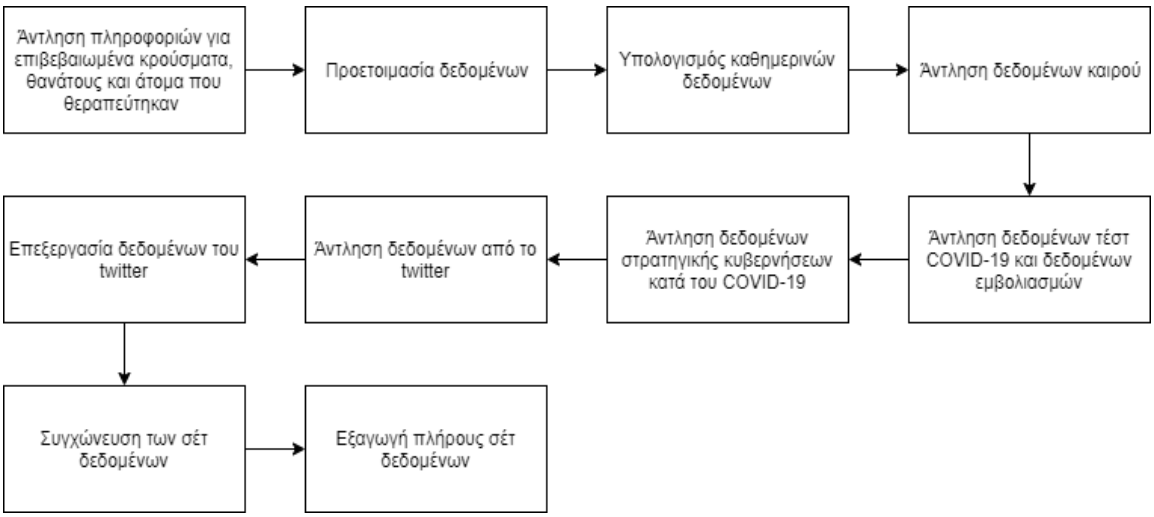
Πιθανώς το πιο συχνά χρησιμοποιούμενο εποχιακό μοντέλο ARIMA είναι το μοντέλο (0,1,1) x (0,1,1) - δηλαδή, ένα μοντέλο MA (1) xSMA (1) με εποχιακή και μη εποχική διαφορά . Αυτό είναι ουσιαστικά ένα μοντέλο «εποχιακής εκθετικής εξομάλυνσης». Όταν τα εποχιακά μοντέλα ARIMA είναι προσαρμοσμένα σε καταγεγραμμένα δεδομένα, είναι σε θέση να παρακολουθούν ένα πολλαπλασιαστικό εποχιακό μοτίβο.

ΚΕΦΑΛΑΙΟ 4

Η συνάρτηση creatCovidDataframe()

Η συνάρτηση creatCovidDataframe() συγκεντρώνει, καθαρίζει, ομαδοποιεί και προετοιμάζει τα δεδομένα για την χώρα που δόθηκε σαν όρισμα καθώς επίσης δημιουργεί και έξτρα διαστάσεις στα δεδομένα συνδυάζοντας υπάρχοντα.

Παρακάτω δίνεται ένα γενικό σχεδιάγραμμα με την ροή εργασίας που ακολουθεί ο αλγόριθμος κατά την εκτέλεση σου



ΕΙΚΌΝΑ 1: Διαδικασία εκτέλεσης αλγορίθμου συνάρτησης

Η διαδικασία που ακολουθείται μπορεί να χωριστεί σε 9 βήματα:

Βήμα 1:

Στο βήμα αυτό πραγματοποιείται άντληση πληροφοριών σχετικά με τα επιβεβαιωμένα κρούσματα, τους θανάτους και τα άτομα που θεραπευτήκαν, από την βάση γνώσης στο GitHub του πανεπιστημίου Johns Hopkins. Κάθε ένα από τα σετ δεδομένων ανατίθενται σε ένα pandas Datframe. Τα σετ αυτά περιέχουν δεδομένα, καθημερινά με το αθροιστικό αριθμό μέχρι εκείνη την ημέρα. Στο τέλος του πρώτου βήματος έχουν δημιουργηθεί 3 σετ δεδομένων με ονόματα confirmed_df, deaths_df και recovered_df.

Βήμα 2:

Τα τρία σετ δεδομένων που έχουν δημιουργηθεί μέχρι τώρα δεν έχουν μορφή που είναι ευκολά αξιοποιήσιμη σε αναλύσεις δεδομένων. Στο δεύτερο αυτό βήμα λοιπόν, στόχος είναι και τα τρία σετ δεδομένων να έρθουν σε μια μορφή χρονοσειράς. Αρχικά όλες οι κενές τιμές αντικαθίστανται από την κενή συμβολοσειρά. Στη συνέχεια συγχωνεύονται οι στήλες «Province/State» και «Country/Region» στην στήλη «Country/Region» και η στήλη «Province/State» διαγράφεται από τα σετ δεδομένων. Έπειτα δημιουργείται ένα επιπλέον σετ δεδομένων με όνομα «coordinates_df». Το σετ αυτό περιέχει τις στήλες «Lat», «Long» και

«Country/Region» και θα λειτουργεί ως αρχείο για τις γεωγραφικές συντεταγμένες των χωρών. Τέλος το σετ δεδομένων «coordinates_df» περιστρέφεται ως προς το πρώτο στοιχείο και τα ονόματα των χωρών γίνονται τα ονόματα των στηλών που πλέον περιέχουν μόνο το γεωγραφικό πλάτος και το γεωγραφικό μήκος

Συνεχίζοντας από τα τρία αρχικά σετ δεδομένων διαγράφονται πλέον οι στήλες «Lat» και «Long» και τα σετ περιστρέφονται ως προς το πρώτο στοιχείο. Στη συνέχεια τα ονόματα των χωρών θέτονται ως ονόματα στις στήλες και ο αριθμός ευρετηρίου γίνεται οι ημερομηνίες που ήταν ονόματα των στηλών πριν την περιστροφή.

Είναι σημαντικό να σημειωθεί ότι τα δεδομένα προέρχονται από διαφορετικές πηγές, με διαφορετικές χώρες να είναι διαθέσιμες σε κάθε μια από αυτές. Για το λόγο αυτό στο συγκεκριμένο σημείο διαγράφονται συγκεκριμένες χώρες από τα σετ δεδομένων. Με σκοπό να προσφέρονται ολοκληρωμένα σετ δεδομένων σαν output της συνάρτησης. Οι χώρες που διαγράφονται είναι επιλεγμένες μετρά από διερεύνηση και σύγκριση με κάθε μια από τις υπόλοιπες βάσεις ξεχωριστά. Επιπλέον μερικές χώρες διαθέτουν δεδομένα μόνο σε επίπεδο νομών. Για αυτές τις χώρες δημιουργείται μια νέα εγγραφή που περιέχει το άθροισμα όλων των νομών της και διαγράφονται οι υπόλοιπες εγγραφές.

Τέλος, σε αυτό το βήμα πραγματοποιείται έλεγχος εάν η χωρά την οποία έχει δώσει ο χρήστης σαν όρισμα υπάρχει στις χώρες που είναι διαθέσιμες προκειμένου να προχωρήσει στα επόμενα βήματα η να ενημερώσει τον χρήστη αναλόγως.

Βήμα 3:

Στο βήμα αυτό πραγματοποιείται υπολογισμών των καθημερινών δεδομένων σχετικά με τα κρούσματα, τους θανάτους και τα άτομα που θεραπευτήκαν. Για τον υπολογισμό αυτό αφαιρείται από κάθε τιμή μιας μέρας η τιμή της προηγουμένης. Με τα δεδομένα που προκύπτουν και με βάση από ποιο σετ δεδομένων υπολογίζονται δημιουργούνται τρία νέα σετ δεδομένων, τα «confirmed_daily_df», «deaths_daily_df» και «recovered_daily_df».

Βήμα 4:

Στο βήμα τέσσερα συλλέγονται τα δεδομένα καιρού για την χωρά που έχει δώσει ο χρήστης σαν όρισμα, γίνεται επεξεργασία τους και δημιουργούνται καθημερινά δεδομένα και δεδομένα μέσου ορού. Τα δεδομένα καιρού τα παρέχει το meteostat και είναι διαθέσιμα ανά ώρα. Με βάση την χωρά που έχει δώσει ο χρήστης σαν όρισμα παρέχεται στο API του meteostat το γεωγραφικό πλάτος και μήκος από το σετ «coordinates_df». Για να ταιριάξουν τα δεδομένα αυτά με τα καθημερινά δεδομένα υπολογίζεται ο μέσος ορός των θερμοκρασιών από τις 24 ώρες της κάθε ημέρας και δημιουργεί μια νέα καταχώρηση για την αντίστοιχη ημέρα σε ένα νέο σετ δεδομένων το «avg_temp_daily_df».

Στη συνέχεια υπολογίζεται η μέση θερμοκρασία 7 ημέρων για δυο εβδομάδες πίσω για κάθε ημέρα. Συγκεκριμένα, για κάθε δεδομένη ημερομηνία υπολογίζεται ο μέσος ορός της θερμοκρασίας από την ημέρα -14 έως -7 και αποθηκεύεται στην αντίστοιχη εγγραφή ημέρας στο σετ «avg_temp_daily_df».

Η ιδιά διαδικασία ακολουθείται και για τα δεδομένα υγρασίας και ανέμου που αποθηκεύονται στο ίδιο σετ δεδομένων με αυτά της θερμοκρασίας.

Τέλος, πραγματοποιείται μια συγχώνευση των σετ «confirmed_df» , «deaths_df», «recovered_df», «confirmed_daily_df», «deaths_daily_df», «recovered_daily_df» και «avg_temp_daily_df» στο σετ «end_rslt_df». Η συγχώνευση πραγματοποιείται πάνω στην ημερομηνία - ευρετήριο που έχει κάθε σετ δεδομένων για τις εγγραφές του.

Βήμα 5:

Στο βήμα πέντε συλλέγονται δεδομένα σχετικά με τα καθημερινά τεστ για COVID-19 και σχετικά με τους εμβολιασμούς που έχουν πραγματοποιηθεί. Από το σετ φιλτράρονται μόνο τα δεδομένα που αφορούν την χωρά που έχει δώσει ο χρήστης σαν όρισμα και δημιουργούν το σετ δεδομένων «testing_full_df» από το οποίο διαγράφονται στήλες που δεν χρειάζονται.

Στη συνέχεια, έχοντας διαθέσιμα δεδομένα σχετικά με τον αριθμό των τεστ που πραγματοποιήθηκαν και τον αριθμό των επιβεβαιωμένων κρουσμάτων είναι εύκολο να υπολογιστεί το καθημερινό ποσοστό θετικότητας των τεστ για COVID-19. Το αποτέλεσμα ωστόσο, καθώς τα δεδομένα είναι από διαφορετικές πηγές και λόγω άλλων παραγόντων, περιέχει άπειρες τιμές. Οι άπειρες τιμές αντικαθιστώντας με την

κενή τιμή και πραγματοποιείται παρεμβολή στην χρονοσειρά για να απομακρυνθούν οι κενές τιμές και να γίνει η χρονοσειρά συνεχής.

Επιπλέον, δημιουργούνται τρεις επιπρόσθετες στήλες που πηγάζουν από τη στήλη με το ποσοστό θετικότητας. Οι στήλες αυτές είναι η στήλη του ποσοστού θετικότητας των τεστ αλλά μετρά από εφαρμογή συντελεστή εξομάλυνσης ίσο με 0.3, 0.5 και 0.8 αντίστοιχα.

Τέλος, συλλέγονται δεδομένα σχετικά με τους εμβολιασμούς κατά του COVID-19. Από το σετ φιλτράρονται μόνο τα δεδομένα που αφορούν την χωρά που έχει δώσει ο χρήστης σαν όρισμα και εισάγονται στο σετ δεδομένων «testing_full_df» από το οποίο διαγράφονται στήλες που δεν χρειάζονται. Τα σετ δεδομένων των τεστ και των εμβολιασμών συγχωνεύονται στο σετ «vaccinations_full_df».

Βήμα 6:

Στο βήμα έξι δημιουργείται ένα νέο σετ δεδομένων με όνομα «gov_index» στο οποίο αποθηκεύονται δεδομένα που συλλέγονται σχετικά με την στρατηγική των κυβερνήσεων απέναντι στον COVID-19. Από το σετ φιλτράρονται μόνο τα δεδομένα που αφορούν την χωρά που έχει δώσει ο χρήστης σαν όρισμα και εισάγονται στο σετ δεδομένων «gov_index». Στη συνέχεια το σετ «gov_index» συγχωνεύεται μέσα στο «vaccinations_full_df».

Βήμα 7:

Στο βήμα επτά συλλέγονται τα δεδομένα με τον αριθμό χρήσης ορών στο twitter και αποθηκεύονται στο σετ δεδομένων «terms_usage_df». Τα δεδομένα για κάθε μέρα είναι σε διαφορετικό αρχείο σε διαφορετικό φάκελο στην βάση στο GitHub. Για το λόγο αυτό δημιουργείται ένας πίνακας με όλες τις ημερομηνίες από την αρχή της καταγραφής των όρων μέχρι και την ημέρα που τρέχει ο αλγόριθμος. Στη συνέχεια αυτές οι ημερομηνίες χρησιμοποιούνται για να δημιουργήσουν τις διαδρομές στους υποφακέλους και να διαβάσουν κάθε αρχείο. Εάν δεν υπάρχει αρχείο σημαίνει ότι μέχρι εκείνη την ημερομηνία έχουν ενημερωθεί τα δεδομένα. Στο τέλος της εκτέλεσης έχει γεμίσει ένα σετ δεδομένων με τα παγκόσμια δεδομένα χρήσης ορών σχετικών με τον COVID-19 στο twitter με όνομα «terms_usage_df».

Βήμα 8:

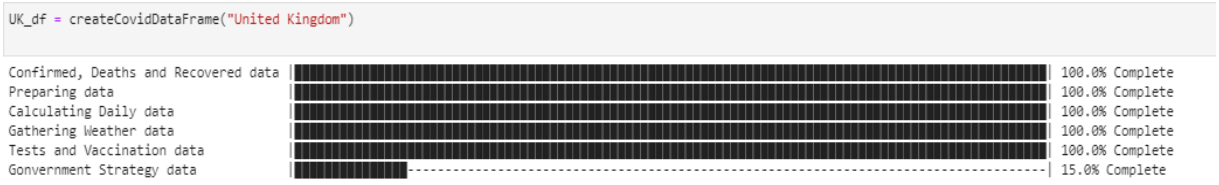
Στο βήμα αυτό δημιουργούνται επιπλέον στήλες/μεταβλητές στη χρονοσειρά βασιζόμενες στα δεδομένα του twitter. Συγκεκριμένα, υπολογίζεται η μέση χρήση ορών 7 ημέρων για δυο εβδομάδες πίσω για κάθε ημέρα. Δηλαδή, για κάθε δεδομένη ημερομηνία υπολογίζεται ο μέσος ορός της θερμοκρασίας από την ημέρα -14 έως -7 και αποθηκεύεται στην αντίστοιχη εγγραφή ημέρας στο σετ «terms_usage_df».

Βήμα 9:

Στο βήμα εννέα γίνεται η τελική συγχώνευση όλων των σετ δεδομένων σε ένα τελικό σετ με όνομα «full_dataset». Το σετ αυτό το επιστρέφει πίσω η συνάρτηση σαν output. Η συγχώνευση πραγματοποιείται πάνω στην ημερομηνία - ευρετήριο που έχει κάθε σετ δεδομένων για τις εγγραφές του.

Παράδειγμα Χρήσης των Δεδομένων

Η συνάρτηση έτρεξε σαν παράδειγμα για την χωρά της Μεγάλης Βρετάνιας. Κατά την διάρκεια της εκτέλεσης η συνάρτηση εκτυπώνει σε ποιο κομμάτι του αλγορίθμου βρίσκεται όπως φαίνεται παρακάτω



ΕΙΚΌΝΑ 2: Output συνάρτησης κατά τη διάρκεια της εκτέλεσης

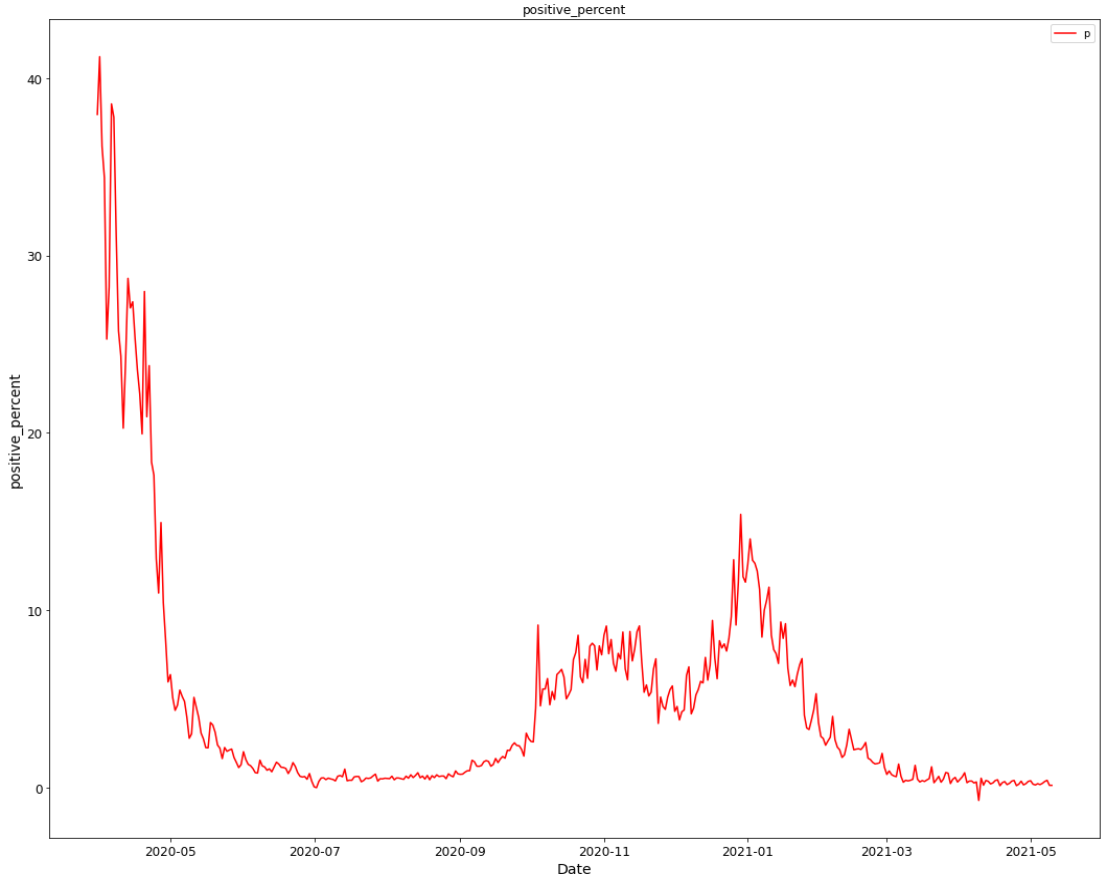
Στο τέλος της εκτέλεσης της επιστρέφει το σετ δεδομένων για τη χωρά που ζητήθηκε. Το σετ περιέχει δεδομένα τα οποία είναι έτοιμα να χρησιμοποιηθούν σε αναλύσεις δεδομένων και σε προβλεπτικά μοντέλα χωρίς ιδιαίτερη επεξεργασία.

Σαν παράδειγμα εφαρμογής επιλέχθηκε η χωρά της Μεγάλης Βρετάνιας για την παραγωγή σετ δεδομένων με σκοπό την εφαρμογή προβλεπτού μοντέλου στα δεδομένα, καθώς επίσης και οπτικοποίηση των μεταβλητών που παράγονται από την συνάρτηση creatCovidDataframe(). Ο λόγος για τον οποίο επιλέχθηκε η χωρά της

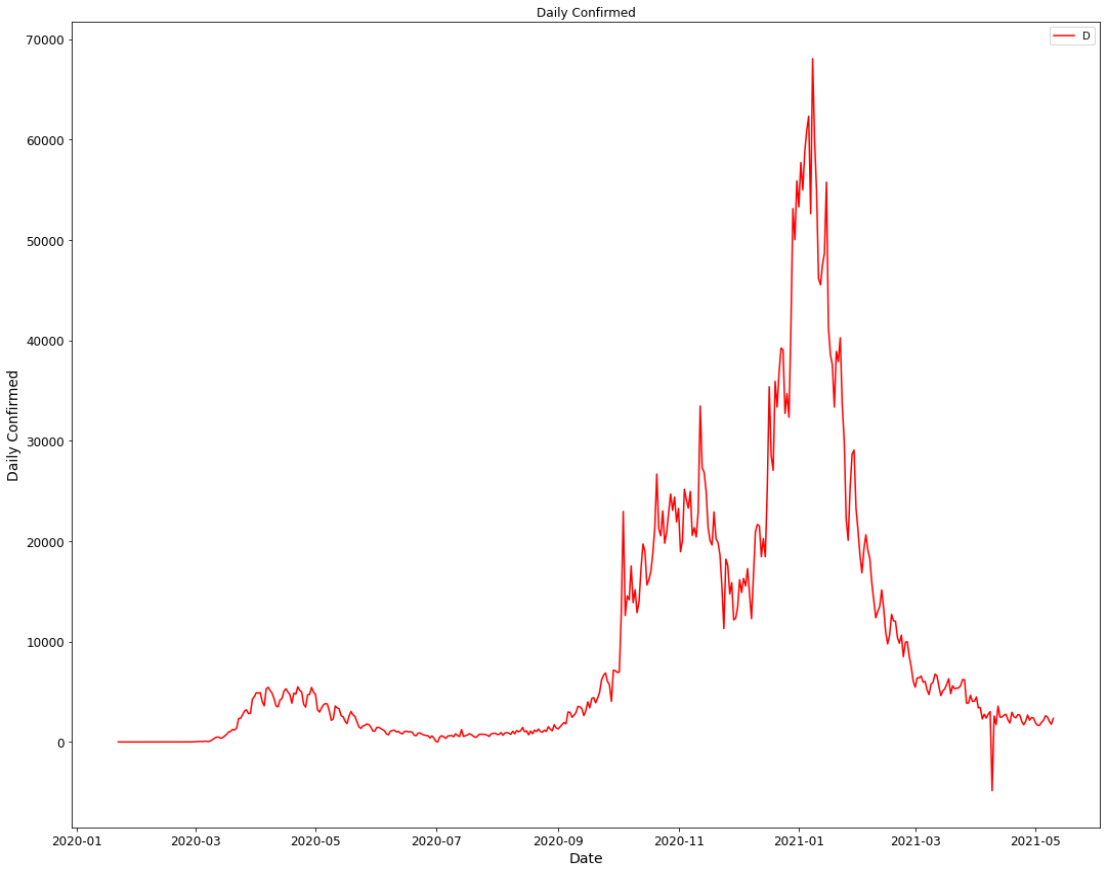
Μεγάλης Βρετάνιας έγκειται στο γεγονός ότι για την χωρά αυτή είναι διαθέσιμα δεδομένα που σχετίζονται με τις εισαγωγές στα νοσοκομεία, δεδομένα που δεν είναι διαθέσιμα σε πολλές χώρες.

Το αποτέλεσμα της συνάρτησης, δηλαδή το σετ δεδομένων, αποθηκεύεται στην μεταβλητή «UK_df». Στη συνέχεια διατρέχοντας όλες τις μεταβλητές που περιέχει το σετ δεδομένων δημιουργούνται σχεδιαγράμματα για κάθε μια από αυτές με σκοπό την οπτικοποίηση τους για κατανόηση της πορείας της πανδημίας για την συγκεκριμένη χωρά. Μερικά από τα πιο ενδιαφέροντα διαγράμματα αναφέρονται παρακάτω.

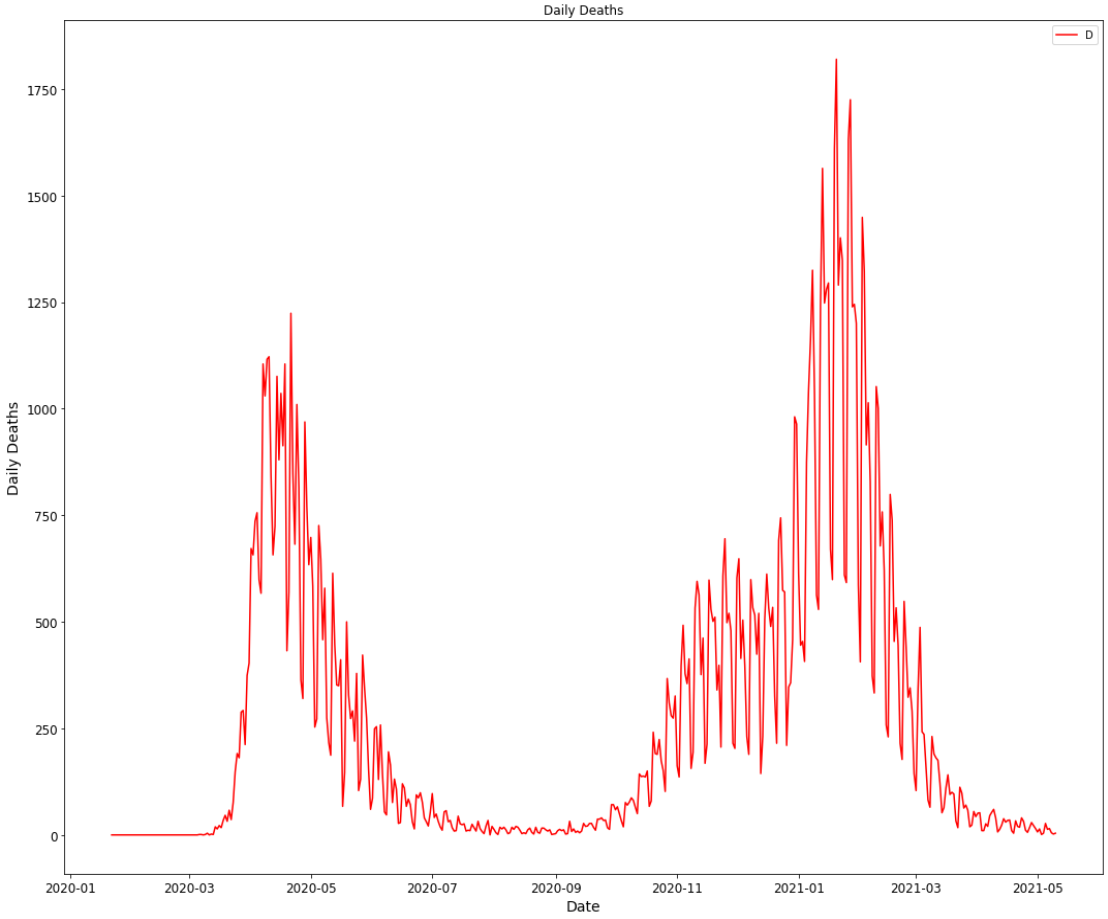
1. Καθημερινό Ποσοστό θετικότητας των τεστ για Covid-19



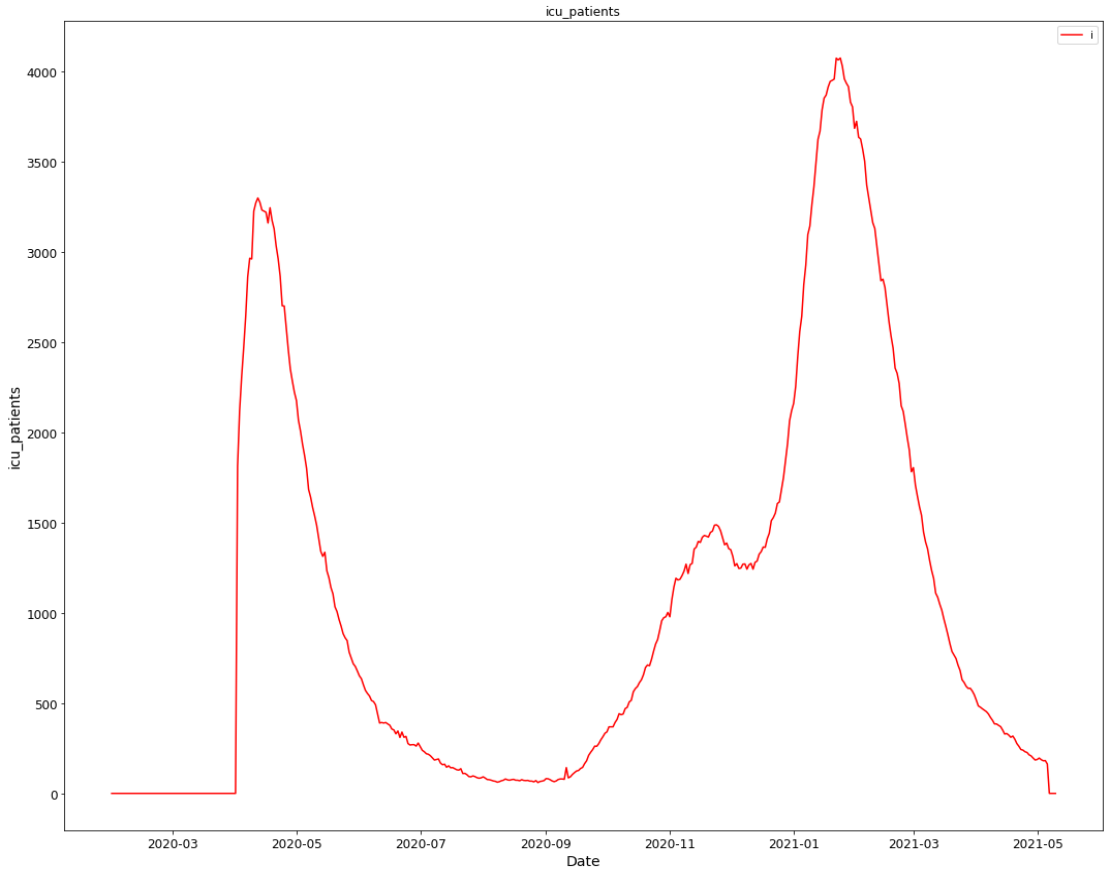
ΕΙΚΌΝΑ 3: Καθημερινό ποσοστό θετικότητας τέστ Covid-19



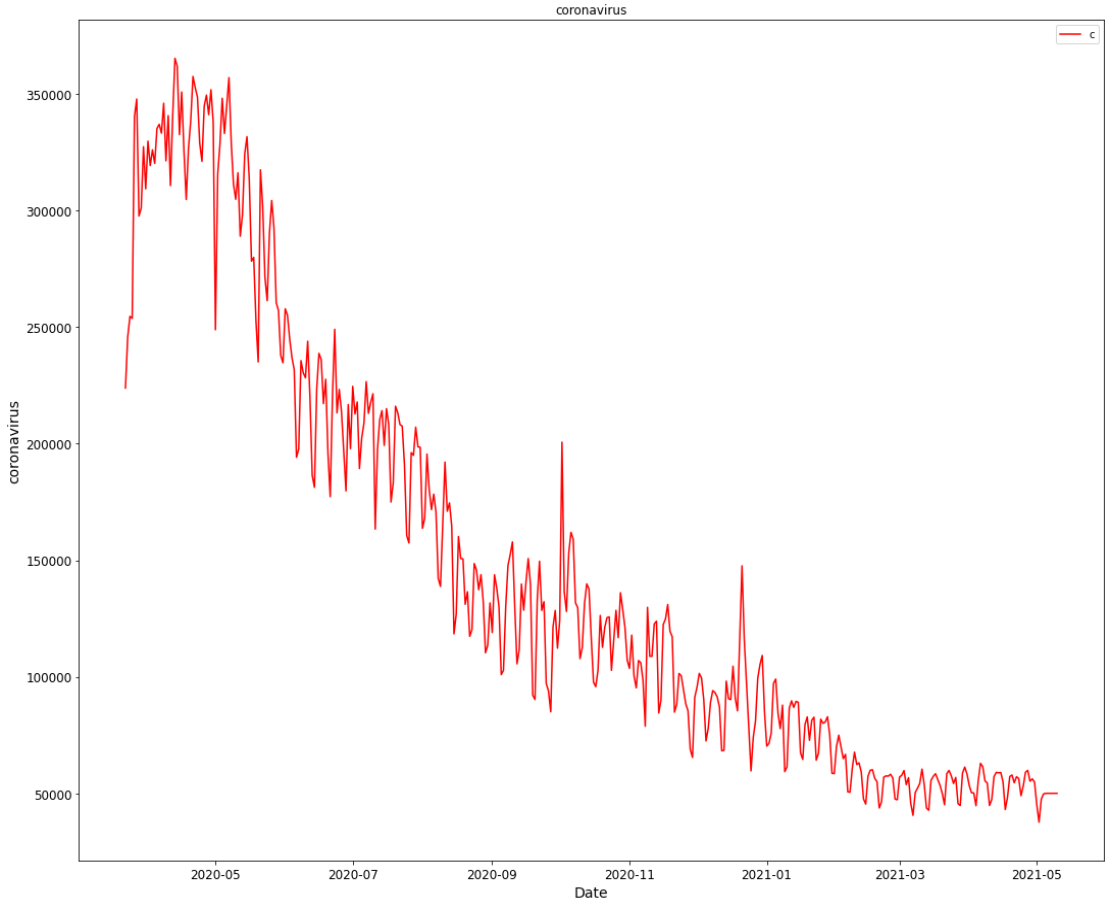
ΕΙΚΌΝΑ 4: Καθημερινός αριθμός επιβεβαιωμένων κρουσμάτων Covid-19



ΕΙΚΌΝΑ 5: Καθημερινός αριθμός επιβεβαιωμένων θανάτων απο Covid-19



ΕΙΚΌΝΑ 6: Καθημερινός αριθμός επιβεβαιωμένων εισαγωγών ασθενών σε Μονάδες Εντατικής Θεραπείας



ΕΙΚΌΝΑ 7: Καθημερινός αριθμός tweets που περιέχουν τον όρο "Coronavirus" (σε παγκόσμιο επίπεδο)

Η μεταβλητή «hosp_patients» επιλέχθηκε για πρόβλεψη από το μοντέλο ARIMA το οποίο δημιουργήθηκε ως παράδειγμα χρήσης των δεδομένων. Ακολουθώντας τις οδηγίες που παρατίθενται στο κεφάλαιο 3, χρησιμοποιώντας τα διαγράμματα ACF και PACF και αξιοποιώντας την συνάρτηση «evaluate_models» αποφασίστηκε ότι το καλύτερο μοντέλο ARIMA(p,d,q) είναι το ARIMA(8,2,2).

Η συνάρτηση «evaluate_models» δημιουργήθηκε για επαλήθευση της απόφασης του «ποιο είναι το καταλληλότερο μοντέλο ARIMA που ταιριάζει στα δεδομένα». Δέχεται τέσσερα ορίσματα τα οποία είναι η χρονοσειρά στην οποία θα εφαρμοστεί το μοντέλο, τις τιμές p θα δοκιμάσει, τις τιμές d που θα δοκιμάσει και τις τιμές q που θα δοκιμάσει. Η συνάρτηση «evaluate_models» δοκιμάζει όλους τους πιθανούς συνδυασμούς των μοντέλων που αποτελούνται από τις τιμές p, d και q που δοθήκαν και τις βαθμολογεί με βάση το Μέσο Τετραγωνικό Σφάλμα. Στο τέλος της εκτέλεσης τυπώνει ποιο μοντέλο έχει το μικρότερο Μέσο Τετραγωνικό Σφάλμα και το προτείνει ως καλύτερο. Το Output της συνάρτησης φαίνεται παρακάτω.

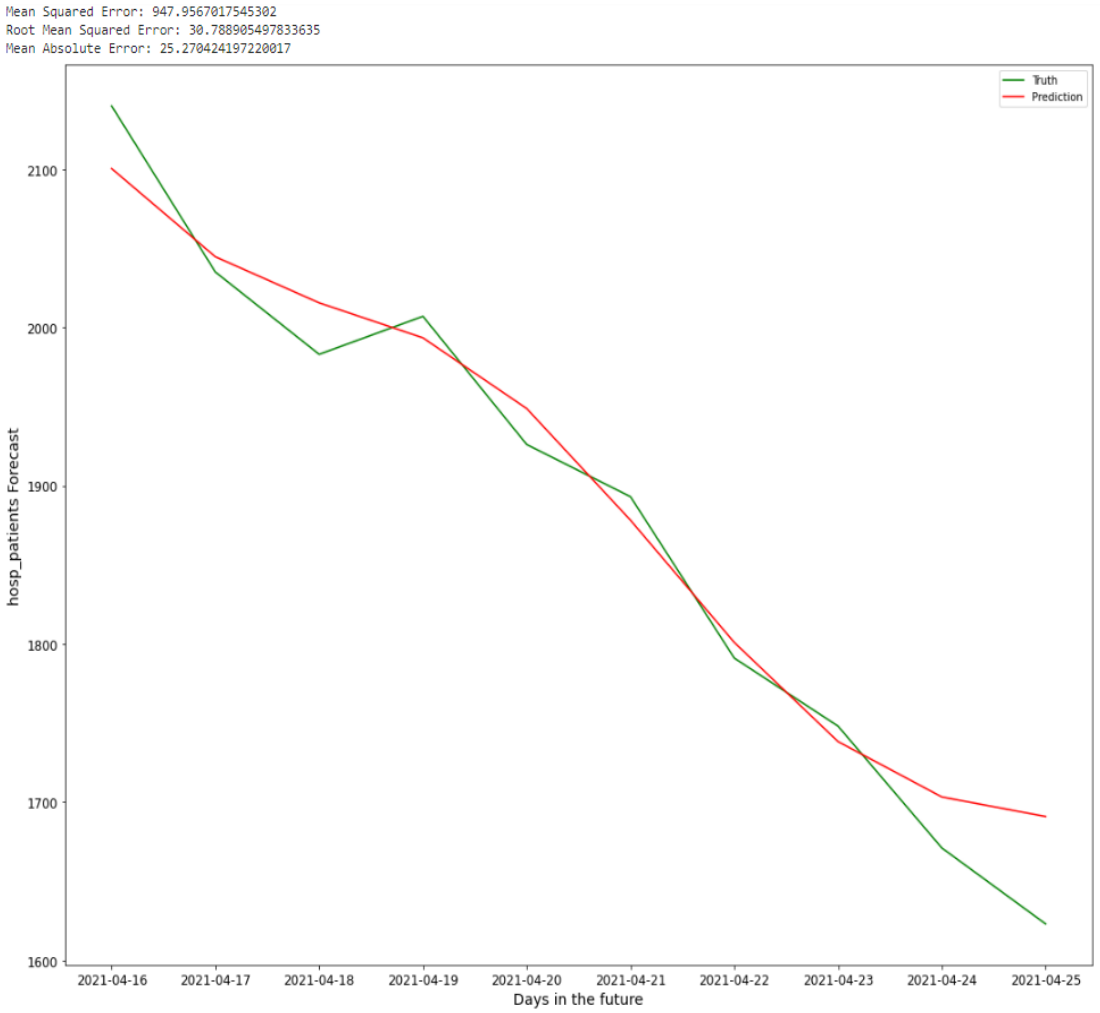
```
ARIMA(6, 0, 0) RMSE=461.884
ARIMA(6, 0, 1) RMSE=461.165
ARIMA(6, 0, 2) RMSE=441.157
ARIMA(6, 1, 0) RMSE=449.375
ARIMA(6, 1, 1) RMSE=430.196
ARIMA(6, 2, 0) RMSE=395.026
ARIMA(6, 2, 1) RMSE=399.107
ARIMA(6, 2, 2) RMSE=385.234
ARIMA(7, 0, 0) RMSE=452.466
ARIMA(7, 0, 1) RMSE=437.619
ARIMA(7, 0, 2) RMSE=444.995
ARIMA(7, 1, 0) RMSE=401.414
ARIMA(7, 1, 1) RMSE=403.152
ARIMA(7, 1, 2) RMSE=388.188
ARIMA(7, 2, 0) RMSE=400.826
ARIMA(7, 2, 1) RMSE=399.199
ARIMA(7, 2, 2) RMSE=381.630
ARIMA(8, 0, 0) RMSE=401.501
ARIMA(8, 0, 1) RMSE=402.023
ARIMA(8, 0, 2) RMSE=391.842
ARIMA(8, 1, 0) RMSE=404.909
ARIMA(8, 1, 1) RMSE=403.122
ARIMA(8, 1, 2) RMSE=384.199
ARIMA(8, 2, 0) RMSE=394.799
ARIMA(8, 2, 1) RMSE=386.356
ARIMA(8, 2, 2) RMSE=377.008
Best ARIMA(8, 2, 2) RMSE=377.008
```

ΕΙΚΌΝΑ 8: Output της συνάρτησης "evaluate_models"

Έπειτα, αφού έχει βρεθεί το πιο αποτελεσματικό μοντέλο για τα δεδομένα, εφαρμόζεται στην χρονοσειρά «hosp_patients» το μοντέλο ARIMA(8,2,2). Το μοντέλο εφαρμόζεται με δυο τρόπους.

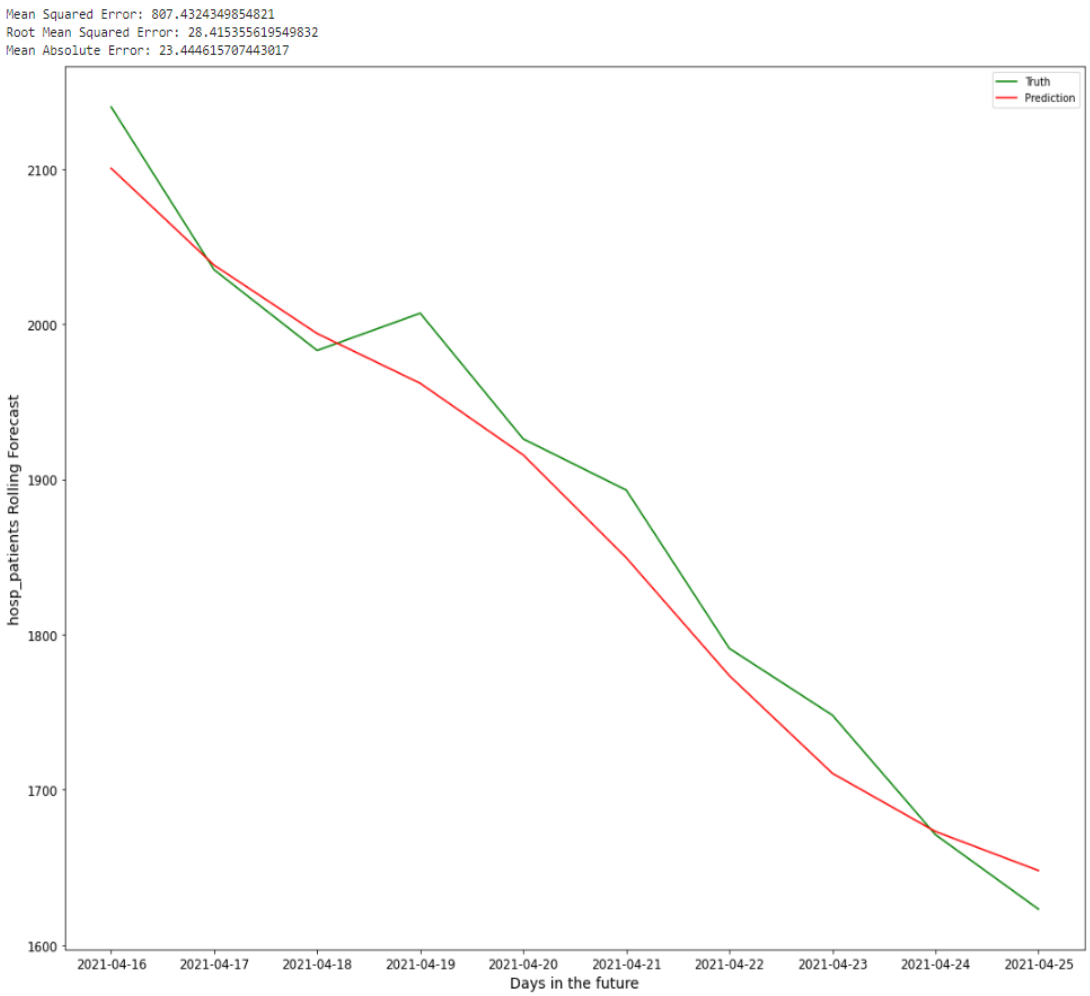
1. Απλή πρόβλεψη δέκα ημέρων
2. Κυλιόμενη πρόβλεψη

Στην απλή πρόβλεψη, αρχικά το μοντέλο εκπαιδεύεται με τα δεδομένα εκπαίδευσης. Μετά την εκπαίδευση του μοντέλου πάνω στα δεδομένα προβλέπονται απευθείας δέκα μέρας μπροστά από την ημερομηνία που τελειώνουν τα δεδομένα χωρίς να εκπαιδευτεί ξανά το μοντέλο. Μετά το πέρας της πρόβλεψης εκτυπώνεται ένα σχεδιάγραμμα με τις πραγματικές τιμές και τις τιμές που προέβλεψε το μοντέλο καθώς και μέτρα αποδοτικότητας του μοντέλου όπως το Μέσο Απολυτό Σφάλμα.



ΕΙΚΌΝΑ 9: Πρόβλεψη μοντέλου δέκα ημέρες στο μέλλον

Στην κυλιόμενη πρόβλεψη οι δέκα ημέρες δεν προβλέπονται απευθείας. Αφού εκπαιδευτεί το μοντέλο με τα δεδομένα εκπαίδευσης την πρώτη φορά, έπειτα, κάθε επανάληψη που προβλέπεται η αμέσως επόμενη ημέρα, η ημέρα αυτή ενσωματώνεται στο μοντέλο, δηλαδή στα δεδομένα εκπαίδευσης και το μοντέλο εκπαιδεύεται ξανά πάνω σε αυτά. Αυτό επαναλαμβάνεται για όλες τις μέρες και κατ' επέκταση απαιτεί περισσότερη υπολογιστική ισχύ σε σχέση με την απλή πρόβλεψη. Μετά το πέρας της πρόβλεψης εκτυπώνεται ένα σχεδιάγραμμα με τις πραγματικές τιμές και τις τιμές που προέβλεψε το μοντέλο καθώς και μετρά αποδοτικότητας του μοντέλου όπως το Μέσο Απολυτό Σφάλμα.



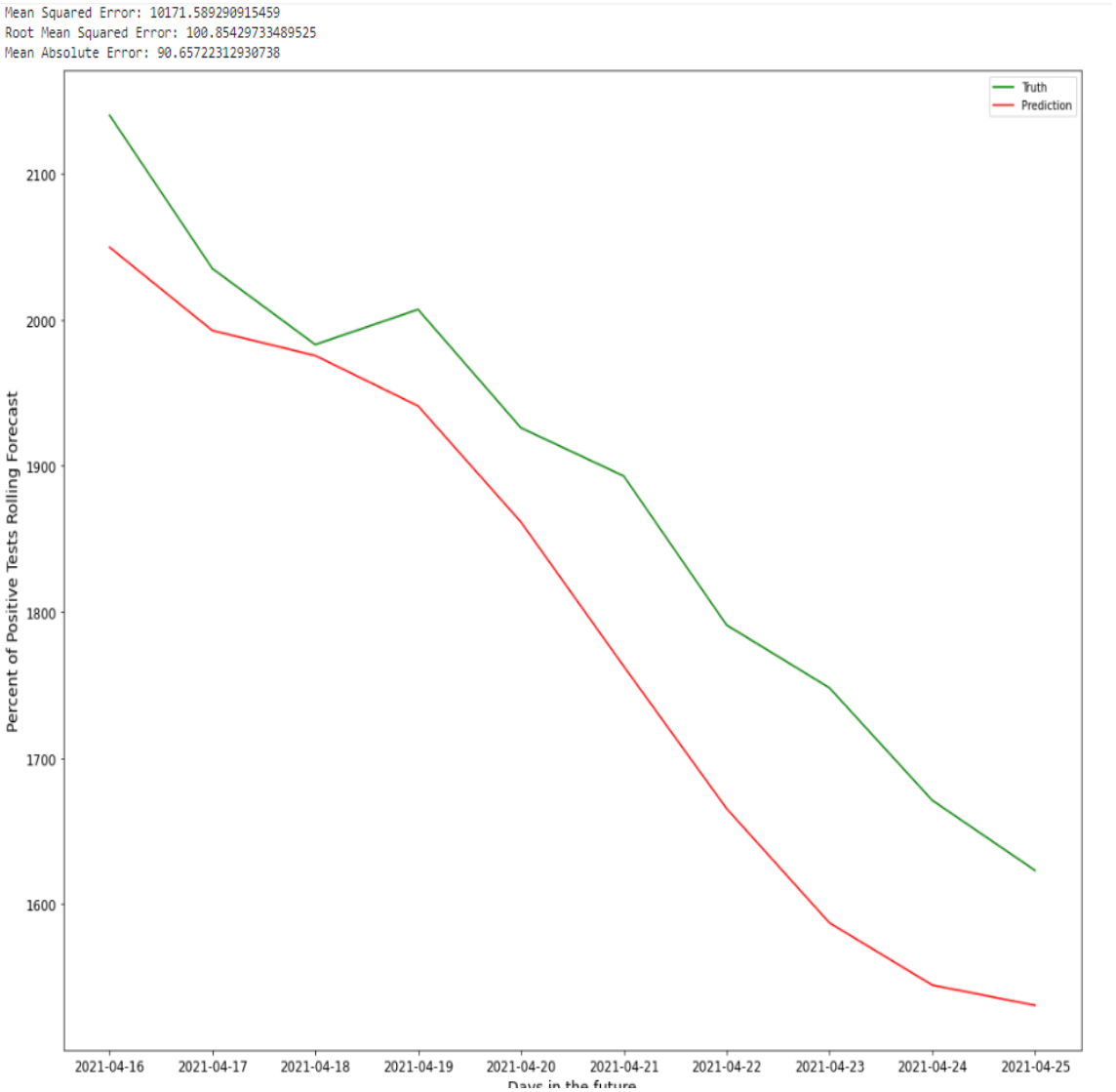
ΕΙΚΌΝΑ 10: **Κυλιώμενη πρόβλεψη δέκα ημερών**

Είναι σημαντικό να σημειωθεί ότι τα δεδομένα της χρονοσειράς επιλέγονται μέχρι μερικές ημέρες πριν την τελευταία ημερομηνία που είναι διαθέσιμη από το εξαγόμενο σετ δεδομένων. Αυτό συμβαίνει διότι καθώς γίνεται παρεμβολή των τιμών κατά την ένωση των σετ δεδομένων, κάποιες μεταβλητές δύναται να μην έχουν ενημερωθεί με αποτέλεσμα να λείπουν μερικές μέρες στο τέλος. Μετά την παρεμβολή τιμών, οι ημέρες αυτές που πριν περιείχαν κενές τιμές , τώρα περιέχουν την τελευταία τιμή με αποτέλεσμα σε κάποιες μεταβλητές οι τελευταίες μέρες να επαναλαμβάνουν (λανθασμένα) την δια τιμή αντί να είναι κενές.

Μέχρι στιγμής η πρόβλεψη με το μοντέλο βασιζόταν μόνο στην ιδιά την χρονοσειρά για την οποία γινόταν η πρόβλεψη. Δίνεται ωστόσο η δυνατότητα να προστεθούν και εξωτερικές χρονοσειρές σαν παράγοντες που επηρεάζουν το μοντέλο στις προβλέψεις του. Οι εξωτερικές χρονοσειρές αυτές είναι σημαντικό να έχουν την

ίδια διάσταση με την χρονοσειρά για την οποία γίνεται η πρόβλεψη, τόσο για τα δεδομένα εκπαίδευσης όσο και για τα δεδομένα επαλήθευσης.

Με βάση τα παραπάνω το μοντέλο ARIMA(8,2,2) εφαρμόζεται στα δεδομένα ξανά μαζί με την χρονοσειρά που είναι η μεταβλητή «stringency_index» του σετ δεδομένων σαν εξωτερική χρονοσειρά. Μετά το πέρας της πρόβλεψης εκτυπώνεται ένα σχεδιάγραμμα με τις πραγματικές τιμές και τις τιμές που προέβλεψε το μοντέλο καθώς και μέτρα αποδοτικότητας του μοντέλου, όπως το Μέσο Απόλυτο Σφάλμα.



ΕΙΚΌΝΑ 11: Πρόβλεψη με εξωγενής παράγοντες

ΑΝΑΦΟΡΕΣ

[1] X. Chen, Y. Cho, and S. Y. Jang, “Crime prediction using twitter sentiment and weather,” in 2015 Systems and Information Engineering Design Symposium. IEEE, 2015, pp. 63–68.

[2] M. S. Gerber, “Predicting crime using twitter and kernel density estimation,” Decision Support Systems, vol. 61, pp. 115–125, 2014.

[3] P. Grover, A. K. Kar, Y. K. Dwivedi, and M. Janssen, “Polarization and acculturation in us election 2016 outcomes–can twitter analytics predict changes in voting preferences,” Technological Forecasting and Social Change, vol. 145, pp. 438–460, 2019.

[4] M. Y. Kabir and S. Madria, “A deep learning approach for tweet classification and rescue scheduling for effective disaster management,” in Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2019, pp. 269–278.

[5] D. Baer, “As sandy became# sandy, emergency services got social,” Fast Company, vol. 9, 2012.

[6] L. Zou, N. S. Lam, S. Shams, H. Cai, M. A. Meyer, S. Yang, K. Lee, S.-J. Park, and M. A. Reams, “Social and geographical disparities in twitter use during hurricane harvey,” International Journal of Digital Earth, vol. 12, no. 11, pp. 1300–1318, 2019.

[7] J. Yang, M. Yu, H. Qin, M. Lu, and C. Yang, “A twitter data credibility frameworkâA˘Thurricane harvey as a use case,” ˇ ISPRS International Journal of Geo-Information, vol. 8, no. 3, p. 111, 2019.

[8] A. Sebastian, W. HighField, S. Brody, and W. Mobley, “Leveraging machine learning and twitter data to identify high hazard areas during hurricane harvey,” 2019.

[9] E. Hirata, M. Giannotti, A. Larocca, and J. Quintanilha, “Flooding and inundation collaborative mapping–use of the crowdmap/ushahidi platform in the city of sao paulo, brazil,” Journal of Flood Risk Management, vol. 11, pp. S98–S109, 2018.

[10] P. S. Earle, D. C. Bowden, and M. Guy, “Twitter earthquake detection: earthquake monitoring in a social world,” Annals of Geophysics, vol. 54, no. 6, 2012.

[11] C. Buntain, J. Golbeck, B. Liu, and G. LaFree, “Evaluating public response to the boston marathon bombing and other acts of terrorism through twitter,” in Tenth International AAAI Conference on Web and Social Media, 2016.

[12] B. G. Southwell, J. Niederdeppe, J. N. Cappella, A. Gaysynsky, D. E. Kelley, A. Oh, E. B. Peterson, and W.-Y. S. Chou, “Misinformation as a misunderstood challenge to public health,” American journal of preventive medicine, vol. 57, no. 2, pp. 282–285, 2019. [13] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, and M. Dredze, “Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate,” American journal of public health, vol. 108, no. 10, pp. 1378–1384, 2018.

[14] S. O. Oyeyemi, E. Gabarron, and R. Wynn, “Ebola, twitter, and misinformation: a dangerous combination?” Bmj, vol. 349, p. g6178, 2014.

[15] Z. Wang, N. S. Lam, N. Obradovich, and X. Ye, “Are vulnerable communities digitally left behind in social responses to natural disasters? an evidence from hurricane sandy with twitter data,” Applied geography, vol. 108, pp. 1–8, 2019.

[16] D. Wladdimiro, P. Gonzalez-Cantergiani, N. Hidalgo, and E. Rosas, “Disaster management platform to support real-time analytics,” in 2016 3rd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM). IEEE, 2016, pp. 1–8.

[17] M. Imran, P. Mitra, and C. Castillo, “Twitter as a lifeline: Humanannotated twitter corpora for nlp of crisis-related messages,” arXiv preprint arXiv:1605.05894, 2016.

[18] R. Nagar, Q. Yuan, C. C. Freifeld, M. Santillana, A. Nojima, R. Chunara, and J. S. Brownstein, “A case study of the new york city 2012- 2013 influenza season with daily geocoded twitter data from temporal and spatiotemporal perspectives,” Journal of medical Internet research, vol. 16, no. 10, p. e236, 2014.

[19] M. Szomszor, P. Kostkova, and E. De Quincey, “# swineflu: Twitter predicts swine flu outbreak in 2009,” in International conference on electronic healthcare. Springer, 2010, pp. 18–26.

[20] M. Odlum and S. Yoon, “What can we learn about the ebola outbreak from tweets?” American journal of infection control, vol. 43, no. 6, pp. 563–571, 2015.

[21] M. Dredze, D. A. Broniatowski, and K. M. Hilyard, “Zika vaccine misconceptions: A social media analysis,” Vaccine, vol. 34, no. 30, p. 3441, 2016.

[22] C. Ordun, S. Purushotham, and E. Raff, “Exploratory analysis of covid19 tweets using topic modeling, umap, and digraphs,” arXiv preprint arXiv:2005.03082, 2020.

[23] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, “Top concerns of tweeters during the covid-19 pandemic: infoveillance study,” Journal of medical Internet research, vol. 22, no. 4, p. e19016, 2020.

[24] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. Vraga, and Y. Wang, “A first look at covid19 information and misinformation sharing on twitter,” arXiv preprint arXiv:2003.13907, 2020.

[25] R. Kouzy, J. Abi Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, and K. Baddour, “Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter,” Cureus, vol. 12, no. 3, 2020.

[26] World Health Organization (WHO): https://www.who.int/

[27] European Centre for Disease Prevention and Control (ECDC): https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases

[28] DXY.cn. Pneumonia. 2020. http://3g.dxy.cn/newh5/view/pneumonia

[29] US CDC: https://www.cdc.gov/coronavirus/2019-ncov/index.html

[30] BNO News: https://bnonews.com/index.php/2020/02/the-latest-coronavirus-cases/

[31] WorldoMeters: https://www.worldometers.info/coronavirus/

[32] 1Point3Arces: https://coronavirus.1point3acres.com/en

[33] COVID Tracking Project: https://covidtracking.com/data. (US Testing and Hospitalization Data. We use the maximum reported value from "Currently" and "Cumulative" Hospitalized for our hospitalization number reported for each state.)

[34] Los Angeles Times: https://www.latimes.com/projects/california-coronavirus-cases-tracking-outbreak/

[35] The Mercury News: https://www.mercurynews.com/tag/coronavirus/

[36] Alaska Department of Health and Social Services: https://alaska-coronavirus-vaccine-outreach-alaska-dhss.hub.arcgis.com/

[37] Washington State Department of Health: https://www.doh.wa.gov/emergencies/coronavirus

[38] Maryland Department of Health: https://coronavirus.maryland.gov/

[39] New York State Department of Health: https://health.data.ny.gov/Health/New-York-State-Statewide-COVID-19-Testing/xdss-u53e/data

[40] New York City Health Department: https://www1.nyc.gov/site/doh/covid/covid-19-data.page

[41] NYC Department of Health and Mental Hygiene: https://www1.nyc.gov/site/doh/covid/covid-19-data.page

[42] NYC Department of Health and Mental Hygiene: https://github.com/nychealth/coronavirus-data

[43] Florida Department of Health Dashboard: https://services1.arcgis.com/CY1LXxl9zlJeBuRZ/arcgis/rest/services/Florida_COVID19_Cases/FeatureServer/0 and https://fdoh.maps.arcgis.com/apps/opsdashboard/index.html#/8d0de33f260d444c852a615dc7837c86

[44] Colorado: https://covid19.colorado.gov/covid-19-data

[45] Virginia: https://www.vdh.virginia.gov/coronavirus/

[46] Northern Mariana Islands CNMI Department of Public Health: https://cnmichcc.maps.arcgis.com/apps/opsdashboard/index.html#/4061b674fc964efe84f7774b7979d2b5

[47] Missouri Department of Health: https://www.arcgis.com/apps/MapSeries/index.html?appid=8e01a5d8d8bd4b4f85add006f9e14a9d

[48] Missouri: Nodaway County: https://www.nodawaypublichealth.org/

[49] Missouri: St. Louis City Department of Health: https://www.stlouis-mo.gov/covid-19/data/#totalsByDate

[50] Missouri: St. Louis County: https://stlcorona.com/resources/covid-19-statistics1/

[51] Massachusetts: https://www.mass.gov/info-details/covid-19-response-reporting

[52] Michigan: https://www.michigan.gov/coronavirus/0,9753,7-406-98163_98173---,00.html

[53] Illinois Department of Public Health: https://dph.illinois.gov/covid19

[54] Indiana State Department of Health: https://hub.mph.in.gov/dataset?q=COVID

[55] Connecticut Department of Public Health: https://data.ct.gov/stories/s/COVID-19-data/wa3g-tfvc/

[56] Ohio Department of Health: https://coronavirus.ohio.gov/wps/portal/gov/covid-19/home

[57] Oregon Health Authority: https://govstatus.egov.com/OR-OHA-COVID-19

[58] Oregon Health Authority (Weekends): https://www.oregon.gov/oha/erd/pages/covid-19-news.aspx

[59] Tennessee Department of Health: https://www.tn.gov/health/cedep/ncov.html

[60] Rhode Island Department of Health: https://ri-department-of-health-covid-19-data-rihealth.hub.arcgis.com/

[61] Wisconsin Department of Health Services: https://www.dhs.wisconsin.gov/covid-19/data.htm

[62] North Carolina City of Greenville GIS: https://www.arcgis.com/apps/opsdashboard/index.html#/7aeac695cafa4065ba1505b1cfa72747

[62] Iowa State Government: https://coronavirus.iowa.gov/

[64] Minnesota Department of Health: https://www.health.state.mn.us/diseases/coronavirus/situation.html

[65] Alabama Public Health: https://www.alabamapublichealth.gov/covid19/

[66] Mississippi State Department of Health: https://msdh.ms.gov/msdhsite/_static/14,0,420.html

[67] Nebraska Department of Health and Human Services: https://experience.arcgis.com/experience/ece0db09da4d4ca68252c3967aa1e9dd

[68] South Carolina Department of Health and Environmental Control: https://scdhec.gov/infectious-diseases/viruses/coronavirus-disease-2019-covid-19/sc-testing-data-projections-covid-19

[69] Nevada Department of Health and Human Services: https://nvhealthresponse.nv.gov/

[70] New Jersey Department of Health: https://covid19.nj.gov/

[71] Delaware Emergency Management Agency: https://coronavirus.delaware.gov/

[72] Utah Department of Health: https://coronavirus-dashboard.utah.gov/#overview

[73] Arizona Emergency Support Function (ESF)/Recovery Support Function (RSF) Operations Dashboard: https://www.arcgis.com/apps/MapSeries/index.html?appid=62e6bfa682a34e6aae9d9255ac865467

[74] Departamento de Salud, Puerto Rico: https://covid19datos.salud.gov.pr/

[75] Arkansas Department of Health: https://experience.arcgis.com/experience/c2ef4a4fcbe5458fbf2e48a21e4fece9

[76] Maine Department of Health and Human Services: https://www.maine.gov/dhhs/mecdc/infectious-disease/epi/airborne/coronavirus/data.shtml

[77] Pennsylvania Department of Health: https://www.health.pa.gov/topics/disease/coronavirus/Pages/Cases.aspx

[78] City of Philadelphia: https://www.phila.gov/programs/coronavirus-disease-2019-covid-19/

[79] Lancaster County: https://covid-19-lancastercountypa.hub.arcgis.com/

[80] Chester County: https://chesco.maps.arcgis.com/apps/opsdashboard/index.html#/975082d579454c3ca7877db0a44e61ca

[81] Louisiana Department of Health: http://ldh.la.gov/coronavirus/

[82] Government of The District of Columbia: https://coronavirus.dc.gov/page/coronavirus-data

[83] North Dakota Department of Health: https://www.health.nd.gov/diseases-conditions/coronavirus/north-dakota-coronavirus-cases

[84] Oklahoma State Department of Health: https://looker-dashboards.ok.gov/embed/dashboards/44

[85] Guam Department of Public Health and Social Services: http://dphss.guam.gov/covid-19/

[86] New Mexico Department of Health: https://cvprovider.nmhealth.org/public-dashboard.html

[87] Kentucky Department of Public Health: https://kygeonet.maps.arcgis.com/apps/opsdashboard/index.html#/543ac64bc40445918cf8bc34dc40e334

[88] Georgia Department of Public Health: https://dph.georgia.gov/covid-19-daily-status-report

[89] State of Hawai'i Department of Health: https://health.hawaii.gov/coronavirusdisease2019/what-you-should-know/current-situation-in-hawaii/

[90] Reno County Health Department: http://reno.maps.arcgis.com/apps/opsdashboard/index.html#/dfaef27aede1414b89abf393b2ccb994

[91] Texas Department of State Health Services: https://dshs.texas.gov/coronavirus/

[92] Texas: City of San Antonio: https://covid19.sanantonio.gov/Home

[93] Texas: Brazioria County: https://www.brazoriacountytx.gov/departments/health-department/brazoria-county-coronavirus-map

[94] Texas: Brazos County Health District: http://www.brazoshealth.org

[95] Texas: Cameron County Public Health: https://www.cameroncounty.us/publichealth/index.php/coronavirus/

[96] Texas: Collin County: https://www.collincountytx.gov/healthcare_services/Pages/news.aspx

[97] Texas: Dallas County: https://www.dallascounty.org/departments/dchhs/2019-novel-coronavirus.php

[98] Texas: Denton County: https://gis-covid19-dentoncounty.hub.arcgis.com/pages/covid-19cases

[99] Texas: Ector County: http://www.co.ector.tx.us/page/ector.CoronavirusCOVID19

[100] Texas: City of El Paso: http://epstrong.org/results.php

[101] Texas: Fayette County: https://www.co.fayette.tx.us/

[102] Texas: Fort Bend County Health & Human Services: https://www.fbchealth.org/ncov/

[103] Texas: Galveston County Health District: https://www.gchd.org/about-us/news-and-events/coronavirus-disease-2019-covid-19/galveston-county-confirmed-covid-19-cases

[104] Texas: Harris County: https://publichealth.harriscountytx.gov/Resources/2019-Novel-Coronavirus

[105] Texas: Hays County: https://hayscountytx.com/covid-19-information-for-hays-county-residents/

[106] Texas: Hidalgo County Health and Human Services: https://www.hidalgocounty.us/2630/Coronavirus-Updates

[107] Texas: Midland County: https://www.midlandtexas.gov/955/Coronavirus-COVID-19

[108] Texas: Montgomery County: https://coronavirus-response-moco.hub.arcgis.com/

[109] Texas: City of Corpus Christi: https://www.cctexas.com/coronavirus

[110] Texas: Amarillo Public Health Department: https://covid-data-amarillo.hub.arcgis.com/

[111] Texas: Tarrant County Public Health: https://www.tarrantcounty.com/en/public-health/disease-control---prevention/coronaviruas.html

[112] Texas: City of Mount Pleasant: https://www.mpcity.net/632/COVID-19-UPDATES

[113] Texas: City of San Angelo: https://www.cosatx.us/departments-services/health-services/coronavirus-covid-19#ad-image-0

[114] Texas: San Angelo Standard-Times: https://www.gosanangelo.com/

[115] Texas: Travis County: https://www.traviscountytx.gov/news/2020/1945-novel-coronavirus-covid-19-information

[116] Texas: City of Laredo: https://www.cityoflaredo.com/coronavirus/

[117] Texas: Williamson County & Cities Health District: http://www.wcchd.org/COVID-19/dashboard.php

[118] California Department of Publich Health: https://www.cdph.ca.gov/covid19

[119] California: Mariposa County: https://www.mariposacounty.org/1592/COVID-19-Information

[120] California: Alameda County Public Health Department: https://covid-19.acgov.org/

[121] California: Fresno County Public Health Department: https://www.co.fresno.ca.us/departments/public-health/covid-19

[122] California: Humboldt County: https://humboldtgov.org/

[123] California: Madera County: https://www.maderacounty.com/government/public-health/corona-virus-covid-19/covid-revised

[124] California: Marin County Health & Human Services: https://coronavirus.marinhhs.org/

[125] California: Mendocino County: https://www.mendocinocounty.org/community/novel-coronavirus/covid-19-case-data

[126] California: Orange County Health Care Agency: https://occovid19.ochealthinfo.com/coronavirus-in-oc

[127] California: Placer County: https://www.placer.ca.gov/coronavirus

[128] California: Riverside County: https://www.rivcoph.org/coronavirus

[129] California: Sacramento County: https://www.saccounty.net/COVID-19/

[130] California: San Francisco Department of Public Health: https://www.sfdph.org/dph/alerts/coronavirus.asp

[131] California: San Benito County Health & Human Services: https://hhsa.cosb.us/publichealth/communicable-disease/coronavirus/

[132] California: San Joaquin County Public Health Services: http://www.sjcphs.org/coronavirus.aspx

[133] California: San Mateo County: https://www.smchealth.org/coronavirus

[134] California: Santa Cruz County Health Services Agency: http://www.santacruzhealth.org/HSAHome/HSADivisions/PublicHealth/CommunicableDiseaseControl/CoronavirusHome.aspx

[135] California: Shasta County: https://www.co.shasta.ca.us/covid-19/overview

[136] California: Solano County: https://www.co.shasta.ca.us/covid-19/overview

[137] California: Sonoma County: https://socoemergency.org/emergency/novel-coronavirus/coronavirus-cases/

[138] California: Stanislaus County Health Services Agency: http://schsa.org/publichealth/pages/corona-virus/

[139] California: Ventura County: https://www.venturacountyrecovers.org/

[140] California: Yolo County: https://www.yolocounty.org/health-human-services/adults/communicable-disease-investigation-and-control/novel-coronavirus-2019/

[141] California: Los Angeles County: http://publichealth.lacounty.gov/media/coronavirus/

[142] California: San Diego County: https://www.sandiegocounty.gov/coronavirus.html

[143] California: Santa Clara County: https://www.sccgov.org/sites/covid19/

[144] California: Imperial County Public Health Department: http://www.icphd.org/health-information-and-resources/healthy-facts/covid-19/

[145] California: San Bernardino County: https://sbcovid19.com/

[146] Montana Department of Public Health and Human Services: https://dphhs.mt.gov/publichealth/cdepi/diseases/coronavirusmt

[147] South Dakota Department of Health: https://doh.sd.gov/news/coronavirus.aspx

[148] Wyoming Department of Health: https://health.wyo.gov/publichealth/infectious-disease-epidemiology-unit/disease/novel-coronavirus/

[149] New Hampshire Department of Health and Human Services: https://www.nh.gov/covid19/dashboard/summary.htm

[150] Idaho Government: https://coronavirus.idaho.gov/

[151] Virgin Islands Department of Health: https://www.covid19usvi.com/

[152] Vermont Department of Health: https://www.healthvermont.gov/response/coronavirus-covid-19/current-activity-vermont

[153] Kansas: Reno County Health Department: https://experience.arcgis.com/experience/9a7d44773e4c4a48b3e09e4d8673961b/page/page_18/

[154] Kansas: Kansas Department Of Health And Environment: https://www.coronavirus.kdheks.gov/160/COVID-19-in-Kansas

[155] Kansas: Douglas County Coronavirus Response and Recovery Hub: https://coronavirus-response-dgco.hub.arcgis.com/

[156] Kansas: Finney County COVID-19 Resource Hub: https://finney-county-coronavirus-response-finneycountygis.hub.arcgis.com/

[157] Kansas: Riley County Corona Virus Response: https://coronavirus-response-rcitgis.hub.arcgis.com/

[158] West Virginia Department of Health & Human Resources: https://dhhr.wv.gov/COVID-19/Pages/default.aspx

[159] National Health Commission of the People's Republic of China (NHC): http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml

[160] China CDC (CCDC): http://weekly.chinacdc.cn/news/TrackingtheEpidemic.htm

[161] Hong Kong Department of Health: https://www.chp.gov.hk/en/features/102465.html

[162] Macau Government: https://www.ssm.gov.mo/portal/

[163] Taiwan CDC: https://sites.google.com/cdc.gov.tw/2019ncov/taiwan?authuser=0

[164] Canada: Government of Canada: https://www.canada.ca/en/public-health/services/diseases/coronavirus.html

[165] Canada: Government of Alberta: https://www.alberta.ca/covid-19-alberta-data.aspx

[166] Canada: Government of British Columbia Centre for Disease Control: https://experience.arcgis.com/experience/a6f23959a8b14bfa989e3cda29297ded

[167] Canada: Government of Manitoba: https://www.gov.mb.ca/covid19/updates/cases.html

[168] Canada: Government of New Brunswick: https://experience.arcgis.com/experience/8eeb9a2052d641c996dba5de8f25a8aa

[169] Canada: Government of Newfoundland and Labrador: https://covid-19-newfoundland-and-labrador-gnl.hub.arcgis.com/

[170] Canada: Government of Northwest Territories: https://www.gov.nt.ca/covid-19/

[171] Canada: Government of Nova Scotia: https://novascotia.ca/coronavirus/data/

[172] Canada: Nunavut Department of Health: https://www.gov.nu.ca/health/information/covid-19-novel-coronavirus

[173] Canada: Government of Ontario: https://covid-19.ontario.ca/data

[174] Canada: Grey Bruce Health Unit: https://www.publichealthgreybruce.on.ca/

[175] Canada: Eastern Ontario Health Unit: https://eohu.ca/en/covid/covid-19-status-update-for-eohu-region

[176] Canada: Windsor-Essex County Health Unit: https://www.wechu.org/cv/local-updates

[177] Canada: Ottawa Public Health: https://www.ottawapublichealth.ca/en/reports-research-and-statistics/daily-covid19-dashboard.aspx

[178] Canada: York Region: https://www.york.ca/wps/portal/yorkhome/health/yr/covid-19/covid19inyorkregion/!ut/p/z1/tZPNcpswFIWfJQuWWBcJjOhOpa6BxNhN4j82HowxKAXkYMXUb1-R0pm20zjtJLAASZx7dPRxQRFaoaiKTzyLJRdVXKj5OhpufDb2Pe8agqlJXWAwZQG2KYwcAy2fBfDCxQBF_1J_QRBdtl-gCEWHhO_Q2qIWdYhj6U6Md7qJnVSnceLoO8PZOhhi9XbfqpNKHmSO1ud6k4hKppXU4Czqr2pylFw-PS_kokzVPY0LmWuQiBPf6YbTjQyHV21FnWYKEwpeO6SiiOuJO8lU1FjmOq_2Aq1-mnajP0xVDX94fIyYCtym_CbR6t0SL1sOv2Ye31ET_EVgs4UxBdMnnQBjc-gZLgTgTSn4n-2Z9Yl6BlzjTnDh0ygsWSG2P7qIVVtC1fnrdJ_WaT14qtVyLuXh-EEDDZqmGWRCZEU6SESpwd9KcnFUDH5XorXqH_tF9rcYLU88bdC8EnWpktz9Z7t40O1gU5d5bAwzuJ_b8GVkm3R4M5nd3Bpv3OGVA_RsT3q1t6Ffe9yv_fvACXxwDdb-Y2REgGHfpR9JQMOwX_Zhv-zDftmH_fb94q1wDuV8XlJiFQ9nAnxWLqmkzf0-z8rNZESsyw92dfUd1P5kcA!!/dz/d5/L2dBISEvZ0FBIS9nQSEh/#.X8UQBqpKi3U

[179] Canada: City of Toronto: https://www.toronto.ca/home/covid-19/covid-19-latest-city-of-toronto-news/covid-19-status-of-cases-in-toronto/

[180] Canada: Region of Peel: https://peelregion.ca/coronavirus/case-status/

[181] Canada: Halton Region: https://www.halton.ca/For-Residents/Immunizations-Preventable-Disease/Diseases-Infections/New-Coronavirus

[182] Canada: Government of Prince Edward Island: https://www.princeedwardisland.ca/en/information/health-and-wellness/pei-covid-19-case-data

[183] Canada: Government of Quebec: https://www.quebec.ca/en/health/health-issues/a-z/2019-coronavirus/situation-coronavirus-in-quebec/

[184] Canada: Government of Saskatchewan: https://dashboard.saskatchewan.ca/health-wellness

[185] Canada: Government of Yukon: https://yukon.ca/en/case-counts-covid-19

[186] Australia Government Department of Health: https://www.health.gov.au/news/coronavirus-update-at-a-glance

[187] COVID Live (Australia): https://www.covidlive.com.au/

[188] Ministry of Health Singapore (MOH): https://www.moh.gov.sg/covid-19

[189] Italy Ministry of Health: http://www.salute.gov.it/nuovocoronavirus

[190] Government of Ireland: https://covid19ireland-geohive.hub.arcgis.com/

[191] Dati COVID-19 Italia (Italy): https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni

[192] Gobierno De El Salvador: https://covid19.gob.sv/

[193] Czechia Ministry of Health: https://onemocneni-aktualne.mzcr.cz/covid-19

[194] French Government: https://dashboard.covid19.data.gouv.fr/ and https://www.data.gouv.fr/en/datasets/donnees-relatives-a-lepidemie-de-covid-19-en-france-vue-densemble/

[195] OpenCOVID19 France: https://github.com/opencovid19-fr

[196] Palestine (West Bank and Gaza): https://corona.ps/details

[197] Israel: https://govextra.gov.il/ministry-of-health/corona/corona-virus/

[198] Israel: https://datadashboard.health.gov.il/COVID-19/general

[199] Indonesia: https://covid19.go.id/peta-sebaran

[200] National Institute of Health of Kosovo: https://corona-ks.info/?lang=en and https://raw.githubusercontent.com/bgeVam/Kosovo-Coronatracker-Data/master/data.json

[201] Berliner Morgenpost (Germany): https://interaktiv.morgenpost.de/corona-virus-karte-infektionen-deutschland-weltweit/

[202] rtve (Spain): https://www.rtve.es/noticias/20200514/mapa-del-coronavirus-espana/2004681.shtml

[203] Ministry of Health, Republic of Serbia: https://covid19.rs/homepage-english/

[204] Chile: https://www.minsal.cl/nuevo-coronavirus-2019-ncov/casos-confirmados-en-chile-covid-19/

[205] Chile: https://www.gob.cl/coronavirus/cifrasoficiales/

[206] Denmark Statens Serum Institut: https://experience.arcgis.com/experience/aa41b29149f24e20a4007a0c4e13db1d

[207] Brazil Ministry of Health: https://covid.saude.gov.br/

[209] Brazil: https://github.com/wcota/covid19br. Data described in DOI: 10.1590/SciELOPreprints.362

[210] Belgium Sciensano: https://datastudio.google.com/embed/reporting/c14a5cfc-cab7-4812-848c-0369173148ab/page/giyUB

[211] Gobierono De Mexico: https://datos.covid-19.conacyt.mx/#DOView

[212] Japan COVID-19 Coronavirus Tracker: https://covid19japan.com/#all-prefectures

[213] Monitoreo del COVID-19 en Perú - Policía Nacional del Perú (PNP) - Dirección de Inteligencia (DIRIN): https://www.arcgis.com/apps/opsdashboard/index.html#/f90a7a87af2548699d6e7bb72f5547c2 and Ministerio de Salud del [35] Perú: https://covid19.minsa.gob.pe/sala_situacional.asp

[214] Colombia National Institute of Health: http://www.ins.gov.co/Noticias/Paginas/Coronavirus.aspx

[215] Russia: https://xn--80aesfpebagmfblc0a.xn--p1ai/information/

[216] Ukraine: https://covid19.rnbo.gov.ua/

[217] Public Health Agency of Sweden: https://experience.arcgis.com/experience/09f821667ce64bf7be6f9f87457ed9aa

[218] Government of India: https://www.mygov.in/covid-19

[219] Lithuania: https://osp.maps.arcgis.com/apps/MapSeries/index.html?appid=c6bc9659a00449239eb3bde062d23caa

[220] Government of Pakistan: http://covid.gov.pk/stats/pakistan

[221] The UK Government: https://coronavirus.data.gov.uk/#category=nations&map=rate

[222] Scottish Government: https://www.gov.scot/publications/coronavirus-covid-19-trends-in-daily-data/

[223] Netherlands National Institute for Health and Environment: https://experience.arcgis.com/experience/ea064047519040469acb8da05c0f100d

[224] Iceland Directorate of Health and Department of Civil Protection and Emergency Management: https://www.covid.is/data

[225] Luxembourg Government: https://data.public.lu/fr/datasets/covid-19-rapports-journaliers/#_

[226] Afghanistan: http://covid.moph-dw.org/#/ and http://www.afghanistantimes.af/

[227] Kazakhstan: https://www.coronavirus2020.kz/

[228] Republic of Turkey Ministry of Health: https://covid19.saglik.gov.tr/EN-69532/general-coronavirus-table.html

[229] Slovakia Ministry of Investment, Regional Development and Information: https://korona.gov.sk/

[230] Switzerland Federal Office Of Public Health: https://www.bag.admin.ch/bag/en/home/krankheiten/ausbrueche-epidemien-pandemien/aktuelle-ausbrueche-epidemien/novel-cov/situation-schweiz-und-international.html

[231] Switzerland Open Government Data Reported By The Swiss Cantons: https://github.com/openZH/covid_19

[232] Thailand Ministry of Public Health: https://ddc.moph.go.th/viralpneumonia/eng/index.php and https://covid19.ddc.moph.go.th/en

[233] JHU CSSE COVID-19 Dataset https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data

[234] The panacealab covid19_twitter counts https://github.com/thepanacealab/covid19_twitter/tree/master/

[235] Data on COVID-19 (coronavirus) by Our World in Data https://github.com/owid/covid-19-data/tree/master/public/data

[236] The panacealab covid19_twitter main branch https://github.com/thepanacealab/covid19_twitter

[237] ARIMA Models https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average

[238] ARIMA for time series forecasting https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

[239] ARIMA tune p,d and q parametershttps://www.kaggle.com/sumi25/understand-arima-and-tune-p-d-q

[240]Jenkins Method https://en.wikipedia.org/wiki/Box%E2%80%93Jenkins_method

[241] Time Series Sesonal ARIMA model in Python http://www.seanabu.com/2016/03/22/time-series-seasonal-ARIMA-model-in-python/

[242] Summary of rules for identifying ARIMA models https://faculty.fuqua.duke.edu/~rnau/Decision411_2007/arimrule.htm

[243] ARIMA time series forecasting https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

[244] PACFs and ACFs in identifying AR and MA terms in ARIMA models https://towardsdatascience.com/identifying-ar-and-ma-terms-using-acf-and-pacf-plots-in-time-series-forecasting-ccb9fd073db8

[245] Meteostat Python API https://github.com/meteostat/meteostat-python

[246] Seaborn documentation https://seaborn.pydata.org/

[247] Pandas Documentation https://pandas.pydata.org/docs/user_guide/index.html

[248] Numpy documentation https://numpy.org/doc/stable/

[249] Matplotlib documentation https://matplotlib.org/stable/tutorials/index.html

[250] Statsmodels documentation
https://www.statsmodels.org/devel/gettingstarted.html#

[251] Scikit-learn documentationhttps://scikit-learn.org/stable/user_guide.html