Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Πανεπιστήμιο Πατρών

Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης

Εργαστηριακή Άσκηση

Εαρινό Εξάμηνο 2022-2023

Πανδής Αναστάσιος

Έτος:7ο ΑΜ: 1056287

Χρυσικόπουλος Γρηγόριος

Έτος:4ο ΑΜ:1066644

Περιβάλλον Υλοποίησης & Βιβλιοθήκες

Για την υλοποίηση της εργασίας χρησιμοποιήσαμε την υπηρεσία google collab για ανάπτυξη μέσω jupyter notebook. Για την διαχείριση των δεδομένων σε dataframes καθώς και για την εξαγωγή στατιστικών μεγεθών χρησιμοποιήθηκε η βιβλιοθήκη pandas dataframes της python. Επιπροσθέτως, αξιοποιήσαμε την βιβλιοθήκη matplotlib για να απεικονίσουμε σε γραφικές παραστάσεις ορισμένα μεγέθη.

Για το clustering χρησιμοποιήσαμε την βιβλιοθήκη sklearn και συγκεκριμένα την μέθοδο KMeans.

Για την απεικόνιση των clusters στον τρισδιάστατο χώρο χρησιμοποιήσαμε βιβλιοθήκες mpl_toolkits.mplot3d και plotly.express.

Για την ανάπτυξη του νευρωνικού δικτύου χρησιμοποιήσαμε τις παρακάτω βιβλιοθήκες:

Για την μετατροπή των δεδομένων σε μία μορφή πιο ευκόλως διαχειρίσιμη από το νευρωνικό πραγματοποιήσαμε εισαγωγή της sklearn.preprocessing και συγκεκριμένα του RobustScaler. Το LSTM μοντέλο αναπτύχθηκε μέσω του tensorflow.keras και συγκεκριμένα:

Sequential για την αρχιτεκτονική του μοντέλου, LSTM και Dense επίπεδα για την κατασκευή του νευρωνικού, Dropout επίπεδο για την αποφυγή overfitting.

Για την συνάρτηση σφάλματος επιλέξαμε το μέσο τετραγωνικό σφάλμα το οποίο χρησιμοποιήσαμε μέσω του keras.metrics.

Επιπλέον, για να δημιουργήσουμε τις επόμενες ημέρες για το dataframe αξιοποιήσαμε την DateOffset της pandas.tseries.offsets.

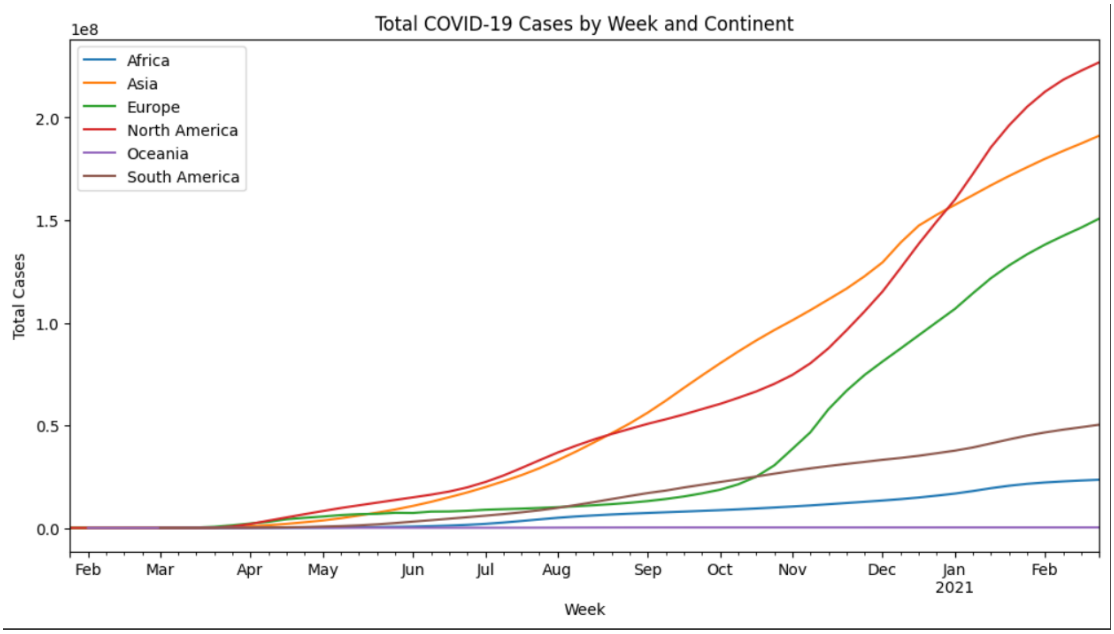
Τέλος από την sklearn.model_selection χρησιμοποιήσαμε το ParameterGrid με ένα διάστημα τιμών για κάθε παράμετρο ώστε να εντοπίσουμε με χρήση «ωμής βίας» τους βέλτιστους συνδυασμούς παραμέτρων.

Για την υλοποίηση του SVM δικτύου η μόνη επιπλέον βιβλιοθήκη ήταν η sklearn.svm και συγκεκριμένα το SVR.

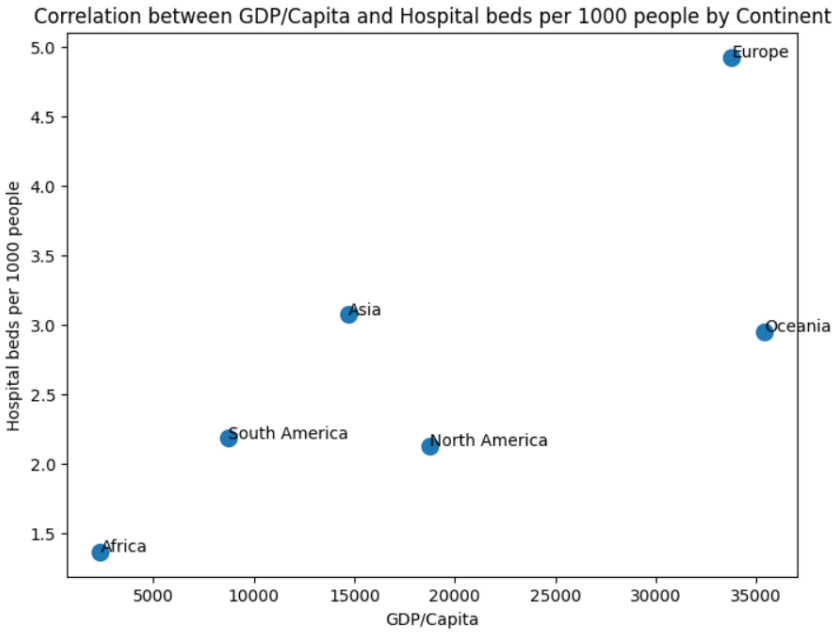
Διαδικασία Υλοποίησης

Αρχικά, για την καλύτερη κατανόηση του συνόλου δεδομένων που μας δόθηκε, υπολογίσαμε κάποια συγκεντρωτικά στατιστικά μεγέθη με τις αντίστοιχες γραφικές απεικονίσεις τους. Παραθέτουμε ενδεικτικά κάποια από αυτά:

Συνολικά κρούσματα ανά εβδομάδα για κάθε Ήπειρο (άξονας y ανά 100.000 άτομα) :

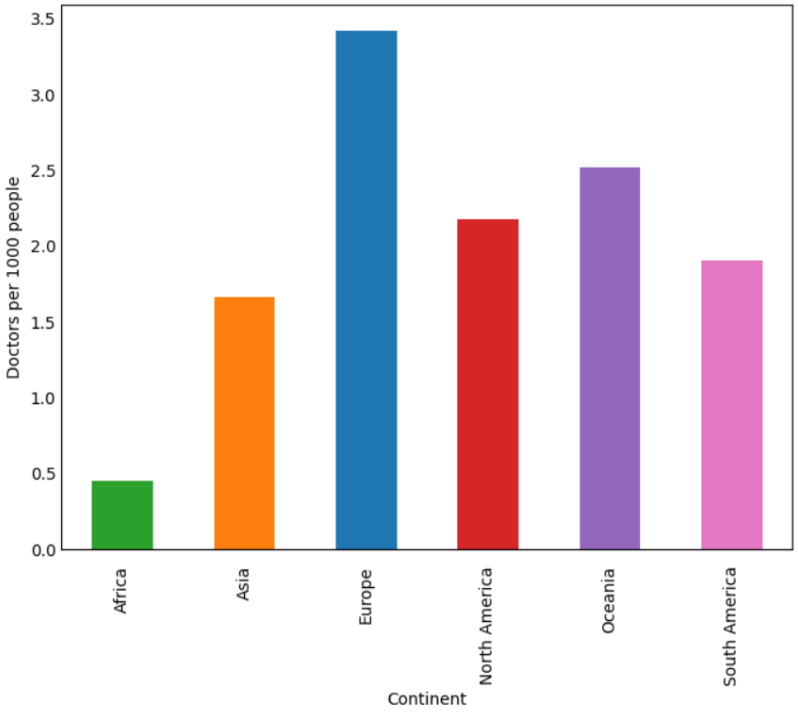


Συσχέτιση μεταξύ ΑΕΠ κατά κεφαλήν και νοσοκομειακά κρεβάτια ανά ήπειρο:

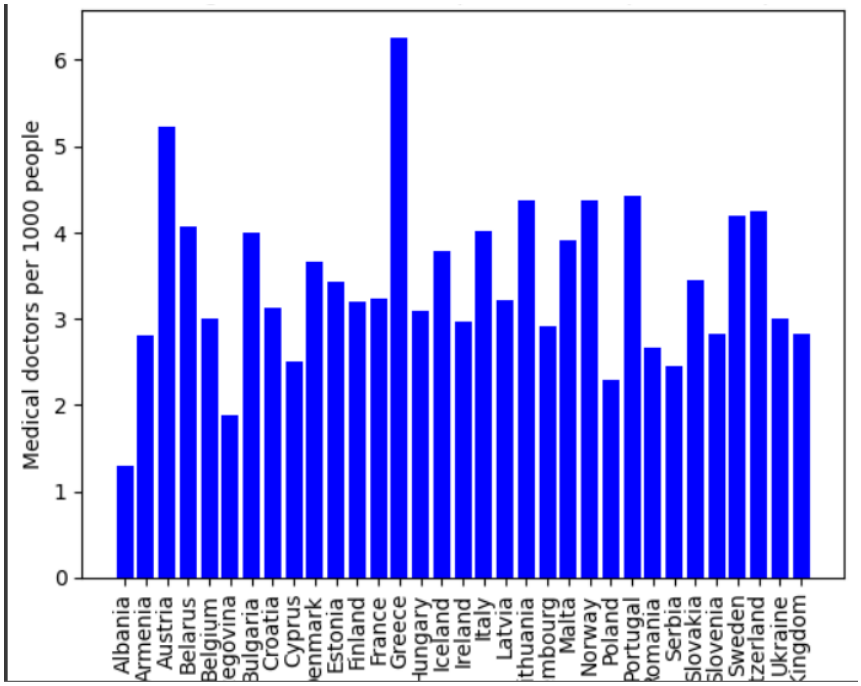


Αναφορικά με το ΑΕΠ κατά κεφαλήν, πρέπει να υπάρχει κάποιο λάθος με τα δεδομένα καθώς το αντίστοιχο μέγεθος για την Βόρεια Αμερική εμφανίζεται αισθητά χαμηλότερο από τα πραγματικά δεδομένα (πηγή: macrotrends).

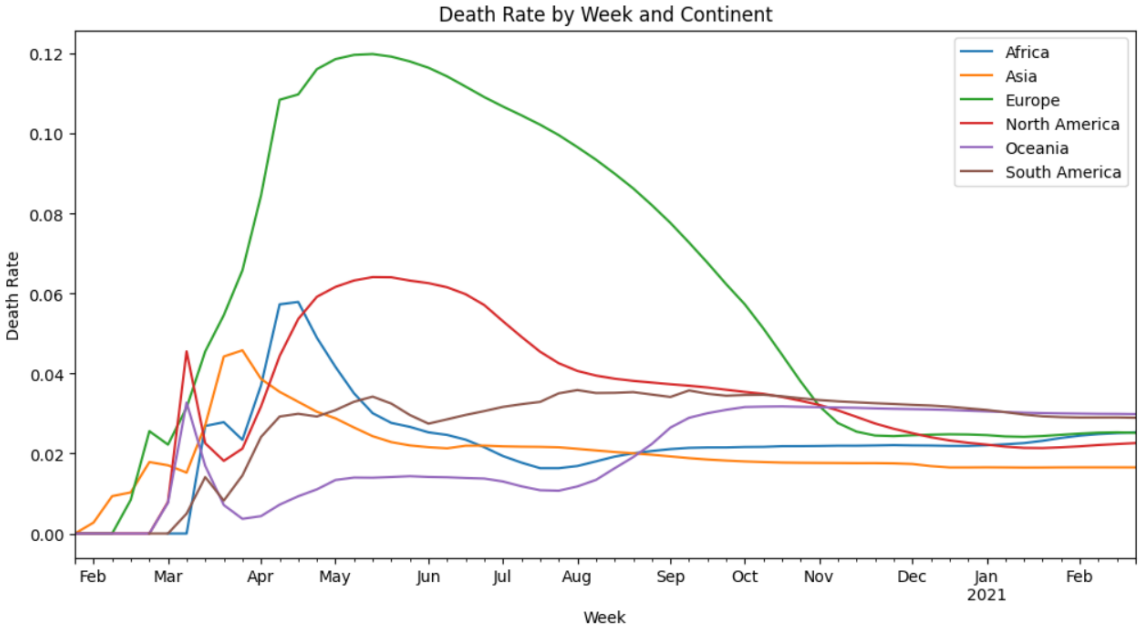
Γιατροί / 1000 ανθρώπου ανά Ήπειρο:



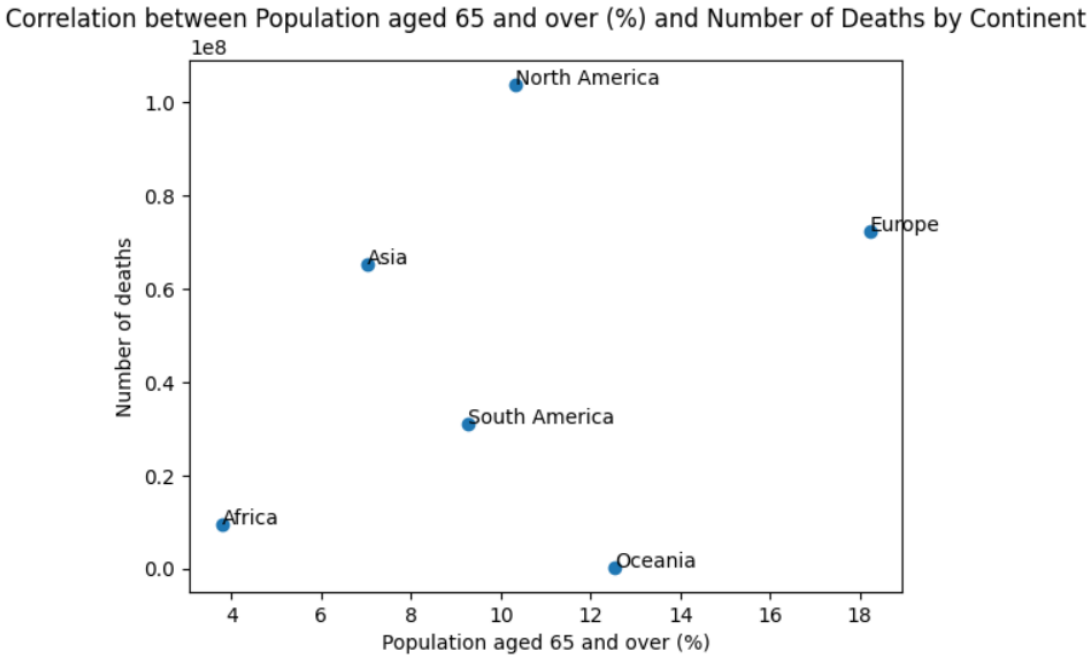
Γιατροί / 1000 ανθρώπους ανά Ευρωπαϊκή χώρα:



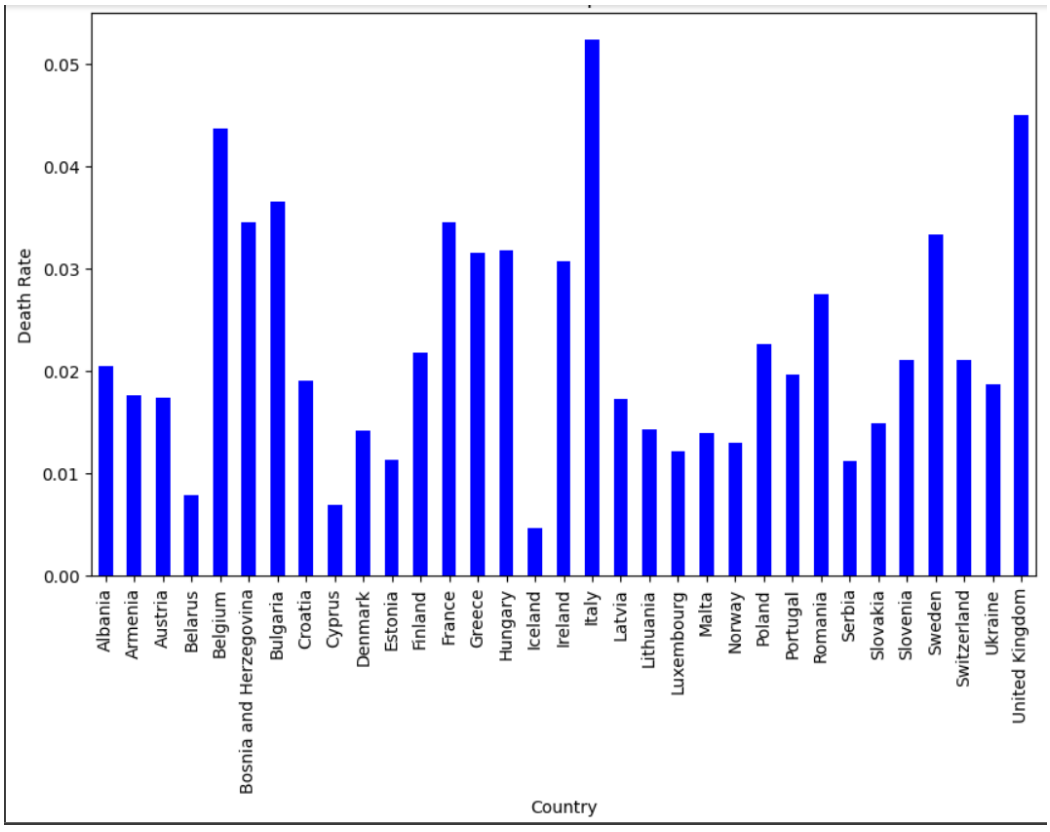
Ρυθμός θανάτων ανά εβδομάδα ανά Ήπειρο



Συσχέτιση μεταξύ των πληθυσμών άνω των 65 και αριθμού θανάτων ανά Ήπειρο.



Ρυθμός θανάτων για τις Ευρωπαϊκές χώρες:



Συμπερασματικά, αποτελέσματα τα όποια αναμέναμε εκ των προτέρων επιβεβαιώθηκαν και από το σύνολο των δεδομένων, πέραν του σημείου που τονίστηκε παραπάνω για το ΑΕΠ της βορείου Αμερικής. Συγκεκριμένα, ο αριθμός των θανάτων φάνηκε να έχει συσχέτιση με το ποσοστό πληθυσμών άνω των 65. Δηλαδή, οι « γηραιότερες» χώρες/ ήπειροι είχαν και υψηλότερα ποσοστά θνησιμότητας. Ακόμη, υπάρχει συσχέτιση μεταξύ του GDP/Capita και του αριθμού των νοσοκομειακών κρεβατιών με εξαίρεση την Βόρεια Αμερική, το οποίο επιβεβαιώνεται από τις αντίστοιχες δημοσιοοικονομικές πολιτικές.

Συσταδοποίηση

Πριν προβούμε στην συσταδοποίηση, πραγματοποιήσαμε κάποιο pre processing στα δεδομένα καθώς δεν ήταν δυνατές κάποιες από τις διαιρέσεις που θα αναφερθούν στη συνέχεια, αφού υπήρχαν κελιά που αφορούσαν για παράδειγμα αριθμούς θανάτων και ημερησίων τεστ τα οποία ήταν κενά. Συγκεκριμένα, γεμίσαμε τα κενά κελιά χρησιμοποιώντας τον μέσο όρο για την κάθε στήλη.

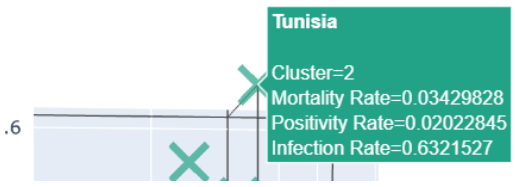
Για το clustering βασιστήκαμε στις επιδόσεις των χωρών στην αντιμετώπιση του ιού με βάση τα παρακάτω κριτήρια:

- Ποσοστό θετικότητας = Συνολικά Κρούσματα / Συνολικός Αριθμός Τεστ
- Ποσοστό θνησιμότητας = Συνολικός Αριθμός Θανάτων / Συνολικά Κρούσματα
- Ποσοστό μολυσματικότητας = Συνολικά Κρούσματα / Πληθυσμός

Θεωρήσαμε τις παραπάνω καλές μετρικές (προτείνονταν και από την εκφώνηση) με σκοπό την αξιολόγηση της αποτελεσματικότητας της κάθε χώρας στην αντιμετώπιση του ιού.

Δημιουργήσαμε έτσι 4 συστάδες τις χώρες όπου η επιτυχία αντιμετώπισης ακολουθεί αύξουσα σειρά (0-> πιο επιτυχής , 4 -> λιγότερο επιτυχής)

Αρνητικά, φαίνεται να ξεχωρίζουν χώρες όπως η Τυνησία και το Περού, οι οποίες βρίσκονται και κοντά μεταξύ τους και αποτελούν outliers.

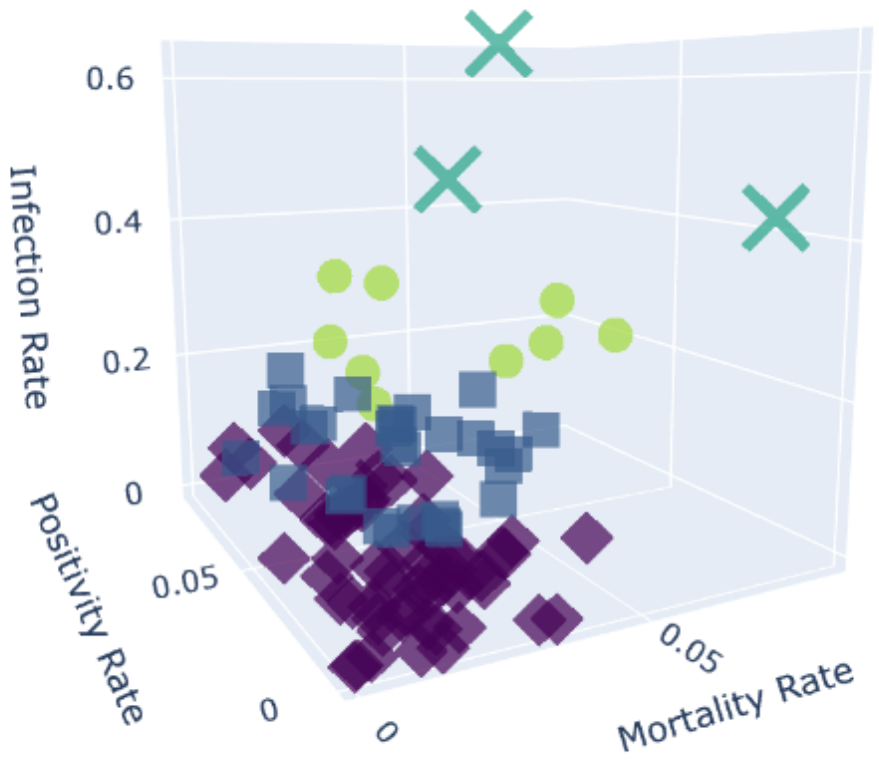


Υπάρχουν από την άλλη ορισμένες χώρες οι οποίες παρουσίαζαν σημαντική έλλειψη δεδομένων με αποτέλεσμα η αντιμετώπισή τους να φαίνεται υποδειγματική, όπως για παράδειγμα η Μογγολία. Χώρες με παρόμοια σημαντική έλλειψη αντιπροσωπευτικών – έμπιστων δεδομένων εμφανίζονται κυρίως στο cluster 0.

Mongolia

Cluster=0
Mortality Rate=677.5068μ
Positivity Rate=959.7981μ
Infection Rate=528.4524μ

Πέραν αυτού τα περισσότερα αποτελέσματα του cluster φαίνεται να είναι τοποθετημένα με ορθό τρόπο. Παρακάτω παραθέτουμε ενδεικτικό screenshot:



Υλοποίηση Νευρωνικών Δικτύων

Αρχικά, για το RNN υλοποιήσαμε LSTM καθώς είναι από τα πιο αποτελεσματικά για πρόβλεψη χρονοσειράς.

Το εκπαιδεύσαμε αρχικά μόνο σε data points που αφορούσαν την Ελλάδα, πράγμα το οποίο δεν ήταν αρκετά ακριβές και οι αποκλίσεις των προβλέψεων του νευρωνικού συγκριτικά με τα πραγματικά δεδομένα ήταν της τάξεως του 40%. Εκτιμούμε ότι αυτό οφειλόταν στον μικρό αριθμό δεδομένων τα οποία χρησιμοποιήθηκαν για το training. Για να πετύχουμε πιο ακριβείς προβλέψεις ωστόσο, εκπαιδεύσαμε το νευρωνικό παρέχοντας του δεδομένα από την Ελλάδα αλλά και από τις υπόλοιπες χώρες, κάτι τέτοιο όμως φαινόταν να περιπλέκει αρκετά την κατάσταση καθώς υπήρχαν πολλές χώρες outliers.Έτσι, αποφασίσαμε να εκπαιδεύσουμε το νευρωνικό με δεδομένα αυτή τη φορά από την Ελλάδα και επιπλέον αυτή τη φορά, από χώρες οι οποίες ήταν εξίσου αποτελεσματικές στην αντιμετώπιση του υιού, με βάση τα αποτελέσματα των μετρικών του clustering και το πώς αυτό διαμορφώθηκε. Πράγματι, κατ' αυτόν τον τρόπο οι προβλέψεις του νευρωνικού σημείωσαν σημαντική βελτίωση και ήταν πολύ πιο κοντά στα μεγέθη του αρχικού dataset.

Για την υλοποίηση, φορτώσαμε το αρχικό αρχείο δεδομένων το οποίο εμπεριέχει επιπλέον πληροφορίες οι οποίες προστέθηκαν κατά την διαδικασία του clustering. Έπειτα, φιλτράραμε το dataframe για να απομονώσουμε τις χώρες ενδιαφέροντος για την εκπαίδευση. Για κάθε χώρα χωριστά, υπολογίσαμε το ημερήσιο ποσοστό θετικότητας και αφαιρέσαμε τις γραμμές που αντιστοιχούν σε ημερομηνίες μεταγενέστερες της 1-1-2021. Τυχόν κενά στα δεδομένα, τα συμπληρώσαμε με μέθοδο bfill. Έπειτα, κάναμε scale τα δεδομένα με χρήση RobustScaler για να είναι τα δεδομένα πιο εύκολα διαχειρίσιμα από το νευρωνικό, ενώ δημιουργούμε τα time series με lookback window 14 για κάθε γκρουπ (χώρα). Μετά, χωρίσαμε τα δεδομένα σε training και testing sets με αναλογία 0.8 – 0.2. Έπειτα δηλώνουμε το είδος του μοντέλου, που είναι sequential, ορίζουμε 5 layers για το μοντέλο τα οποία αποτελούνται από LSTM με 64 νευρώνες, Dropout με πιθανότητα 0.2 για αποφυγή overfitting, άλλο ένα LSTM layer και 2 Dense Layers, ένα relu και ένα linear για την παραγωγή της εξόδου. Ως loss function χρησιμοποιήσαμε το μέσο τετραγωνικό σφάλμα καθώς φάνηκε να ταιριάζει με τα δεδομένα μας. Προπονούμε το μοντέλο, το κάνουμε evaluate και εν συνεχεία Τα αποτελέσματα του evaluation εκτυπώνονται, δείχνοντας την τιμή του loss στο σύνολο δοκιμών. Αυτό, μας παρέχει μια ένδειξη της απόδοσης του μοντέλου όσον αφορά την πρόβλεψη του ποσοστού θετικότητας. Για να κάνουμε την πρόβλεψη για τις 3 μέρες μετά την 1-1-2021 για την Ελλάδα, δίνουμε ως είσοδο στο νευρωνικό τις 14 τελευταίες μέρες για να προβλέψει την επόμενη. Αυτό βγάζει ως έξοδο την εκτιμώμενη θετικότητα την επόμενη ημέρα, δηλαδή την 2-1-2021. Για να προβλέψουμε τις 2 επόμενες ημέρες, αποθηκεύουμε το αποτέλεσμα της προηγούμενης πρόβλεψης στην λίστα των δεδομένων και στην λίστα lookback, επαναλαμβάνουμε 2 φορές και έχουμε εν τέλει τις προβλέψεις για τις 3 ζητούμενες ημέρες.

Παρόμοια διαδικασία ακολουθήθηκε για την υλοποίηση του SVM. Το SVM που χρησιμοποιήσαμε ήταν το SVR. Για την επεξεργασία των δεδομένων η διαδικασία είναι η ίδια. Το μοντέλο το αρχικοποιούμε με παραμέτρους (kernel='rbf', C=10, epsilon=0.1). Επιλέξαμε rbf Kerner καθώς φαίνεται να είναι πιο ευέλικτο στις προβλέψεις του, το C=10 το μοντέλο ταιριάζει στο training set αποφεύγοντας ωστόσο το overfitting και το epsilon = 0.1 για ένα λογικό περιθώριο σφάλματος στις προβλέψεις του δικτύου. Με παρόμοιο τρόπο αξιολογούμε την αποδοτικότητα του μοντέλου κάνοντας χρήση του μέσου τετραγωνικού σφάλματος αλλά και του απολύτου. Τελικώς, κάνουμε τις προβλέψεις για τις 3 ημέρες μετά την 1-1-2021 που μας ζητείται με τρόπο όμοιο με προηγουμένως.

Παρατηρήσαμε ότι το SVM βγάζει μικρότερο τετραγωνικό σφάλμα από το LSTM. Ενδεχομένως, το μεγαλύτερο τετραγωνικό σφάλμα του LSTM να οφείλεται στο μέγεθος συνόλου δεδομένων το οποίο μπορεί να θεωρηθεί σχετικά μικρό για ένα πολύπλοκο μοντέλο όπως είναι το LSTM όπου με περιορισμένο αριθμό δειγμάτων, μπορεί να δυσκολεύεται να γενικεύσει καλά γεγονός που μπορεί να οδηγήσει σε υψηλότερο MSE σε σύγκριση με απλούστερα μοντέλα όπως τα SVM. Συγκεκριμένα το LSTM μετά από 12 epochs έχει MSE ~=10 και στο validation set ~=8. Το SVM από την άλλη έχει MSE = 0.4.

Τα αποτελέσματα του LSTM για lookback=14 LST layer =14 και activation : tanh dropout 0.2 lstm 32 activation: tanh dense 15 activation: relu dense 1 activation: linear compiled με MSE και optimizer Adan για 3 epochs και batch size=7 προβλέπει τις 3 επόμενες ημέρες

Θετικότητα: 4.3%, 4.3%, 4.4%

Τα αποτελέσματα του SVR για Rbf ,c=10,e=0.1

Θετικότητα: 3.1% , 3.8%, 3.9%