ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

Παράδοση Project 2023

❖ **Η εργασία εκπονήθηκε από τους:**

Σφήκας Θεόδωρος ΑΜ:1072550 CEID 4° έτος

Ανδρέας Μοραΐτης ΑΜ:1057736 ΗΜ/ΤΥ 5° έτος

❖ Συνοπτικοί σχολιασμοί:

Το project υλοποιήθηκε με την χρήση της python και συγκεκριμένα με στόχο την ανάλυση των αποτελεσμάτων σε κάθε βήμα και την βηματική επεξήγηση του κώδικα χρησιμοποιήσαμε ένα Jupyter Notebook.

Όλα τα στοιχεία και τα ερωτήματα παραθέτονται μέσα στο αρχείο 'project.ipynb' που εμπεριέχεται στο zip αρχείο που παραδώθηκε. Καθώς τόσο ο κώδικας όσο και τα συμπεράσματα από αυτόν αναλύονται καλύτερα και εκτενώς μέσα στο Notebook θα προσπαθήσουμε στην αναφορά μας να είμαστε πιο περιληπτικοί αναφέροντας τα κεντρικά στοιχεία της σκέψης και των διαδικασιών που ακολουθήσαμε. Μέσα στο Notebook συχνά αναφέρουμε διαφορετικές προσεγγίσεις που θα μπορούσαμε να επιλέξουμε και πιθανούς τρόπους βελτιστοποίησης των αποτελεσμάτων, οι οποίοι δεν ακολουθήθηκαν λόγω περιορισμένων χρονικών δυνατοτήτων.

❖ Κεντρικές βιβλιοθήκες και εργαλεία:

- **Matplotlib**
- **Seaborn**
- **Plotly**
- **Scipy**
- **Pandas**
- **Numpy**
- **Scikit-learn**
- **Tensorflow**
- **Keras**
- **Nbformat**
- **Ipykernel**

Παραδώθηκε μαζί με τον κώδικα το αρχείο 'requirements.txt' το οποίο μπορεί να χρησιμοποιηθεί για να εγκαταστήσει αυτόματα όλες τις απαραίτητες βιβλιοθήκες με την χρήση της εντολής 'pip install –r /path/to/requirements.txt'. Προτείνεται η δημιουργία και το sourcing ενός virtual environments πρωτού εκτελεστεί η παραπάνω εντολή, ωστόσο αυτό δεν είναι αναγκαίο (python –m venv ./venv).

Μοναδική δυσκολία υπάρχει περίπτωση να παρουσιάσει η βιβλιοθήκη Plotly που χρησιμοποιείται για 3D γραφήματα μέσα σε ένα jypyter Notebook και να μην εμφανίζει τα γραφήματα λόγω παλιότερου nbformat στο Notebook. Η εκτέλεση της εντολής 'pip install --upgrade jupyter notebook nbformat' και επανεκκίνηση του Notebook διορθώνει τυχόν πρόβλημα.

❖ Συνοπτική ανάλυση της διαδικασίας και συμπεράσματα:

1. Preprocessing:

Το πρώτο βήμα της διαδικασίας αποτελεί η κατανόηση των στοιχείων του dataset και ανάλυση των βασικών στατιστικών του στοιχείων όπως υποδεικνύεται και στην εκφώνηση.

Με στόχο την ουσιαστική οπτικοποίηση του dataset επιλέξαμε αρχικά να παραθέσουμε ορισμένα ιστογράμματα με βάση τα διαφορετικά features του dataset ώστε να διαθέτουμε ένα frequency analysis των τιμών τους.

Στην συνέχεια επιλέξαμε να αθροίσουμε τις μετρήσεις όλων των χωρών στα features των test του Covid-19, των κρουσμάτων και των θανάτων που καταγράφθηκαν για κάθε ημερομηνία του dataset. Η κλίμακα που χρησιμοποιήσαμε είναι λογαριθμική καθώς ο τεράστιος αριθμός των tests σε σύγκριση με τις άλλες τιμές δυσκόλευε την αναγνωσιμότητα του γραφήματος. Επίσης για τα tests στο τέλος των ημερομηνιών παρατηρούμε μια απότομη πτώση που πιθανώς ευθύνεται στις ελλιπές τιμές του dataset.

Με το τρίτο γράφημα επιλέγουμε να αναδείξουμε πως δεν έχουν όλες οι χώρες μετρήσεις για όλες τις ημέρες, γεγονός που είναι σημαντικό στοιχείο ανομοιογένειας των δεδομένων σε κάθε χώρα. Στην συνέχεια θα χρειαστεί να αντιμετωπίσουμε παρόμοια προβλήματα για την ορθότερη εξόρυξη αποτελεσμάτων.

Στην παράγραφο "First observation from a basic statistical analysis" του Notebook παραθέτουμε τα πρώτα συμπεράσματα μας παίρνοντας την θέση ενός data analyst. Τα οποία στην συνέχεια μας οδηγούν και στην δημιουργία ενός correlation matrix μεταξύ των features του dataset για την ανάδειξη των συσχετίσεων που θα μας απασχολήσουν αργότερα.

Τέλος επιλέξαμε βάση ορισμένων συγκεντρωτικών στατιστικών στοιχείων να κατατάξουμε τις χώρες πάνω στην αντιμετώπιση της πανδημίας του Covid-19 με την χρήση boxplots. Καθώς οι υπολογισμένες τιμές δεν είναι normalized με βάση τον πληθυσμό κάθε χώρας και την διασπορά των μετρήσεων, οι παραγόμενοι outliers των boxplots δεν αποτελούν κάποια σοβαρή ένδειξη των αποτελεσμάτων που θα ακολουθήσουν σε επόμενα ζητούμενα.

Έχοντας υλοποιήσει βασικές εφαρμογές οπτικοποίησης, συνεχίζουμε στην προετοιμασία του dataset και στο γέμισμα πεδίων με διάφορες μεθόδους. Συγκεκριμένα πρέπει να επιλέξουμε ορισμένες χώρες τις οποίες λόγω συντριπτικού αριθμού ελλιπών πεδίων κρίνουμε unusable. Θεωρήσαμε πως μετά από ένα ποσοστό του 50% σε ανύπαρκτα πεδία τα οποία δεν μπορούν να συμπληρωθούν, οι συγκεκριμένες χώρες είναι άχρηστες για την ανάλυση μας και αφαιρούνται. Έπειτα μπορούμε να παρακολουθήσουμε πως τις πρώτες εβδομάδες οι περισσότερες χώρες παρουσιάζουν σημαντικό αριθμό ελλιπών στοιχείων και μάλιστα σε αλληλουχία. Όπως παρατηρήσαμε και με το τρίτο γράφημα που

υλοποιήσαμε, για την συγκεκριμένη χρονική περιόδο παρατηρούμε πως πολλές χώρες δεν διαθέτουν καν records όλων των ημερομηνίών. Ως εκτούτου επιλέξαμε να τις αφαιρέσουμε, μειώνοντας σημαντικά τον αριθμό των κενών στοιχείων στο dataset. Για όσες λίγες τμές έλλειπαν αναφέρουμε πολλούς τρόπους γεμίσματων των πεδίων ωστόσο επιλέξαμε να κάνουμε interpolation βάση των γειτονικών του τιμών. Η συγκεκριμένη διαδικασία της προετοιμασίας του dataset αποτελεί ένα από τα βασικότερα βήματα οποιασδήποτε ανάλυσης, καθώς οτιδήποτε ακολουθεί είναι άρρηκτα συνδεδεμένο με την προ-επεξεργασία. Προσπαθήσαμε να κάνουμε evaluate την διαδικασία της προεπεξεργασίας ξανά-υπολογίζοντας τα στατιστικά στοιχεία και το correlation matrix του dataset, προσέχοντας να μην έχουν γίνει πιθανές αλλαγές στις συσχετίσεις και στις κατανομές των features, που μπορεί να υποδεικνύουν την καταστροφή των αρχικών μας στοιχείων.

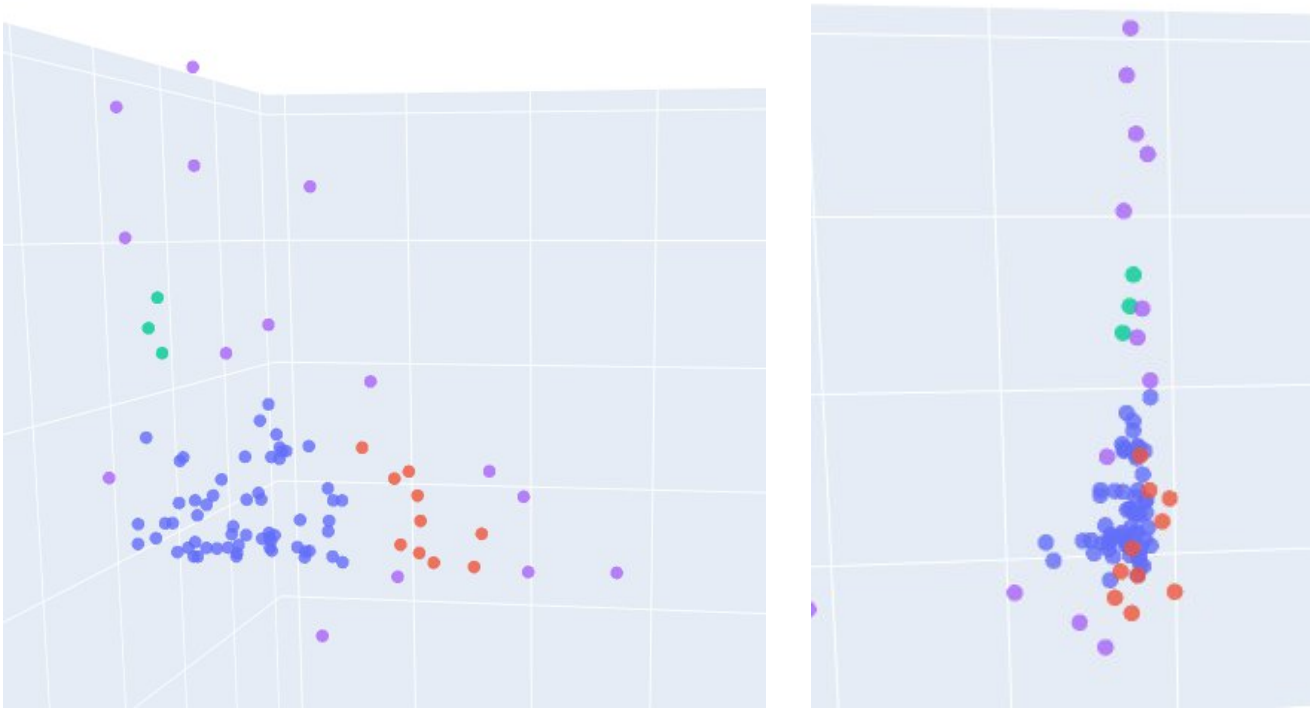
2. Clustering

Προσπαθώντας να αναλύσουμε την απόδοση των χωρών πάνω στην αντιμετώπιση τους για την πανδημία, όπως καταγράφουμε και στο Notebook, πρέπει να επιλέξουμε τα features σύμφωνα με τα οποία θα κάνουμε την κατηγοριοποίηση και έπειτα να επιλέξουμε πως θα τα κάνουμε scale στον ίδιο range. Σύμφωνα και με τις υποδείξεις τις εκφώνησης επιλέξαμε τα possitivity rate, fatality rate and the rate of tests/population. Προσπαθήσαμε να χρησιμοποιήσουμε άλλους παράγοντες όπως τον δείκτη των νεκρών κάθε χώρας με βάση τις διαθέσιμες κλινικές ή τους γιατρούς που διαθέτουν,. Ωστόσο καθώς δεν ήμασταν σίγουροι πως τέτοιες ενδείξεις παρατηρούν την γενική απόδοση στην πανδημία ή αν υποδεικνύουν στοιχεία όπως η σημασία κλινικών ΜΕΘ στην θνησιμότητα, δεν συμπεριλήφθηκαν.

Για το Scaling των features επιλέξαμε συνδιασμό των μεθόδων του RobustScaler και του MinMaxScaler για την επεξεργασία των features ώστε να αποφευχθεί πιθανό dominance μιας παραμέτρου κατά την διαδικασία του clustering. Για το clustering των αποτελεσμάτων επιλέξαμε την χρήση του αλγορίθμου DBSCAN, έχοντας δοκιμάσει επίσης και τον αλγόριθμο του Kmeans. Η επιλογή των υπερπαραμέτρων του μοντέλου επιλέχθηκε με βάση ένα Density Reachability Plot. Η βασική επιλογή έγινε λόγω της Density based φύσης του DBSCAN καθώς και την προσπάθεια αποφυγής του oversampling που απαιτούσε ο αλγόριθμος του Kmeans για να παράξει καλύτερα αποτελέσματα, ωστόσο και οι δύο αλγόριθμοι παρουσιάζουν θετικά και αρνητικά. Και με τις δύο διαδικασίες καταλήξαμε σε 4 clusters ως την καλύτερη ομαδοποίηση των εισόδων, ωστόσο στον DBSCAN ένας εξ αυτών είναι ο 'cluster' -1 ο οποίος συμπεριλαμβάνει τις χώρες που δεν ομαδοποιήθηκαν πουθενά. Οπτικοποιώντας τα αποτελέσματα του clustering (χρησιμοποιώντας την plotly) παρατηρούμε πως τα αποτελέσματα του DBSCAN φαίνονται ορθά, ακόμα και αν πολλές χώρες έχουν κατηγοριοποιηθεί ως outliers.

Παρατηρούμε πως η μεγαλύτερη πλειοψηφία των χωρών εμφανίζει παρόμοια στατιστικά για το positivity rate ενώ διαφοροποιούνται στους παράγοντες της θνησιμότητας και στο

αναλογικό πλήθος των τέστ βάση του πληθυσμού (tests / population). Μπορούμε να παρατηρήσουμε χώρες που τα πήγαν καλά στην αντιμετώπιση τους, οι οποίες διακρίνονται από πολλά test/population και λίγους θανάτους (πράσινες), σε χώρες που παρουσιάζουν μέτρια αποτελέσματα (μεγαλύτερη πλειοψηφία των χωρών - μπλε) και σε αυτές που είχαν μεγάλη θνησιμότητα ενώ είχαν ίδιο βαθμό θετικότητας και tests με την πλειοψηφία των χωρών (πορτοκαλί – εμπεριέχεται και η Ελλάδα).



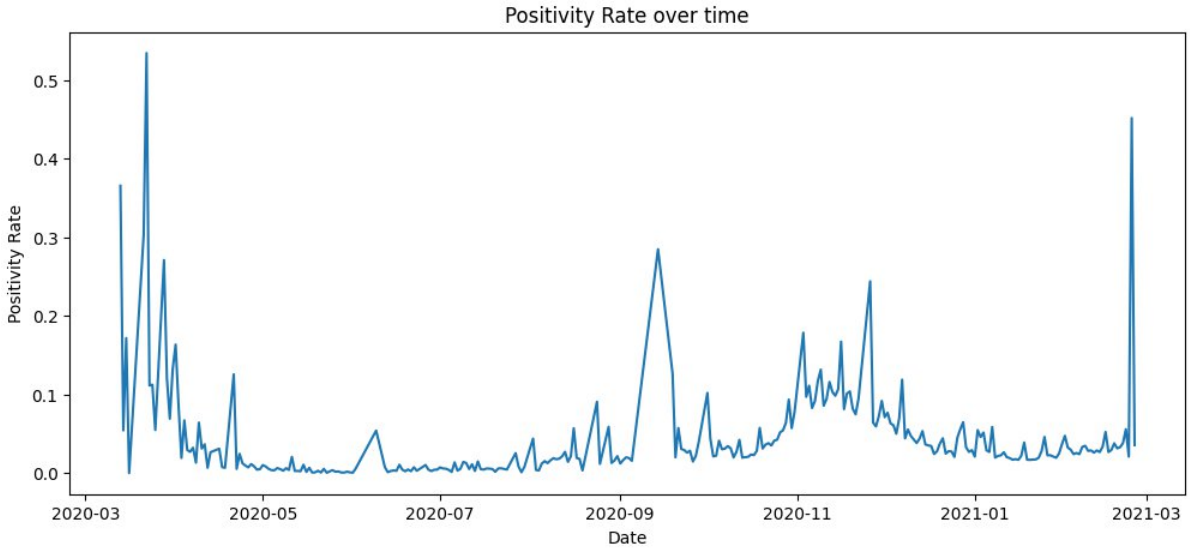
Στην πρώτη εικόνα βλέπουμε *"en face"* το fatality των χωρών ενώ στην δεύτερη το positivity rate. Είναι σημαντικό να παρατηρηθεί πως η πλειοψηφία των χωρών έχουν βρεθεί στον ίδιο υπόχωρο υποδεικνύοντας την παρόμοια αντιμετώπιση της πανδημίας. Ωστόσο μπορούμε να παρατηρήσουμε πως οι outliers παρά το scaling, επηρέασαν αρκετά τα δεδομένα ώστε να μην μπορούμε να πάρουμε βέλτιστα αποτελέσματα. Συγκεκριμένα καθώς υπάρχουν outliers που φτάνουν στα όρια του δυνατού range τιμών, και μάλιστα με μεγάλη απόκλιση σε σύγκριση με άλλες χώρες, τα υπόλοιπα data points "συμπιέζονται". Πιθανώς διαφοροποιήσεις του DBSCAN όπως ο HDBSCAN να μπορούσαν να ομαδοποιήσουν καλύτερα τα δεδομένα μας. Παρουσιάζεται λοιπόν η μεγάλη αξία του scaling. Δοκιμάσαμε να αντικαταστήσουμε επίσης τον Robust Scaler με Standardization ωστόσο δεν μετέβαλλε τα αποτελέσματα.

Παραδειγματικά, οι χώρες πάνω αριστερά στην πρώτη εικόνα, αν και μπορούμε να πούμε πως συγκριτικά με τις υπόλοιπες τα πήγαν πολύ καλύτερα στην αντιμετώπιση της πανδημίας, διαφέρουν αρκετά πολύ μεταξύ τους για να τις ομαδοποιήσει ο DBSCAN, γεγονός που μπορεί να είναι θεμιτό ή αρνητικό. Στην δεύτερη περίπτωση ο Kmeans μπορεί να παράξει πιο επιθυμητά αποτελέσματα, ομαδοποιούσε ωστόσο τους μπλε και πορτοκαλί clusters. Στα πλαίσια 'Conclusions from the clustering process' στο Notebook αναφέρουμε αναλυτικότερα μερικά εκ των παραπάνω συμπερασμάτων καθώς και κοινά χαρακτηριστικά των χωρών βάση γενικότερων υποθέσεων.

Αναλυτικότερα μπορούμε να παρατηρήσουμε την αντίστροφη συσχέτιση μεταξύ του fatality και των tests/population που δεν είχε διακριθεί προηγουμένως και καταλήγουμε πως η καλύτερη γενίκευση που μπορεί να παρατηρηθεί είναι πως αναπτυσσόμενες χώρες, με πολύ χαμηλό GPT και αδύναμο σύστημα υγείας, όπως χώρες της λατινικής Αμερικής και της Αφρικής φάνηκαν αδύνατα να αντιμετωπίσουν την πανδημία και βρίσκονται ως outliers. Αντίθετα οι χώρες των άλλων τριών clusters δεν φαίνεται να παρουσιάζουν κάποια συσχέτιση ούτε στην γεωγραφική τους κατανομή, ούτε στις οικονομικές τους δυνατότητες, υποδεικνύοντας πως τα μέτρα που πάρθηκαν ή μη ευθύνονται σε μεγάλο βαθμό στην κατάταξη τους.

3. Prediction of Possitivity Rate in Greece

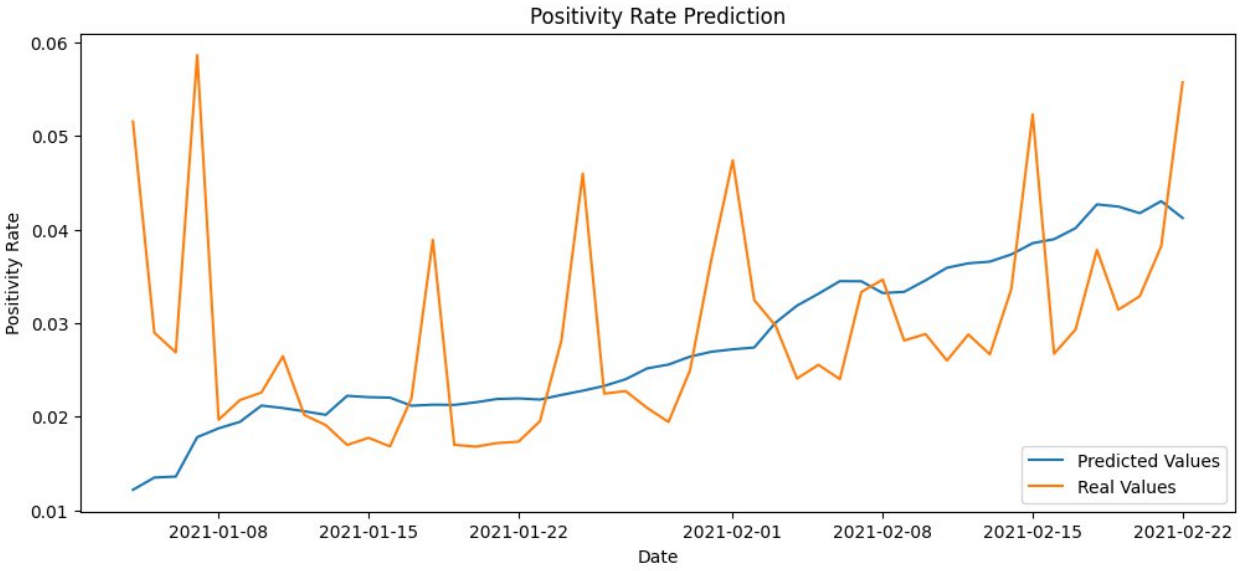
Προσπαθώντας να προβλέψουμε το possitivity rate αρχικά πρέπει να επαναλάβουμε τα βήματα που είχαμε υλοποιήσει στην προεπεξεργασία, γεμίζουμε ή αφαιρούμε καταλλήλως records και features που δεν μπορούν να βοηθήσουν στην διαδικασία της πρόβλεψης. Με στόχο να μπορούμε να χρησιμοποιήσουμε το πεδίο των ημερομηνιών πρέπει είτε να το ορίσουμε ως η απόσταση από ένα reference day ή ως την απόσταση σε δευτερόλεπτα από ένα reference timestamp. Οι τιμές του positivity rate παρουσιάζονται παρακάτω:



Στις αρχές και τέλη του dataset παρατηρούμε τιμές που θα μπορούσαν να κατηγοριοποιηθούν ως noise λόγω της ρητής φύσης του positivity rate. Τα απότομα spikes που παρατηρούμε υποδεικνύουν μια σημαντική δυσκολία στην εκπαίδευση των μοντέλων μας. Επίσης ειναι σημαντικό να αναφερθεί πως τα δεδομένα δεν είναι αρκετά για να εκπαιδεύσουμε διάφορα μοντέλα και ίσως μπορούμε να καταφύγουμε σε oversampling και interpolation μεταξύ των πραγματικών τιμών.

Στην διαδικασία του SVR, για την εύρεση των υπερ-παραμέτρων του μοντέλου χρησιμοποιήσαμε ένα GridSearch εκπαιδεύοντας το μοντέλο με όλες τις τιμές έως την 1/1/2021. Οφείλουμε επίσης να ακολουθήσουμε την προσπάθεια μίας online εκπαίδευσης, καθώς περνάει ο χρόνος με τις νέες γνωστές τιμές του positivity rate επανεκπαιδεύσουμε το μοντέλο. Ωστόσο όπως επεξηγούμε και πιο αναλυτικά στο Notebook, οι συγκεκριμένες

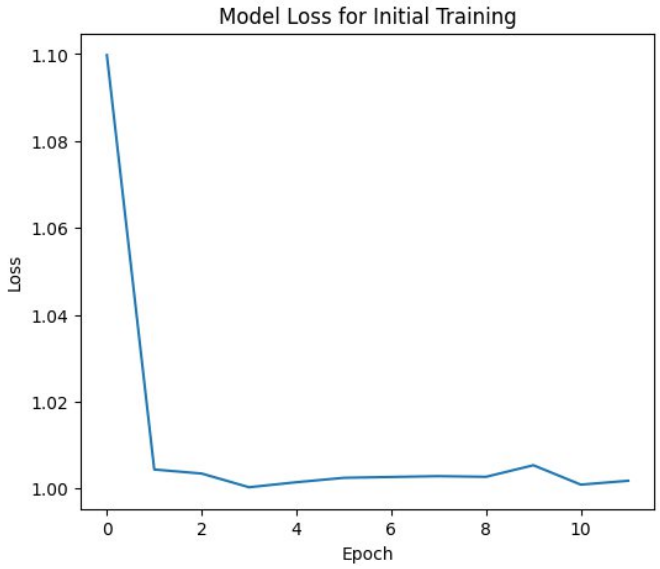
τιμές δεν είναι αναγκαστικά και οι βέλτιστες για τα δεδομένα που θέλουμε να κάνουμε predict. Με το gamma ορισμένο σε 'scaled' και μεγαλύτερο C από ότι συνηθίζεται (αν και μεγαλώνουμε τον κίνδυνο του overfitting στο μοντέλο μας) μπορούμε να ακολουθήσουμε καλύτερα τις ελαφριές εναλλαγές τιμών στο prediction stage. Δίνουμε μεγαλύτερο βάρος δηλαδή στις νέες τιμές του μοντέλου προσπαθώντας να ακολουθήσουμε πιο πιστά μικρές αλλαγές του positivity rate. Πρέπει επίσης να αναλογιστούμε εάν μπορούμε να χρησιμοποιήσουμε features όπως το fatality rate και το tests/population, με στόχο να τα συσχετίσουμε με μελοντικό possitivity rate. Ωστόσο συγκεκριμένες διαδικασίες μπορούν να προβούν και σε χειρότερα αποτελέσματα ανάλογα με την υλοποίηση, την σχεδίαση και την φύση του προβλήματος. Επιλέξαμε να επικεντρωθούμε μετά από δοκιμές αποκλειστικά στο temporal relationship του possitivity rate για το training του μοντέλου.

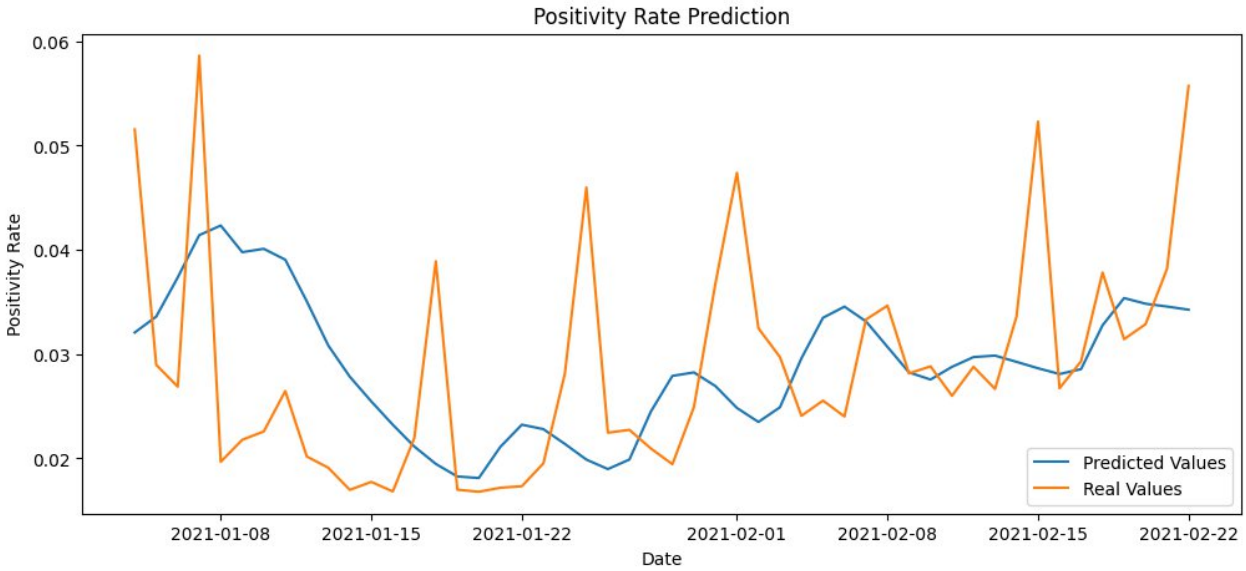


Για το evaluation, κρίνουμε καλύτερο να παρατηρήσουμε πόσο καλά ακολουθεί το μοντέλο μας τις πραγματικές τιμές ή το R2 score, καθώς κοινές μετρικές όπως το MSE λόγω των χαμηλών τιμών του MSE < 1 δεν είναι τόσο αντιπροσωπευτικές των αποτελεσμάτων, αν και στο Notebook υπολογίζονται σε κάθε βήμα. Παρατηρούμε πως δυστυχώς δεν έχουμε καταφέρει να ορίσουμε ένα μοντέλο αρκετά περίπλοκο ώστε να προβλέπει το temporal sequence των spikes αλλά απλά ακολουθεί ορθά το γενικό trend των πραγματικών τιμών.

Η εκπαίδευση και η ορθή σχεδίαση ενός νευρωνικού δικτύου παρουσιάστηκε δυσκολότερη διαδικασία από το SVR. Η διαδικασία που ακολουθείται για την εκπαίδευση των δύο μοντέλων επιλέξαμε να είναι η ίδια. Θα μπορούσαμε πιθανώς να χρησιμοποιήσουμε και διαφορετικά features για το training όπως το fatality rate και τα tests/population, τα οποία σε συνδιασμό με την χρήση ενός πιο περίπλοκου μοντέλου να οδηγούσαν σε καλύτερα αποτελέσματα όπως προαναφέραμε, αλλά κάτι τέτοιο δεν υλοποιήθηκε ούτε στο RNN. Κατά την σχεδίαση της αρχιτεκτονικής έπρεπε να επιλέξουμε τόσο τις υπερ-παραμέτρους της εκπαίδευσης όσο και τις υπερ-υπερ-παραμέτρους του μοντέλου. Για το RNN επιλέξαμε ένα LSTM layer καθώς χρησιμοποιούμε και μόνο ένα feature για το training. Πραγματοποιήσαμε πολλές δοκιμές και είδαμε στην πράξη πως η άστοχη περιπλοκή της αρχιτεκτονικής οδηγεί συχνά σε χειρότερα αποτελέσματα. Ενώσαμε το LSTM layer με ένα fully connected dense layer ώστε να επιτύχουμε ουσιαστικά την λειτουργία του regressor

καθώς ορίσαμε το output dimension με 1. Η επιλογή των epochs της εκπαίδευσης έγινε με βάση ένα trial and error approach χρησιμοποιώντας παραδείγματα και το history της εκπαίδευσης. Η online εκπαίδευση ενός νευρωνικού έχει φυσικά πολύ μεγαλύτερη λογική στο υπολογιστικό κόστος από ότι στο SVM. Σε κάθε μέρα επανεκπαιδεύουμε το μοντέλο με το πραγματικό positivity rate και ύστερα κάνουμε την πρόβλεψη 3 ημερών μετά. Μεγαλώνοντας τα epochs σε αυτή την διαδικασία, δίνουμε πάλι μεγαλύτερη βαρύτητα στις νέες τιμές του μοντέλου για τους ίδιους λόγους που προαναφέραμε στην διαδικασία του SVM. Κατανοούμε πως έχουμε overfitting, όπως φαίνεται στο γράφημα που ακολουθεί λογω του overreaction του μοντέλου στις νέες τιμές. Παρόλα αυτά παρατηρούμε πως το Μέσο τετραγωνικό λάθος σε αυτό το διάστημα προβλέψεων μειωνόταν (χάνουμε ωστόσο το generalization).





Μπορούμε να παρατήσουμε πως τα διάφορα spikes επηρεάζουν το μοντέλο με μια μικρή χρονοκαθυστέρηση, η απότομη φύση τους ωστόσο δεν βοηθάει τα αποτελέσματα καθώς το positivity rate έχει πέσει σε φυσιολογικά ποσοστά πρωτού το μοντέλο μπορέσει να αξιοποιήσει την ραγδαία αλλαγή (overreaction). Παρά τις προηγούμενες παρατηρήσεις το μοντέλο κάνει καλύτερες προβλέψεις σε σχέση με τον SVR. Ένας ακόμα τρόπος που θα μπορούσε να βελτιώσει το μοντέλο είναι η αλλαγή του input layer σε διάστημα συνεχόμενων 6 ημερομηνιών και η παραγογή μοναδικής τιμής εξόδου για 3 μέρες μετά. Με αυτό τον τρόπο ίσως το μοντέλο να μπορούσε να προβλέψει το επόμενο spike στην τιμή του possitivity rate, ωστόσο οι τιμές και τα spikes είναι κατά βάση irregular.

Φυσικά μπορούμε να συμπεριλάβουμε πως τόσο στο Clustering όσο και στις διαδικασίες των Predictors τα αποτελέσματα θα μπορούσαν να εκφράζουν σε πολύ καλύτερο βαθμό και να είχαμε δοκιμάσει περισσότερες τεχνικές, οι οποίες λόγω περιορισμένου χρονικού περιθωρίου δεν αναλύθηκαν περαιτέρω.