



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

**ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ**

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

**Ανάλυση δεδομένων κορονοϊού και πρόβλεψη μελλοντικών
κρουσμάτων
Covid data analysis and prediction of future cases**

**Πέτρος Βενιέρης
Α.Μ: Ε18023**

**Επιβλέπουσα Καθηγήτρια:
Μαρία Χαλκίδη, Αναπληρώτρια Καθηγήτρια**

Εργασία υποβληθείσα στο Τμήμα Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς
ΠΕΙΡΑΙΑΣ ,Σεπτέμβριος 2022

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ανάλυση δεδομένων κορονοϊού και πρόβλεψη μελλοντικών κρουσμάτων
Covid data analysis and prediction of future cases

Πέτρος Βενιέρης

A.M.: E18023

ΠΕΡΙΛΗΨΗ

Τα τελευταία 2 χρόνια με την εμφάνιση του κορονοϊού στον κόσμο ανέβηκε απότομα η ανάγκη διαχείρισης δεδομένων τέτοιου τύπου και οι κινήσεις των χωρών βασίζονται πάνω σε αυτά. Ταυτόχρονα ζούμε στην εποχή των big data και τα δεδομένα σιγά σιγά γίνονται <<χρυσός>> , όλες οι εταιρίες , χώρες κτλ βασίζονται σε δεδομένα για να κάνουν κινήσεις και επιχειρηματικές επιλογές, έτσι η ανάγκη για ανάλυση δεδομένων καθώς και την διαχείριση τους και αξιοποίηση τους κατάλληλα ώστε να βοηθήσουν στις μελλοντικές αυτές επιλογές όλο και αυξάνεται. Η συγκεκριμένη εργασία θα βασιστεί στον καθαρισμό , ανάλυση δεδομένων , παρουσίαση τους με γραφήματα . έπειτα θα φτιαχτούν προβλεπτικά μοντέλα με χρήση : polynomial regression, svm(support vector machine), holt models , arima models, sarima , prophet ώστε να βρούμε την καλύτερη δυνατή προβλεπτική λύση καθημερινών αλλά και συνολικών κρουσμάτων. Επίσης φτιάχνεται μια εφαρμογή με χρήση pickle, html , flask app που θα μπορεί οποιοσδήποτε χρήστης απλώς να βάλει δεδομένα κορονοϊού ενός συγκεκριμένου τύπου και να διαλέξει το μοντέλο που επιθυμεί και να δει τα μελλοντικά αποτελέσματα.

Για κάθε αποτέλεσμα κάθε μοντέλου γίνεται έλεγχος με χρήση root mean square error (mse) μεταξύ των δεδομένων που έχουμε και τον δεδομένων που προβλέψαμε.

Στην εφαρμογή χρησιμοποιούνται τα 2 πιο αποτελεσματικά προβλεπτικά μοντέλα και όχι όλα.

Η εργασία ακουμπάει περισσότερο την θεματική έννοια του data science και machine learning καθώς και εστιάζει περισσότερο στην πρόβλεψη των κρουσμάτων και την δημιουργία της εφαρμογής. Η έννοια του data science δεν αφορά απλώς την χρήση αλγορίθμων αλλά βασίζεται στην κατανόηση στόχων , επιλογή αλγορίθμων και στην σωστή διαχείριση και μετατροπή των δεδομένων ώστε να μπορούν να διαβαστούν να επεξεργαστούν και να περαστούν σε κάθε ένα μοντέλο ξεχωριστά καθώς για κάθε μοντέλο απαιτείται διαφορετική προσέγγιση, επίσης απαιτείται η κατανόηση και η σωστή μοντελοποίηση τους ώστε να μπορούν να αξιοποιηθούν. Είναι δηλαδή ένας τομέας που απαιτεί τόσο γνώσεις όσο και κατανόηση προβλήματος και επιχειρήσεων όσο και δημιουργική σκέψη και problem solving skills

Τα συμπεράσματα μετά την εκπόνηση της εργασίας είναι ότι παρά τον μεγάλο αριθμό δεδομένων (120 ευρωπαϊκά συνολικά εκατομμύρια κρούσματα πχ) είναι δυνατόν με τη σωστή διαχείριση να προβλεφθούν με μικρό error και τα συνολικά και τα καθημερινά κρούσματα , καθώς και να χρησιμοποιηθεί αυτή η πληροφορία ή τα γραφήματα και η ανάλυση δεδομένων για κατάλληλες επιχειρηματικές κινήσεις και βλέπουμε το πώς συμπεριφέρεται κάθε μοντέλο ξεχωριστά.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Data science, ανάλυση δεδομένων και πρόβλεπτικά μοντέλα

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: ανάλυση δεδομένων,data science,μεγάλα δεδομένα, μηχανική μάθηση, προβλεπτική ανάλυση

Covid data analysis and prediction of future cases

ABSTRACT

In the last 2 years with the emergence of the coronavirus in the world, the need for data management of this type has risen sharply and countries' movements are based on it. At the same time we are living in the era of big data and data is slowly becoming <<gold>>, all companies, countries etc rely on data to make moves and business choices, so the need for data analysis as well as managing and utilizing data appropriately to help in these future choices is increasing. This paper will be based on cleaning, analyzing data, presenting it with graphs. Then predictive models will be built using: polynomial regression, svm, holt models, arima models, sarima, prophet to find the best possible predictive solution of daily but also total counts. Also an app is being made using pickle, html, flask app that any user can just put coronavirus data of a particular type and choose the model they want and see the future results.

For each result of each model a check is done using mean square error (mse) between the data we have and the predicted data.

In the application the 2 most effective predictive models are used, not all of them.

The paper touches more on the thematic concept of data science and machine learning as well as focuses more on predicting the cases and creating the application. The understanding objectives, selection of algorithms and proper management and transformation of data so that it can be read to process and passed to each model individually as each model requires a different approach, it also requires understanding and proper modeling so that it can be utilized. In other words, it is an area that requires both knowledge and understanding of the problem and business as well as creative thinking and problem solving.

The conclusions after the work is that despite the large amount of data (120 European total million cases) it is possible with the right management to predict with a small error both the total and the daily cases, and to use this information or the graphs and data analysis for appropriate business moves and we also see how each model act on our data.

TOPIC AREA: Data science, Data analysis and predictive models

KEYWORDS: data analysis, data science, big data, machine learning, predictive analysis

ΕΥΧΑΡΙΣΤΙΕΣ

Η ολοκλήρωση της πτυχιακής αυτής εργασίας θα ήταν αδύνατη χωρίς την πολύτιμη υποστήριξη της καθηγήτριάς μου, Αν. Καθηγήτριας του Πανεπιστημίου Πειραιώς., Κα Μαρία Χαλκίδη που μετά από ανταλλαγή πολλών μηνυμάτων και βιντεοκλήσεων μου παρείχε απαραίτητη υποστήριξη και καθοδήγηση και στο θέμα και πάνω στον τρόπο και χρόνο ολοκλήρωσης.

Έπειτα θα ήθελα να ευχαριστήσω τους φίλους μου εντός και εκτός πανεπιστημίου που έδειξαν υπομονή στην πιεστική αυτή διάρκεια της εκπόνησης της εργασίας που με στήριξαν και βοήθησαν με τον δικό τους τρόπο καθώς και την διάρκεια όλης της σχολής κατανόησαν και κατάλαβαν τις υποχρεώσεις μου και τον χρόνο που χρειάστηκε να αφιερώσω σε αυτή .

Τέλος θα ήθελα να ευχαριστήσω την οικογένεια μου που όλα αυτά τα χρόνια με στηρίζει και μου παρέχει ότι χρειάζεται και ελπίζω να μπορώ να τους το επιστρέψω.

Πίνακας περιεχομένων

ΠΡΟΛΟΓΟΣ	8
1. ΕΙΣΑΓΩΓΗ	9
Αντικείμενο της εργασίας	9
Δομή εργασίας	10
2. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ	11
Θεωρία ανάλυσης δεδομένων	11
Η python στην ανάλυση δεδομένων	11
Προεπεξεργασία και καθαρισμός των δεδομένων	12
Οπτικοποίηση των δεδομένων	14
Συσταδοποίηση	17
Προβλεπτική ανάλυση	17
Polynomial regression	18
Support vector machine	19
Holt's models	19
AR,MA,ARIMA,SARIMA	21
Facebook's prophet	23
Θεωρία αξιολόγησης αποτελεσμάτων	24
Pickle , html , flask	25
3. ΔΕΔΟΜΕΝΑ- BIG DATA	27
Δεδομένα εργασίας	27
Επεξεργασία των δεδομένων	28
Μορφοποίηση των δεδομένων	29
4. ΑΝΑΛΥΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ	30
Γραφήματα	30
Συμπεράσματα γραφημάτων	33
Αποκωδικοποίηση των δεδομένων	34
Clustering	36
Συμπεράσματα ανάλυσης	37
5. ΠΡΟΒΛΕΨΗ ΜΕΛΛΟΝΤΙΚΩΝ ΚΡΟΥΣΜΑΤΩΝ	39
Polynomial regression	39
Support vector machine	41
Holt's Models	42
AR,MA,ARIMA,SARIMA	44
Prophet model	47
Αξιολόγηση-συμπεράσματα	50
Δοκιμή σε μια μόνο χώρα	52
6. Εφαρμογή	53
Αποθήκευση των μοντέλων	53
HTML templates	54
Python flask	55
7. ΣΥΜΠΕΡΑΣΜΑΤΑ-ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΤΑΣΕΙΣ	57
Συμπεράσματα	57

Μελλοντικές προτάσεις	59
8. ΕΙΚΟΝΕΣ, ΠΙΝΑΚΕΣ, ΣΧΗΜΑΤΑ, ΟΡΟΛΟΓΙΕΣ, ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ.....	61
Παράθεση εικόνων	61
Παράθεση Πινάκων.....	61
Παράθεση Σχημάτων.....	61
Πίνακας Ορολογίας-Συντομογραφιών.....	62
ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ	64

ΠΡΟΛΟΓΟΣ

Ο τομέας της επεξεργασίας των δεδομένων , του data science και του big data είναι ίσως ο πιο ανεβασμένος τομέας της πληροφορικής τα τελευταία χρόνια λόγω της πληροφορίας που περιέχουν τα δεδομένα και ένας τομέας που με ενδιαφέρει άμεσα, παλιά οι άνθρωποι απλώς τα αποθήκευαν και τα ανέλυαν ώστε να συζητήσουν τρόπους να τα βελτιώσουν και να πάρουν αποφάσεις, τώρα με έννοιες όπως το machine learning και το AI και την συνεχόμενη εξέλιξη που ακόμα υπάρχει βρίσκονται τρόποι άμεσης αξιοποίησης των δεδομένων με χρήση μαθηματικών εννοιών , στατιστικής και φυσικών επιστημών, δηλαδή ακριβής αξιοποίηση των δεδομένων χρησιμοποιώντας μαθηματικά ακριβές μοντέλα και υπολογισμούς με αριθμούς και αποδείξεις. Ένα τεράστιο <<όπλο>> στα χέρια κάθε επιχείρησης και οργανισμού. Επίσης με την δημιουργία μιας εφαρμογής μπορεί οποιοσδήποτε χωρίς ειδικές γνώσεις υπολογιστών απλώς εισάγοντας δεδομένων και επιλέγοντας κουτάκια να πάρει ένα σύνολο προβλεπτικών αποτελεσμάτων κάτι που κάνει ακόμα πιο δυνατό το <<όπλο>> των δεδομένων .

Η πτυχιακή αυτή εργασία πραγματοποιήθηκε στο τμήμα ψηφιακών συστημάτων του πανεπιστημίου Πειραιώς στην κατεύθυνση συστήματα λογισμικού και δεδομένων και συγκεκριμένα στην διαχείρισης δεδομένων, μια κατεύθυνση βασισμένη στον διάσημο τα τελευταία χρόνια τομέα των δεδομένων ,μεγάλων δεδομένων, μηχανικής μάθησης και διαχείρισης τους καθώς και βάσεων δεδομένων κ.α. Βάση των εξελίξεων και του σημαντικού θέματος του covid στην εποχή μας και την ξαφνική ανάγκη για μοντελοποίηση ιατρικών δεδομένων καθώς και χρήση των δεδομένων αυτών επιλέχτηκε ένα θέμα που βασίζεται σε δεδομένα κορονοϊού από την ευρωπαϊκή ένωση , την επεξεργασία τους, τον σωστό καθαρισμό τους την ανάλυση των δεδομένων αυτών καθώς και την εξαγωγή πληροφορίας ώστε να σχηματιστούν και δοκιμαστούν προβλεπτικά μοντέλα για τα μελλοντικά κρούσματα και τους μελλοντικούς θανάτους.

Ο σκοπός της εργασίας αυτής και το αποτέλεσμα που θέλει να δείχτει είναι ότι με την σωστή προσέγγιση και τη σωστή επεξεργασία δεδομένων μπορούμε έως και σε ένα τεράστιο παγκόσμιο θέμα με τεράστιο αριθμό δεδομένων να βγάλουμε συμπεράσματα , προβλέψεις και με μια εύκολη για χρήστη εφαρμογή να το κάνουμε άνετα προσπελάσιμο από οποιοδήποτε χωρίς γνώσεις πληροφορικής.

1. ΕΙΣΑΓΩΓΗ

Αντικείμενο της εργασίας

εκπόνηση μιας εύχρηστης εφαρμογής. Το αντικείμενο της εργασίας αφορά την συλλογή δεδομένων από το site της ευρωπαϊκής ένωσης σχετικά με τον covid(κορονοϊός) με την χρήση python μέσα στο περιβάλλον του ισότοπου kaggle τον καθαρισμό τους (κενές τιμές κ.α) την επεξεργασία τους ώστε να τα δείξουμε καλύτερα (ομαδοποίηση ανά χώρα, κρούσματα και θάνατοι συνολικά και καθημερινά) , εύρεση κάποιας σχέσης μεταξύ των χωρών μέσα από διαγράμματα που δείχνουν την εξέλιξη κατά τον χρόνο καθώς και clustering βάση του mortality $((\text{deaths/cases}) * 100)$ έπειτα φέρνοντας τα κρούσματα στην κατάλληλη θέση κτίζονται τα προβλεπτικά μοντέλα με train και test :

1) polynomial regression

2) svm

3) holt's linear model

4) holt's winter model

5) AR model

6) MA model

7) ARIMA model

8) SARIMA model

9) prophet model

Και γίνεται έλεγχος βάση του mse τους

Τέλος με χρήση και αποθήκευση των μοντέλων sarima και prophet χρήση pickle , και html φτιάχνεται μια flask app που μπορεί ένας χρήστης να βάλει δεδομένα και να του βγάλει τα μελλοντικά κρούσματα.

Σκοπός της εργασίας είναι να αποδείξουμε ότι με τα σημερινά μέσα που έχουμε μπορούμε με απλό κώδικα και ακολουθώντας βασικά μοντέλα να προβλέψουμε αποτελεσματικά πάνω σε μια βάση δεδομένων ένα τόσο περίπλοκο θέμα όπως ο covid να δούμε πως συμπεριφέρεται κάθε μοντέλο ξεχωριστά , τρόπους βελτίωσης , να αναλύσουμε τα δεδομένα μας καθώς και να χρησιμοποιήσουμε αυτά τα αποτελέσματα για την

Δομή εργασίας

Το πρώτο κεφάλαιο είναι εισαγωγικό και αναφέρει περιληπτικά τι θα δούμε στην συνέχεια, Το δεύτερο κεφάλαιο είναι ο πυρήνας της εργασίας μας , το μεγαλύτερο κεφάλαιο και αφορά όλο το θεωρητικό υπόβαθρο που θα χρειαστούμε σε όλα τα υπόλοιπα κεφάλαια , το τρίτο κεφάλαιο αφορά τα δεδομένα μας και τα χαρακτηριστικά τους , το 4^ο και το 5^ο είναι η διαδικασία της ανάλυσης των δεδομένων μας καθώς και έπειτα η προβλεπτική ανάλυση με λεπτομέρειες και αναφορά επιμέρους συμπερασμάτων , το 6^ο κεφάλαιο αναφέρει αναλυτικά την διαδικασία εκπόνησης της εφαρμογής. Το 7^ο είναι τα συμπεράσματα που καταλήξαμε καθώς και μελλοντικές προτάσεις για τους επόμενους ερευνητές , το 8ο κεφάλαιο είναι ο κατάλογος πινάκων , σχημάτων , εικόνων, ορολογιών , συντομογραφιών και τέλος αναγράφονται οι βιβλιογραφικές αναφορές.

2. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

Σε αυτό το κεφάλαιο θα αναπτυχθούν όλο το θεωρητικό υπόβαθρο των μεθόδων που θα χρησιμοποιήσουμε για ανάλυση δεδομένων καθώς και τον σχηματισμό των προβλεπτικών μοντέλων και της εφαρμογής.

Θεωρία ανάλυσης δεδομένων

Για να αρχίσουμε να κατανοούμε το τι βλέπουμε και τα δεδομένα που έχουμε δεν αρκεί η εμπειρία του προγραμματιστή και μια ιδέα , απαιτείται να απαντήσουμε σε ερωτήσεις όπως το τι δεδομένα έχουμε, πόσα , τι μας δείχνουν , τι πληροφορία περιέχουν και αυτό απαιτείται να γίνει σαφές σε οποιοδήποτε χρήστη, έτσι με τη χρήση παρουσίασης και γραφημάτων που περιέχει η ρύθμιση και της ποικιλίας που περιέχει μπορούμε να δούμε και να καταλάβουμε καλύτερα το τι βλέπουμε και να βγάλουμε συμπεράσματα που θα χρησιμοποιηθούν από επιχειρήσεις και φορείς για την ύπαρξη σημαντικών αποφάσεων. Τα δεδομένα αυξάνονται πολύ πιο γρήγορα στον σημερινό κόσμο. Δεν είναι έτσι εύκολη η χειροκίνητη επεξεργασία των δεδομένων. Η Ανάλυση δεδομένων και Τα προγράμματα οπτικοποίησης επιτρέπουν να φτάσουμε σε ακόμα πιο βαθιά κατανόηση.

Οι αναλύσεις δεδομένων επιτρέπουν στις επιχειρήσεις να κατανοήσουν τη δική τους αποτελεσματικότητα και απόδοση, και τελικά βοηθά την επιχείρηση λήψη πιο τεκμηριωμένων αποφάσεων. Για παράδειγμα, σε ένα ηλεκτρονικό εμπόριο η εταιρεία μπορεί να ενδιαφέρεται να αναλύσει τα χαρακτηριστικά των πελατών προκειμένου να προβάλλονται στοχευμένες διαφημίσεις για τη βελτίωση των πωλήσεων. Η ανάλυση μπορεί να εφαρμοστεί σχεδόν σε οποιαδήποτε πτυχή μιας επιχείρησης εάν κατανοεί κανείς τα διαθέσιμα εργαλεία για την επεξεργασία πληροφοριών.

Η python στην ανάλυση δεδομένων

Φτιαγμένη 1989 μια γλώσσα υψηλού επιπέδου από τις πρώτες και πιο παλιές εδραιωμένες γλώσσες προγραμματισμού του κόσμου , παρά της ηλικίας της και του πόσο καιρό πριν τις σημερινές τάσεις είναι φτιαγμένη, η ευκολία χρήσης της , η απλότητα της η αμεσότητα τις και οι συνεχόμενα νέες βιβλιοθήκες της την κάνουν στους κυρίαρχες γλώσσες για τους περισσότερους data scientists και data analysts.

Η γλώσσα προγραμματισμού Python, με την Αγγλικές εντολές και εύκολη στην παρακολούθηση σύνταξη, προσφέρει ένα εκπληκτικά ισχυρή (και δωρεάν!) εναλλακτική λύση ανοιχτού κώδικα παραδοσιακές τεχνικές και εφαρμογές.

Παρά της <<ηλικίας>> της από το 2010 και μετά οι μεγαλύτεροι φορείς όπως intel, IBM, NASA, Pixar, Netflix, Facebook, JP Morgan Chase, Spotify, χρησιμοποιούν python για διαδικασίες που αφορούν τα δεδομένα και την επεξεργασία τους λόγω της απλότητας της και της συνεχόμενης βελτίωσης βάση των τελευταίων τάσεων και αναγκών στην βιομηχανία της πληροφορικής

και του προγραμματισμού. Επίσης μπορεί να διαχειριστεί σχεδόν κάθε διαδικασία και ανάγκη , και δεν είναι μια γλώσσα με έναν συγκεκριμένο σκοπό, όπως θα δούμε πχ μπορεί το ίδιο εύκολα με την ανάλυση και επεξεργασία δεδομένων να διαχειριστεί έναν εντελώς διαφορετικό τομέα όπως η δημιουργία εφαρμογής με χρήση html και δημιουργία μιας flask app , η python είναι πιθανότερα η πιο γενική και ικανή γλώσσα στο να διαχειριστεί οποιοδήποτε πρόβλημα .

Έχοντας βιβλιοθήκες όπως η **numpy** για την δημιουργία και επεξεργασία πινάκων και την δυνατότητα πολλών αριθμητικών πράξεων χωρίς μεγάλη πίεση στην μνήμη ανεξαρτήτως μεγέθους (μιας που μιλάμε και για big data βασικό στοιχείο) και σε γρήγορο χρόνο (κάτι που μας ενδιαφέρει σίγουρα). Έπειτα έχουμε την **pandas** την βασική βιβλιοθήκη επεξεργασίας δεδομένων της python , πάνω κάτω είναι ο βασικός παράγοντας σχεδόν όλων των σημαντικών διαδικασιών και των πιο απαιτητικών στο θέμα των δεδομένων , όπως το πέρασμα δεδομένων ανεξαρτήτως τύπου στην python (πχ excel) , επεξεργασία και ευελιξία των δεδομένων (η pandas είναι χτισμένη πάνω στην numpy που αναφέρθηκε πριν και μπορεί να χρησιμοποιήσει όλες τις λειτουργίες της) καθώς και χάρη στην pandas μπορούν να γίνουν στατιστικά μοντέλα πρόβλεψης πάνω στα δεδομένα μας καθώς χωρίς αυτή δεν θα μπορούσαμε να έχουμε τα κατάλληλα δεδομένα σχεδόν ποτέ. Η **scikit learn** επιτρέπει βασικές διαδικασίες της ανάλυσης δεδομένων όπως συσταδοποίηση , ομαδοποίηση και άλλους βασικούς αλγόριθμους. Έπειτα για το βασικό στοιχείο της παρουσίασης των δεδομένων υπάρχει η βιβλιοθήκη της **matplotlib** που επιτρέπει την οπτικοποίηση των δεδομένων μας με ποικίλους τρόπους βάση των αναγκών μας.

Προεπεξεργασία και καθαρισμός των δεδομένων

Μιλήσαμε ήδη για επεξεργασία δεδομένων και διάφορες έννοιες της ανάλυσης δεδομένων όπως συσταδοποίηση . Πάμε να εξηγήσουμε μερικές βασικές έννοιες.

Βάση Jason W. Osborne ² και Μαρία Χαλκίδη – Μ.Βαζιργίνης ¹³, για αρχή έχουμε την προ επεξεργασία δεδομένων , Τα δεδομένα που θα έχουμε ανεξάρτητος περιεχομένου , αφού ελέγξουμε την αξιοπιστία τους και αναλύσουμε την ποιότητα τους θα έχει αρχίσει η διαδικασία της ανάλυσης τους. Το πιθανότερο ειδικά στην επεξεργασία των big data θα είναι ότι τα δεδομένα θα παρουσιάζουν διάφορα προβλήματα.

Απώλεια δεδομένων είναι το συχνότερο θέμα σε μεγάλα δεδομένα και συνήθως ή θα υπάρχουν η εντελώς κενές τιμές ή τιμές με την ύπαρξη ενός χαρακτήρα που δηλώνει απώλεια. Υπάρχουν 3 Βασικές προσεγγίσεις των αναλυτών :

- Διαγραφή της σειράς: Σε μεγάλη ποικιλία δεδομένων που η μια γραμμή δεν έχει μεγάλη πληροφοριακή σημασία απλώς διαγράφεται η γραμμή , αλλά δεν συνιστάται σε περίπτωση πολλών γραμμών με απώλειες .

- Μέση τιμή: Αντικατάσταση της τιμής που λείπει με την μέση τιμή της στήλης ή κάποιων στηλών πχ επόμενης και προηγούμενης ώστε να μπει μια λογική ισορροπημένη τιμή που δεν θα επηρεάσει τα δεδομένα μας.
- Αντικατάσταση με την πραγματική τιμή ή μια σταθερή τιμή : Το πιο δύσκολο και απίθανο σενάριο να πηγαίναμε να βρούμε την πραγματική τιμή , η να αντικαταστήσουμε με μια προκαθορισμένη λογική τιμή.

Μέσω python όλα αυτά γίνονται με χρήση της pandas με διάφορες διαδικασίες που επιτρέπουν όλα όσα προαναφέρθηκαν όπως το `isnull()` που μας δείχνει αν υπάρχουν κενές τιμές και ποιες και διαδικασίες για την επίτευξη κάθε από τους προαναφερόμενους τρόπους.

Επόμενο μεγάλο θέμα που απασχολεί είναι τα λεγόμενα δεδομένα με **θόρυβο**. Εκτός της απώλειας τι θα συμβεί αν κάποιος πήγε να γράψει 100 και έγραψε 1? Ή αν κάποιος χαρακτήρας παραμορφώθηκε? Θα είχε πολύ κακό αποτέλεσμα στα δεδομένα μας καθώς και σε ότι έρχεται μετά. Το να βρεθούν τα δεδομένα με θόρυβο 1-1 ειδικά σε big data είναι πρακτικά.

αδύνατον , οπότε χρησιμοποιούνται γενικές τεχνικές σε όλα τα δεδομένα που έχουν ως αποτέλεσμα την μείωση έστω του θορύβου , βέβαια το αρνητικό αυτών των τεχνικών είναι η αλλαγή των δεδομένων μας:

Τα διαστήματα μπορεί να είναι ίσου πλάτους ή ίσης συχνότητας.

- Αντικατάσταση μέσο όρων : διαλέγουμε ένα διάστημα / όριο πχ ανά 3 και κάνουμε μέσο όρο ώστε να ισορροπηθούν όλες οι τιμές.
- Αντικατάσταση με οριακή τιμή : Σε αυτό το διάστημα κάθε τιμή αντικαθιστάται με την μικρότερη ή την μεγαλύτερη τιμή ανάλογα που είναι πιο κοντά .
- Εντοπισμός ακραίων τιμών και διαλέγουμε τι θα τις κάνουμε.

Έπειτα υπάρχουν τεχνικές όπως η **κανονικοποίηση** που μετατρέπει τις τιμές σε ένα όριο $[-1,1]$ ή κάποιες πιο <<κατάλληλες>> τιμές , **η μείωση των διαστάσεων** όταν έχουμε πολλά περιττά πεδία και δεδομένα , διαλέγοντας τα πιο κατάλληλα ή και ακριβώς το αντίθετο , **δηλαδή να χρειαστεί να προσθέσουμε πεδία και διαστάσεις** σε περίπτωση που δεν υπάρχει κάποια κατάλληλη για τα δεδομένα μας.

Οπτικοποίηση των δεδομένων

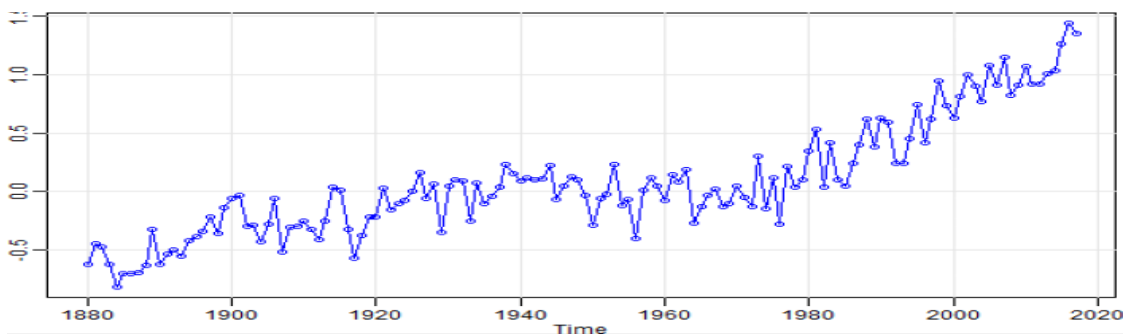
Όπως αναφέρεται από Μαρία Χαλκίδη – Μ.Βαζιργίνης ¹³(, pp 291-344,)

Η οπτικοποίηση δεδομένων είναι η αναπαράσταση δεδομένων μέσω της χρήσης κοινών γραφικών, όπως γραφήματα, γραφήματα, infographics, ακόμη και κινούμενα σχέδια. Αυτές οι οπτικές εμφανίσεις πληροφοριών επικοινωνούν πολύπλοκες σχέσεις δεδομένων και γνώσεις που βασίζονται σε δεδομένα με τρόπο που είναι εύκολο να κατανοηθεί.

Είναι βασικό μέσο των data analysts και data scientists και θα μπορούσαμε να πούμε ότι χωρίς την οπτικοποίηση θα ήταν σχεδόν αδύνατον να γίνουν αντιληπτά τα δεδομένα καθώς και τα αποτελέσματα που θα υπάρξουν.

Ας αναλύσουμε μερικούς από τους πιο βασικούς τρόπους οπτικοποίησης και της χρησιμότητας τους και περισσότερο θα επικεντρωθούμε στην οπτικοποίηση με χρήση python.

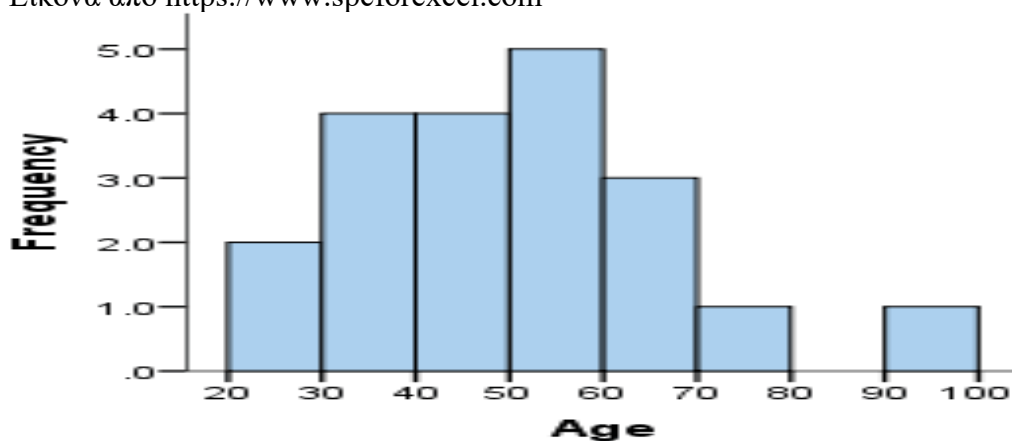
- ✓ **Line Charts:** Απλά γραφήματα δείχνοντας απλώς μια χρονοσειρά συνήθως, περιέχει ένα x και ένα y και χρησιμοποιείται για να βλέπουμε συνήθως την πορεία κάποιας τιμής στην πάροδο του χρόνου.



Σχήμα 1 Line chart *

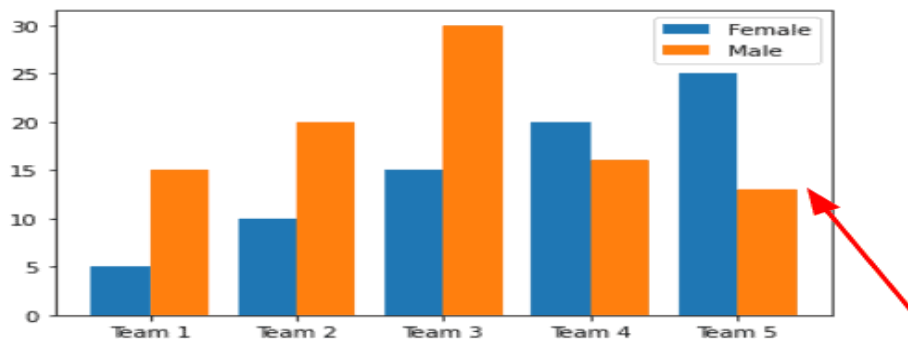
- ✓ **Ιστογράμματα(histograms):** το πιο γνωστό γράφημα δεδομένων , ορίζει ένα όριο στον y άξονα και δείχνει την τιμή και την ποσότητα στον x άξονα.

*Εικόνα από <https://www.spcforexcel.com>



Σχήμα 2 Ιστόγραμμα *

- ✓ γραφήματα μπάρας(barplot): γραφήματα με χρήση μπάρας , συνήθως χρησιμοποιούνται σε προβλήματα κατηγοριοποίησης για να δείξει τον αριθμό κάθε τύπου στο σύνολο των δεδομένων , αλλά μπορεί να χρησιμοποιηθεί εύκολα και για άλλες χρήσεις.

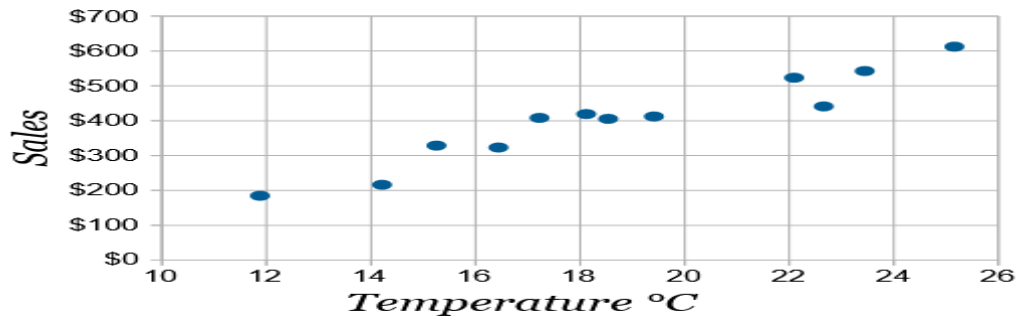


Σχήμα 3 Γράφημα Μπάρας*

- ✓ Scatter plot: όταν θέλουμε να κάνουμε plot δεδομένα χωρίς κάποιο συγκεκριμένο range σαν βάση , συνήθως αφορά πάνω από 2 μεταβλητές.

*Εικόνα από <https://statistics.laerd.com/statistical-guides/understanding-histograms.php>

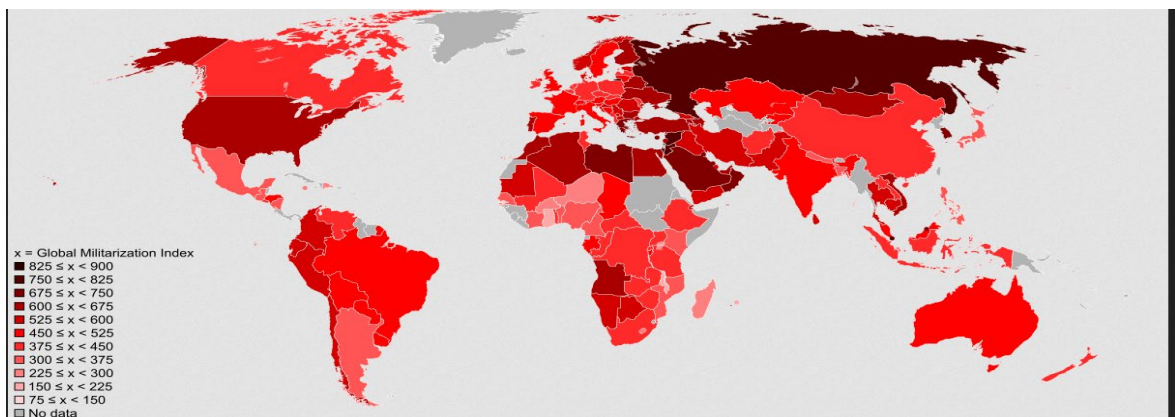
*Εικόνα από <https://pythonguides.com/matplotlib-multiple-bar-chart/>



Σχήμα 4 Scatter plot*

Αυτά ήταν κάποια βασικά plot που χρησιμοποιείται στην επιστήμη των δεδομένων, ας δούμε κάποια πιο περίπλοκα plot .

- ✓ Χάρτης χωροπλήθους (choropleth map): ένας χάρτης όπου δείχνει βάση μιας κλίμακας τα μέρη με το μεγαλύτερο πλήθος με ένα συγκεκριμένο χρώμα.



Εικόνα 1 Χάρτης χωροπλήθους*

Υπάρχουν δυνατότητες να φτιάξουμε συνεχόμενα plot και να παίξουμε μίνι βίντεο από plots όταν θέλουμε να δείξουμε την πάροδο του χρόνου παραδείγματος χάρη καθώς και μπορούμε να δείξουμε πολλά plot μαζί στην ίδια εικόνα

*Εικόνα από <https://www.mathsisfun.com/data/scatter-xy-plots.html>

*Εικόνα από https://en.wiktionary.org/wiki/choropleth_ma

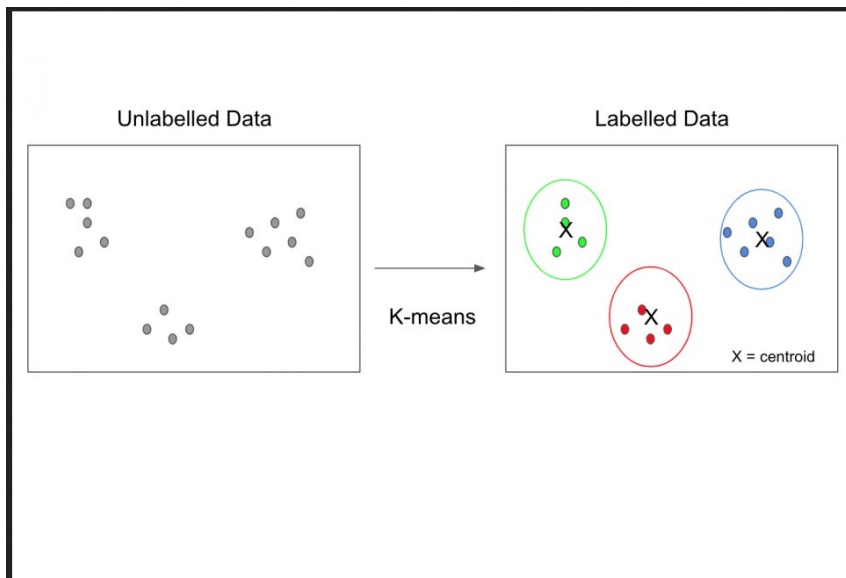
Συσταδοποίηση

Βάση της βιβλιογραφίας Mohammed j.Zaki , Wagner ¹ βλέπουμε και αναφέρεται ότι σκοπός όλων των clustering αλγορίθμων αποτελεί ο διαχωρισμός

ενός πλήθους σημείων ή αντικειμένων σε ομοειδείς ομάδες (clusters).

αυτή χρησιμοποιούνται συναρτήσεις οι οποίες υπολογίζουν την απόσταση μεταξύ των σημείων. Με βάση την απόσταση μεταξύ των σημείων δημιουργούνται ομάδες από αντικείμενα τα οποία εμφανίζουν τη μικρότερη δυνατή απόσταση. Δηλαδή 2 αντικείμενα που ανήκουν σε διαφορετικά clusters θα πρέπει να εμφανίζουν μεγαλύτερη απόσταση σε σχέση με αυτά που βρίσκονται στο ίδιο το δικό τους cluster.

Υπάρχουν πολλοί αλγόριθμοι και προσέγγισης clustering όπως ο k means που φτιάχνει τυχαία κέντρα και βάζει τα πιο κοντινά δεδομένα σε εκείνη την ομάδα, και έπειτα επαναλαμβάνει με τον μέσο όρο ή αλγόριθμοι ιεραρχικοί που ξεκινάνε από ένα σύνολο και σιγά σιγά το κόβουν σε μικρότερες ομάδες βάση μιας τιμής ή αλγόριθμοι πυκνότητας , που βάζουν σχηματικά κοντά δεδομένα μαζί .



Εικόνα 2 Clustering*

Προβλεπτική ανάλυση

Όπως αναφέρεται και από Dipti Theng ³ Από το 1600 σχεδόν οι άνθρωποι άρχισαν να ψάχνουν πώς να προβλέψουν μαθηματικά το μέλλον, στο peak των φυσικών τεχνών και με καλύτερη αντίληψη των εννοιών της άρχισαν να προσπαθούν να βγάλουν μαθηματικά αποτελέσματα για κάθε είδος δραστηριότητα. Στην πάροδο των χρόνων φτιάχτηκαν πολλές θεωρίες πρόβλεψης μη αποδεδειγμένες και διάφορες <<

*Εικόνα από <https://datascience.eu/machine-learning/clustering-algorithms-and-their-significance-in-machine-learning/>

τεχνικές>> όπως πρόβλεψη σε παιχνίδια τζόγου , χαρτιών , πρόβλεψη καθημερινών δραστηριοτήτων κ.α

Το 1940 με την αρχή των υπολογιστών και αφού απάντησαν πρώτα με τη χρήση επεξεργαστών και υπολογιστών στο τι έγινε και στο τι γίνεται άρχισε η ερώτηση :

Τι θα γίνει? Βασισμένη σε μαθηματικά ήδη υπάρχοντα μοντέλα άρχισαν να χτίζονται προβλεπτικοί αλγόριθμοι και προβλήματα εύρεσης μονοπατιών , γραμμικών και μη προβλημάτων κ.α

Τα τελευταία 10 χρόνια με την ύπαρξη τεράστιων βάσεων δεδομένων και την ευκολία εύρεσης τους , αντιλαμβάνοντος την ύπαρξη πληροφοριών που περιέχουν αυτά τα δεδομένα η προβλεπτική ανάλυση έγινε ένας τεράστιος ξεχωριστός κλάδος που κάθε εταιρία θέλει και μπορεί να χρησιμοποιηθεί από εύρεση χρονοσειρών μέχρι προβλήματα κατηγοριοποίησης από εταιρίες και κάθε φορέα.

Ας αρχίσουμε με τη βασική λογική των αλγορίθμων πριν πάμε στους αλγόριθμους .

Όπως αναφέρεται από Ζιουάλμα Μαρία, «Ανάλυση χρονοσειρών για την πρόβλεψη επιχειρήσεων» , ¹⁰

Κάθε αλγόριθμοι έχει τις δικές του απαιτήσεις και θέλει τα δεδομένα σε συγκεκριμένο τύπο , πχ κάποιοι αλγόριθμοι χρησιμοποιούνται σε δεδομένα **χωρίς τάση ή εποχικότητα , δηλαδή τα δεδομένα μας να είναι όπως λέγεται <<στατικά(stationary)>>** , η τάση αφορά απότομες αλλαγές της μέσης τιμής βάση χρόνου , πχ τιμές της βενζίνης που αυξάνονται απότομα , και η εποχικότητα αφορά ακριβώς αυτό που λέει , τα Χριστούγεννα τα έλατα θα έχουν τις περισσότερες πωλήσεις , αυτό είναι αποτέλεσμα της εποχικότητας . Επίσης τα δεδομένα μας πρέπει να είναι << καθαρά>> χωρίς outliers(ακραίες τιμές) και μηδενικές τιμές.

Polynomial regression

Φτιαγμένη από το 1800 σχεδόν , όταν δεν υπάρχει γραμμική σχέση μεταξύ των δεδομένων , είναι ένα μοντέλο optimize που προσπαθεί να βάλει την καλύτερη μεταβλητή b_0-b_n .

Στην εξίσωση

$$y=b_0+b_1x_1+b_2x_1^2+b_3x_1^3+...b_nx_1^n$$

Η πολυωνυμική παλινδρόμηση ταιριάζει σε μια μη γραμμική σχέση μεταξύ της τιμής του x και του αντίστοιχου υπό όρους μέσου όρου του y , που συμβολίζεται με $E(y | x)$. Αν και η πολυωνυμική παλινδρόμηση ταιριάζει με ένα μη γραμμικό μοντέλο στα δεδομένα, ως στατιστική εκτίμηση, το πρόβλημα είναι γραμμικό, με την έννοια ότι η συνάρτηση παλινδρόμησης $E(y | x)$ είναι γραμμική στις άγνωστες παραμέτρους που υπολογίζονται από τα δεδομένα . Για το λόγο αυτό, η πολυωνυμική παλινδρόμηση θεωρείται μια ειδική περίπτωση πολλαπλής γραμμικής παλινδρόμησης .

Αρνητικά:

Είναι επιρρεπής στα outliers και στις απότομες αλλαγές πράγμα που την κάνει επιρρεπής στα δεδομένα μας ,Η προσαρμοσμένη καμπύλη από την πολυωνυμική παλινδρόμηση λαμβάνεται με καθολική εκπαίδευση. Δηλαδή,

χρησιμοποιούμε όλο το εύρος τιμών του προγνωστικού για να χωρέσουμε στην καμπύλη. Αυτό μπορεί να είναι προβληματικό: εάν λάβουμε νέα δείγματα από μια συγκεκριμένη υποπεριοχή του προγνωστικού, αυτό μπορεί να αλλάξει το σχήμα της καμπύλης σε άλλες υποπεριοχές! Στην ιδανική περίπτωση, θα θέλαμε η καμπύλη να προσαρμόζεται τοπικά εντός υποπεριοχών που σχετίζονται με την ανάλυση.

Θετικά:

Δουλεύει σε οποιαδήποτε δεδομένα.

Δουλεύει ανεξαρτήτως μεγέθους δεδομένων.

Support vector machine

Από τα πιο ευρέως γνωστά μοντέλα λόγω της ευελιξίας επιλογής πυρήνα, φτιαγμένος το 1963 για γραμμικά δεδομένα και με βελτίωση το 1992 για επιλογή πυρήνα και μετατροπές ώστε να μπορεί να διαχωριστεί κάθε είδος δεδομένων.

Πάλι ένας αλγόριθμος βελτιστοποίησης θεωρητικά αλγόριθμος κατηγοριοποίησης, πρακτικά αλγόριθμος λύσης οποιουδήποτε προβλήματος. Στα μη γραμμικά δεδομένα μας με χρήση ενός polynomial kernel προσπαθεί μια πολυωνυμική εξίσωση να βρει την καλύτερη λύση στη χρονοσειρά προσπαθώντας να χωρέσει όσα περισσότερα δεδομένα μπορεί σε ένα όριο e που το δίνουμε εμείς και επιστρέφει την καλύτερη δυνατή λύση με το μικρότερο error.

$$K(X_1, X_2) = (a + X_1^T X_2)^b$$

Εικόνα 3 polynomial svm*

Όπου b ο βαθμός που θα δώσουμε εμείς και a μια σταθερά.

Το κύριο χαρακτηριστικό και **πλεονέκτημα** του svm εκτός της ευελιξίας του και της εύκολης χρήσης του είναι ότι είναι πολύ εύκολο να κάνουμε fit δεδομένα ανεξάρτητος βαθμού.

Από την άλλη μπορεί συχνά να μπλέξει με overfitting και σε μεγάλα δεδομένα με πολλές μεταβλητές και σε μεγάλο βαθμό να είναι αρκετά αργός.

Holt's models

holt-Η πρόβλεψη Winters είναι ένας τρόπος για να μοντελοποιήσουμε και να προβλέψουμε τη συμπεριφορά μιας ακολουθίας τιμών με την πάροδο του χρόνου μιας χρονοσειράς. Το Holt-Winters είναι μια από τις πιο δημοφιλείς τεχνικές πρόβλεψης χρονοσειρών. Είναι παλιό δεκαετίες, αλλά εξακολουθεί να είναι πανταχού παρόν σε πολλές εφαρμογές, συμπεριλαμβανομένης της παρακολούθησης, όπου χρησιμοποιείται για σκοπούς όπως η ανίχνευση ανωμαλιών και ο σχεδιασμός χωρητικότητας

*Εικόνα από

<https://www2.cs.sfu.ca/~oschulte/teaching/726/spring11/slides/mychapter6.pdf>

Η μέθοδος Holt-Winters χρησιμοποιεί εκθετική εξομάλυνση για να κωδικοποιήσει πολλές τιμές από το παρελθόν και να τις χρησιμοποιήσει για να προβλέψει «τυπικές» τιμές για το παρόν και το μέλλον. Η εκθετική εξομάλυνση αναφέρεται στη χρήση ενός εκθετικά σταθμισμένου κινητού μέσου όρου για την «εξομάλυνση» μιας χρονοσειράς.

Το holt's linear model βασίζεται σε δεδομένα με τάση και είναι μια εξίσωση τύπου

$$(1) \text{Forecast } p_{t+h|t} = \underbrace{l_t}_{\text{level}} + \underbrace{hb_t}_{\text{trend}}$$

$$(2) l_t = \alpha y_t + (1 - \alpha)l_{t-1} \quad (\alpha = \text{smoothing level})$$

$$(3) b_t = \beta(l_t - l_{t-1}) + (1 - \beta)b_{t-1} \quad (\beta = \text{smoothing slope})$$

Εικόνα 4 Holt's linear model*

Εμείς διαλέγουμε τα α και β .

Σε αντίθεση το holt's winter model περιέχει και άλλη μια μεταβλητή ώστε να μπορεί να διαχειριστεί δεδομένα με εποχικότητα.

Δηλαδή το forecast $y(t)$ γίνεται

$$\hat{y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)}$$

$$l_t = \alpha(y_t - s_{t-m}) + (1 - \alpha)(l_{t-1} + b_{t-1})$$

$$b_t = \beta^*(l_t - l_{t-1}) + (1 - \beta^*)b_{t-1}$$

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1 - \gamma)s_{t-m},$$

Εικόνα 5 Holt's Winter model*

Αρνητικά

η μέθοδος Holt-Winter εξακολουθεί να έχει ελλείψεις. Ένας σημαντικός περιορισμός αυτού του αλγορίθμου είναι το πολλαπλασιαστικό χαρακτηριστικό της εποχικότητας. Το θέμα της πολλαπλασιαστικής εποχικότητας είναι πώς αποδίδει το μοντέλο όταν έχουμε χρονικά πλαίσια με πολύ χαμηλά ποσά. Ένα χρονικό πλαίσιο με σημείο δεδομένων 10 ή 1 μπορεί να έχει πραγματική διαφορά 9, αλλά υπάρχει σχετική διαφορά περίπου 1000%, επομένως η εποχικότητα, η οποία εκφράζεται ως σχετικός όρος θα μπορούσε να αλλάξει δραστικά και θα πρέπει να ληφθεί μέριμνα για κατά την κατασκευή του μοντέλου.

Θετικά

Το μοντέλο χρονοσειράς του holt είναι ένας πολύ ισχυρός αλγόριθμος πρόβλεψης παρόλο που είναι ένα από τα απλούστερα μοντέλα. Μπορεί να χειριστεί την εποχικότητα στο σύνολο δεδομένων υπολογίζοντας απλώς την

*Εικόνα από <https://otexts.com/fpp2/holt.html>

*Εικόνα από <https://otexts.com/fpp2/holt-winters.html>

κεντρική τιμή και στη συνέχεια προσθέτοντας ή πολλαπλασιάζοντάς την στην κλίση και την εποχικότητα. Απλώς πρέπει να φροντίσουμε να συντονιστούμε στο σωστό σύνολο παραμέτρων.

AR,MA,ARIMA,SARIMA

Δύο από τα πιο κοινά μοντέλα σε χρονοσειρές είναι τα μοντέλα Autoregressive (AR) και τα μοντέλα Moving Average (MA).

Το αυτοπαλινδρομικό μοντέλο χρησιμοποιεί παρατηρήσεις από προηγούμενα χρονικά βήματα ως είσοδο σε εξισώσεις παλινδρόμησης για να προβλέψει την τιμή στο επόμενο βήμα. Το μοντέλο AR παίρνει ένα όρισμα, το p , το οποίο καθορίζει πόσα προηγούμενα χρονικά βήματα θα εισαχθούν.

Το μοντέλο κινητού μέσου όρου είναι ένα μοντέλο χρονοσειρών που αντιπροσωπεύει πολύ βραχυπρόθεσμη αυτοσυσχέτιση. Βασικά δηλώνει ότι η επόμενη παρατήρηση είναι ο μέσος όρος κάθε προηγούμενης παρατήρησης.

Το ARIMA(autoregressive integrated moving average) model είναι ένας συνδιασμός των πάνω .

Το autoregressive αναφέρεται ως p , το moving average ως q αυτό το integrated σημαίνει ότι σε περίπτωση που υπάρχει τάση στα δεδομένα μας χρησιμοποιείται για να μειώσει αυτή την τάση Ως όρος d ,

$$\begin{aligned}\phi(B)(1-B)^d X_t &= \theta(B) Z_t, \\ \phi(B) &= 1 - \phi_1 B - \dots - \phi_p B^p \\ \theta(B) &= 1 + \theta_1 B + \dots + \theta_q B^q\end{aligned}$$

Η πρώτη εξίσωση είναι η εξίσωση της του arima ,

Για κάθε μοντέλο που έχουμε απαιτείται να βρούμε τα καλύτερα p , q , d για να λυθεί το καλύτερο δυνατότερο η εξίσωση , για αυτό χρησιμοποιείται το `auto_arima` στην `rython` , δοκιμάζει διάφορες τιμές για τα (p,q,d) και τελικά επιλέγει αυτό που έχει το καλύτερο δυνατό αποτέλεσμα με το μικρότερο `akaike information criterion` ένας εκτιμητής του σφάλματος πρόβλεψης και συνεπώς της σχετικής ποιότητας των στατιστικών μοντέλων για ένα δεδομένο σύνολο δεδομένων. Δεδομένης μιας συλλογής μοντέλων για τα δεδομένα, η AIC εκτιμά την ποιότητα κάθε μοντέλου, σε σχέση με καθένα από τα άλλα μοντέλα. Έτσι, το AIC παρέχει ένα μέσο για την επιλογή μοντέλου.

Αρνητικά

μερικές από τις παραδοσιακές τεχνικές αναγνώρισης μοντέλων είναι υποκειμενικές και οι η αξιοπιστία του επιλεγμένου μοντέλου μπορεί να εξαρτάται από την ικανότητα και την εμπειρία του προγραμματιστή

Δεν είναι ενσωματωμένο σε κανένα υποκείμενο θεωρητικό μοντέλο ή δομή σχέσης. Επομένως, η οικονομική σημασία του επιλεγμένου μοντέλου δεν είναι Σαφή. Επιπλέον, δεν είναι δυνατή η εκτέλεση προσομοιώσεων πολιτικής με την ARIMA Τα μοντέλα ARIMA είναι ουσιαστικά «κοιτάζοντας προς τα πίσω» τύπου, είναι γενικά ανεπαρκής στην πρόβλεψη των σημείων καμπής δηλαδή

των σημείων που αλλάζει η κλίση των δεδομένων μας και υπάρχει απότομη αύξηση ή μείωση.

Θετικά

Το κύριο πλεονέκτημα της πρόβλεψης ARIMA είναι ότι απαιτεί δεδομένα για τις χρονοσειρές που μας νοιάζουν μόνο . αυτό το χαρακτηριστικό είναι πλεονεκτικό εάν κάποιος προβλέπει μεγάλο αριθμό των χρονοσειρών. έτσι αποφεύγεται ένα πρόβλημα που παρουσιάζεται μερικές φορές με τα μοντέλα με πολλές μεταβλητές. Για παράδειγμα, σκεφτείτε ένα μοντέλο που περιλαμβάνει μισθούς, τιμές και χρήματα. Είναι πιθανό ότι μια σταθερή σειρά χρημάτων είναι διαθέσιμη μόνο για μικρότερο χρονικό διάστημα από τις άλλες δύο σειρές, περιορίζοντας τη χρονική περίοδο κατά την οποία μπορεί να βρίσκεται το model . με πολυμεταβλητά μοντέλα, η επικαιρότητα των δεδομένων μπορεί να είναι πρόβλημα. Αν κατασκευάζει ένα μεγάλο δομικό μοντέλο που περιέχει μεταβλητές που δημοσιεύονται μόνο με μεγάλη καθυστέρηση, όπως τα στοιχεία μισθών, τότε οι προβλέψεις που χρησιμοποιούν αυτό το μοντέλο είναι υπό όρους προβλέψεις που βασίζονται σε προβλέψεις των μη διαθέσιμων παρατηρήσεων, προσθέτοντας μια επιπλέον πηγή αβεβαιότητας για τις προβλέψεις.

Το Arima είναι πολύ ευαίσθητο σε δεδομένα σε εποχικά δεδομένα , και τα αποτελέσματα όταν υπάρχει εποχικότητα δεν είναι απλώς πιο μακριά αλλά είναι εντελώς λάθος , έτσι φτιάχτηκαν εκτός των προαναφερομένων p q d όπως αναφέρεται από A.K.Dubeya,A.Kumara, , "Study and analysis of SARIMA and LSTM in forecasting time series data" ⁴

Υπάρχουν τέσσερα εποχιακά στοιχεία που δεν αποτελούν μέρος του ARIMA και πρέπει να διαμορφωθούν. αυτά είναι:

P: Εποχιακή αυτοπαλινδρομική σειρά.

D: Σειρά εποχικής διαφοράς.

Q: Εποχιακή παραγγελία κινητού μέσου όρου.

m: Ο αριθμός των χρονικών βημάτων για μια μεμονωμένη εποχική περίοδο

SARIMA(p,d,q)(P,D,Q)m

$$\begin{array}{ccccccc} (1 - \phi_1 B) & (1 - \Phi_1 B^4) & (1 - B) & (1 - B^4) y_t = & (1 + \theta_1 B) & (1 + \Theta_1 B^4) e_t. \\ \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ \left(\begin{array}{c} \text{Non-seasonal} \\ \text{AR}(1) \end{array} \right) & \left(\begin{array}{c} \text{Non-seasonal} \\ \text{difference} \end{array} \right) & & & \left(\begin{array}{c} \text{Non-seasonal} \\ \text{MA}(1) \end{array} \right) & & \\ & \left(\begin{array}{c} \text{Seasonal} \\ \text{AR}(1) \end{array} \right) & \left(\begin{array}{c} \text{Seasonal} \\ \text{difference} \end{array} \right) & & & \left(\begin{array}{c} \text{Seasonal} \\ \text{MA}(1) \end{array} \right) \end{array}$$

Εικόνα 6 AR,MA,ARIMA,SARIMA*

*Εικόνα από <https://stats.stackexchange.com/questions/129901/sarima-model-equation>

Και στις 2 περιπτώσεις μπορούμε να υπολογίσουμε μόνοι μας τις μεταβλητές , αλλά πάλι με το auto_sarima βρίσκεται αυτόματα ο καλύτερος συνδυασμός καθώς η εναλλακτική θα έπαιρνε πολύ χρόνο.

Facebook's prophet

Βλέπουμε με οδηγό το Prophet: forecasting at scale By: Sean J. Taylor, Ben Letham⁵. ότι το prophet είναι ένα εργαλείο πρόβλεψης διαθέσιμο σε Python και R. Η πρόβλεψη είναι μια εργασία επιστήμης δεδομένων που είναι κεντρική σε πολλές δραστηριότητες μέσα σε έναν οργανισμό. Για παράδειγμα, μεγάλοι οργανισμοί όπως το Facebook πρέπει να συμμετάσχουν σε σχεδιασμό χωρητικότητας για την αποτελεσματική κατανομή των σπάνιων πόρων και τον καθορισμό στόχων προκειμένου να μετρηθεί η απόδοση σε σχέση με μια βασική γραμμή. Η παραγωγή προβλέψεων υψηλής ποιότητας δεν είναι εύκολο πρόβλημα ούτε για τις μηχανές ούτε για τους περισσότερους αναλυτές. Παρατηρήσαμε δύο βασικά θέματα στην πρακτική δημιουργίας ποικίλων επιχειρηματικών προβλέψεων:

Οι εντελώς αυτόματες τεχνικές πρόβλεψης μπορεί να είναι εύθραυστες και συχνά είναι πολύ άκαμπτες για να ενσωματώσουν χρήσιμες υποθέσεις ή ευρετικές μεθόδους.

Οι αναλυτές που μπορούν να παράγουν προβλέψεις υψηλής ποιότητας είναι αρκετά σπάνιοι επειδή η πρόβλεψη είναι μια εξειδικευμένη δεξιότητα επιστήμης δεδομένων που απαιτεί σημαντική εμπειρία.

Δεν μπορούν να λυθούν όλα τα προβλήματα πρόβλεψης με την ίδια διαδικασία. Το Prophet έχει βελτιστοποιηθεί για τις εργασίες πρόβλεψης επιχειρήσεων που έχουμε συναντήσει στο Facebook, οι οποίες συνήθως έχουν οποιοδήποτε από τα ακόλουθα χαρακτηριστικά:

- ωριαίες, ημερήσιες ή εβδομαδιαίες παρατηρήσεις με τουλάχιστον μερικούς μήνες (κατά προτίμηση ένα χρόνο) ιστορικό
- ισχυρές πολλαπλές εποχικότητες «ανθρώπινης κλίμακας»: ημέρα της εβδομάδας και ώρα του χρόνου
- σημαντικές διακοπές που συμβαίνουν σε ακανόνιστα διαστήματα που είναι γνωστά εκ των προτέρων (π.χ. το Super Bowl)
- λογικό αριθμό παρατηρήσεων που λείπουν ή μεγάλες ακραίες τιμές
- αλλαγές ιστορικών τάσεων, για παράδειγμα λόγω λανσαρίσματος προϊόντων ή αλλαγών καταγραφής
- τάσεις που είναι μη γραμμικές καμπύλες ανάπτυξης, όπου μια τάση αγγίζει ένα φυσικό όριο ή κορεστεί

Οι προεπιλεγμένες ρυθμίσεις του <<Προφήτη>> για την παραγωγή προβλέψεων που είναι συχνά ακριβείς όπως αυτές που παράγονται από ειδικευμένους data scientists, με πολύ λιγότερη προσπάθεια.

$$y(t) = g(t) + s(t) + h(t) + e(t)$$

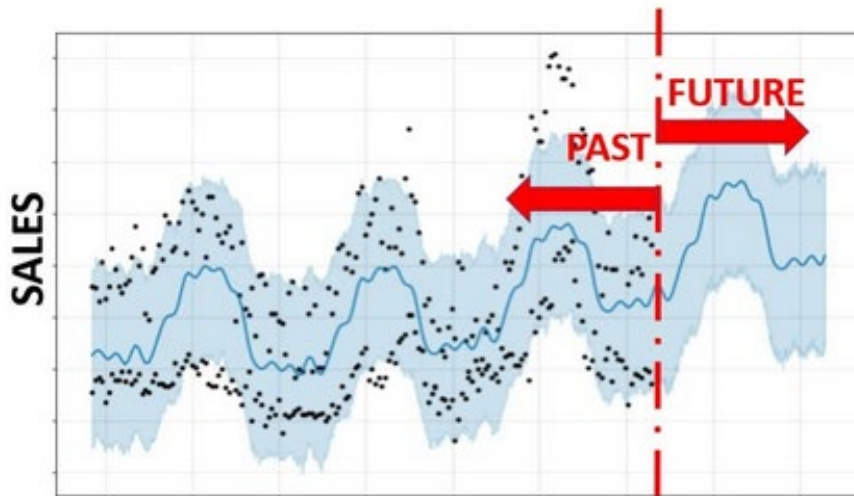
όπου $g(t)$ το trend $s(t)$ το seasonality $h(t)$ οι επιδράσεις των γιορτών, και $e(t)$ οποιαδήποτε αλλαγή δεν καλύπτεται από τα άλλα πχ θόρυβος, outliers κ.α

Θετικά

- Υπάρχουν ήδη πολλοί πόροι για την παλινδρόμηση
- Καταγράφει τη συνολική τάση
- Μπορεί να υποστηρίξει πολλαπλές διαστάσεις εισόδου

Αρνητικά

- Η εμπιστοσύνη του μοντέλου υποβαθμίζεται με την πάροδο του χρόνου
- Δεν καταγράφει τη μεταβλητότητα εντός της περιόδου



Εικόνα 7 prophet model*

οι τελείες είναι τα αληθινά δεδομένα , η γραμμή η πρόβλεψη και το μπλε περίγραμμα ένα διάστημα εμπιστοσύνης που μπορεί να κυμαίνονται τα δεδομένα , οτιδήποτε εκτός θεωρείται outlier .

Θεωρία αξιολόγησης αποτελεσμάτων

Η αξιολόγηση ενός μοντέλου είναι το βασικό μέρος της δημιουργίας ενός επιτυχημένου μοντέλου Μηχανικής Μάθησης. Ο σκοπός της αξιολόγησης ενός μοντέλου είναι να συγκριθούν οι προβλέψεις του εκπαιδευμένου μοντέλου με τα πραγματικά δεδομένα. Μας βοηθά να συνειδητοποιήσουμε την απόδοση του μοντέλου μας και σας διευκολύνει να παρουσιάσουμε το μοντέλο σας στο κοινό. Χρησιμοποιούνται διαφορετικές μετρήσεις σφάλματος για διαφορετικά είδη μοντέλων μηχανικής εκμάθησης. Υπάρχουν τρεις μετρήσεις αξιολόγησης που χρησιμοποιούνται συχνά για την αξιολόγηση της απόδοσης ενός μοντέλου παλινδρόμησης. Αυτοί είναι:

- Μέσο τετράγωνο σφάλμα (MSE)
- Σφάλμα ρίζας μέσου τετραγώνου (RMSE)
- Μέσο απόλυτο σφάλμα (MAE)

*Εικόνα από <https://www.youtube.com/channel/UC76VWNgXnU6z0RSPGwSkNIg>

Το μέσο τετράγωνο σφάλμα (MSE) ορίζεται ως ο μέσος όρος ή ο μέσος όρος του τετραγώνου της διαφοράς μεταξύ πραγματικών και εκτιμώμενων τιμών. Αυτό σημαίνει ότι το MSE υπολογίζεται με το τετράγωνο της διαφοράς μεταξύ των προβλεπόμενων και των πραγματικών μεταβλητών στόχου, διαιρούμενο με τον αριθμό των σημείων δεδομένων. Είναι πάντα μη αρνητικές τιμές και κοντά στο μηδέν είναι καλύτερες.

Το Root Mean Square Error (RMSE) χρησιμοποιείται επίσης ως μέτρο για την αξιολόγηση του μοντέλου. Είναι η τετραγωνική ρίζα του μέσου τετραγώνου σφάλματος (MSE). Αυτό είναι το ίδιο με το μέσο τετράγωνο σφάλμα (MSE), αλλά η ρίζα της τιμής λαμβάνεται υπόψη κατά τον προσδιορισμό της ακρίβειας του μοντέλου.

Το μέσο απόλυτο σφάλμα (MAE) είναι το άθροισμα της απόλυτης διαφοράς μεταξύ πραγματικών και προβλεπόμενων τιμών. Απόλυτη διαφορά σημαίνει ότι αν το αποτέλεσμα έχει αρνητικό πρόσημο, αγνοείται.

Τα αποτελέσματα των τριών μετρήσεων αξιολόγησης (MSE, RMSE και MAE) είναι τα ίδια και στις δύο μεθόδους. Μπορούμε να χρησιμοποιήσουμε οποιαδήποτε μέθοδο (χειροκίνητη ή sklearn) ανάλογα με την άνεσή μας στην Ανάλυση παλινδρόμησης.

Pickle , html , flask

Τώρα πια θέλουν οι αναλυτές δεδομένων , και οι data scientists να μπορούν να κάνουν τα αποτελέσματα τους χρήσιμα , καθώς και εύκολα προσβάσιμα από οποιονδήποτε είτε έχει γνώσεις του συγκεκριμένου αντικειμένου είτε όχι , παραδείγματος χάριν φτιάξαμε ένα μοντέλο το οποίο δουλεύει κάπως αποτελεσματικά , και θέλουμε τώρα να το χρησιμοποιήσει κάποιος τρίτος , αυτό ο μόνος αποδοτικός τρόπος να γίνει είναι μέσω ενός interface , μιας εφαρμογής που θα μπορεί εύκολα κάποιος χρήστης να χρησιμοποιήσει.'

Αναλύοντας τα , Flask Web Development: Developing Web Applications with Python 1st Edition, 2014 ⁸ και Ravindra Sharma, "model deployment using flask" ⁹ έχουμε τις συγκεκριμένες πληροφορίες.

pickle

Το Pickle στην Python χρησιμοποιείται κυρίως για τη σειριοποίηση και την αποσειριοποίηση μιας δομής αντικειμένου Python. Με άλλα λόγια, είναι η διαδικασία μετατροπής ενός αντικειμένου Python σε ροή byte για την αποθήκευση του σε ένα αρχείο/βάση δεδομένων, τη διατήρηση της κατάστασης του προγράμματος σε συνεδρίες ή τη μεταφορά δεδομένων μέσω του δικτύου. Η ροή byte pickled μπορεί να χρησιμοποιηθεί για την εκ νέου

δημιουργία της αρχικής ιεραρχίας αντικειμένων, αφαιρώντας τη ροή. Όλη αυτή η διαδικασία είναι παρόμοια με τη σειριοποίηση αντικειμένων σε Java ή .Net. Με αυτή τη διαδικασία μπορούμε να αποθηκεύσουμε τα μοντέλα που φτιάχνουμε στα προηγούμενα δεδομένα και να τα χρησιμοποιήσουμε όπου θέλουμε με μια διαφορετική βάση δεδομένων .

Html

Η HTML είναι ένα τυποποιημένο σύστημα για την προσθήκη ετικετών σε αρχεία κειμένου που δημιουργεί τη δομή σχεδόν για κάθε σελίδα που βρίσκουμε και χρησιμοποιούμε στον Ιστό. Είναι η HTML που προσθέτει αλλαγές σελίδας, παραγράφους, έντονους χαρακτήρες, πλάγιους χαρακτήρες και πολλά άλλα. Η HTML λειτουργεί για τη δημιουργία αυτής της δομής χρησιμοποιώντας ετικέτες που λένε στα προγράμματα περιήγησης τι να κάνουν με το κείμενο.

Για παράδειγμα, για να κάνουμε μια λέξη να φαίνεται έντονη, βάζουμε τη λέξη ανάμεσα στις παρακάτω ετικέτες `έντονη`. Η πρώτη ετικέτα υποδηλώνει την αρχή της λέξης(ών) που θέλουμε να γράψουμε έντονη γραφή και η ετικέτα κλεισίματος(/) δείχνει πού θέλουμε να σταματήσει η έντονη γραφή. Είναι η βάση για σχεδόν κάθε σελίδα στον Ιστό.

Με την χρήση html μπορούμε να κάνουμε πολύπλοκες επεξεργασίες και να φτιάξουμε π.χ μια σελίδα που δέχεται δεδομένα και εμφανίζει ένα αποτέλεσμα βάση της επιλογής του χρήστη , πάνω κάτω η html είναι ο βασικός παράγοντας οποιασδήποτε ιστοσελίδας.

Flask app

Το Flask είναι ένα πλαίσιο web, είναι μια λειτουργική μονάδα Python που επιτρέπει να αναπτύσσουμε εύκολα εφαρμογές Ιστού. Έχει έναν μικρό και εύκολο στην επέκταση πυρήνα: είναι ένα μικροπλαίσιο που δεν περιλαμβάνει ORM (Object Relational Manager) ή τέτοια χαρακτηριστικά.

Έχει πολλά ωραία χαρακτηριστικά, όπως δρομολόγηση url, μηχανή προτύπων. Και κάνει εύκολη τη διαδικασία χρήσης κάποιας θύρας στον υπολογιστή , την επικοινωνία server – client με εύκολα επεξεργάσιμα responses .

Python flask app using html and pickle

Αν βάλουμε όσα προαναφέρθηκαν μαζί μπορούμε να φτιάξουμε ένα πολύπλοκο μοντέλο με κώδικα html και python όπου μέσω του flask θα ανοίγουμε έναν server με χαρακτηριστικά που παίρνουμε από html και τα αποθηκευμένα με pickle μοντέλα και να διαχειριστούμε κάθε response , δηλαδή κάθε ενέργεια χρήστη κατάλληλα , όλα αυτά μαζί φτιάχνουν έναν σύνθετο πολύπλοκο οργανισμό που αποτελείται πρώτα από έναν κώδικα που το αποτέλεσμα του αποθηκεύεται από την pickle έπειτα από μια εφαρμογή που αρχίζει έναν server μέσω flask , html και χρήση του pickle και αντιδράει ανάλογα.

3. ΔΕΔΟΜΕΝΑ- BIG DATA

Όπως αναφέρθηκε στο κεφαλαίο 2 στην προεπεξεργασία μας καθώς και από τους Pang-Ning tan , Michael Steinbach, Anuj Karpatne, Vipin Kumar¹² μπορούμε να αρχίσουμε την υλοποίηση της εργασίας.

Δεδομένα εργασίας

Στην εργασία επιλέχθηκαν δεδομένα από την επίσημη σελίδα της ευρωπαϊκής ένωσης European Centre for Disease Prevention and Control . Τα δεδομένα την στιγμή που τα επέλεξα εγώ περιέχουν δεδομένα από 30 χώρες από Ιανουάριου του 2020 έως τον Μάρτιο του 2022 , τα δεδομένα περιείχαν κάποιες μηδενικές τιμές και κάποια outliers που διορθώθηκαν κατά την διάρκεια της εργασίας .

Τα δεδομένα αποτελούνται από την εξής μορφολογία:

- dateRep: ημερομηνία
- day
- month
- year
- cases
- deaths
- countriesAndTerritories: η χώρα
- geold: κωδικός χώρας
- countryTerritoryCode: κωδικός χώρας
- popdata2020: σύνολο κρουσμάτων
- continentExpr: ήπειρος δηλαδή ευρώπη

Πίνακας 1 Δεδομένα

	dateRep	day	month	year	cases	deaths	countriesAndTerritories	geold	countryterritoryCode
0	06/03/2022	6	3	2022	25705.0	8.0	Austria	AT	AUT
1	05/03/2022	5	3	2022	27879.0	21.0	Austria	AT	AUT
2	04/03/2022	4	3	2022	32180.0	21.0	Austria	AT	AUT
3	03/03/2022	3	3	2022	32644.0	26.0	Austria	AT	AUT

Επεξεργασία των δεδομένων

Έχοντας βάσεις με εκατομμύρια δεδομένα όπως τα συγκεκριμένα είναι σχεδόν βέβαιο ότι θα υπάρχουν αλλοιώσεις στα δεδομένα , κάποια θα λείπουν ,κάποια ίσως έχουν τυχαίες τιμές που χαλάνε το αποτέλεσμα την μέση τιμή και την ισορροπία των δεδομένων μας, έτσι απαιτείται ο κατάλληλος έλεγχος και επεξεργασία ώστε τα δεδομένα που θα χρησιμοποιήσουμε να είναι ρεαλιστικά , σωστά και καθαρά.

Τα δεδομένα μας αποτελούνται από 11 κατηγορίες , που είναι πολλές και δεν προσφέρουν πληροφορία στα δεδομένα μας , έχουμε πολύ θόρυβο και άχρηστα δεδομένα , στην συγκεκριμένη στιγμή έχουμε 3 κατηγορίες (geoid,countryterritorycode,counteisandtettiroties) να δείχνουν το ίδιο πράγμα , δηλαδή την χώρα, το continentExp που δείχνει την Ευρώπη που ούτως η άλλως είναι μοναδική και το popData που δεν μας αφορά άμεσα, έτσι απαιτείται ο καθαρισμός των περιττών δεδομένων για να έχουμε καθαρότερη εικόνα.

Πίνακας 2 καθαρά δεδομένα

	dateRep	day	month	year	cases	deaths	countriesAndTerritories
0	06/03/2022	6	3	2022	25705.0	8.0	Austria
1	05/03/2022	5	3	2022	27879.0	21.0	Austria
2	04/03/2022	4	3	2022	32180.0	21.0	Austria
3	03/03/2022	3	3	2022	32644.0	26.0	Austria
4	02/03/2022	2	3	2022	32160.0	25.0	Austria
...
22186	08/02/2020	8	2	2020	0.0	0.0	Sweden
22187	07/02/2020	7	2	2020	0.0	0.0	Sweden
22188	06/02/2020	6	2	2020	0.0	0.0	Sweden
22189	05/02/2020	5	2	2020	0.0	0.0	Sweden
22190	04/02/2020	4	2	2020	1.0	0.0	Sweden

Έπειτα σε έναν τόσο μεγάλο αριθμό δεδομένων είναι σχεδόν σίγουρο ότι θα υπάρχουν λανθασμένες κενές τιμές (null) , συγκεκριμένα έχουμε 127 κενές τιμές στα 22190 δεδομένα μας . λόγω του μικρού αριθμού των μηδενικών τιμών (0,5%) μπορούμε απλώς να παραλείψουμε τις γραμμές με μηδενικές τιμές ώστε να μην μπλέξουμε με περεταίρω διαδικασίες και αλλοιωθούν οι άλλες τιμές.

Μορφοποίηση των δεδομένων

Πέρα από το κομμάτι των καθαρών σωστών δεδομένων , όταν έχουμε έναν τεράστιο αριθμό δεδομένων και καθόλου εύκολα επεξεργάσιμων όπως στο συγκεκριμένο dataset(έχουμε δεδομένα ανά ημερομηνία για κάθε μια χώρα ξεχωριστά με χιλιάδες εγγραφές) απαιτείται η σωστή μορφοποίηση ώστε να είναι αναγνώσιμα και κατανοητά , υπάρχουν πολλές τεχνικές και τρόποι για να γίνει αυτό και είναι από τα κύρια skills των σύγχρονων προγραμματιστών να χρησιμοποιήσουν τις κατάλληλες που ταιριάζουν στα δεδομένα, η δύναμη και η αξία ενός προγραμματιστή ειδικά του συγκεκριμένου κλάδου όπου όλοι οι αλγόριθμοι είναι αυτοματοποιημένα είναι το να ξέρουν που και ποτέ να τους χρησιμοποιούν και να τους κατανοούν.

Η python και η βιβλιοθήκη της pandas επιτρέπει ευέλικτες πράξεις στα δεδομένα μας , αφού μας ενδιαφέρουν και τα καθημερινά δεδομένα και τα συνολικά κάθε χώρας χωρίσουμε σε group βάση της χώρας και διαχωρισμό βάση κρουσμάτων ή θανάτων, τα αποτελέσματα θα φανούν πιο καθαρά στο επόμενο κεφάλαιο που αφορά την παρουσίαση των δεδομένων μας.

Περεταίρω επεξεργασία των δεδομένων

Για να γίνει η κατάλληλη ανάλυση που θα δείξουμε στο επόμενο κεφάλαιο χρειάστηκε να γίνουν μετατροπές στα δεδομένα όπως να πάρουμε μεμονωμένες χώρες και να τις χωρίσουμε από τα υπόλοιπα δεδομένα καθώς να ορίσουμε και σαν index , δηλαδή σαν όρο που θα είναι η βάση της χρονοσειράς μας την ημερομηνία, ώστε να έχουμε σωστή κατανομή των αξόνων στα plot μας , δηλαδή να βλέπουμε στον x άξονα την ημερομηνία και στον ψ τα κρούσματα ώστε να γίνεται κατανοητή εύκολα κάθε χρονοσειρά.

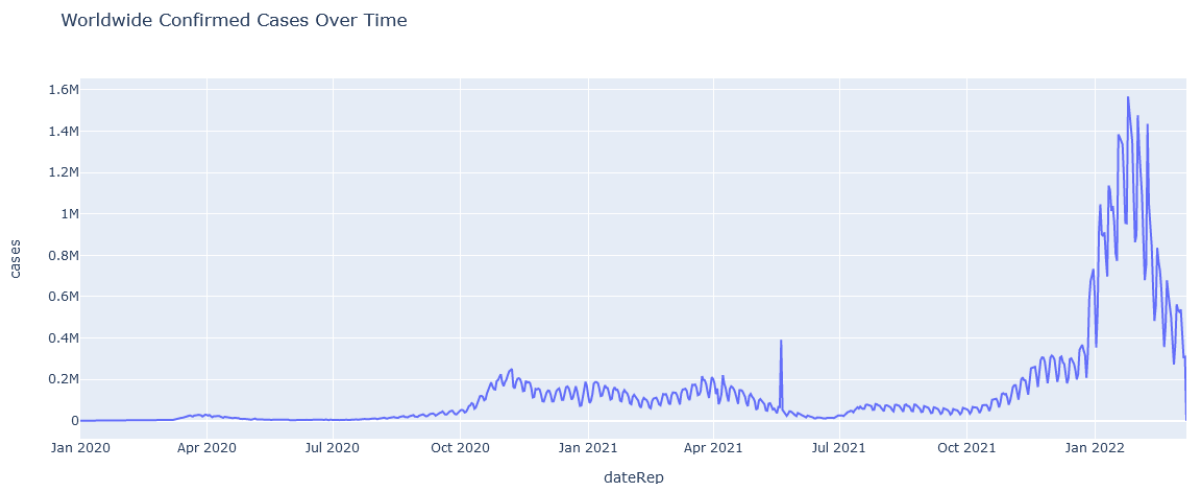
4. ΑΝΑΛΥΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Για να αρχίσουμε να κατανοούμε το τι βλέπουμε και τα δεδομένα που έχουμε δεν αρκεί η εμπειρία του προγραμματιστή και μια ιδέα , απαιτείται να απαντήσουμε σε ερωτήσεις όπως το τι δεδομένα έχουμε, πόσα , τι μας δείχνουν , τι πληροφορία περιέχουν και αυτό απαιτείται να γίνει σαφές σε οποιοδήποτε χρήστη, έτσι με τη χρήση παρουσίασης και γραφημάτων που περιέχει η ρύθμιση και της ποικιλίας που περιέχει μπορούμε να δούμε και να καταλάβουμε καλύτερα το τι βλέπουμε και να βγάλουμε συμπεράσματα.

Θα βασιστούμε στην θεωρία της ανάλυσης δεδομένων που προαναφέραμε στο κεφάλαιο 2.1 καθώς και στις διπλωματικές εργασίες Γεώργιου Ζέρβα, “DATA SCIENCE AND BUSINESS ANALYTICS – DEVELOPMENT OF A MODEL FOR PREDICTING FUEL RETAIL SALES”⁶ Και Θωμά Σαμαρά, «Ανάλυση δεδομένων και μοντέλα πρόβλεψης στην ασφαλιστική αγορά»⁷

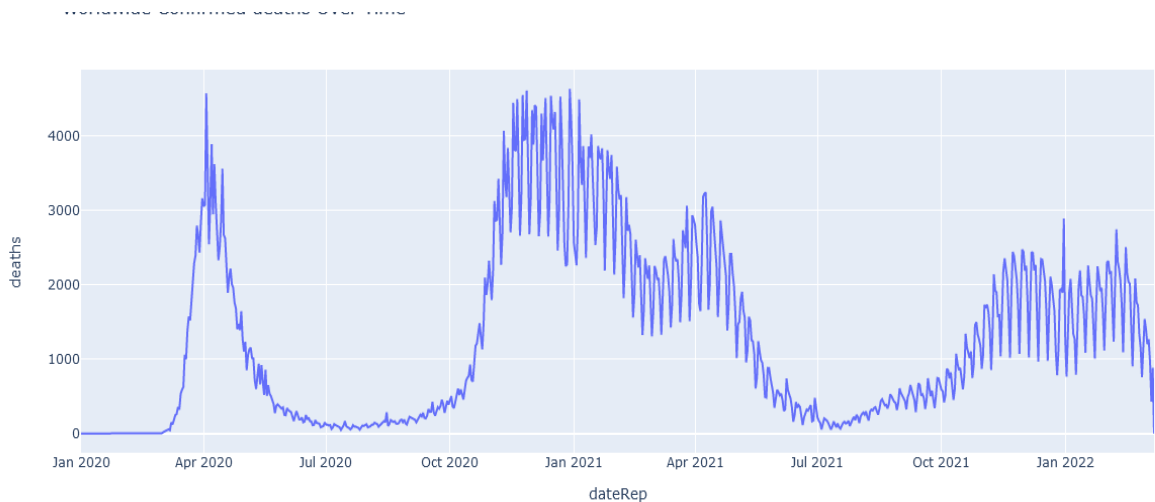
Γραφήματα

Μετά την επεξεργασία που προαναφέρθηκε μπορούμε στα δεδομένα μας να δούμε τι έχουμε , έχοντας ως κύριο παράγοντα την ημερομηνία βλέπουμε την πορεία του κορονοϊού στην Ευρώπη .

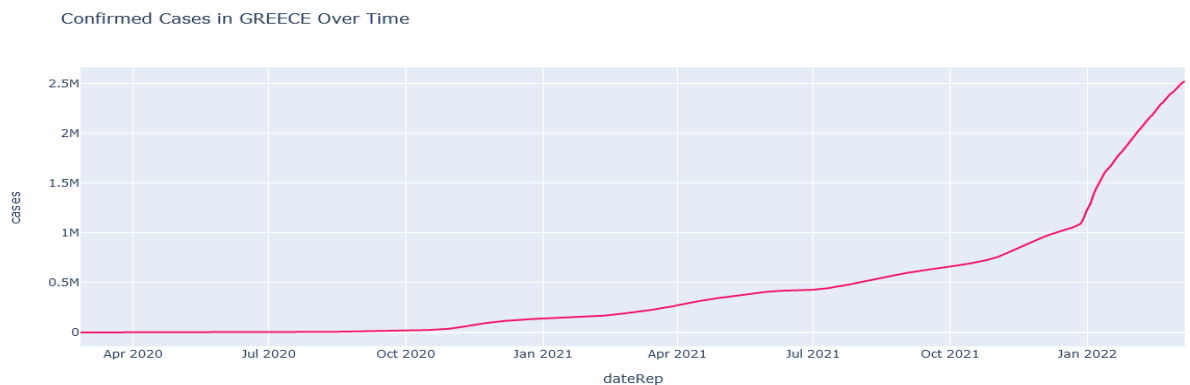


Σχήμα 5 Καθημερινά κρούσματα κορονοϊού

Με την ίδια διαδικασία βλέπουμε και τους θανάτους καθώς και διάφορες χώρες τα συνολικά κρούσματα. Έπειτα απλώς αθροίζοντας τα δεδομένα βλέπουμε και τα συνολικά κρούσματα
Παραδείγματα:



Σχήμα 6 Συνολικού θάνατοι

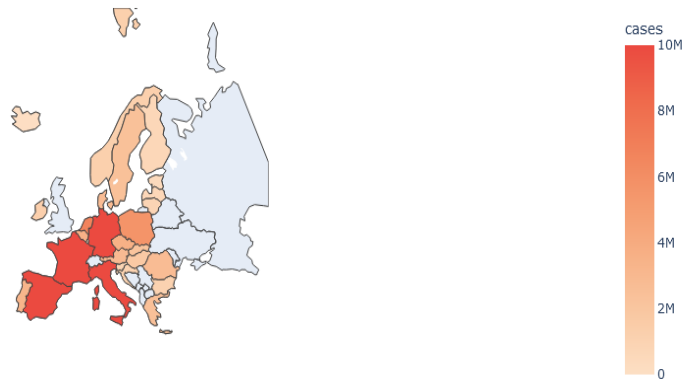


Σχήμα 7 Συνολικά κρούσματα Ελλάδα

Τώρα μπορούμε να δούμε καθαρά ότι τα δεδομένα μας είναι καθαρές χρονοσειρές που ανάλογα του τι θέλουμε να δείξουμε μπορούν με βάση την ημερομηνία να αναπαρασταθούν και να μας δώσουν πληροφορίες για τα κρούσματα και τους θανάτους λόγω covid ανά χώρα . Κάνοντας για διάφορες γειτονικές χώρες και μακρινές χώρες και καταλήγουμε στο ότι όλες οι χώρες παρόλο που αλλάζουν σε κλίμακα και ποσοστό ακολουθούν παρόμοια εκθετική αύξηση των κρουσμάτων με δείγματα εποχικότητας .

Μιας που μιλάμε για την Ευρώπη μιλάμε για έναν εικονικό χάρτη που αποτελείται από κρούσματα ανά χώρα έως την συγκεκριμένη χρονική στιγμή. Μια πιο αποτελεσματική εικονοποίηση φαίνεται στο παρακάτω σχήμα.

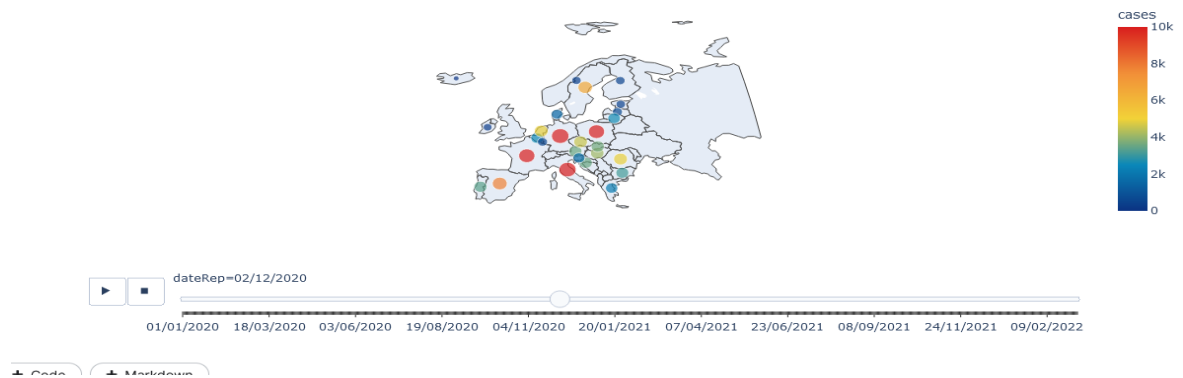
Countries with Confirmed Cases



Εικόνα 8 Χάρτης κρουσμάτων Ευρώπης

Βλέποντας αυτό το γράφημα εγείρονται όπως γιατί η νοτιοδυτική Ευρώπη έχει τα περισσότερα κρούσματα? Υπάρχει συσχέτιση μεταξύ των χωρών – κρουσμάτων? Πχ γειτονικές χώρες έχουν παρόμοια πορεία? Από τι εξαρτάται η κατανομή αυτή? Επίσης να αναφερθεί ότι ο χάρτης των θανάτων είναι σχεδόν ολόιδιος με μόνη διαφορά την Πολωνία που παρά τον μικρότερο αριθμό κρουσμάτων έχει πολλούς θανάτους.

Για να καταλάβουμε και την καθημερινή διασπορά του covid στην Ευρώπη φτιάχτηκε ένα γράφημα που δείχνει τα καθημερινά κρούσματα από την αρχή έως το τέλος σε όλη την Ευρώπη που θα μας οδηγήσει σε πολλά συμπεράσματα.



Εικόνα 9 Καθημερινά κρούσματα Ευρώπης

Αυτό το σχετικά απλό γράφημα περιέχει πολλές πληροφορίες για την Ευρώπη τα οποία θα αναφερθούν παρακάτω στα συμπεράσματα των γραφημάτων μας.

Συμπεράσματα γραφημάτων

Ας αρχίσουμε από το τελευταίο γράφημα που περιείχε την περισσότερη πληροφορία. Παρόλο που τα πρώτα δεδομένα δεν είναι πολύ αξιόπιστα καθώς μπορεί να μην καταγράφονταν σωστά η εξάπλωση άρχισε στις μεγαλύτερες χώρες της Ευρώπης νοτιοδυτικά στην Γαλλία, Ιταλία, Ισπανία, Γερμανία οι οποίες πιάστηκαν απροετοίμαστες και γρήγορα γέμισαν πολλά κρούσματα, οι κεντρικές και ανατολικές χώρες δεν πρόλαβαν να αντιδράσουν και παρόλο που αργότερα γέμισαν και αυτές κρούσματα και βάση του πληθυσμού τους είχαν πολλά κρούσματα, ενώ σε αντίθεση οι βόρειες και λόγω του ότι ήταν οι πιο απομακρυσμένες και λόγω της πολιτικής τους δεν είχαν σχεδόν κανένα θέμα να περιορίσουν τα κρούσματα και σχεδόν να έχουν το 1/10 όλων των άλλων χωρών με παρόμοιο πληθυσμό. Πάντως μέσα σε μόλις έναν μήνα όλες οι χώρες είχαν τουλάχιστον 1 κρούσμα κάτι που δείχνει την δύναμη της εξάπλωσης του covid.

Μετά από έναν χρόνο με χρήση lockdown, αλλαγή μεταλλάξεων και <<συνήθειας>> σταθεροποιήθηκαν τα κρούσματα και μειώθηκαν, σχεδόν εξαφανίστηκαν σε πολλές χώρες και ξαναείχαμε αύξηση τον επόμενο Ιανουάριο του 21 αλλά όχι σε μεγάλο βαθμό που αντιμετωπίστηκε παρόμοια από όλες τις χώρες με παρόμοια πορεία και λιγότερα κρούσματα πιθανότερα λόγω εμβολίων και χαμηλότερης μετάλλαξης.

Παρά την καλύτερη πορεία τους τελευταίους μήνες και τους πρώτους μήνες του 22 λόγω της πιο μεταδοτικής χαμηλής σε θνητότητα μετάλλαξης και λόγω της κούρασης των χωρών στα μέτρα και την μείωση επίδρασης των εμβολίων υπάρχει τεράστια αύξηση σε όλες τις χώρες της Ευρώπης και η πιο μεγάλη έξαρση γίνεται στις βόρειες τις προηγούμενα σχετικά ασφαλής χώρες, και στις κεντρικές όπως η Πολωνία συγκεκριμένα έχουμε 3 φορές περισσότερα κρούσματα από ότι είχαμε ποτέ 1,6 εκατομμύρια τη μέρα από 0,6 που ήταν το

προηγούμενο max αλλά το κύριο στοιχείο είναι ότι είχαμε μισούς λιγότερους θανάτους σε τριπλάσια κρούσματα από όσους είδαμε το 20-21.

Στο τέλος φτιάχνεται ο τελικός χάρτης των συνολικών κρουσμάτων που δείχνει την κατάσταση τον Φεβρουάριο του 22 . Βλέποντας τα γραφήματα καθημερινών και συνολικών κρουσμάτων ανά ημερομηνία (εικόνα 1,2) παρατηρούμε την ίδια <<ιστορία>> που μας είπε και τα 2 τελευταία γραφήματα Τώρα πρέπει να απαντήσουμε στις ερωτήσεις που αναφέρθηκαν πριν, αν υπάρχει σχέση μεταξύ των χωρών , τελικά ποια είχε την καλύτερη αντιμετώπιση και από τι εξαρτιόταν η εξάπλωση βάση των κρουσμάτων μας?

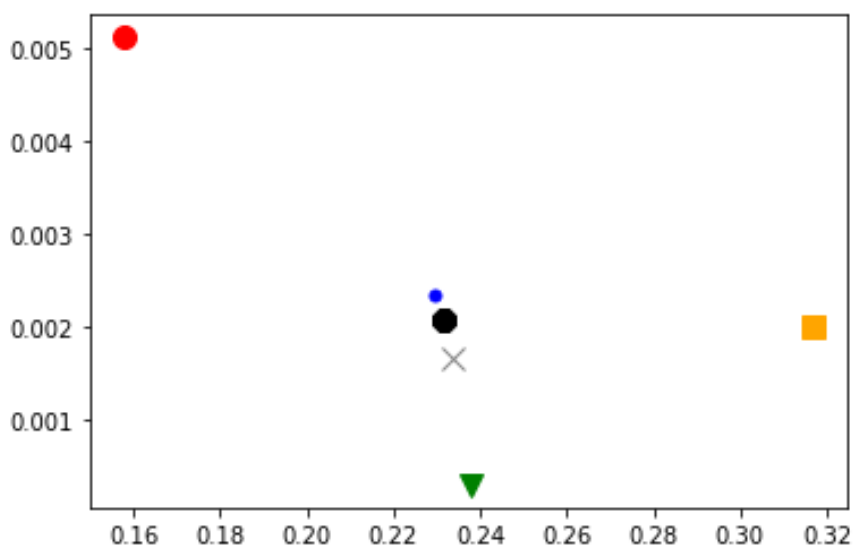
Αποκωδικοποίηση των δεδομένων

Το να βρεις και καθαρίσεις κάποια δεδομένα είναι το πρώτο βήμα , ίσως και το πιο σημαντικό για μια καλή δουλειά, καθώς εκεί στηρίζεται οτιδήποτε κάνεις έπειτα, μετά η δοκιμασία που έχει κάθε προγραμματιστής είναι ένα δημιουργικό κομμάτι που βασίζεται περισσότερο στην φαντασία και του τι θέλει να δείξει κάποιος. Συγκεκριμένα μιλάω για την αναζήτηση πληροφοριών σε καθαρισμένα δεδομένα και στο τι θέλεις να βρεις και να αποδείξεις μέσα σε αυτά .

Έως τώρα χρησιμοποιήσαμε πολλά γραφήματα για να δείξουμε πιο καθαρά τα δεδομένα μας, συγκεκριμένα δείξαμε χρονοσειρές , chloropleth (ο χάρτης με τα κρούσματα) και συνεχόμενα scatter geo plots που δείχνουν στο χάρτη την πορεία του κορονοϊού . Τώρα το επόμενο βήμα θα είναι να αναζητήσουμε πληροφορία σε αυτά και να βγάλουμε συμπεράσματα περεταίρω με αριθμητικές αποδείξεις. Ας αρχίσουμε με ένα plot που στον χ άξονα έχει τα κρούσματα/ τον πληθυσμό της χώρας ενώ στον ψ το θανάτων/ πληθυσμό χώρας συγκεκριμένα έχουμε:

2 δυτικές , 2 βόρειες και 2 νοτιοανατολικές.

Με κόκκινο την Βουλγαρία με μπλε την Ελλάδα με πράσινο την Νορβηγία με γκρι την Σουηδία με μαύρο την Ισπανία και με πορτοκαλί την Πορτογαλία



Εικόνα 10 θάνατοι/κρούσματα

Βλέπουμε ότι το 60% των τυχαίων αυτών επιλεγμένων χωρών , η Ελλάδα , η Νορβηγία, η Ισπανία και η Σουηδία είναι στο κέντρο με την Νορβηγία να έχει κάτω το 0.001 ποσοστού θανάτου βάση του πληθυσμού της .

Η Βουλγαρία παρά τον μικρότερο δείκτη κρουσμάτων από όλες έχει και το μεγαλύτερο ποσοστό θανάτων και η Πορτογαλία ενώ έχει σταθερό δείκτη θανάτων με τις άλλες έχει μεγαλύτερο δείκτη κρουσμάτων κατά πολύ.

Από αυτό βλέπουμε ότι γειτονικές χώρες δεν σημαίνει ότι θα έχουν παρόμοια χαρακτηριστικά εκτός των βόρειων που γνωρίζουμε ήδη ότι είχαν καλή αντιμετώπιση από την αρχή .

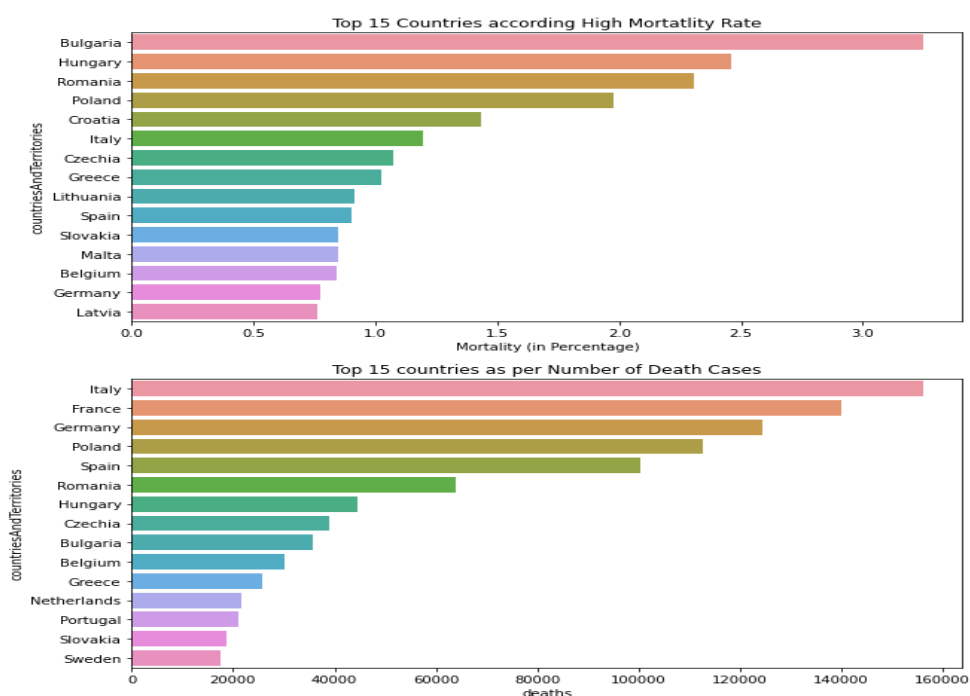
Επειδή έχουμε και κρούσματα και θανάτους και τα 2 είναι πολύ βασικά για εμάς και είναι δύσκολο να τα βλέπουμε σαν ξεχωριστές οντότητες ενώ προφανώς συσχετίζονται καθώς δεν μας νοιάζει και το πόσα κρούσματα έχουμε ή το πόσοι πέθαναν αλλά περισσότερο το πόσοι πέθαναν βάση αριθμό κρουσμάτων θα εισάγουμε έναν νέο όρο :

$$\text{MORTALITY} = (\text{DEATHS} / \text{CASES}) * 100$$

Θα είναι ο δείκτης που βάση αυτού θα συνεχίσουμε τα συμπεράσματά μας.

Χρησιμοποιώντας γραφήματα και την ανάλυση όπως αναφέρεται από M. Rubaiyat Hossain Mondal, Subrato Bharati, Prajoy Podder, Priya Podder ¹¹

Θα χρησιμοποιήσουμε τώρα 2 barplot που θα μας δώσουν μεγάλο όγκο πληροφορίας και θα κάνουν καθαρότερη μια πληροφορία:



Εικόνα 11 Barplots

Βλέπουμε τις 15 χώρες με το μεγαλύτερο mortality rate και τον περισσότερο αριθμών κρουσμάτων στην Ευρώπη , και από αυτό παίρνουμε κάτι σημαντικό: Οι θάνατοι είναι περισσότεροι στις χώρες με τον μεγαλύτερο πληθυσμό κάτι που είναι λογικό αλλά δεν το είχαμε ξεκαθαρίσει έως τώρα , και είναι κάτι που δεν έχει να κάνει τόσο με την αντιμετώπιση της χώρας αλλά τον τεράστιο αριθμό πληθυσμού. Σε αντίθεση με αυτό το mortality rate μας δείχνει καθαρά την αντιμετώπιση της χώρας και μας απαντάει στην ερώτηση , σχετίζονται γειτονικές χώρες? Οι 11 από τις 15 χώρες αφορούν κεντρικές και νοτιοανατολικές χώρες που παρά τον μικρότερο αριθμό κρουσμάτων της κατά πολύ από άλλες είχαν πολύ μεγαλύτερο ποσοστό θανάτων, στα συμπεράσματα της ανάλυσης στη συνέχεια θα αναφερθούν περισσότερα του τι μας δείχνει αυτό.

clustering

Το clustering (συσταδοποίηση) είναι η διαδικασία που ένα σύνολο δεδομένων διαχωρίζονται σε λογικές μονάδες.

Στο σημείο που είμαστε το clustering είναι αρκετά απλό καθώς θα γίνει βάση του mortality rate και θα φτιάξουμε 3 clusters χρησιμοποιώντας έναν k- means αλγόριθμο :

k- means: αρχίζει με τυχαία ανάθεση των k ομάδων και ανάθεση των κοντινότερων σημείων στην k ομάδα. Έπειτα βρίσκει τον μέσο όρο και επαναλαμβάνει μέχρι να μην αλλάξουν καθόλου οι ομάδες .

	cases	deaths	Mortality	Clusters
countriesAndTerritories				
Germany	16024707.00	124447.00	0.78	1.00
Italy	13049070.00	156079.00	1.20	1.00
Spain	11115711.00	100231.00	0.90	1.00
Czechia	3642514.00	38986.00	1.07	1.00
Belgium	3592201.00	30296.00	0.84	1.00
Greece	2522829.00	25843.00	1.02	1.00
Sweden	2455966.00	17479.00	0.71	1.00
Slovakia	2203137.00	18704.00	0.85	1.00
Croatia	1063985.00	15234.00	1.43	1.00
Lithuania	934309.00	8528.00	0.91	1.00
Slovenia	904428.00	6363.00	0.70	1.00
Latvia	704134.00	5345.00	0.76	1.00
Malta	72052.00	609.00	0.85	1.00
Poland	5685903.00	112551.00	1.98	2.00
Romania	2771449.00	63993.00	2.31	2.00
Hungary	1805898.00	44436.00	2.46	2.00
Bulgaria	1103785.00	35832.00	3.25	2.00
France	23771052.00	139949.00	0.59	0.00
Netherlands	6771643.00	21612.00	0.32	0.00
Portugal	3327430.00	21184.00	0.64	0.00
Austria	2919373.00	14444.00	0.49	0.00
Denmark	2779384.00	4370.00	0.16	0.00
Ireland	1337093.00	6610.00	0.49	0.00
Norway	1307219.00	1668.00	0.13	0.00
Finland	694142.00	2571.00	0.37	0.00
Estonia	505735.00	2293.00	0.45	0.00
Cyprus	341228.00	878.00	0.26	0.00
Luxembourg	187971.00	999.00	0.53	0.00

Εικόνα 12 Clustering

Οι ομάδες που φτιάχτηκαν είναι οι Πολωνία Βουλγαρία Ρουμανία Ουγγαρία σαν ένα cluster μόνοι τους με mortality 1.98++
Μετά χώρες με μεγαλύτερο του 0.7 mortality το 1^ο cluster και οι υπόλοιπες σε αυτό που απομένει.

Συμπεράσματα ανάλυσης

Δεν θα αναφέρω πάλι τα συμπεράσματα των διαγραμμάτων αλλά του clustering και του mortality rate. Ο κορονοϊός παρά του ότι ήταν ένα πολύ άτυχο και κακό γεγονός σε όλο τον κόσμο αλλά είχε σαν θετικό αποτέλεσμα την επαγρύπνηση των συστημάτων υγείας κάθε χώρας και την βελτίωση του υγειονομικού συστήματος κάθε χώρας, καθώς και της εμφάνισης των προβλημάτων που υπάρχουν. Όπως είπαμε οι θάνατοι είναι αναπόφευκτο να μην υπάρχουν στις χώρες με τον μεγαλύτερο πληθυσμό αφού θα έχουν εκατομμύρια κρούσματα σε μια τόσο μεταδοτική ασθένεια. Αλλά με την εισαγωγή του mortality rate μπορούμε να δούμε καθαρότερα την αντίδραση κάθε χώρας. Χώρες που βρίσκονται στο << κόκκινο>> παρά τον πολύ μικρότερο αριθμό κρουσμάτων από τις άλλες είχαν υπερβολικά μεγαλύτερους θανάτους από τις άλλες. Η συσχέτιση σε αυτές τις 4 χώρες δεν έχει να κάνει με τη γεωγραφική τους θέση ή τον πληθυσμό τους αλλά με το κακό σύστημα υγείας και την χαμηλότερη οικονομική τους δυνατότητα και ίσως τις κακές αποφάσεις που πάρθηκαν για την αντιμετώπιση του covid. Μετά στις 13

χώρες που βρίσκονται στο επόμενο cluster είναι μια μίξη χωρών που έχουν τεράστιο πληθυσμό και έτσι αναπόφευκτα λόγω συμφόρησης του συστήματος υγείας (όπως είχε η Ιταλία) και άλλων << μεσαίων >> χωρών που δεν αντιμετώπισαν σωστά τον covid. Αυτό που κάνει εντύπωση είναι ότι οι χώρες με μικρό mortality rate είχαν υπερβολικά μικρότερο mortality rate παρά τα κρούσματα τους. Η Γαλλία και η Ολλανδία παρά τον τεράστιο αριθμό κρουσμάτων τους είχαν απίστευτα καλή αντιμετώπιση και υπάρχουν και χώρες με πολύ λίγα κρούσματα και θανάτους.

Όλα αυτά μας δείχνουν ότι έως και σε ένα θέμα που αφορά την υγεία και μια τόσο μεταδοτική αρρώστια εκτός από την ατομική ικανότητα του καθενός να προσέχει τον εαυτό του και τους γύρω του παίζει τον μεγαλύτερο ρόλο η συνολική δυνατότητα και αντιμετώπιση και ετοιμότητα κάθε χώρας για ένα τέτοιο περιστατικό και το σχέδιο της. Με λιγότερα λόγια το ποσοστό θανάτων ήταν σχεδόν ανεξάρτητο από έκταση και πληθυσμό και αφορούσε περισσότερο ετοιμότητα και διαδικασίες αντιμετώπισης.

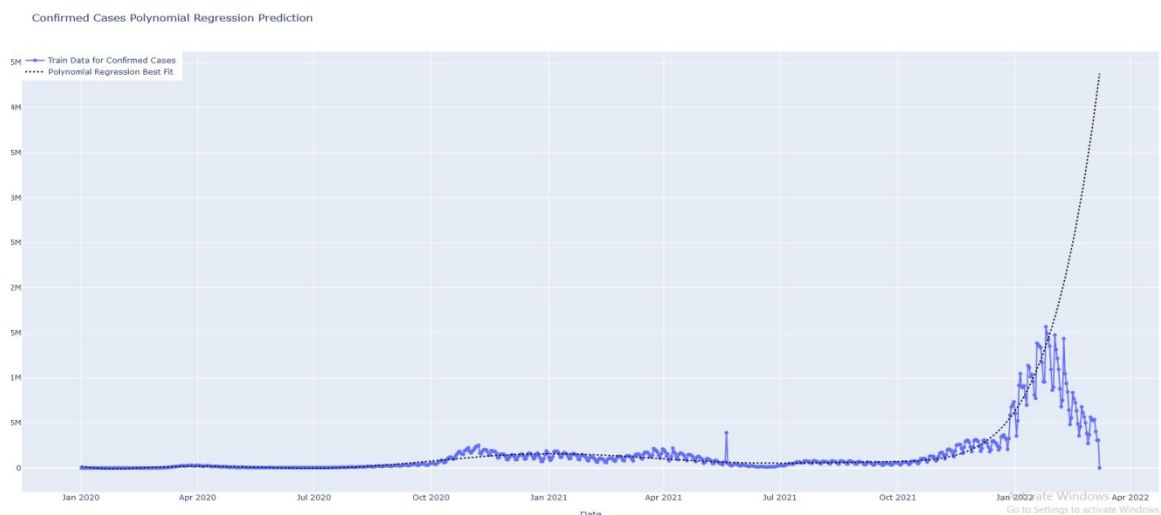
5. ΠΡΟΒΛΕΨΗ ΜΕΛΛΟΝΤΙΚΩΝ ΚΡΟΥΣΜΑΤΩΝ

Βασισμένοι στην θεωρία της προβλεπτικής ανάλυσης που αναπτύχθηκε στο κεφάλαιο 2.2 θα αναπτύξουμε τα μοντέλα και στα δικά μας δεδομένα παίρνοντας πληροφορίες και βγάζοντας συμπεράσματα , προβλέψεις και προσπαθώντας να τις σχολιάσουμε βάση της απόδοσης τους.

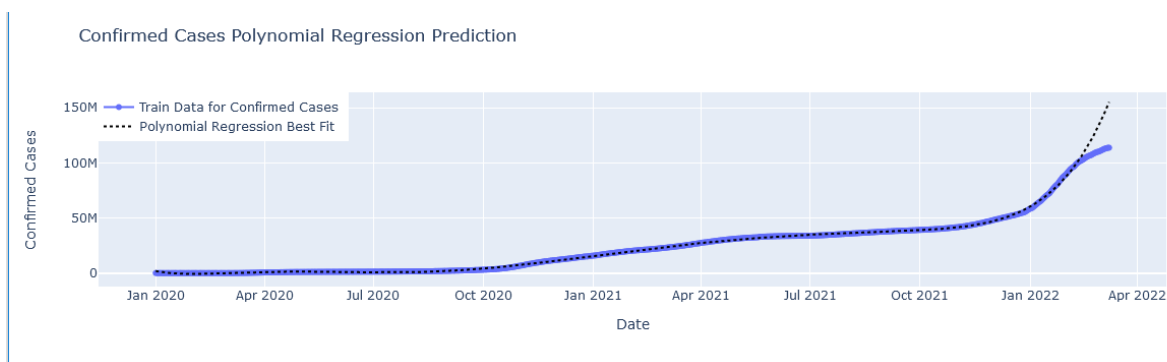
Κανονικά βάση του μοντέλου θα απαιτούταν η κατάλληλη προεπεξεργασία όπως απαλοιφή εποχικότητας και τάσης , και να φέρουμε τα δεδομένα στη βάση των αλγορίθμων αυτών , αλλά ο σκοπός της εργασίας και της συγκεκριμένης βάσης είναι να δείξουμε πως σε μια απότομη έξαρση όπως αυτή του κορονοϊού που είναι από μόνο του μια μεγάλη τάση , έχει αυξομειώσεις στα δεδομένα και παρουσιάζει απότομη περίεργη εποχικότητα , τέτοια δεδομένα πρέπει να λαμβάνουμε υπόψη εμβόλια , μεταλλάξεις και ανάλογα τη χώρα αντιμετώπιση , κάθε χώρα έχει δικιά της πορεία οπότε το να δούμε όλης της Ευρώπης μαζί πρακτικά δεν θα είναι <<σωστό>> . Θέλω να δείξω πως γίνεται σε τέτοια ανορθόδοξα δεδομένα τι αποτελέσματα θα δείξουν διάφορα μοντέλα και κατά πόσο θα είναι αποτελεσματικά. Πάνω κάτω να αναλύσουμε τους αλγορίθμους και την επίδρασή τους καθώς και πως μπορούμε να τα βελτιώσουμε μελλοντικά.

Polynomial regression

Όπως προαναφέρθηκε στο θεωρητικό υπόβαθρο που αναλύσαμε το polynomial regression θα δούμε τώρα την πρακτική υλοποίησή του πάνω στα δεδομένα μας



Σχήμα 8 Polynomial daily cases



Σχήμα 9 Polynomial total cases

Έχουμε ανάλυση των error σε επόμενο κεφάλαιο οπότε θα επικεντρωθώ στην εμφάνιση των αποτελεσμάτων.

Βλέπουμε ότι στα καθημερινά κρούσματα ακολουθείται η πορεία της καμπύλης χωρίς ίχνος μειώσεις και η απότομη αύξηση χαλάει εντελώς την συνέχεια της καμπύλης και αστοχεί να προβλέψει την πορεία της που πολύ πιθανό στην συνέχεια να αλλάζει και άλλο στα συνολικά κρούσματα παρόλο που οπτικά το αποτέλεσμα φαίνεται καλύτερο συμβαίνει ακριβώς το ίδιο πράγμα, Στο δεύτερο σχήμα μιλάμε για αριθμό 120 εκατομμυρίων κρουσμάτων, οπότε αυτή η οπτικά μικρή αλλαγή κλήσης πρόβλεψης είναι ρεαλιστικά 20+ εκατομμύρια κρούσματα εκτός. Αριθμός τεράστια εσφαλμένος.

Support vector machine

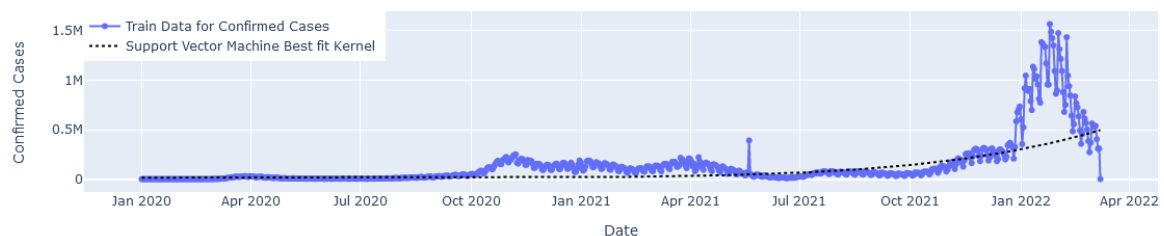
Πάλι ένας αλγόριθμος βελτιστοποίησης θεωρητικά αλγόριθμος κατηγοριοποίησης, πρακτικά αλγόριθμος λύσης οποιουδήποτε προβλήματος. Στα μη γραμμικά δεδομένα μας με χρήση ενός polynomial kernel προσπαθεί μια πολυωνυμική εξίσωση να βρει την καλύτερη λύση στη χρονοσειρά προσπαθώντας να χωρέσει όσα περισσότερα δεδομένα μπορεί σε ένα όριο e που το δίνουμε εμείς και επιστρέφει την καλύτερη δυνατή λύση με το μικρότερο error.

Όπου b ο βαθμός που θα δώσουμε εμείς και a μια σταθερά.

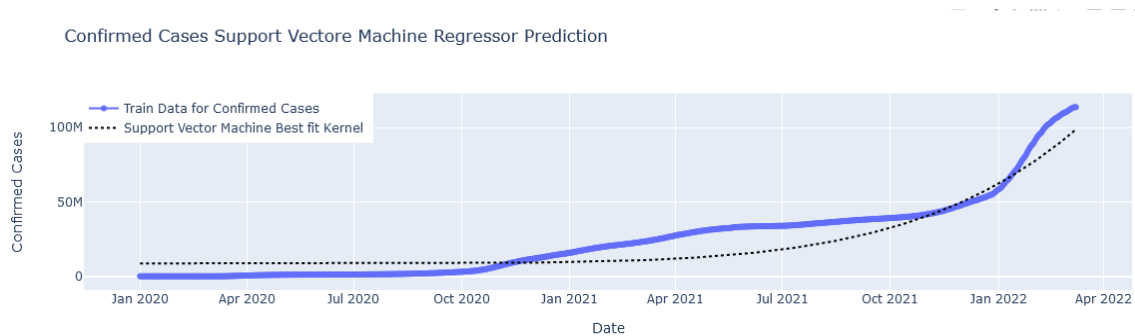
Το κύριο χαρακτηριστικό και πλεονέκτημα του svm εκτός της ευελιξίας του και της εύκολης χρήσης του είναι ότι είναι πολύ εύκολο να κάνουμε fit δεδομένα ανεξάρτητος βαθμού.

Από την άλλη μπορεί συχνά να μπλέξει με overfitting και σε μεγάλα δεδομένα με πολλές μεταβλητές και μεγάλο βαθμό να είναι αρκετά αργός.

Στο συγκεκριμένο παράδειγμα βάλαμε όριο margin $e = 0.01$, και βαθμό 6 μετά από δοκιμές και σταθερά 1!



Σχήμα 10 SVM daily cases

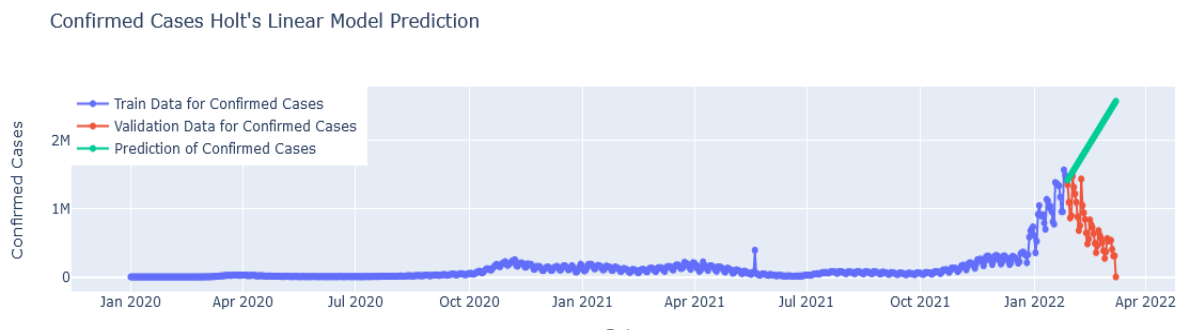


Σχήμα 11 SVM total cases

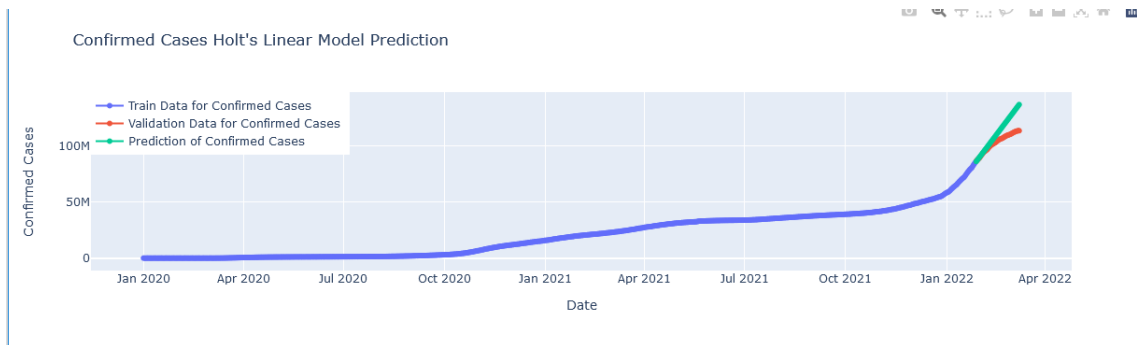
Από την πορεία της καμπύλης μπορούμε να καταλάβουμε ότι στα συνολικά κρούσματα αποτυγχάνει πλήρως να προβλεφθεί και θα δοθούν πολλά ψεύτικα δεδομένα, τώρα στα καθημερινά κρούσματα βλέπουμε ένα παράξενο φαινόμενο που συμβαίνει και δείχνει την επίδραση της ξαφνικής αύξησης και το πώς το support vector machine είναι εύθραυστο σε αυτή, στην μεγάλη αύξηση του Ιανουαρίου δεν επηρεάζεται σχεδόν καθόλου η πρόβλεψη εκτός από μια μικρή αύξηση, που μπορεί στο μέλλον όμως να ισορροπηθεί και να μην υπάρχει μεγάλη απόκλιση στα ρεαλιστικά δεδομένα αν σταθεροποιηθούν.

Holt's Models

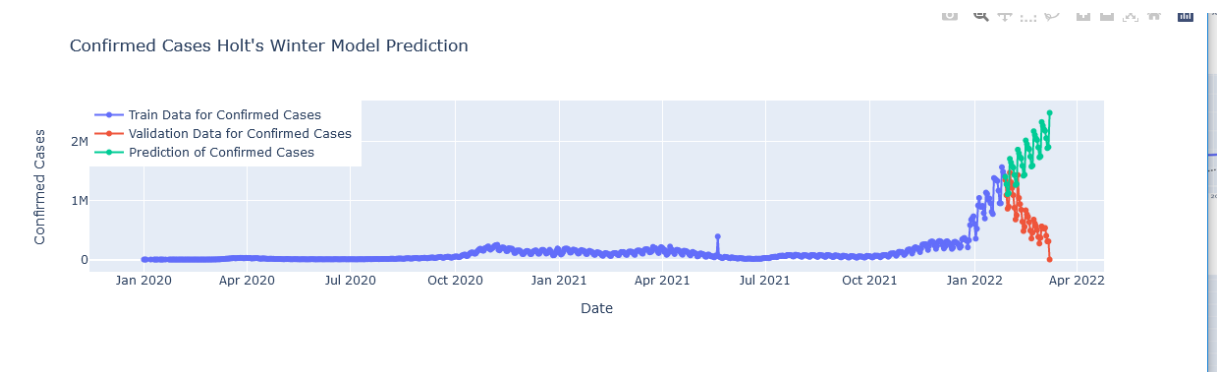
Έπειτα έχουμε τα 2 μοντέλα του hold, αυτό με την διπλή εξομάλυνση το linear model, και το holt's winter model που θεωρητικά παίρνει υπόψη του και την εποχικότητα και την τάση όπως προαναφέρθηκε.



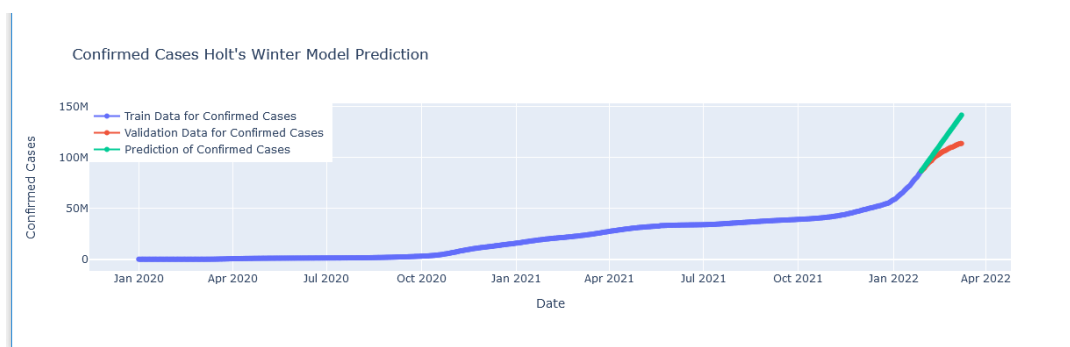
Σχήμα 12 Linear Holt daily cases



Σχήμα 13 Linear Holt total cases



Σχήμα 14 Holt's Winter daily cases



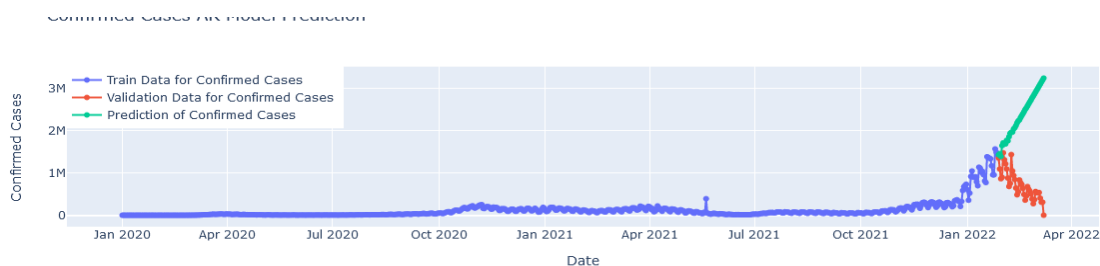
Σχήμα 15 Holt's Winter Total cases

Βλέπουμε ότι όλα τα μοντέλα σε όλες τις περιπτώσεις έχουν πολύ κακά αποτελέσματα καθώς η απότομη αύξηση μπερδεύει πολύ την πολυωνυμική λύση των holt's model και αδυνατούν να προβλέψουν την τεράστια απότομη

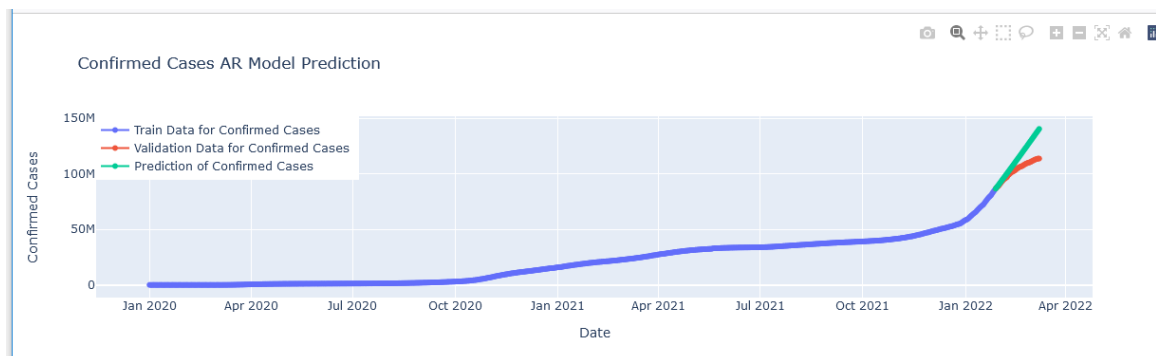
αύξηση που συμβαίνει , είτε με την ύπαρξη εποχικότητας είτε με το γραμμικό μοντέλο.

AR,MA,ARIMA,SARIMA

Τώρα έχουμε τα 4 επόμενα μοντέλα τα οποία θα τα δείξουμε ένα ένα ξεχωριστά ώστε να δείξουμε την επίδραση του καθενός και το πώς διαφέρουν μεταξύ τους , καθώς και τα αποτελέσματα τους.

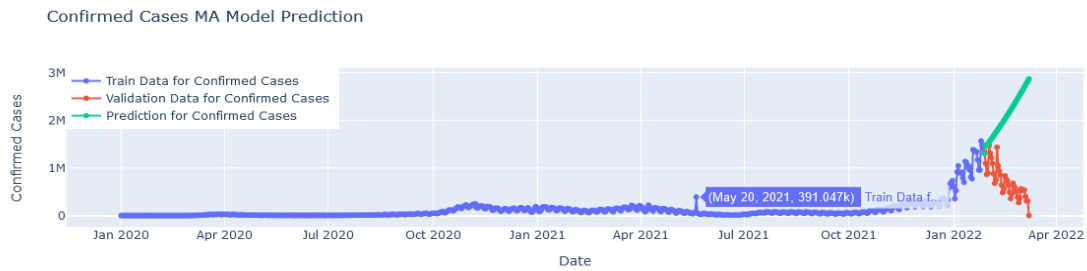


Σχήμα 16 AR daily cases

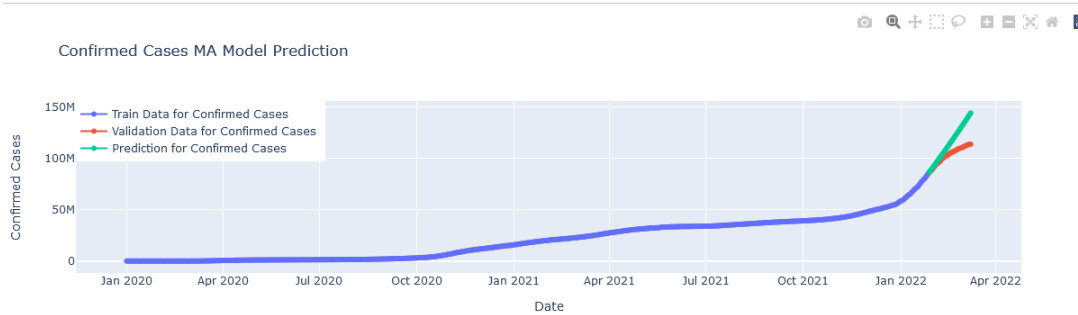


Σχήμα 17 AR total cases

Τα μοντέλα αυτοπαλινδρόμησης θεωρούν από τη μεριά τους γραμμικές σχέσεις ανάμεσα στην παρατήρηση της χρονοσειράς που εξετάζεται και στις προηγούμενες τιμές αυτής. Οπότε είναι λογικό να μην μπορούν να προβλέψουν τάσεις και εποχικότητες καθώς βασίζονται καθαρά στα προηγούμενα αριθμητικά δεδομένα , είναι ένα απλό μοντέλο που απλώς μπορεί να προβλέψει λογικά γραμμικά μοντέλα .



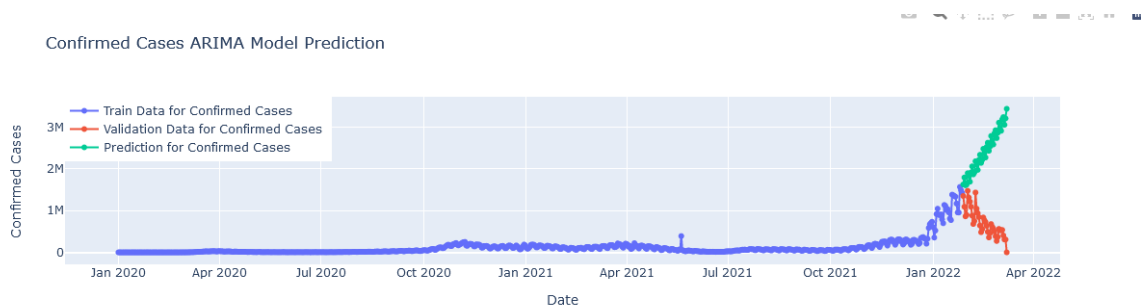
Σχήμα 18 MA daily cases



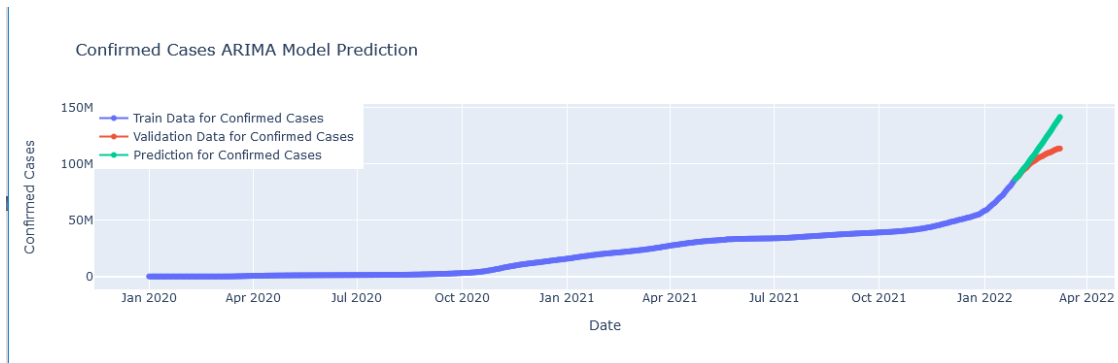
Σχήμα 19 MA total cases

Τα μοντέλα κινητού μέσου όρου θεωρούν γραμμικές σχέσεις ανάμεσα στην παρατήρηση της χρονοσειράς που εξετάζεται και στα σφάλματα που εμφάνισε το μοντέλο MA σε προηγούμενες περιόδους.

Προσπαθεί βάση των προηγούμενων error στις τιμές που βρήκε και στον κινητό μέσο όρο να προβλέψει το μέλλον αλλά όταν τα δεδομένα είναι μη στάσιμα δηλαδή περιέχουν εποχικότητα και απότομες αυξομειώσεις αδυνατεί να γίνει αυτό αποτελεσματικά αφού το error αυξομειώνεται απότομα.

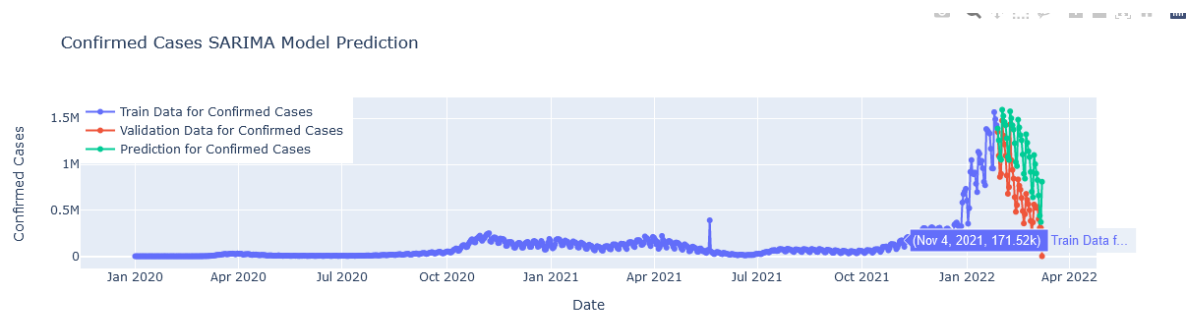


Σχήμα 20 Arima daily cases

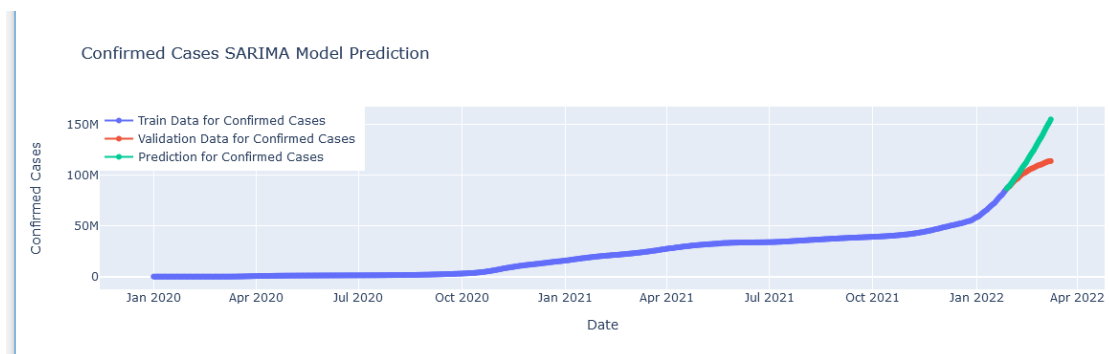


Σχήμα 21 Arima total cases

Τα μοντέλα AR και MA μπορούν να συνδυαστούν αποδοτικά για την ανάλυση και πρόβλεψη στάσιμων χρονοσειρών. Έτσι, εισάγοντας στην εξίσωση και τα μοντέλα διαφορίσης για τη διασφάλιση της στασιμότητας, προκύπτουν τα μοντέλα ARIMA(p,d,q), όπου p,d,q η τάξη του αντίστοιχου μοντέλου. Εδώ που δεν είναι στάσιμα τα δεδομένα μας βλέπουμε κατά πόσο μη αποτελεσματικό είναι το μοντέλο και αστοχεί πλήρως τα μελλοντικά κρούσματα.



Σχήμα 22 Sarima daily cases

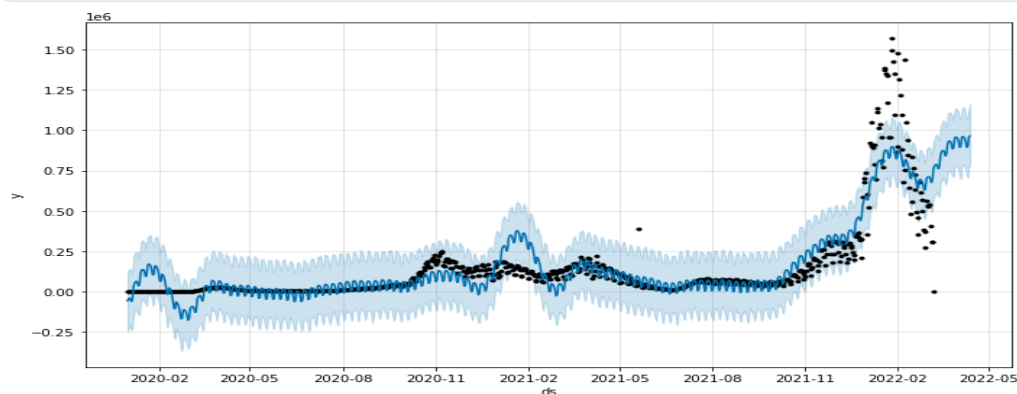


Σχήμα 23 Sarima total cases

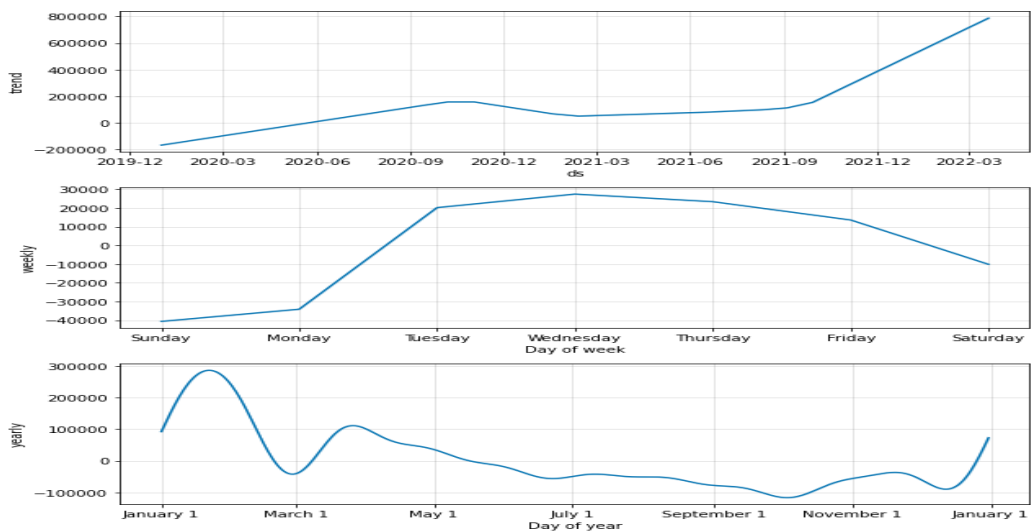
Στα καθημερινά κρούσματα βλέπουμε το πρώτο **ΟΠΤΙΚΑ** αποτελεσματικό μοντέλο. Τα δεδομένα μας σαφώς δεν είναι σταθεροποιημένα και υπάρχουν τάσεις και εποχικότητα μέσα , αλλά με τα καθημερινά κρούσματα βλέπουμε ότι οι μεταβλητές προσαρμόζονται πολύ καλά και προβλέπει πολύ καλά την πορεία των κρουσμάτων , σε αντίθεση τα συνολικά κρούσματα αστοχούν κατά αρκετά εκατομμύρια και ας είναι καλύτερο από τα προηγούμενο και αστοχεί και εντελώς την κλήση της.

Prophet model

Εδώ θα κάνουμε ξεχωριστή ανάλυση για τα καθημερινά γιατί υπάρχει πολύ πληροφορία , όπως αναφέρθηκε στην θεωρία το prophet είναι ένα πιο σύνθετο εποχιακό μοντέλο που γίνεται καθαρά σε εποχιακά δεδομένα.



Σχήμα 24 Prophet daily cases



Εικόνα 13 Prophet daily trend and seasonality

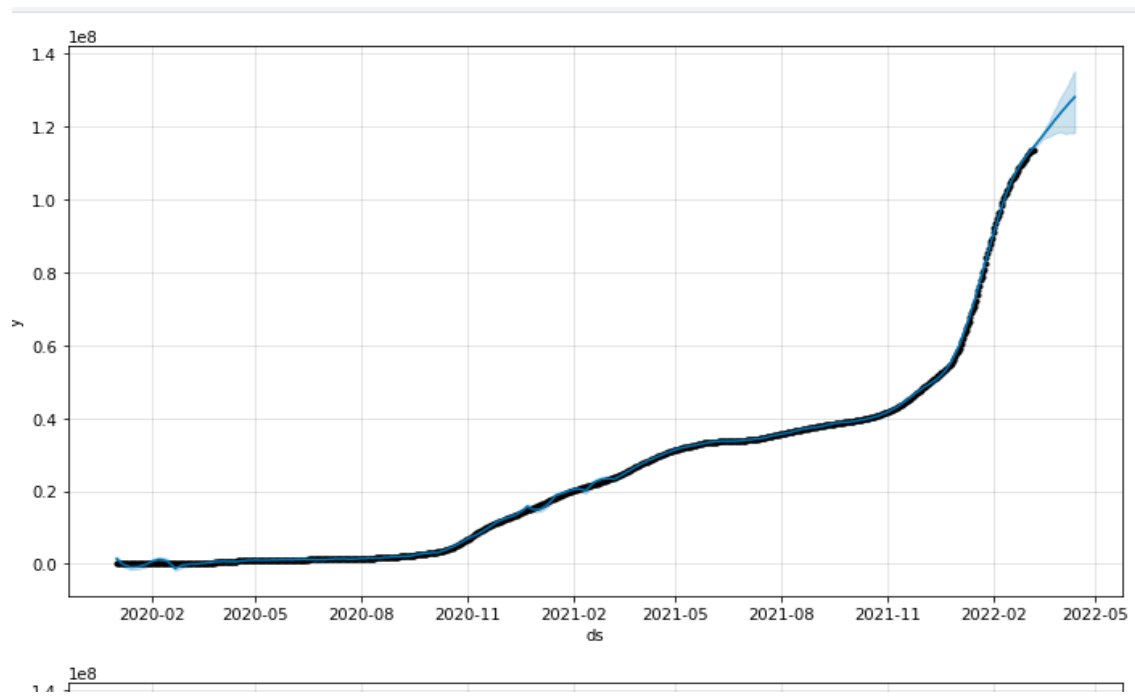
Οι τελείες είναι τα κανονικά μας δεδομένα και το περίγραμμα το διάστημα εμπιστοσύνης που κυμαίνονται τα δεδομένα μας. Η καμπύλη εξισορροπείται και παίρνοντας υπόψη την εποχικότητα, την τάση τις διακοπές-αργίες και τις μη αναγνωρισμένες αυξομειώσεις πχ εμβόλια φτιάχνει μια καμπύλη πάνω στα δεδομένα που όπως βλέπουμε συνεχίζει και μετά το τέλος των υπαρχόντων δεδομένων.

Παρόλο που το `sarima` φαίνεται ότι ακολουθεί καλύτερα τα δεδομένα μας η καμπύλη αυτή και το `prophet` υπολογίζει πιο αποτελεσματικά τα κρούσματα αν και αποτυγχάνει να προβλέψει την απότομη μείωση των κρουσμάτων.

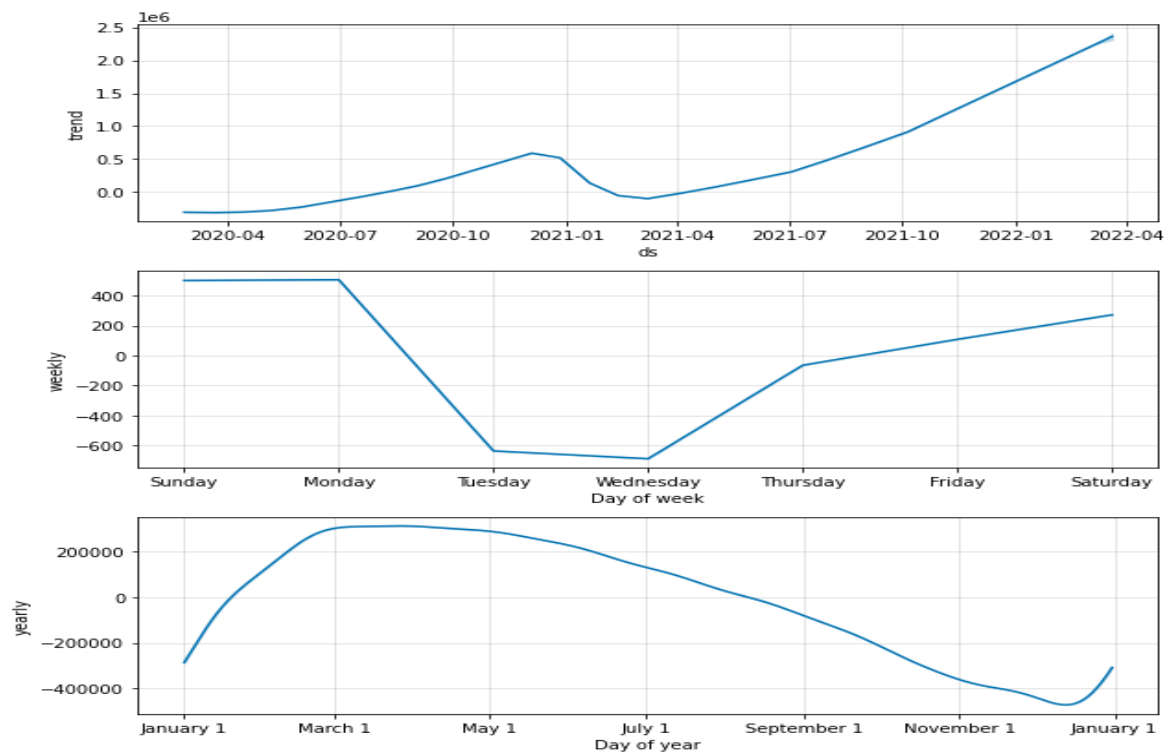
Έπειτα γίνεται ένα διάγραμμα που μας αποκαλύπτει την πορεία ανά ημέρα της εβδομάδας, μήνα του χρόνου.

Το `ds` μας δείχνει την τάση στην καμπύλη του χρόνου.

Πάμε να δούμε και το διάγραμμα των συνολικών κρουσμάτων.



Σχήμα 25 Prophet total cases



Εικόνα 14 Prophet total trend and seasonality

Εδώ έχουμε το μοναδικό αποτελεσματικό μοντέλο για τα συνολικά κρούσματα , βλέπουμε πόσο προσεγγμένα και με μικρό διάστημα ακολουθούνται τα δεδομένα .

Το άλλο αξιοσημείωτο είναι ότι ενώ τα καθημερινά και τα συνολικά κρούσματα παρουσιάζουν ίδια τάση η εποχικότητα βάση μέρας και χρόνου είναι πολύ διαφορετική , στην μέρα παρουσιάζεται π.χ αύξηση στην Τετάρτη ενώ στα συνολικά μείωση στην Πέμπτη πράγμα περίεργο .

Αξιολόγηση-συμπεράσματα

Για κάθε μοντέλο που γίνεται εκτός από τις οπτικές εικόνες που παίρνουμε και τα οπτικά μοντέλα που αναλύουμε , πρέπει να έχουμε και αριθμητικά , στατιστικά αποτελέσματα που δικαιολογούν τα ευρήματά μας , Για αρχή θα χρησιμοποιήσουμε το root mean square error που αναφέρθηκε στην ενότητα 2.2.6 για τα καθημερινά μας κρούσματα.

Επίσης να τονίσω ένα βασικό χαρακτηριστικό , το prophet ως ένα μοντέλο που κάνει fit στα δεδομένα μας έγινε έλεγχος του rmse καθ' όλη την διάρκεια της καμπύλης από την αρχή έως το τέλος.

model,rmse

-----	-----
Polynomial Regression	2.31117e+06
Support Vectore Machine	463541
Holt's Linear Model	1.44296e+06
Holt's Winter Model	1.20022e+06
AR Model	1.8295e+06
MA Model	1.57851e+06
ARIMA Model	1.89087e+06
SARIMA Model:	452434
Prophet Model	110031

Όπως βλέπουμε τα περισσότερα απλά μοντέλα που δεν ελέγχουν τον τύπο των δεδομένων και θεωρούν ότι είναι στάσιμα μοντέλα δηλαδή δεν έχουν τάση και εποχικότητα και ακολουθούν μια λογική πορεία , συγκεκριμένα τα μοντέλα :

Polynomial Regression	2.31117e+06
Holt's Linear Model	1.44296e+06
Holt's Winter Model	1.20022e+06
AR Model	1.8295e+06
MA Model	1.57851e+06
ARIMA Model	1.89087e+06

Έχουν τραγικά άσχημα αποτελέσματα και πέφτουν εντελώς έξω και στην πορεία και στην κατάσταση.

Το error είναι πολύ μεγάλο και μας δείχνει ότι έχουμε πέσει εκατομμύρια κρούσματα έξω παρόλο που όταν μιλάμε για ευρωπαϊκή κλίμακα μπορεί να μην φαίνεται τραγικό , πρακτικά αν δούμε και την καμπύλη σε αυτά τα μοντέλα και την κλήση της φεύγει εντελώς εκτός,
Το Αριμα έπειτα έχει την χειρότερη απόδοση από όλα ,

Ενώ το sarima παρόλο που φαίνεται να ακολουθά καλά την πορεία και κλήση των καθημερινών κρουσμάτων αστοχεί πλήρως τον αριθμό των κρουσμάτων.

Η μεγάλη έκπληξη είναι το svm που ακολουθά το μοντέλο , την πορεία του και την μέση τιμή του απόλυτα σωστά και καλά .

Τέλος το prophet έχει τον αριθμό 110031 . Φαίνεται ένας τεράστιος αριθμός όταν σκεφτόμαστε ότι το rmse πρέπει να τείνει στο 0 , αλλά αν πάρουμε υπόψη ότι έχουμε 750++ μέρες περίπου καθημερινών κρουσμάτων που τείνουν στο 1,4 εκατομμύρια ο αριθμός είναι σχετικά μικρός , μιλάμε για λίγα εκατομμύρια ίσως και λιγότερο αριθμό εκτός σε πορεία κρουσμάτων όλης της Ευρώπης που πιθανώς σε αυτές τις 36 μέρες να είναι πάνω από 300 εκατομμύρια. Βέβαια η πορεία των κρουσμάτων αποτυγχάνεται σχεδόν πλήρως καθώς μειώνονται απότομα τα κρούσματα και το prophet τα κρατάει σταθερά πάνω. Θα δείξουμε στο επόμενο κεφάλαιο την αλλαγή κλίμακας όταν δοκιμάσουμε το μοντέλο για μια χώρα μόνο.

Τώρα στα συνολικά κρούσματα έχουμε τα εξής αποτελέσματα:

	model,rmse
-----	-----
Polynomial Regression	1.7886e+07
Support Vectore Machine	1.67728e+07
Holt's Linear Model	1.11994e+07
Holt's Winter Model	1.39142e+07
AR Model	1.3021e+07
MA Model	1.49665e+07
ARIMA Model	1.3271e+07
SARIMA Model:	1.96183e+07
Prophet Model	352260

Όλα τα μοντέλα εκτός του prophet είναι μη αποτελεσματικά και αδυνατούν να προβλέψουν σωστά τα κρούσματα αλλά εδώ βλέπουμε κάτι εντυπωσιακό .

Το rmse του prophet είναι 352260

Αριθμός που αντιστοιχεί σε λίγα εκατομμύρια κρούσματα ,

Αλλά τώρα οι αριθμοί που έχουμε αφορούν 120 εκατομμύρια κρούσματα συνολικά , δηλαδή η πορεία που ελέγχεται σε τεράστιο βαθμό και ακολουθείται κατάλληλα. Φαίνεται και από το διάγραμμα που αναφέρθηκε στην προηγούμενη ενότητα που δείχνουμε τα αληθινά κρούσματα και τα προβλεπόμενα και οι γραμμές συμπίπτουν σχεδόν.

Τα συμπεράσματα της στατιστικής μελέτης μας δείχνουν ότι το prophet model είναι το πιο αποτελεσματικό και χωρίς πολλές πληροφορίες μπορεί να <<αγκαλιάσει>> την καμπύλη μας και να την ακολουθήσει σε αποτελεσματικό βαθμό.

Από την τελική εικόνα που παίρνουμε είναι ότι το sarima υπολογίζει καλύτερα την πορεία των καθημερινών κρουσμάτων και ως αποτυγχάνει να προβλέψει τον αριθμό , ενώ το prophet είναι φανερά το καλύτερο μοντέλο για τα συνολικά κρούσματα.

Δοκιμή σε μια μόνο χώρα

Για να γίνουν ακόμα καλύτερα αντιληπτά τα αποτελέσματα και να γίνει πιο καθαρή η εικόνα από τις τεράστιες κλίμακες δοκιμάσαμε το sarima και το prophet στην Ελλάδα ,

Τα αποτελέσματα ήταν

Για καθημερινά κρούσματα :

Root Mean Square Error for SARIMA Model: : 5306.7348547530755

Root Mean Squared Error for Prophet Model : 2974.7496728542455

Για συνολικά κρούσματα

Root Mean Square Error for SARIMA Model: 67689.36734391864

Root Mean Squared Error for Prophet Model: 11115.799417503564

Τα συνολικά κρούσματα ανέρχονται σε αριθμό άνω του ένα εκατομμύριο και τα καθημερινά σε αριθμό 40 χιλιάδες , έτσι βλέπουμε ότι σε μικρότερη κλίμακα τα error που υπάρχουν είναι πολύ μικρά ειδικά όσο αφορά το prophet model .

6. Εφαρμογή

Ένα κύριο χαρακτηριστικό της ανάλυσης που κάναμε έως τώρα και αυτό που θέλουμε να χρησιμοποιήσουμε και να πάρουμε από αυτή είναι η αποθήκευση των μοντέλων που χρησιμοποιήσαμε για μελλοντική χρήση και η δημιουργία μιας online διεπαφής χρήστη που θα μπορεί εύκολα να διαχειριστεί και να βάλει δεδομένα και να πάρει αποτελέσματα ,

Ας κάνουμε όμως κάποιες βασικές παρατηρήσεις πριν ξεκινήσουμε να αναλύουμε την εφαρμογή .

- Τα δεδομένα πρέπει να είναι στην ίδια μορφή που ήταν και τα αρχικά μας δεδομένα, δηλαδή στην μορφή.

Πίνακας 3 Μορφή δεδομένων

date Rep	d ay	mo nth	ye ar	cas es	dea ths	countriesAndT erritories	geo Id	countryterrit oryCode	popData 2020	continen tExp
-------------	---------	-----------	----------	-----------	------------	-----------------------------	-----------	--------------------------	-----------------	------------------

Τα δεδομένα αυτά μπορούν να βρεθούν στην ιστοσελίδα European Centre for Disease Prevention and Control , Αν κάποιος θέλει να βάλει δικά του δεδομένα θα πρέπει να είναι σε μορφή csv και να είναι σε αυτήν ακριβώς την μορφή επεξεργάσιμα

- Τα μοντέλα που χρησιμοποιούνται θα είναι το καθημερινό sarima , το καθημερινό prophet και το συνολικό sarima και prophet.
- Η ιστοσελίδα που φτιάχτηκε είναι σε βασική μορφή χωρίς χρήση JavaScript και πολυπλοκότητας , περισσότερο θέλει να δείχτεί η χρησιμότητα της και η ευκολία της, καθώς και ο τρόπος δημιουργίας και χρήσης παρά η εμφάνιση της και η εμφάνιση των αποτελεσμάτων .
- Το αποτέλεσμα θα είναι σε μορφή εικόνας και επιστρέφονται οι επόμενες 40 μέρες από τα δεδομένα που έβαλε ο χρήστης . Το κατά πόσο είναι κοντά στα πραγματικά δεν μπορούμε να το γνωρίζουμε μέχρι να περάσουν οι μέρες χωρίς μελέτη.

Αποθήκευση των μοντέλων

Όπως αναφέρθηκε στην ενότητα 2.3.1 η αποθήκευση των μοντέλων θα γίνει με τη χρήση της pickle , η pickle επιτρέπει την σειριοποίηση δεδομένων.

Η διαδικασία σειριοποίησης είναι ένας τρόπος μετατροπής μιας δομής δεδομένων σε μια γραμμική μορφή που μπορεί να αποθηκευτεί ή να μεταδοθεί μέσω ενός δικτύου.

Στην Python, η σειριοποίηση σας επιτρέπει να πάρετε μια σύνθετη δομή αντικειμένου και να τη μετατρέψετε σε μια ροή byte που μπορεί να αποθηκευτεί

σε έναν δίσκο ή να σταλεί μέσω ενός δικτύου. Μπορεί επίσης να δείτε αυτή τη διαδικασία που αναφέρεται ως marshalling. Η αντίστροφη διαδικασία, η οποία παίρνει μια ροή από byte και τη μετατρέπει ξανά σε μια δομή δεδομένων, ονομάζεται deserialization ή unmarshalling.

Έτσι χρησιμοποιώντας την pickle για να φορτώσουμε τα μοντέλα των καθημερινών sarima, prophet και των συνολικών sarima, prophet μπορούμε να αποθηκεύσουμε την προσαρμογή που κάνουν τα μοντέλα .

Η προσαρμογή μοντέλου είναι ένα μέτρο του πόσο καλά γενικεύεται ένα μοντέλο μηχανικής μάθησης σε παρόμοια δεδομένα με αυτά στα οποία εκπαιδεύτηκε. Ένα μοντέλο που είναι καλά τοποθετημένο παράγει πιο ακριβή αποτελέσματα. Ένα μοντέλο που είναι υπερπροσαρμοσμένο ταιριάζει πολύ με τα δεδομένα. Ένα μοντέλο που είναι ανεπαρκές δεν ταιριάζει αρκετά.

Έτσι μπορούμε να χρησιμοποιήσουμε το <<μέτρο>> αυτό για να προβλέψουμε οποιοδήποτε dataset δεδομένης της κατάλληλης βάσης. Τα μοντέλα αυτά αποθηκεύονται με ένα όνομα που εμείς δίνουμε στον φάκελο που εμείς ορίζουμε και είναι έτοιμα για χρήση μέσω python από οποιοδήποτε python πρόγραμμα έχει access σε αυτά.

Σε εμάς τα μοντέλα αποθηκεύονται με το όνομα prophet για prophet daily , sarimad για sarima daily , prophet για prophet total και sarimat για sarima total.

HTML templates

Έπειτα έχουμε τα templates που θα χρησιμοποιήσουμε , με τον όρο templates εννοούμε τις σελίδες που θα περιέχει η εφαρμογή μας , ονομάσαμε έναν φάκελο templates και μέσα θα έχουμε κάθε html σελίδα που θα χρειαστούμε , η συγκεκριμένη εφαρμογή θα έχει μια ιστοσελίδα που θα φτιαχτεί με χρήση βασικών εντολών html .

Θα περιέχει τα εξής:

- Μια φόρμα εισαγωγής δεδομένων όπου ο χρήστης θα πρέπει να να τοποθετήσει τα δεδομένα του στην μορφή που προαναφέρθηκε προηγουμένως .
- Μια φόρμα επιλογής μοντέλου όπου περιέχονται οι επιλογές :daily sarima , daily prophet , total sarima ,total prophet .
- Ένα κουμπί predict που όταν έχουν επιλεγεί το μοντέλο και τοποθετηθεί το αποτέλεσμα θα επιστρέψει τα μελλοντικά κρούσματα σε μορφή εικόνας.

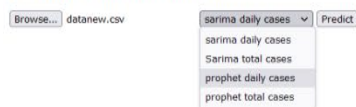
Python flask

Όσα προαναφέρθηκαν είναι απλώς η δημιουργία οπτικά της σελίδας , για να γίνονται οι παραπάνω ιδιότητες πρακτικά χρειάζεται κωδικοποίηση μέσω python , θα φτιάξουμε ένα πρόγραμμα με το όνομα app που θα μας παρέχει τις ιδιότητες αυτές και επίσης θα ανοίγει έναν server ώστε να μπορεί οποιοσδήποτε να τρέξει και να χρησιμοποιήσει το πρόγραμμα αυτό.

Για αρχή φορτώνονται τα pickle μοντέλα που αποθηκεύτηκαν πριν , και φορτώνεται η σελίδα που φτιάχτηκε με χρήση html , έπειτα γίνεται σύνδεση της σελίδας με το πρόγραμμα και φτιάχνουμε RESPONSE ανάλογα την χρήση.

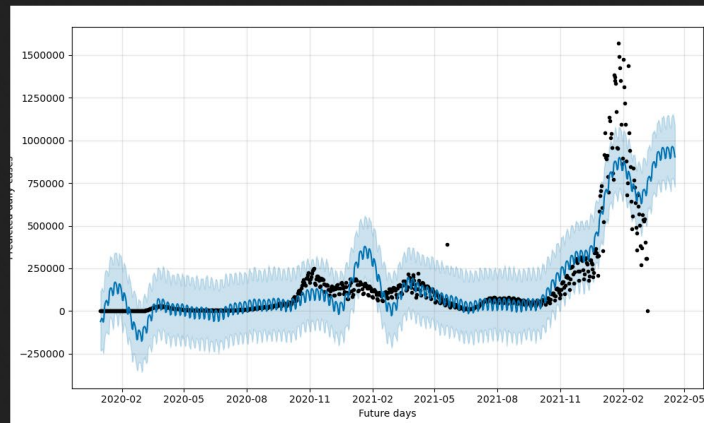
Φορτώνονται τα δεδομένα που έβαλε ο χρήστης στην φόρμα των δεδομένων , δηλαδή το csv και γίνονται οι κατάλληλες επεξεργασίες στα δεδομένα όπως καθαρισμός αχρείαστων column , διαγραφή των κενών δεδομένων , γενικά γίνεται η απαραίτητη προεπεξεργασία. Έπειτα φορτώνονται τα δεδομένα και ανάλογα με την επιλογή του χρήστη στο ποιο μοντέλο θέλει να επιλέξει επιστρέφεται η κατάλληλη εικόνα που δείχνει την πρόβλεψη των 40 επόμενων μερών

Predict covid cases



The screenshot shows a web application interface titled "Predict covid cases". It features a "Browse..." button next to the filename "datanew.csv". To the right is a dropdown menu currently showing "sarima daily cases", with a list of options: "sarima daily cases", "sarima total cases", "prophet daily cases", and "prophet total cases". A "Predict" button is located to the right of the dropdown menu.

Εικόνα 15 Εφαρμογή



Εικόνα 16 Αποτέλεσμα Εφαρμογής

Βλέπουμε το πόσο εύκολο είναι κάποιος αποτελέσματα χωρίς καμιά γνώση προγραμματισμού απλώς έχοντας στην κατοχή του δεδομένα της μορφής που προαναφέρθηκε έως κάποιο συγκεκριμένο χρονικό διάστημα, μπορούμε να χρησιμοποιούμε και να βελτιώνουμε συνεχώς τα μοντέλα μέχρι να φτάσουμε σε ένα καλό προβλεπτικό αποτέλεσμα.

7. ΣΥΜΠΕΡΑΣΜΑΤΑ-ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΤΑΣΕΙΣ

Συμπεράσματα

Σκοπός της εργασίας ήταν να δείξουμε ότι σε ένα τεράστιο θέμα όπως αυτό της κλίμακας του κορονοϊού με εκατομμύρια δεδομένα , μπορούμε να αναλύσουμε , να εξάγουμε συμπεράσματα , να προβλέψουμε και να χρησιμοποιήσουμε αυτές τις προβλέψεις. Με ένα dataset που περιείχε πολύ λίγη πληροφορία . Τα επιμέρους συμπεράσματα ανά επιμέρους τομέα είναι τα εξής.

Όσο αφορά την ανάλυση των δεδομένων καταφέραμε να εξάγουμε συμπεράσματα της πορείας της πανδημίας όπως ποιες χώρες άρχισαν πρώτες να έχουν και περιείχαν τα περισσότερα κρούσματα , καθώς και την πορεία των κρουσμάτων , από τα αποτελέσματα είναι φανερό το πόσο στην αρχή πιάστηκαν απροετοίμαστες οι περισσότερες χώρες και αδυνατούσαν να αντιμετωπίσουν έγκαιρα την πανδημία , αλλά και το πόσο η αντιμετώπιση κάθε χώρας ξεχωριστά είχε διαφορετικά αποτελέσματα στο μέγεθος της έξαρσης , για αυτό και είδαμε χώρες που συνορεύουν και είναι κυριολεκτικά δίπλα η μια στην άλλη να έχουν εντελώς διαφορετική κλίμακα κρουσμάτων. Επίσης έγινε φανερή η σύνδεση πολλών κρουσμάτων με μεγάλο πληθυσμό καθώς και ότι οι χώρες που <<χτυπήθηκαν>> πρώτες ήταν αυτές που άργησαν να μειώσουν τα κρούσματά τους.

Είναι φανερό ειδικά από το clustering των χωρών και τα διαγράμματα , χωρίς να έχουμε αριθμό εμβολίων ή τρόπο αντιμετώπισης κάθε χώρας ότι χώρες που δεν είχαν νωρίς πολλά κρούσματα και ήταν πιο έτοιμες μέσα από αποφάσεις και πλάνο αντιμετώπισης είχαν πολύ αποτελεσματική και καλή πορεία .

Ένα άλλο συμπέρασμα είναι ότι στην αρχή της πανδημίας με λιγότερα κρούσματα είχαμε περισσότερους θανάτους ενώ στο τέλος με πολύ περισσότερα 1 εκατομμύριο είχαμε πολύ λιγότερους θανάτους πράγμα που δεν μπορεί να εξηγηθεί βάση των δεδομένων που έχουμε στη διάθεση μας αλλά πιθανώς να οφείλεται στην μετάλλαξη του ιού που είναι πιο μεταδοτική αλλά λιγότερο θανάσιμη ή στην ετοιμότητα και <<συνήθεια>> των κέντρων υγείας κάθε χώρας.

Τώρα όσο αφορά την προβλεπτική ανάλυση δοκιμάσαμε πολλά μοντέλα χωρίς την απαραίτητη προεπεξεργασία για καθένα από αυτά ώστε να δούμε την απόδοση τους σε ένα τέτοιο dataset με καθαρά απότομες αυξομειώσεις και μη λογική σειρά, όσο αφορά τα καθημερινά κρούσματα είδαμε ότι τα περισσότερα μοντέλα χωρίς την στασιμότητα του dataset αδυνατούν κατά πολύ να προβλέψουν αν όχι τον αριθμό των κρουσμάτων την πορεία τους, η μεγάλη έκπληξη ήταν το svm που προσπάθησε να κάνει fit στην καμπύλη των δεδομένων και την ακολούθησε σε πολύ καλό βαθμό . Το prophet είχε το μικρότερο error αλλά αν δούμε και οπτικά το μοντέλο γνωρίζοντας από το test μας ότι υπάρχει απότομη μείωση των κρουσμάτων δεν ακολουθεί καλά την πορεία των καθημερινών κρουσμάτων. Οπότε θα μπορούσαμε να πούμε ότι το πιο καλό μοντέλο σε θέμα error ήταν το prophet αλλά το πιο καλό μοντέλο σε

θέμα πορείας της πανδημίας που παίρνει την σωστή κλήση αλλά αστοχεί τον αριθμό των κρουσμάτων ήταν το sarima.

Όσο αφορά τα συνολικά κρούσματα το Facebook's prophet είναι με διαφορά το καλύτερο μοντέλο και ακολουθεί πολύ κοντά την καμπύλη των κρουσμάτων , είναι το μόνο που ακολουθεί σωστά την κλίση από όλα τα μοντέλα και επίσης έχει ελάχιστα μικρό error όταν μιλάμε για εκατομμύρια κρούσματα , όλα αυτά αποδεικνύονται και όταν χρησιμοποιούμε το παράδειγμα της Ελλάδας και βλέπουμε σε πολύ μικρότερη κλίμακα τους αριθμούς αυτούς, στα καθημερινά έχει διπλάσιο error το sarima αλλά η κλίση ακολουθάτε πολύ πιο αποτελεσματικά και τα 2 είναι σχετικά μικρά ενώ στα συνολικά κρούσματα το sarima είναι 6 φορές χειρότερο από ότι το prophet και αδυνατεί να ακολουθήσει και καλά την καμπύλη.

Τονίζοντας πάντα ότι δεν έχει γίνει πλήρης προεπεξεργασία για κάθε μοντέλο στα δεδομένα μας εκτός των βασικών ενεργειών που προαναφέρθηκαν στο κεφάλαιο 3 που καθαρίσαμε και φτιάξαμε τα δεδομένα το καλύτερο μοντέλο για καθημερινά κρούσματα φαίνεται το sarima και το καλύτερο μοντέλο για συνολικά το prophet.

Τέλος είδαμε ότι όλα αυτά μπορούν να χρησιμοποιηθούν και περαστούν σε μια εύκολη ευέλικτη εφαρμογή η οποία είναι πολύ πρακτική στην χρήση της και μπορεί εύκολα με λίγο κώδικα να επιστρέψει σημαντικές πληροφορίες , αυτή την εφαρμογή μπορεί να την χρησιμοποιήσει μια εταιρία για να προβλέπει τα κρούσματα προς δικό της όφελος , μια οργάνωση , μια χώρα , καθώς και μπορεί εύκολα να μεταβληθεί να δείχνει κάτι άλλο πχ κρούσματα συγκεκριμένης χώρας ή κρούσματα όλου του κόσμου.

Με όλα όσα προαναφέρθηκαν καταλήγουμε στο ότι όλα αυτά τα ξεχωριστά θέματα που μπορούσαν να είναι μια ανάλυση σκέτα από μόνα τους , όταν τα συνδυάσουμε και τα χρησιμοποιήσουμε μαζί φτιάχνουν ένα αποτελεσματικό , χρήσιμο σύνολο που μπορεί να χρησιμοποιηθεί , να βελτιωθεί και να επεξεργαστεί από εταιρίες , προγραμματιστές και διάφορους οργανισμούς που θα ωφεληθούν από αυτό , την πρόβλεψη των μελλοντικών κρουσμάτων ή την ανάλυση των δεδομένων μας και την εφαρμογή που φτιάχτηκε.

Σε σύνολο η έρευνα αυτή ανέλυσε βασικά πράγματα για τους αλγόριθμους προβλεπτικής ανάλυσης και το πώς αυτοί συμπεριφέρονται , την ανάλυση και οπτικοποίηση των δεδομένων μιας ξαφνικής πανδημίας από την αρχή της έως το τωρινό χρονικό σημείο , και πως τα αποτελέσματα αυτά μπορούν να χρησιμοποιηθούν αποτελεσματικά και να εξαχθεί η πληροφορία που θέλει κάποιος .

Τα δεδομένα μπορεί να είναι ο μελλοντικός χρυσός , αλλά εκτός από το μεγάλο θέμα που είναι η συλλογή τους , η επεξεργασία τους και η εξαγωγή τους είναι και το τι θα κάνουμε με αυτά. Όλη αυτή η έρευνα ήταν ένα δυνατό παράδειγμα του πως με βασικούς αλγόριθμους ανάλυσης δεδομένων και κάποιες γνώσεις ο χρυσός μπορεί να αξιοποιηθεί και εξηγεί το γιατί ο περισσότερος κόσμος πιστεύει ότι στα επόμενα χρόνια η αποθήκευση , ανάλυση , πρόβλεψη των δεδομένων και η εξαγωγή πληροφοριών αυτά θα είναι κάτι βασικό που θα έχει κύριο ρόλο στον κόσμο

Όπως αναφέρεται από την allied market research παγκόσμια δύναμη ανάλυσης δεδομένων της αγοράς και εξαγωγής συμπερασμάτων:
<<Allied Market Research reports the big data and business analytics market hit \$193.14 billion in 2019, and estimates it will grow to \$420.98 billion by 2027 at a compound annual growth rate of 10.9%.>>

Μελλοντικές προτάσεις

Όπως κάθε έρευνα και εργασία πάντα υπάρχει περιθώριο βελτίωσης και αλλαγής σε κάθε υπάρχον δουλειά , προς τους επαγγελματίες του χώρου , τους φοιτητές και οποιοδήποτε θέλει να ασχοληθεί με ένα παρόμοιο θέμα θα αναφέρω πάλι ανά τομέα το πώς θα μπορούσε να βελτιωθεί η να γίνει μια διαφορετική δοκιμή.

πανδημία έχει πολλές ακόμα σημαντικές μεταβλητές όπως , ποια στάση ακολούθησε κάθε χώρα για την πανδημία? Ηλικίες ανθρώπων που νόσησαν και τυχόν προβλήματα υγείας για να γίνει και ανάλυση ανά χώρα προς αυτόν τον τομέα , πόσα εμβόλια είχε κάνει ο πληθυσμός της χώρας στην πορεία του covid και πως συνδέονται όλα αυτά με τα κρούσματα και τους θανάτους? Υπάρχουν πολλές σημαντικές μεταβλητές που δεν μπορούσαμε να έχουμε πρόσβαση με αυτό το dataset και μπορούν να χρησιμοποιηθούν για την καλύτερη απόδοση και μελέτη των χρονοσειρών , καθώς και σε πιο πολλά συμπεράσματα στην ανάλυση των δεδομένων και χρονοσειρών ανά χώρα κ.α. , επίσης θα μπορούσε να διερευνηθεί και παγκόσμια το φαινόμενο και όχι μόνο στην Ευρώπη ή να διερευνηθεί σε πιο μικρή κλίμακα για να βγουν πιο συγκεκριμένα πορίσματα.

Έπειτα υπάρχουν πολλών ειδών διαγράμματα που θα μπορούσαν ναδειχθούν με την μεγάλη γκάμα των pandas και θα μπορούσε να αναλυθεί περισσότερο η συσχέτιση των χωρών μεταξύ τους όσο αφορά την πορεία αυτής της νόσου ή να χρησιμοποιηθούν και άλλα πιο πολύπλοκα διαγράμματα.

Μετά όσο αφορά την προβλεπτική ανάλυση οι μελλοντικοί αναλυτές θα μπορούσαν να χρησιμοποιήσουν ως βάση τα αποτελέσματα των προβλεπτικών αυτών μοντέλων για να δουν πως μπορούν να επεξεργαστούν καλύτερα τα δεδομένα τους για κάθε μοντέλο ξεχωριστά, στα περισσότερα φαίνεται η ανάγκη για stationary δεδομένα και ότι αν κάποιος θέλει να κάνει αποτελεσματική πρόβλεψη με κάποιο μοντέλο όπως πχ το arima η το holt model θα πρέπει πρώτα να απάλειψη τυχόν εποχικότητα και τάση και έπειτα να κάνει αντιστροφή στο αποτέλεσμα ώστε να δει την σωστή και αποτελεσματική πρόβλεψη με αυτά τα μοντέλα , εδώ περισσότερο δόθηκε βάση όχι στην αποτελεσματικότητα αλλά στο πως αντιδρούν αυτά τα μοντέλα σε τέτοια δεδομένα , καθαρά μεν όμως μη επεξεργάσιμα πάνω στο κάθε μοντέλο και τις ανάγκες του. Έτσι θα μπορούσε ανά μοντέλο να γίνει η κατάλληλη προεπεξεργασία .

Η όλη έρευνα βάση των δεδομένων που δίνονται θα μπορούσε να γενικευτεί και σε οποιαδήποτε μορφή πανδημίας ή απότομης εξάπλωσης βάση του dataset που χρησιμοποιείται και θα μπορούσαν να παραχθούν πολλών ειδών αποτελέσματα βάση την φαντασία και τον σκοπό του προγραμματιστή ή του οργανισμού που παράγει αυτή την έρευνα , θα μπορούσε να δοθεί βάση προς μια συγκεκριμένη κατεύθυνση εκτός μόνο των κρουσμάτων και των θανάτων , θα μπορούσε να είναι μια έρευνα του κατά πόσο αποτελεσματικά είναι τα εμβόλια , ή του κατά πόσο η σημερινή Ευρώπη είναι ικανή να καταπολεμήσει κάτι τέτοιο με την τωρινή ιατρική δύναμη , πόσο βελτιώθηκε η ιατρική δύναμη ανά χώρα κατά την διάρκεια της πανδημίας , ποια μέθοδος αντιμετώπισης φάνηκε η πιο αποτελεσματική και πολλά άλλα παραδείγματα .

Όσο αφορά την εφαρμογή είναι το κύριο αποτέλεσμα γιατί κάνει την έρευνα από μια απλή έρευνα , να έχει ένα αποτέλεσμα που μπορεί να χρησιμοποιήσει πρακτικά οποιοσδήποτε , η εφαρμογή βασίζεται στα μοντέλα οπότε αν κάποιος θέλει να φτιάξει μια τέτοια αποτελεσματική εφαρμογή θα πρέπει πρώτα να τελειοποιήσει το αποτέλεσμα των μοντέλων του . Επίσης εδώ φτιάχτηκε ένα βασικό template με βασικές html εντολές καθαρά για να δείχτεί ότι είναι λειτουργικό και δουλεύει , μια τέτοια εφαρμογή ένας μελλοντικός ερευνητής θα μπορούσε να την κάνει πολύ πιο εξεζητημένη με χρήση JavaScript , css πιο περίπλοκες χρήσεις και ευφάνταστες , πιο όμορφα χρώματα και οδηγίες για την χρήση της .

8. ΕΙΚΟΝΕΣ, ΠΙΝΑΚΕΣ, ΣΧΗΜΑΤΑ, ΟΡΟΛΟΓΙΕΣ, ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

Παράθεση εικόνων

Εικόνα 1 : Χάρτης χωροπλήθους

Εικόνα 2 : clustering

Εικόνα 3 : polynomial svm

Εικόνα 4: holt's linear model

Εικόνα 5 : Holt's winter model

Εικόνα 6: AR,MA,ARIMA,SARIMA

Εικόνα 7 : Prophet model

Εικόνα 8 : Χάρτης κρουσμάτων Ευρώπης

Εικόνα 9 : Καθημερινά κρούσματα Ευρώπης

Εικόνα 10: Θάνατοι/κρούσματα

Εικόνα 11: barplots

Εικόνα 12: clustering

Εικόνα 13 :prophet daily trend and seasonality

Εικόνα 14 prophet total trend and seasonality

Εικόνα 15 Εφαρμογή

Εικόνα 16 Αποτέλεσμα εφαρμογής

Παράθεση Πινάκων

Πίνακας 1 δεδομένα

Πίνακας 2 καθαρά δεδομένα

Πίνακας 3 Μορφή δεδομένων

Πίνακας 4 Πίνακας ορολογίας

Πίνακας 5 Συντομογραφίες

Παράθεση Σχημάτων

Σχήμα 1: line chart

Σχήμα 2 : ιστόγραμμα

Σχήμα 3: Γράφημα μπάρας

Σχήμα 4 : scatter plot

Σχήμα 5: Καθημερινά κρούσματα κορονοϊού

Σχήμα 6 : Συνολικού θάνατοι

Σχήμα 7 : Συνολικά κρούσματα Ελλάδας

Σχήμα 8 : polynomial daily cases

Σχήμα 9 : polynomial total cases

Σχήμα 10 : svm daily cases

Σχήμα 11: svm total cases

Σχήμα 12 : Linear holt daily cases

Σχήμα 13 : Linear holt total cases

Σχήμα 14 : Holt's Winter daily cases

Σχήμα 15 : Holt's Winter total cases

Σχήμα 16 : AR daily cases

Σχήμα 17 : AR total cases

Σχήμα 18 : MA daily cases

Σχήμα 19 : MA total cases

Σχήμα 20 : ARIMA daily cases

Σχήμα 21 : Arima total cases

Σχήμα 22 : Sarima daily cases

Σχήμα 23 : Sarima total cases

Σχήμα 24: Prophet daily cases

Σχήμα 25 : Prophet total cases

Πίνακας Ορολογίας

COVID	Κορονοϊός
Clustering	Συσταδοποίηση
Stationary data	Στατικά δεδομένα
outliers	Ακραίες τιμές
ιστόγραμμα	histogram
Γράφημα μπάρας	barplot
Χάρτης χωροπλήθους	Chloropleth map
Χρονοσειρές	Time series
Αυτοπαλινδρόμηση	Autoregression

Συντομογραφίες

SVM	Support vector machine
MSE	Mean square error
RMSE	Root mean square error

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

- [1] Mohammed j.Zaki , Wagner , " Data Mining and Machine Learning: Fundamental Concepts and Algorithms," p. 357-385, 2014.
- [2] Jason W. Osborne, Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data , 2013, p 105-132
- [3] Dipti Theng "Machine Learning Algorithms for Predictive Analytics: A Review and New Perspectives" conference : High technology letters scopus Vol 26 Issue 2 ,2020
- [4] A.K.Dubeya,A.Kumara, , "Study and analysis of SARIMA and LSTM in forecasting time series data" volume 47 october 2021
- [5] Prophet: forecasting at scale By: Sean J. Taylor, Ben Letham.
<https://research.facebook.com/blog/2017/02/prophet-forecasting-at-scale/> 2017
- [6] Γεώργιου Ζέρβα, "DATA SCIENCE AND BUSINESS ANALYTICS – DEVELOPMEN OF A MODEL FOR PREDICTING FUEL RETAIL SALES", Διπλωματική εργασία, Πανεπιστήμιο Αιγαίου Τμήμα Μηχανικών Πληροφοριακών και επικοινωνιακών συστημάτων, 2018
- [7] Θωμά Σαμαρά, «Ανάλυση δεδομένων και μοντέλα πρόβλεψης στην ασφαλιστική αγορά», Μεταπτυχιακή Διπλωματική Εργασία», Σχολή Θετικών Επιστημών και Τεχνολογίας Μεταπτυχιακή Εξειδίκευση στα Πληροφοριακά Συστήματα, Ελληνικό ανοιχτό πανεπιστήμιο, 2021.
- [8] Miguel Grinberg , Flask Web Development: Developing Web Applications with Python 1st Edition, 2014
- [9] Ravindra Sharma, "model deployment using flask " ,
<https://towardsdatascience.com/model-deployment-using-flask-c5dcbb6499c9> , 2021
- [10] Ζιουάλα Μαρία, «Ανάλυση χρονοσειρών για την πρόβλεψη επιχειρήσεων» , πτυχιακή εργασία , τεχνολογικό εκπαιδευτικό ίδρυμα δυτικής Ελλάδας τμήμα σχολή διοίκησης και οικονομίας , 2018
- [11] M. Rubaiyat Hossain Mondal,Subrato Bharati, Prajoy Podder,Priya Podder , "data analytics for novel coronavirus disease", Volume 20, 2020
- [12] Pang-Ning tan , Michael Steinbach, Anuj Karpatne,Vipin Kumar , « εισαγωγή στην εξόρυξη δεδομένων 2^η έκδοση» pp17-73 ,2019
- [13] Μαρία Χαλκίδη – Μ.Βαζιργινης´, «εξόρυξη γνώσης από βάσεις δεδομένων και τον παγκόσμιο ιστό 2^η έκδοση» , pp 171-230 291-344