ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΜΕΛΕΤΗ ΙΑΤΡΙΚΩΝ ΚΑΙ ΚΟΙΝΩΝΙΟΛΟΓΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΠΟΥ ΑΦΟΡΟΥΝ ΤΟΝ ΝΕΟ ΚΟΡΟΝΟΪΟ COVID-19 ΜΕ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Διπλωματική Εργασία

ΝΙΚΟΣ ΠΛΕΣΣΑΣ-ΑΥΓΗΤΙΔΗΣ

Επιβλέπουσα: ΕΛΕΝΗ ΤΟΥΣΙΔΟΥ

Βόλος 2021

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΜΕΛΕΤΗ ΙΑΤΡΙΚΩΝ ΚΑΙ ΚΟΙΝΩΝΙΟΛΟΓΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΠΟΥ ΑΦΟΡΟΥΝ ΤΟΝ ΝΕΟ ΚΟΡΟΝΟΪΟ COVID-19 ΜΕ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Διπλωματική Εργασία

ΝΙΚΟΣ ΠΛΕΣΣΑΣ-ΑΥΓΗΤΙΔΗΣ

Επιβλέπουσα: ΕΛΕΝΗ ΤΟΥΣΙΔΟΥ

Βόλος 2021

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

A STUDY ON MEDICAL AND SOCIOLOGICAL DATA REGARDING CORONAVIRUS COVID-19 USING DATA MINING ALGORITHMS

Diploma Thesis

NIKOS PLESSAS-AVGITIDIS

Supervisor: ELENI TOUSIDOU

Volos 2021

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπουσα **ΕΛΕΝΗ ΤΟΥΣΙΔΟΥ**
ΕΔΙΠ, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **ΜΙΧΑΗΛ ΒΑΣΙΛΑΚΟΠΟΥΛΟΣ**
Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **ΠΑΝΑΓΙΩΤΑ ΤΣΟΜΠΑΝΟΠΟΥΛΟΥ**
Αναπληρώτρια Καθηγήτρια, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ημερομηνία έγκρισης: 25-2-2021

Ευχαριστίες

Φθάνοντας στην ολοκλήρωση αυτής της διπλωματικής εργασίας που σηματοδοτεί και το τέλος των σπουδών μου, κατ'αρχήν θέλω να ευχαριστήσω το Δημόσιο Πανεπιστήμιο για το γεγονός ότι δίνει δεύτερες ευκαιρίες σε ανθρώπους που άργησαν να ολοκληρώσουν τις σπουδές τους για τους όποιους λόγους. Εύχομαι να μην αλλάξει αυτό.

Ευχαριστώ ιδιαίτερα την επιβλέπουσα Καθηγήτρια κα Ελένη Τουσίδου για την άψογη συνεργασία μας και τις συμβουλές/διορθώσεις που μου υπέδειξε κάθ' όλη την διάρκεια της εκπόνησης της διπλωματικής. Χωρίς αυτές τις καίριες παρατηρήσεις και την αμεσότητά τους, θα ήταν αδύνατη η έγκαιρη ολοκλήρωση της εργασίας. Ευχαριστώ επίσης τα υπόλοιπα μέλη της τριμελούς επιτροπής, κο Βασιλακόπουλο και κα Τσομπανοπούλου.

Μεγάλη χάρη χρωστάω σε αρκετούς ανθρώπους που με βοήθησαν με διάφορους τρόπους στο να ανταπεξέλθω στο ότι ζούσα στη Θεσσαλονίκη ενώ η σχολή ήταν στο Βόλο - με φιλοξένησαν, μου έδωσαν σημειώσεις κτλ. Χωρίς την βοήθεια τους δεν θα τα κατάφερνα.

Χωρίς την συμπαράσταση και την στήριξη των γονιών μου, πρακτική και ψυχολογική επίσης. Ένα ιδιαίτερο ευχαριστώ στην Κωνσταντίνα που με άντεξε όλα αυτά τα χρόνια και στάθηκε δίπλα μου.

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

«Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής».

Ο/Η Δηλών/ούσα

ΝΙΚΟΣ ΠΛΕΣΣΑΣ-ΑΥΓΗΤΙΔΗΣ

19-2-2021

Περίληψη

Στα τέλη του 2019 εμφανίστηκε στην πόλη Wuhan της Κίνας ένας νέος, άγνωστος και εξαιρετικά μεταδοτικός ιός. Ο ιός αυτός ονομάστηκε από τους επιστήμονες Covid-19 και ανακυρήχθηκε ως πανδημία από τον Παγκόσμιο Οργανισμό Υγείας τον Μάρτιο του 2020. Έναν χρόνο αργότερα ο αντίκτυπός του στον κόσμο είναι τεράστιος, έχοντας μολύνει πάνω από 89.000.000 ανθρώπους και προκαλώντας περίπου 2.000.000 θανάτους.

Σε αυτήν την εργασία θα παρουσιαστεί μια σφαιρική μελέτη της πανδημίας Covid-19 κάνοντας χρήση αλγορίθμων και μεθόδων εξόρυξης δεδομένων. Πιο συγκεκριμένα, εισαγωγικά θα παρουσιαστεί μια στατιστική μελέτη των επιδημιολογικών δεδομένων αρχικά παγκόσμια και, στη συνέχεια, με έμφαση στην Ευρώπη. Έπειτα θα αξιοποιηθούν οι δυνατότητες των αλγορίθμων πρόβλεψης τόσο για να προβλεφθεί η εξέλιξη των κρουσμάτων με χρήση αλγορίθμων παλινδρόμησης όσο και για να προβλεφθεί η πορεία της υγείας ορισμένων κατηγοριών ασθενών με τη βοήθεια αλγορίθμων κατηγοριοποίησης. Τέλος θα μελετηθούν δεδομένα σε μορφή κειμένου (ειδησεογραφικά άρθρα και tweets) με χρήση αλγορίθμων εξόρυξης κειμένου με σκοπό την ανάδειξη θεματικών ενοτήτων της δημόσιας συζήτησης για την πανδημία και ανάλυσης συναισθημάτων. Η υλοποίηση του προγραμματιστικού μέρους της εργασίας έγινε σε γλώσσα Python.

Λέξεις Κλειδιά:

Εξόρυξη Δεδομένων, Μηχανική Μάθηση, Covid-19, Εξόρυξη Κειμένου, Python

Abstract

In late 2019 a new, unknown and extremely contagious virus emerged in the city of Wuhan, China. The virus was named by the scientists as Covid-19 and was declared a pandemic by the World Health Organisation in March 2020. One year later the pandemic has a huge impact in the world, having infected more than 89.000.000 people, while being responsible for about 2.000.000 deaths.

In this diploma thesis an overall study of Covid-19's medical and sociological data will be presented using data mining algorithms and techniques. Specifically, at first, a statistical study of the epidemiological data will be introduced, regarding global and European data. Subsequently we will make use of predictive algorithms to forecast the growth of confirmed cases using regression algorithms and predict the recovery/positivity of individual patients using classification algorithms. Finally, textual data (news articles and tweets) regarding Covid-19 will be analysed with the help of text mining methods. The aim is to discover sub-topics in the texts and perform sentiment analysis. The implementation of the programming part for the above analysis was done in Python language

Keywords:

Data mining, Machine Learning, Covid-19, Text Mining, Python

Πίνακας περιεχομένων

Κατάλογος σχημάτων

Συντομογραφίες

βλ.	βλέπε
κτλ.	και τα λοιπά
κ.ο.κ	και ούτω καθεξής
κ.α.	και άλλα
ΑΚΣ	Ανάλυση Κύριων Συνιστωσών
ΜΕΘ	Μονάδα Εντατικής Θεραπείας
PCA	Principal Component Analysis
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
NLP	Natural Language Procesing
TFIDF	Term Frequence Inverse Document Frequence
KNN	K-Nearest Neighbors
SVM	Support Vector Machines
EDA	Exploratory Data Analysis

Κεφάλαιο 1

Εισαγωγή

Με την ραγδαία ανάπτυξη της πληροφορικής, του διαδικτύου και των κοινωνικών δικτύων, προκύπτουν ολοένα και μεγαλύτεροι όγκοι δεδομένων. Τα δεδομένα αυτά στην ακατέργαστη (raw) μορφή τους δεν προσφέρουν ιδιαίτερα χρήσιμη πληροφορία στους επιστήμονες. Το επιστημονικό πεδίο της Εξόρυξης Δεδομένων (Data Mining) έχει ως αντικείμενο τον μετασχηματισμό των δεδομένων αυτών σε Γνώση, ανακαλύπτοντας - μη προφανή - χρήσιμα μοτίβα από δεδομένα μεγάλου όγκου, για αυτό και εναλλακτικά αποκαλείται "Εξόρυξη Γνώσης από Δεδομένα" [1]. Το πεδίο της Εξόρυξης Δεδομένων βρίσκει εφαρμογή σε διάφορους κλάδους, όπως στον κλάδο των επιχειρήσεων, των ιατρικών εφαρμογών, στις τηλεπικοινωνίες κ.α.

1.1 Αντικείμενο της διπλωματικής

Η παρούσα διπλωματική εργασία έχει ως σκοπό να χρησιμοποιήσει αλγορίθμους και τεχνικές της Εξόρυξης Δεδομένων για να μελετήσει τόσο ιατρικές όσο και κοινωνιολογικές πτυχές της πανδημίας Covid-19 που βρίσκεται ακόμα σε πλήρη εξέλιξη. Στόχος είναι να παρουσιάσει στον αναγνώστη μια ολοκληρωμένη μελέτη που θα περιλαμβάνει παρουσίαση των επιδημιολογικών δεδομένων σε διάφορες χώρες, προβλέψεις για την εξέλιξή τους,ανάλυση ιατρικών δεδομένων ασθενών καθώς και ανάλυση ειδήσεων και tweets σχετικών με τον Covid-19, υπό το πρίσμα της Εξόρυξης Δεδομένων.

Η Covid-19 είναι μια λοίμωξη του αναπνευστικού συστήματος που προκαλείται από τον ιο SARS-Cov-2 ο οποίος ανήκει στην οικογένεια τον κορονοϊών και πρωτοεμφανίστηκε στα τέλη του Δεκέμβρίου 2019 στην πόλη Wuhan της Κίνας [2]. Έναν χρόνο αργότερα έχει

προκαλέσει περίπου 89.000.000 μολύνσεις παγκοσμίως και θεωρείται υπεύθυνη για περίπου 1.900.000 θανάτους με βάση τα στοιχεία της ιστοσελίδας Worldometer (τέλη Ιανουαρίου 2021), ενώ έχει προκαλέσει μεγάλο αντίκτυπο στον τρόπο ζωής των ανθρώπων λόγω των σκληρών περιοριστικών μέτρων που έχουν επιβληθεί σε πάρα πολλές χώρες για την ανάσχεσή της. Οι περισσότεροι ασθενείς εμφανίζουν ήπια έως μέτρια συμπτώματα και ανακάμπτουν χωρίς ιδιαίτερη ιατρική βοήθεια. Δεν συμβαίνει το ίδιο όμως με τους ηλικιωμένους, τους πάσχοντες απο χρόνια νοσήματα του αναπνευστικού, τους καρδιοπαθείς και τους καρκινοπαθείς οι οποίοι κινδυνεύουν περισσότερο από σοβαρή λοίμωξη και θάνατο.Τα πιο κοινά συμπτώματα είναι ο πυρετός, ο ξηρός βήχας και η κούραση, ενώ συμπτώματα όπως δύσπνοια και πόνος στο στήθος χαρακτηρίζονται ως συμπτώματα σοβαρής νόσου [3].

Αρχικά, θα γίνει προσπάθεια να περιγραφεί η επιδημιολογική κατάσταση που επικρατούσε στα τέλη του Σεπτεμβρίου 2020 σε επίπεδο κρουσμάτων, θανάτων και θνησιμότητας μελετώντας την εξέλιξη της πανδημίας τόσο παγκόσμια όσο και σε συγκεκριμένες χώρες. Έπειτα, στο δεύτερο μέρος της εργασίας, θα μελετηθεί τόσο το κομμάτι της πρόβλεψης της πορείας των κρουσμάτων/θανάτων όσο και η πρόβλεψη της εξέλιξης της νόσου σε ασθενείς βάσει ιατρικών τους εξετάσεων με την βοήθεια αλγορίθμων εξόρυξης δεδομένων. Τέλος, στο τρίτο μέρος της εργασίας θα εξεταστούν σύνολα δεδομένων που αφορούν ειδησεογραφικά άρθρα και tweets, σε μια προσπάθεια να εξεταστεί η επιρροή της πανδημίας στην δημόσια σφαίρα των ΜΜΕ και των κοινωνικών δικτύων, κάνοντας χρήση τεχνικών εξόρυξης κειμένου.

1.2 Συνεισφορά της Εργασίας

Η συνεισφορά της διπλωματικής συνοψίζεται στα παρακάτω:

1. Χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python καθώς και πολλές βιβλιοθήκες της γλώσσας για την υλοποίηση των πειράματων της εργασίας.

2. Πραγματοποιήθηκε σε βάθος ανάλυση των επιδημιολογικών δεδομένων σε διάφορες χώρες και μελετήθηκαν συσχετίσεις που αφορούν τα κρούσματα με δημογραφικές και ιατρικές μεταβλητές.

3. Υλοποιήθηκαν τρεις αλγόριθμοι παλινδρόμησης για την πρόβλεψη της εξέλιξης των κρουσμάτων καθώς και πέντε αλγόριθμοι ταξινόμησης για την πρόγνωση του αποτε-

λέσματος των ασθενών.

4. Αξιολογήθηκε η επίδοση των αλγορίθμων και διαπιστώθηκε ότι καλύτερη απόδοση είχε ο αλγόριθμος της πολυωνυμικής παρεμβολής για τις προβλέψεις συνεχούς μεταβλητής. Ο αλγόριθμος Adaboost σημείωσε την καλύτερη απόδοση στα δεδομένα με τις κλινικές πληροφορίες των ασθενών, ενώ ο αλγόριθμος Τυχαίου Δάσους πέτυχε την καλύτερη απόδοση στην πρόβλεψη της κατάληξης της υγείας του ασθενούς.

5. Μελετήθηκαν μέθοδοι επεξεργασίας φυσικής γλώσσας και εξόρυξης κειμένου ώστε να βγουν συμπεράσματα από κείμενα που αφορούν την πανδημία.

1.3 Οργάνωση του τόμου

Η εργασία οργανώνεται σε κεφάλαια. Στο Κεφάλαιο 2 παρουσιάζονται άρθρα συγγενικά με το αντικείμενο της διπλωματικής εργασίας. Στο Κεφάλαιο 3 γίνεται αναφορά στο θεωρητικό υπόβαθρο των αλγορίθμων και μεθόδων που χρησιμοποιούνται στην παρούσα εργασία. Το Κεφάλαιο 4 αναφέρεται σε μια στατιστική μελέτη και εξερεύνηση των επιδημιολογικών δεδομένων με έμφαση στις Ευρωπαϊκές χώρες, παρουσιάζοντας διαγράμματα που αφορούν την εξέλιξη των κρουσμάτων/θανάτων, της θνησιμότητας κτλ. στον αναγνώστη. Εξετάζονται επίσης συσχετίσεις (correlations) μεταξύ των προαναφερθέντων μεταβλητών και ιατρικών/δημογραφικών μεταβλητών κάθε χώρας. Στα Κεφάλαια 5 και 6 παρουσιάζονται τα αποτελέσματα των μελλοντικών προβλέψεων για την εξέλιξη των κρουσμάτων και πρόγνωσης θετικότητας/αποτελέσματος της νόσου σε ασθενείς αντίστοιχα. Πιο συγκεκριμένα, στο Κεφάλαιο 5 γίνεται χρήση αλγορίθμων παλινδρόμησης (regression) για την πρόβλεψη των κρουσμάτων στο μέλλον ενώ στο Κεφάλαιο 6 μελετώνται σύνολα δεδομένων με εξειδικευμένες ιατρικές πληροφορίες (ιατρικές εξετάσεις) με στόχο να προβλεφθεί το αν ο ασθενής είναι θετικός στον Covid-19 και το αν θα επιβιώσει ή όχι από την ασθένεια. Στο επόμενο Κεφάλαιο (7) γίνεται χρήση τεχνικών εξόρυξης κειμένου σε σύνολα δεδομένων τα οποία περιλαμβάνουν ειδησεογραφικά άρθρα και tweets σχετικά με τον κορονοϊό. Με χρήση αλγορίθμων συσταδοποίησης (clustering) και μεθόδων ανάλυσης συναισθημάτων (sentiment analysis) γίνεται προσπάθεια να χωριστούν τα άρθρα/tweets σε θεματικές ενότητες και να συσχετιστούν με συναισθήματα. Στο Κεφάλαιο 8 παρατίθενται οι τεχνικές λεπτομέρειες της υλοποιήσης των πειραμάτων (προγραμματιστικό μέρος).Τέλος, το Κεφάλαιο 9 αναφέρεται στα συμπεράσματα που προέκυψαν από την εκπόνηση της διπλωματικής εργασίας και σε

ιδέες για την επέκτασή της. Στο Παράρτημα υπάρχουν σύνδεσμοι για τον κώδικα και τα δεδομένα που χρησιμοποιήθηκαν.

Κεφάλαιο 2

Συναφείς εργασίες

Σε αυτό το κεφάλαιο αναφέρονται συνοπτικά άρθρα και έρευνες που παρουσιάζουν συνάφεια με το αντικείμενο αυτής της διπλωματικής εργασίας και βοήθησαν στην υλοποίησή της. Οι εργασίες που παρουσιάζονται χωρίζονται σε τρείς κατηγορίες, σε αντιστοιχία με τα κεφάλαια της εργασίας οι οποίες είναι: Προβλέψεις, Πρόγνωση Ασθενούς και Επεξεργασία Φυσικής Γλώσσας.

2.1 Προβλέψεις

Η δυνατότητα να προβλέψουμε την πορεία μιας αριθμητικής μεταβλητής είναι εξαιρετικά σημαντική, ιδιαίτερα όταν αυτή η μεταβλητή αντικατοπτρίζει τα κρούσματα μιας τόσο μεταδοτικής ασθένειας όσο η Covid-19. Στην μάχη για την ανάσχεση της πανδημίας τα μοντέλα προβλέψεων παίζουν πολύ σημαντικό ρόλο.

Στο άρθρο των Cooper, Mondal και Antonopoulos [4] χρησιμοποιείται το επιδημιολογικό μοντέλο SIR για να μοντελοποιηθεί η εξέλιξη της Covid-19 σε διάφορες χώρες.Οι Ardabilie et al. στο άρθρο τους [5], παρουσιάζουν μια σειρά μοντέλων μηχανικής μάθησης με σκοπό την πρόβλεψη της εξέλιξης των κρουσμάτων της Covid-19, ως εναλλακτική στο επιδημιολογικό μοντέλο SIR, χρησιμοποιώντας Εξελικτικούς Αλγόριθμους (Evolutionary Algorithms) και μοντέλα Νευρωνικών Δίκτυων. Οι Wang et al. [6] χρησιμοποιούν ενα υβριδικό μοντέλο που συνδυάζει το απλό λογιστικό μοντέλο (Logistic Model) με το μοντέλο χρονοσειρών FBprophet της Facebook, με σκοπό να επιτευχθούν έγκυρες μακροπρόθεσμες προβλέψεις, ενώ οι Chaurasia και Pal [7] χρησιμοποιούν μια σειρά μοντέλων χρονοσειρών με σκοπό να προβλέψουν τα μελλοντικά κρούσματα σε παγκόσμιο επίπεδο.

2.2 Πρόγνωση Ασθενούς

Στην συγκεκριμένη ενότητα ο στόχος είναι να δημιουργηθούν συστήματα τα οποία θα μπορούν να προβλέπουν το αν ο ασθενής είναι σε κίνδυνο να νοσήσει βαριά ή να πεθάνει, βασισμένα σε δεδομένα όπως ιατρικές εξετάσεις του ασθενούς, υποκείμενα νοσήματα, φύλο, ηλικία κτλ. Παρουσιάζονται επίσης μελέτες που προσπαθούν να διακρίνουν το αν ο ασθενής είναι θετικός ή όχι στον Covid-19 με βάση κλινικές και απεικονιστικές εξετάσεις του ασθενούς. Τέτοια συστήματα πρόγνωσης είχαν ιδιαίτερη σημασία στα πρώτα στάδια της πανδημίας όπου τα ειδικα μοριακά τεστ (PCR) βρίσκονταν σε μεγάλη έλλειψη.

Οι Muhammad et al. στο άρθρο τους [8] χρησιμοποιούν κλινικά δεδομένα από ασθενείς και εφαρμόζουν κλασσικούς αλγόριθμους ταξινόμησης όπως Δέντρα Απόφασης, Naive Bayes, K-κοντινότεροι γείτονες κτλ. με σκοπό να προβλέψουν το αν ο ασθενείς θα ανακάμψει ή θα πεθάνει από την Covid-19. Συμπεραίνουν ότι ο αλγόριθμος των Δέντρων Απόφασης πετυχαίνει καλύτερα ποσοστά ακρίβειας από τους υπόλοιπους, ενώ σημειώνουν ότι η ηλικία είναι καθοριστικός παράγοντας για την ανάκαμψη του ασθενούς. Στο άρθρο [9] των Osi et al., γίνεται χρήση παρόμοιων αλγορίθμων κατηγοριοποίησης, ενώ δίνεται ιδιαίτερο βάρος στην συννοσηρότητα των ασθενών σε μια προσπάθεια να βρεθεί ποια υποκείμενα νοσήματα επιδεινώνουν την πρόγνωση του ασθενούς. Σε μια διαφορετική προσέγγιση οι Khanday et al. [10] χρησιμοποιούν έναν συνδυασμό Επεξεργασίας Φυσικής Γλώσσας και αλγόριθμων κατηγοριοποίησης σε μια προσπάθεια να αναπτύξουν ένα σύστημα που θα ανιχνεύει την παρουσία του Covid-19 στον ασθενή. Χρησιμοποιούν ως είσοδο στα μοντέλα τους κλινικές αναφορές που συνοδεύουν τις ιατρικές εξετάσεις των ασθενών. Αντίστοιχα οι Khuzani, Heidari και Shariati [11] δημιουργούν ένα σύστημα πρόβλεψης θετικότητας του ασθενούς που έχει ως είσοδο απεικονιστικές εξετάσεις (ακτινογραφίες και μαγνητικές θώρακος) και χρησιμοποιούν Νευρωνικά Δίκτυα για την κατηγοριοποίηση των εικόνων.

2.3 Επεξεργασία Φυσικής Γλώσσας

Σε αυτήν την ενότητα παρουσιάζονται εργασίες που μελετούν τον αντίκτυπο της πανδημίας στους πολίτες εστιάζοντας σε posts σε μέσα κοινωνικής δικτύωσης, ειδησεογραφικά άρθρα κτλ. Πρώτος στόχος αυτής της μελέτης είναι να βρεθούν ποια θέματα κυριαρχούν στην δημόσια συζήτηση γύρω από τον Covid-19 και με τι συναισθήματα σχετίζονται σε μια προσπάθεια κατανοήσης του πως η πανδημία επηρεάζει τις ζωές των ανθρώπων. Καθώς ο

τεράστιος όγκος των ειδήσεων κάνει δύσκολη την διάκριση μεταξύ πραγματικών και ψευδών ειδήσεων, ένας δεύτερος στόχος είναι, σε μία τόσο σημαντική συγκυρία όπως αυτή, να εντοπιστούν οι ψευδείς ειδήσεις που οδηγούν στην παραπληροφόρηση και εξαπάτηση των πολιτών.

Στο άρθρο [12] των Xue et al. εξετάζεται ένα σύνολο δεδομένων που αποτελείται από tweets. Γίνεται διάκριση θεματικών ενοτήτων (topic labeling) και ανάλυση συναισθημάτων των tweets του dataset. Στα ευρήματά τους επισημαίνεται το γεγονός ότι τα περισσότερα tweets κυριαρχούνται από το συναίσθημα του φόβου. Οι Zhang et al. [13] χρησιμοποιούν μεθόδους βαθιάς μάθησης (deep learning) για να ανιχνεύσουν τις τάσεις κατάθλιψης στο twitter και την συσχέτισή τους με την εξέλιξη της πανδημίας. Με βάση τα ευρήματά τους διαπιστώνουν, ότι όσο η πανδημία προχωρά, τόσο αυξάνονται τα μηνύματα στα κοινωνικά δίκτυα που υποδηλώνουν τάσεις κατάθλιψης και άλλων ψυχικών ασθενειών. Τέλος οι Patwa et al. [14] δημιουργούν έναν ταξινομητή ψευδών ειδήσεων με στόχο την αυτοματοποιημένη διάκριση των παραπλανητικών ειδήσεων.

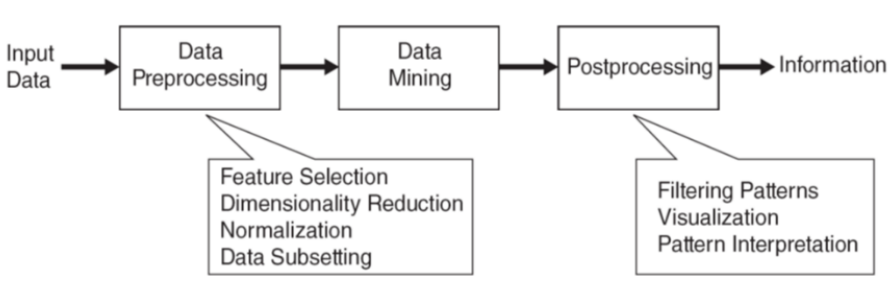
Κεφάλαιο 3

Θεωρητικό Υπόβαθρο

Σε αυτό το κεφάλαιο παρουσιάζεται το θεωρητικό υπόβαθρο της Εξόρυξης Δεδομένων και γίνεται μια συνοπτική παρουσίαση των αλγορίθμων και τεχνικών που χρησιμοποιούνται στην Διπλωματική Εργασία.

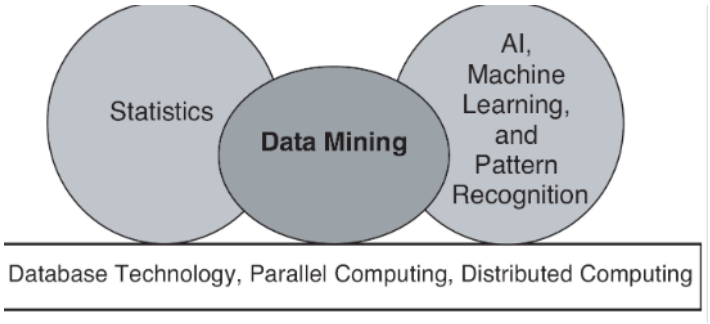
3.1 Εξόρυξη Δεδομένων

Οι Tan, Steinbach και Kumar στο βιβλίο τους [15], ορίζουν την Εξόρυξη Δεδομένων ως την "μη τετριμμένη εξαγωγή χρήσιμων και προηγούμενα άγνωστων πληροφοριών από δεδομένα" ή εναλλακτικά ως "την εξερεύνηση και ανάλυση με χρήση αυτοματοποιημένων ή ημι-αυτοματοποιημένων μέσων από μεγάλους όγκους δεδομένων με σκοπό την ανακάλυψη χρήσιμων και ερμηνεύσιμων μοτίβων". Στην εικόνα 3.1 παρουσιάζεται σχηματικά ο μετασχηματισμός των δεδομένων εισόδου σε χρήσιμη πληροφορία. Η Εξόρυξη Δεδομένων, ως



Σχήμα 3.1: Μετατροπή δεδομένων σε γνώση [15]

επιστημονικό πεδίο, αντλεί ιδέες από τα αντικείμενα της στατιστικής, της μηχανικής μάθησης, της τεχνητής νοημοσύνης και των βάσεων δεδομένων. Οι δύο βασικές κατηγορίες αλ-



Σχήμα 3.2: Εξόρυξη δεδομένων και συναφή επιστημονικά πεδία [15]

γορίθμων Εξόρυξης Δεδομένων είναι η επιβλεπόμενη και η μη-επιβλεπόμενη μάθηση [16]. Η κύρια διαφορά τους έγκειται στο γεγονός ότι στους αλγορίθμους επιβλεπόμενης μάθησης ένα κομμάτι του συνόλου δεδομένων, η μεταβλητή - στόχος (target variable), είναι ήδη γνωστή (labeled). Οι αλγόριθμοι εκπαιδεύονται στο γνωστό σύνολο δεδομένων (training dataset) και προσπαθούν να προβλέψουν την μεταβλητή-στόχο σε νέα, άγνωστα σύνολα δεδομένων. Χαρακτηριστικά παραδείγματα επιβλεπόμενης μάθησης είναι η Κατηγοριοποίηση (Classification) και η Παλινδρόμηση (Regression). Οι αλγόριθμοι επιβλεπόμενης μάθησης είναι αλγόριθμοι πρόβλεψης και οι εφαρμογές τους χρησιμοποιούνται στον ιατρικό κλάδο, στον τραπεζικό/χρηματιστηριακό, στις υπηρεσίες πρόβλεψεις καιρού κτλ. Στους αλγόριθμους κατηγοριοποίησης η μεταβλητή-στόχος είναι διακριτή και μπορεί να ανήκει σε δύο ή περισσότερες κλάσεις, ενώ στους αλγόριθμους παλινδρόμησης είναι συνεχής.

Στους αλγόριθμους μη-επιβλεπόμενης μάθησης δεν υπάρχει μεταβλητή-στόχος και σκοπός είναι η αναγνώριση μοτίβων στα δεδομένα. Οι αλγόριθμοι κατασκευάζουν μοντέλα χωρίς να γνωρίζουν από πριν το επιθυμητό αποτέλεσμα. Χαρακτηριστικά παραδείγματα είναι η Συσταδοποίηση (Clustering) και οι Κανόνες Συσχέτισης (Association Rules). Με την εφαρμογή αλγορίθμων μη-επιβλεπόμενης μάθησης τα δεδομένα μπορούν να μετατραπούν από unlabeled σε labeled και να χρησιμοποιηθούν ως είσοδος σε αλγορίθμους προβλέψεων. Χαρακτηριστικά παραδείγματα χρήσης μη-επιβλεπόμενης μάθησης είναι η συσταδοποίηση εγγράφων για την εύρεση παρόμοιων εγγράφων καθώς και η ανάλυση συσχετίσεων στις αγορές των πελατών ενός καταστήματος.

3.1.1 Σύνολο Εκπαίδευσης και Σύνολο Ελέγχου

Το σύνολο δεδομένων που χρησιμοποιείται ως είσοδος στον αλγόριθμο πρέπει να χωριστεί σε σύνολο εκπαίδευσης (training dataset) και σύνολο ελέγχου (test dataset). Το σύ-

νολο εκπαίδευσης, με γνωστές τις κλάσεις των μεταβλητών-στόχων, θα χρησιμοποιηθεί για να εκπαιδεύσει το μοντέλο, το οποίο στην συνέχεια δέχεται ως είσοδο το άγνωστο σύνολο (test dataset). Η συνήθης αναλογία για να χωριστεί το συνολο δεδομένων σε σύνολο εκπαίδευσης και ελέγχου είναι 80/20 χωρίς αυτό να είναι δεσμευτικό. Αναλόγως με τα ιδιαίτερα χαρακτηριστικά του συνόλου δεδομένων, διαφορετικοί χωρισμοί μπορούν να οδηγήσουν σε μεγαλύτερη ακρίβεια.

3.1.2 Υπερ-προσαρμογή και Υπό-προσαρμογή

Δύο κλασσικά προβλήματα στα μοντέλα μηχανικής μάθησης είναι αυτά της υποπροσαρμογής (underfitting) και υπερ-προσαρμογής (overfitting). Το πρόβλημα της υπερ-προσαρμογής εμφανίζεται όταν ένα μοντέλο έχει προσαρμοστεί υπερβολικά στο σύνολο εκπαίδευσης και δυσκολεύεται να γενικεύσει σε άγνωστα δεδομένα. Αντίθετα η υπο-προσαρμογή εμφανίζεται όταν τα δεδομένα του συνόλου εκπαίδευσης δεν επαρκούν για την σωστή εκπαίδευση του μοντέλου.

3.1.3 Η κατάρα των πολλών διαστάσεων

Όταν το σύνολο δεδομένων περιέχει πολύ μεγάλο αριθμό γνωρισμάτων, τότε ο χώρος αναζήτησης αυξάνεται και μαζί του αυξάνεται η πιθανότητα το μοντέλο να ανακαλύψει λανθασμένα πρότυπα τα οποία δεν ισχύουν. Αυτό είναι γνωστό ως “Η κατάρα των πολλών διαστάσεων” (curse of dimensionality) [15]. Παράλληλα, αυξάνονται οι απαιτήσεις τόσο σε υπολογιστική ισχύ όσο και σε χώρο για την εκπαίδευση του μοντέλου. Μια αρκετά αποτελεσματική αντιμετώπιση του προβλήματος είναι η επιλογή ορισμένων μόνο γνωρισμάτων του συνόλου δεδομένων. Στόχος είναι να επιλέξουμε τα γνωρίσματα εκείνα τα οποία είναι περισσότερο σημαντικά και να μειώσουμε τον θόρυβο. Γνωστές τεχνικές μείωσης διαστάσεων είναι η Ανάλυση Κύριων Συνιστωσών (Principal Components Analysis - PCA) και η Ανάλυση σε Ιδιάζουσες τιμές (Singular Value Decomposition - SVD).

3.2 Κατηγοριοποίηση

Η Κατηγοριοποίηση είναι η συχνότερα εφαρμοζόμενη τεχνική σε προβλήματα εξόρυξης δεδομένων. Χρησιμοποιεί ένα σύνολο δεδομένων όπου οι εγγραφές του είναι ήδη προκατηγοριοποιημένες για να εκπαιδεύσει μοντέλα τα οποία θα βρίσκονται σε θέση να ταξι-

νομήσουν νέα, άγνωστα δεδομένα. Μοντέλα κατηγοριοποίησης χρησιμοπούνται σε πάρα πολλά πρακτικά προβλήματα, όπως η διάκριση ανάμεσα σε επιθυμητά και ανεπιθύμητα (spam) μηνύματα ηλεκτρονικού ταχυδρομείου, στον ιατρικό τομέα, για παράδειγμα στην διάκριση μεταξύ καλοήθων ή κακοήθων όγκων από ιατρικές ή απεικονιστικές εξετάσεις, στην πρόληψη φυσικών καταστροφών κτλ. [17] [15]. Πιο συγκεκριμένα, η διαδικασία της κατηγοριοποίησης όριζεται ως:

Ορισμός 3.1 (Κατηγοριοποίηση)**.** *Δεδομένου ενός συνόλου εγγραφών: κάθε εγγραφή χαρακτηρίζεται από μια πλειάδα* (x, y) *όπου* x *είναι το σύνολο γνωρισμάτων και* y *είναι η κλάση. Στόχος είναι η δημιουργία ενός μοντέλου το οποίο θα είναι σε θέση να αντιστοιχήσει κάθε σύνολο γνωρισμάτων* x *σε μία κλάση* y*.*

Μερικοί από τους πιο γνωστούς αλγόριθμους κατηγοριοποίησης είναι οι παρακάτω:

- Δέντρα Απόφασης
- Bayesian Ταξινομητές
- Ταξινομητές Κοντινότερου Γείτονα
- Μηχανές Διανυσματικής Υποστήριξης
- Νευρωνικά Δίκτυα
- Μετα-αλγόριθμοι (ensemble) όπως Boosting, Τυχαία Δάση κτλ.

3.2.1 Μέτρα απόδοσης

Για να εκτιμήσουμε την απόδοση ενός μοντέλου κατηγοριοποίησης χρειάζεται να μετρήσουμε πόσες από τις εγγραφές του συνόλου ελέγχου κατηγοριοποιήθηκαν στην σωστή κλάση και πόσες κατηγοριοποιήθηκαν λάθος. Σε ένα πρόβλημα δυαδικής ταξινόμησης με δύο κλάσεις (θετική, αρνητική) οι εγγραφές που κατηγοριοποιήθηκαν σωστά ως θετικές ονομάζονται *Αληθώς Θετικές - True Positive (TP)* ενώ αυτές που κατηγοριοποίηθηκαν σωστά ως αρνητικές ονομάζονται *Αληθώς Αρνητικές - True Negative (TN)*. Αντίστοιχα οι εγγραφές που ταξινομούνται λανθασμένα λέγονται *Ψευδώς Θετικές - False Positive (FP)* και *Ψευδώς Αρνητικές - False Negative (FN)*. [18]

Τα αποτελέσματα τοποθετούνται σε έναν πίνακα ο οποίος ονομάζεται “Πίνακας Σύγχυσης” (Confusion Matrix). Η ανάλυση του πίνακα μας δίνει τα διάφορα μέτρα απόδοσης των αλγορίθμων. Στο σχήμα 3.3 παρουσιάζεται ένας τέτοιος πίνακας.

	Actual Positive Class	**Actual Negative Class**
Predicted Positive Class	True positive (*tp*)	False negative (*fn*)
Predicted Negative Class	False positive (*fp*)	True negative (*tn*)

Σχήμα 3.3: Παράδειγμα πίνακα σύγχυσης [16]

Accuracy

Το μέτρο της ακρίβειας (accuracy) δηλώνει τον αριθμό των εγγραφών που ταξινομήθηκαν σωστά προς τον αριθμό των συνολικών εγγραφών. Δηλαδη:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

. Το μέτρο της ακρίβειας είναι αυτό που χρησιμοποείται συχνότερα, ιδιαίτερα όταν η κατανομή κλάσεων των δεδομένων ειναι ισορροπημένη.

Precision

Το Precision αναφέρεται μόνο στην θετική κλάση του συνόλου δεδομένων και δείχνει τον αριθμό θετικών εγγραφών που ταξινομήθηκαν σωστά προς τον συνολικό αριθμό θετικών εγγραφών.

$$Prec = \frac{TP}{TP + FP}$$

Recall

Το Recall εκφράζει το πόσες θετικές εγγραφές ταξινομήθηκαν πραγματικά ως τέτοιες.

$$Rec = \frac{TP}{TP + FN}$$

Χρησιμοποειαι όταν μας ενδιαφέρει ιδαίτερα το να μην ταξινομήσουμε μια θετική εγγραφή ως αρνητική.

F1 Score

Το F1-Score είναι ο αρμονικός μέσος των Precision και Recall και εκφράζεται ως

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

Το F1-Score είναι το πιο αξιόπιστο μέτρο σε σύνολα δεδομένων που οι κλάσεις τους δεν ειναι ισορροπημένες.

3.3 Παλινδρόμηση

Η παλινδρόμηση (regression) είναι κομμάτι της επιβλεπόμενης μάθησης όπου η μεταβλητή-στόχος είναι συνεχής. Πιο συγκεκριμένα ορίζεται ως:

Ορισμός 3.2 (Παλινδρόμηση)**.** *Η ανάλυση παλινδρόμησης (regression analysis) είναι ένα στατιστικό εργαλείο που χρησιμοποιείται για να μοντελοποιήσει την σχέση μεταξύ μιας εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. [15]*

Μερικοί από τους πιο γνωστούς αλγόριθμους παλινδρόμησης είναι οι παρακατω:

- Γραμμική Παλινδρόμηση
- Πολυωνυμική Παλινδρόμηση
- Παλινδρόμηση Διανυσματικής Υποστήριξης (Support Vector Regression)
- Παλινδρόμηση με χρήση Δεντρων Απόφασης
- Λογιστική Παλινδρόμηση

3.3.1 Μέτρα Απόδοσης

Τα πιο συχνά χρησιμοποιούμενα μέτρα απόδοσης για τους αλγόριθμους παλινδρόμησης είναι τα εξής:

Μέσο Τετραγωνικό Σφάλμα

Το Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error - MSE) υπολογίζει το άθροισμα της διαφοράς των τετραγώνων μεταξύ των πραγματικών τιμών της μεταβλητής-στόχου και αυτών που προέβλεψε το μοντέλο παλινδρόμησης

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i')^2$$

Μέσο Απόλυτο Σφάλμα

Το Μέσο Απόλυτο Σφάλμα (Mean Absolute Error - MAE) υπολογίζει το άθροισμα της διαφοράς των πραγματικών τιμών της μεταβλητής και αυτών που προέβλεψε το μοντέλο

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(Y_i - Y_i')|$$

3.4 Συσταδοποίηση

Η συσταδοποίηση (clustering) ανήκει στις μεθόδους μη-επιβλεπόμενης μάθησης. Το σύνολο δεδομένων δεν απαιτείται να ειναι προ-κατηγοριοποιημένο ενώ σκοπός είναι να χωριστούν τα δεδομένα σε συστάδες, τέτοιες ώστε τα αντικείμενα κάθε συστάδας να είναι σχετικά μεταξύ τους και μη σχετιζόμενα με τα αντικείμενα των άλλων συστάδων. Πρακτικά αυτό σημαίνει ότι στόχος είναι να ελαχιστοποιηθούν οι αποστάσεις των αντικειμένων εντος των συστάδων και να μεγιστοποιηθούν οι αποστάσεις των συστάδων μεταξύ τους [15]. Ο αριθμός των συστάδων δεν είναι γνωστός εκ των προτέρων και συνήθως προκύπτει με δοκιμές, έχοντας κριτήριο το να μειωθεί το σφάλμα. Κύριες μέθοδοι συσταδοποίησης είναι οι:

- Διαμεριστική Συσταδοποίηση με πιο γνωστό αλγόριθμο τον K-means
- Ιεραρχική Συσταδοποίηση
- Συσταδοποίηση με βάση την πυκνότητα με πιο χαρακτηριστικό αλγόριθμο τον DBSCAN

3.5 Εξόρυξη Κειμένου

Η εξόρυξη κειμένου (text mining) είναι μια υποκατηγορία της εξόρυξης δεδομένων όπου τα δεδομένα δεν βρίσκονται σε κλασσική μορφή πίνακα αλλά βρίσκονται σε μορφή κειμένου. Χρησιμοποιεί μεθόδους Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing - NLP) για να μετατρέψει το κείμενο σε δομημένα δεδομένα κατάλληλα για ανάλυση [19]. Η εξόρυξη κειμένου είναι ένα σχετικά νεο πεδίο της εξόρυξης δεδομένων που γνωρίζει ιδιαίτερη άνθηση τα τελευταία χρόνια ιδιαίτερα με την έκρηξη του διαδικτύου και των κοινωνικών δικτύων όπου ο όγκος των δεδομένων σε μορφή κειμένου αυξάνεται διαρκώς. Η φυσική γλώσσα ενώ είναι πολύ εύκολα κατανοητή στον άνθρωπο, όμως δυσκολεύει αρκετά τους υπολογιστές για αυτό χρησιμοποιούνται μια σειρά από μέθοδοι επεξεργασίας ώστε τα δεδομένα να μετατραπούν σε μορφή κατανοητή από τους υπολογιστές. Πολλές τεχνικές της επεξεργασία φυσικής γλώσσας αντλούν την προέλευσή τους από την επιστήμη της Γλωσσολογίας [20]. Μετά την προπεξεργασία και την διανυσματοποίηση μπορούν να εφαρμοστούν οι κλασσικοί αλγόριθμοι εξόρυξης δεδομένουν για να εφαρμοστούν τεχνικές συσταδοποίησης, κατηγοριοποιήσης, ανάλυσης συναισθημάτων κτλ.

3.5.1 Προεπεξεργασία Κειμένου

Μερικές από τις πιο συνηθισμένες τεχνικές στην προεπεξεργασία του κειμένου είναι οι παρακάτω [21] [20]:

- Μετατροπή όλων των χαρακτήρων σε πεζούς (lowercase).
- "Καθάρισμα" του κειμένου από links, hashtags, HTML tags κτλ.
- Αφαίρεση ειδικών χαρακτήρων.
- Χωρισμός σε λεκτικές μονάδες (tokenization): Είναι η διαδικασία κατά την οποία το κείμενο χωρίζεται σε λεκτικές μονάδες (tokens). Είναι απαραίτητη για να εφαρμοστούν όλες οι παρακάτω τεχνικές.
- Αφαίρεση stop-words: Ως stop-words ορίζονται οι λέξεις εκείνες οι οποίες είναι πολύ συχνές και συνήθως δεν προσφέρουν επιπλέον πληροφορία στο νόημα του κειμένου. Παράδειγματα stop-words στην Αγγλική γλώσσα είναι οι λέξεις "the, it, at, which, I, we, about".
- Λημματοποίηση: Η Λημματοποίηση (lemmatization) είναι η διαδικασία στην οποία κάθε λέξη αντιστοιχίζεται στην απλούστερη μορφή της (λήμμα) αναλόγως με το τι μέρος του λόγου είναι (ουσιαστικό, ρήμα, επίθετο κτλ). Παραδείγματος χάριν μετά την λημματοποίηση οι λέξεις troubles,troubling,troubled μετασχηματίζονται σε trouble.
- Αποκοπή καταλήξεων: Η αποκοπή καταλήξεων (stemming) αφαιρεί όλες τις καταλήξεις απο συγγενείς λέξεις και τις μετασχηματίζει σε μια πηγαία μορφή (stem) η οποία πολλές φορές δεν είναι πραγματική λέξη. Για παράδειγμα οι λέξεις troubles, troubling, troubled μετά το stemming μετασχηματίζονται σε troubl. Θεωρείται πιο "βίαιη" μέθοδος από την λημματοποίηση και μπορεί να οδηγήσει σε αλλοίωση του νοήματος των λέξεων. Για παράδειγμα η λέξη "caring" μετά την αποκοπή κατάληξης θα γίνει "car" πράγμα που θα αλλοιώσει τελείως την πρότερη σημασία της.

Όλες οι παραπάνω τεχνικές αποσκοπούν στο να μειωθεί ο αριθμός των λέξεων του εγγράφου και κατά συνέπεια οι διαστάσεις του συνόλου δεδομένων, χωρίς να χαθεί σημαντική πληροφορία.

3.5.2 Διανυσματοποίηση

Μετά την προεπεξεργασία του κειμένου τα δεδομένα πρέπει να μετασχηματιστούν σε μορφή τέτοια ώστε να μπορούν να δοθούν ως είσοδος σε αλγορίθμους. Να μετατραπούν δηλαδή από την αρχική μορφή του κειμένου σε αριθμητική μορφή. Μια τέτοια μετατροπή επιτυγχάνεται με την διαδικασία της διανυσματοποίησης (vectorization). Οι κυριότερες μέθοδοι διανυσματοποίησης περιγράφονται συνοπτικά παρακάτω [22].

Bag Of Words

Αποτελεί την πιο απλή μέθοδο διανυσματοποίησης. Η συγκεκριμένη μέθοδος αγνοεί την σειρά των λέξεων στο κείμενο και μας πληροφορεί αν μια λέξη είναι παρούσα στο έγγραφο. Όλες οι λέξεις από όλα τα έγγραφα της συλλογής καταγράφονται σε ένα λεξικό (corpus). Έπειτα για κάθε έγγραφο δημιουργείται ένα δυαδικό διάνυσμα που κάθε τιμή του (0/1) αντιπροσωπεύει κάθε λέξη του λεξικού και έχει τιμή 1 αν η λέξη υπάρχει στο έγγραφο και τιμή 0 αν δεν υπάρχει. Κάθε τιμή του διανύσματος αποτελεί και ένα γνώρισμα (attribute). Αυτά τα διανύσματα τοποθετούνται σε έναν αραιό πίνακα (sparse matrix) και αποτελούν την μετάφραση των εγγράφων σε αριθμητική τιμή.

TF-IDF

Η μέθοδος TF-IDF είναι και η πιο συχνά χρησιμοποιούμενη μέθοδος σήμερα για την διανυσματοποίηση κειμένων. Ο όρος TF σημαίνει Term Frequence (συχνότητα λέξης) και ο όρος IDF σημαίνει Inverse Document Frequence (αντίστροφη συχνότητα εγγράφου). Η διαφορά του με την προηγούμενη μέθοδο είναι ότι μας δίνει την σημασία/βάρος μια λέξης στο νοήμα του κειμένου. Η συχνότητα λέξης (TF) προκύπτει από τον διαίρεση του αριθμού των εμφανίσεων της λέξης στο έγγραφο προς τον συνολικό αριθμό των λέξεων του εγγράφου.

$$TF = \frac{\text{Αριθμός εμφανίσεων της λέξης στο έγγραφο}}{\text{Συνολικός αριθμός λέξεων στο έγγραφο}}$$

Η αντίστροφη συχνότητα εγγράφου (IDF) μας δείχνει την σημασία της λέξης και βασίζεται στην ιδέα ότι οι πιο σπάνιες λέξεις συχνά προσφέρουν μεγαλύτερη πληροφορία. Προκύπτει από τον δεκάδικό λογάριθμο της διαίρεσης του αριθμού των εγγράφων προς τον αριθμό των εγγράφων που η συγκεκριμένη λέξη εμφανίζεται.

$$IDF = log_{10} * \frac{\text{Συνολικός αριθμός εγγράφων}}{\text{Αριθμός εγγράφων όπου η λέξη εμφανίζεται}}$$

Για σπάνιες λέξεις η τιμή του IDF είναι μεγάλη ενώ για συχνές λέξεις είναι μικρή. Το τελικό νούμερο TFIDF κάθε λέξης προκύπτει από τον πολλαπλασιασμό του όρου TF μέ τον όρο IDF. Το TFIDF μας δείχνει πόσο σχετική είναι κάθε λέξη με το νόημα του κειμένου. Έτσι παρόμοια κείμενα θα παρουσιάζουν παρόμοια διανύσματα TFIDF.

3.6 Περιγραφή Αλγορίθμων

3.6.1 Δέντρα Απόφασης

Ο αλγόριθμος των Δέντρων Απόφασης (Decision Trees) είναι από τους πρώτους αλγορίθμους μηχανικής μάθησης/εξόρυξης δεδομένων και παρουσιάστηκε πρώτη φορά το 1986 [23]. Είναι ένας από το τους πιο απλούς αλλά ευρέως διαδομένους αλγορίθμους. Τα πλεονεκτήματα του είναι ότι μπορεί να εφαρμοστεί τόσο σε προβλήματα κατηγοριοποίησης όσο και σε προβλήματα παλινδρόμησης, η ικανότητα του να χειρίζεται διάφορους τύπους δεδομένων (ονομαστικά,αριθμητικά κτλ.). Επίσης απαιτεί μικρή υπολογιστική ισχύ συγκριτικά με άλλους αλγόριθμους αλλά και είναι κατάλληλος για τον χειρισμό μεγάλου όγκου δεδομένων. Τα Δέντρα Απόφασης είναι ιεραρχικές δομές και αποτελούνται από κόμβους και ακμές.

- Ο κόμβος-ρίζα του δέντρου δεν διαθέτει εισερχόμενες ακμές παρά μόνο εξερχόμενες.
- Οι εσωτερικοί κόμβοι διαθέτουν μία εισερχόμενη ακμή και δύο ή περισσότερες εξερχόμενες.
- Οι τερματικοί κόμβοι - φύλλα του δέντρου διαθέτουν μόνο μια εισερχόμενη ακμή.

Κάθε εσωτερικός κόμβος αναπαριστά έναν υπολογισμό που αναφέρεται σε κάποιο στοιχείο του συνόλου γνωρισμάτων ενώ κάθε φύλλο αναπαριστά μια κλάση. Ο αλγόριθμος κατασκευής του δέντρου απόφασης [16] παρατίθεται στην συνέχεια. Το γράμμα D αναφέρεται στο σύνολο δεδομένων, το στο δέντρο ενώ τα x, y συμβολίζουν τα γνωρίσματα και την μεταβλητή στόχο αντίστοιχα. Προκειμένου να αποφασιστεί το γνώρισμα βάσει του οποίου θα γίνει ο διαχωρισμός σε κάθε κόμβο χρησιμοποιούνται τα μέτρα καθαρότητας (node impurity measures) όπως η Εντροπία, το Gini και το Misclassification Error [15]. Στους επόμενους τύπους το $p_i(t)$ συμβολίζει την συχνότητα της κλάσης i στον κόμβο t, ενώ το c συμβολίζει τον συνολικό αριθμό των κλάσεων.

Algorithm 1.1: Decision Tree

Protocol *DT Inducer* (D, x, y)

1. $T = Tree\ Growing\ (D, x, y)$
2. Return Tree Pruning (D, T)

Method Tree Growing (D, x, y)

1. Create a tree T
2. **if** at least one of the Stopping Criteria is satisfied **then**;
3. label the root node as a leaf with the most frequent value of y in D as the correct class.
4. **else**;
5. Establish a discrete function f(x) of the input variable so that splitting D according to the functions outcomes produces the best splitting metric
6. **if** the best metric is greater or equal to the threshold **then**;
7. Mark the root node in T as f(x)
8. **for** each outcome of f(x) at the node **do**;
9. $Subtree = Tree\ Growing\ (\delta_{f(x)=t_1}, D, x, y)$
10. Connect the root of T to Subtree and label the edge t_1
11. **end for**
12. **else**
13. Label the root node T for a leaf with the frequent value of y in D as the assigned class
14. **end if**
15. **end if**
16. Return T

Protocol *Tree Pruning* (D, T, y)

1. **repeat**
2. Select a node t in T to maximally improve pruning evaluation procedure
3. **if** $t \neq 0$ **then**;
4. $T = pruned\ (T, t)$
5. **end if**
6. **until** $t = 0$
7. Return T

Σχήμα 3.4: Αλγόριθμος κατασκευής Δέντρου Απόφασης [16]

$$Entropy = -\sum_{i=1}^{c-1} log_2 p_i(t)$$

$$Gini = 1 - \sum_{i=1}^{c-1} p_i(t)^2$$

$$Misclassification\_error = 1 - max[p_i(t)]$$

Η καθαρότητα των κόμβων υπολογίζεται με ένα από τα παραπάνω μέτρα πριν τον διαχωρισμό (P) και μετά (M) και τελικά επιλέγεται για τον διαχωρισμό το γνώρισμα με το μεγαλύτερο κέρδος (Gain).

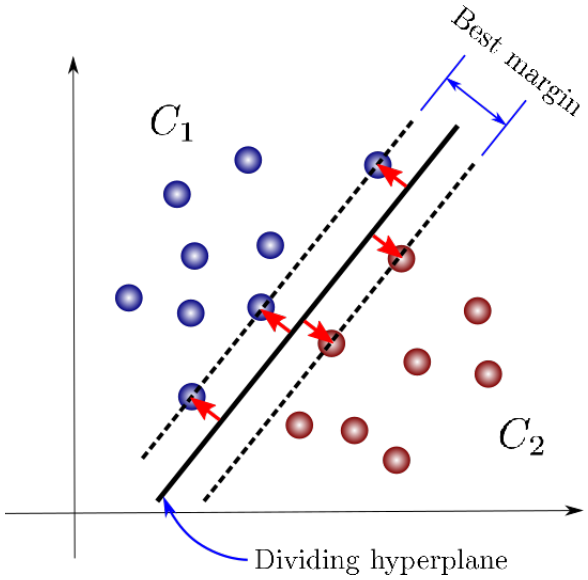
$$Gain = P - M$$

Για την αποφυγή του φαινομένου της υπερ-προσαρμογής μπορεί να εφαρμοστεί κλάδεμα (pruning) στο δέντρο. Με βάση τον κανόνα του "Ocam's Razor" [24] όταν δύο μοντέλα

αποδίδουν παρόμοια προτιμάται το απλούστερο μοντέλο. Το κλάδεμα του δέντρου μπορεί να γίνει κατά την δημιουργία του (pre-pruning) σταματώντας τον αλγόριθμο προτού ολοκληρώσει το δέντρο ή μετά την δημιουργία του (post-pruning).

3.6.2 Μηχανές Διανυσματικής Υποστήριξης

Ο αλγόριθμος των Μηχανών Διανυσματικής Υποστήριξης (Support Vector Machines-SVM) χρησιμοποιείται εκτός από προβλήματα κατηγοριοποίησης και σε προβλήματα παλινδρόμησης. Ο SVM χωρίζει τα δεδομένα σε κλάσεις ορίζοντας ενα υπερ-πεδίο (hyperplane) το οποίο στον δισδιάστατο χώρο είναι μια ευθεία μεταξύ των κλάσεων προσπαθώντας να μεγιστοποιήσει τα περιθώρια μεταξύ τους. Στο σχήμα 3.5 βλέπουμε μια σχηματική αναπαράσταση των παραπάνω. Υπάρχουν πολλοί τρόποι να χωριστούν οι 2 κλάσεις και το πλεο-



Σχήμα 3.5: Το βέλτιστο υπερπεδίο και τα διανύσματα υποστήριξης [25]

νέκτημα του SVM έγκειται στην εύρεση της βέλτιστης γραμμής που χωρίζει τις κλάσεις. Οι δύο γραμμές που που ορίζουν το υπερπεδίο ονομάζονται συνοριακές γραμμές, ενώ η γραμμή στο μέσον τους ονομάζεται βέλτιστο υπερπεδιο (optimal hyperplane).Τα διανύσματα υποστήριξης είναι τα σημεία εκείνα τα οποία βρίσκονται κοντινότερα στο πεδίο, ισαπέχουν από το βέλτιστο υπερπεδίο και αν μετακινηθούν, μετακινείται μαζί τους και το υπερπεδίο. Παρακάτω παρουσιάζεται ο αλγόριθμος των SVM σε ψευδοκώδικα. Αναλυτικότερες πληροφορίες βρίσκονται στα άρθρα [26] και [25].

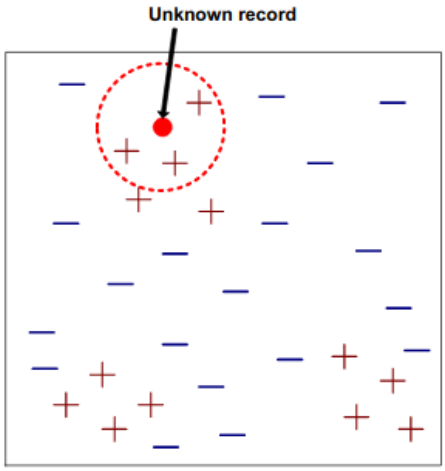
Algorithm 1.3: Support Vector Machine

Input: S, λ, T, k
Initialize: Choose w_1 such that $\|w_1\| \leq \sqrt{\lambda}$
FOR $t = 1,2 \ldots ,T$
Select $A_t \subseteq S$, *in which* $|A_t| = k$
Set $A_t^+ = \{(x, y) \in A_t : y\,(w_t, x) < 1\}$
Set $\delta_t = \frac{1}{\lambda t}$
Set $w_{t+0.5} = (1 - \delta_t \lambda)\, w_t + \frac{\delta_t}{k} \sum_{(x,y)\in A_t^+} yx$
Set $w_{t+1} = \left\{1, \frac{1/\sqrt{\lambda}}{\|w_{t+0.5}\|}\right\} w_{t+0.5}$
Output: w_{T+1}

Σχήμα 3.6: Ψευδοκώδικας αλγορίθμου SVM [16]

3.6.3 Κ-κοντινότεροι Γείτονες

Ο αλγόριθμος των Κ-κοντινότερων Γειτόνων (K-nearest Neightbors - KNN) [27] είναι ενας αλγόριθμος ταξινόμησης ο οποίος χρησιμοποιεί μια συνάρτηση απόστασης για να κατηγοριοποιήσει ένα άγνωστο σημείο ανάλογα με την πλειοψηφούσα κλάση των Κ κοντινότερων γειτόνων του, όπως βλέπουμε στο σχήμα 3.7. Η τιμή του Κ είναι ιδιαίτερα κρίσιμη



Σχήμα 3.7: Κ-κοντινότεροι γείτονες[15]

καθώς αν επιλεγεί πολύ μικρή τιμή τότε ο αλγόριθμος είναι επιρρεπής στον θόρυβο ενώ αν επιλεγεί πολύ μεγάλη, η γειτονιά που ορίζεται γύρω από το άγνωστο σημείο μπορεί να περιλαμβάνει σημεία από άλλες κλάσεις. Οι πιο συνηθισμένες συναρτήσεις απόστασης είναι η Ευκλείδια απόσταση, η απόσταση Manhattan και η απόσταση Minkowski [15].

$$Euclidean\_distance = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

$$Manhattan\_distance = \sum_{i=1}^{k} |x_i - y_i|$$

$$Minkowski\_distance = (\sum_{i=1}^{k} (|x_i - y_i|)^q)^{\frac{1}{q}})$$

Στο σχημα 3.8 παρουσιάζεται ο ψευδοκώδικας του αλγορίθμου.Περισσότερες λεπτομέρειες για τον αλγόριθμο των K-κοντινότερων γειτόνων στο άρθρο [28].

Algorithm 1: Brute force kNN Algorithm

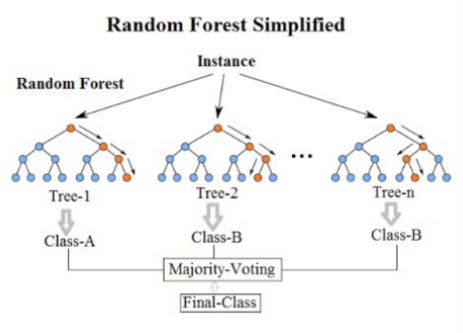
Input : $\mathcal{Q}$, a set query points and $\mathcal{R}$, a set of reference point;
Output: A list of k *reference* points for each *query* point;

```
foreach query point q ∈ Q do
    compute distances between q and all r ∈ R;
    sort the computed distances;
    select k-nearest reference points corresponding to k smallest distances;
```

Σχήμα 3.8: Ψευδοκώδικας αλγορίθμου KNN [28]

3.6.4 Τυχαίο Δάσος

Ο αλγόριθμος του Τυχαίου Δάσους (Random Forest) είναι ένας αλγόριθμος που δεν βασίζεται σε έναν ταξινομητή αλλά σε μια ομάδα ταξινομητών (ensemble). Στην προκειμένη περίπτωση το Δάσος αποτελεί την ομάδα ταξινομητών η οποία απαρτίζεται από πολλά μεμονωμένα Δέντρα Αποφάσεων τα οποία δεν έχουν κλαδευτεί (unpruned). Για την εκπαίδευση κάθε ξεχωριστού δέντρου χρησιμοποιείται ένα τυχαίο τμήμα των δεδομένων (bootstrapping) ενώ ο διαχωρισμός σε κάθε κόμβο των δέντρων γίνεται με τυχαίο τρόπο. Έτσι τα δέντρα που προκύπτουν είναι μη σχετισμένα μεταξύ τους (uncorrelated). Η τελική απόφαση παίρνεται με βάση την πλειοψηφία των αποφάσεων των δέντρων σε προβλήματα κατηγοριοποίησης ή με βάση τον μέσο όρο των δέντρων σε προβλήματα παλινδρόμησης [15] [29]. Στην εικόνα 3.9 βλέπουμε μια σχηματική αναπαράσταση ενός τυχαίου δάσους. Ο ακριβής αλγόριθμος παρουσιάζεται στο άρθρο [30].



Σχήμα 3.9: Τυχαίο Δάσος[29]

3.6.5 Adaboost

Ο αλόριθμος Adaboost είναι και αυτός ένας ensemble αλγόριθμος. Είναι ένας επαναληπτικός αλγόριθμος όπου σε κάθε γύρο επιλέγονται τυχαία δείγματα του συνόλου δεδομένων και οι ταξινομητές εκπαιδεύονται πάνω σε αυτά. Αρχικά όλα τα δείγματα έχουν τα ίδια βάρη, όμως στους επόμενους γύρους του αλγορίθμου τα βάρη των δειγμάτων στα οποία οι ταξινομητές δεν απέδωσαν καλά και παρουσιάζουν ρυθμό σφάλματος μεγαλύτερο του 50%, αυξάνονται έτσι ώστε να αυξηθεί η πιθανότητα να επιλεχθούν σε μεταγενέστερους γύρους. Οι ταξινομητές συνήθως είναι δέντρα απόφασης [15] [31]. Στο σχήμα 3.10 παρουσιάζεται ο ψευδοκώδικας του αλγορίθμου.

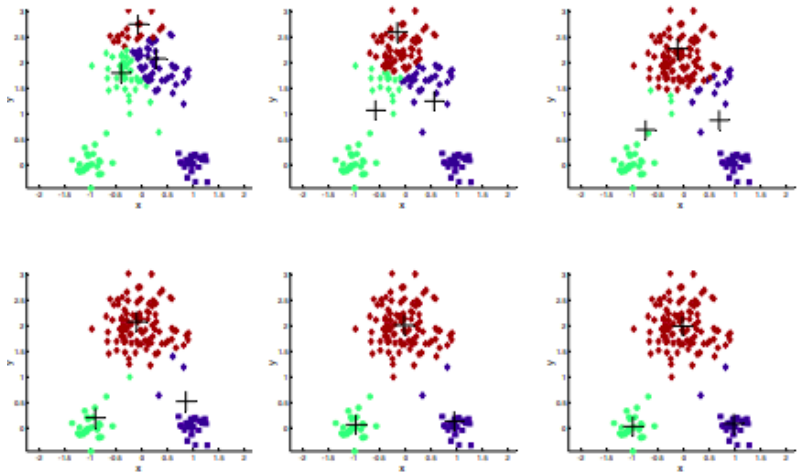
Algorithm 4.6 AdaBoost algorithm.

1: $\mathbf{w} = \{w_j = 1/N \mid j = 1, 2, \ldots, N\}$. {Initialize the weights for all N examples.}
2: Let k be the number of boosting rounds.
3: **for** $i = 1$ to k **do**
4: Create training set D_i by sampling (with replacement) from D according to $\mathbf{w}$.
5: Train a base classifier C_i on D_i.
6: Apply C_i to all examples in the original training set, D.
7: $\epsilon_i = \frac{1}{N}\left[\sum_j w_j \, \delta\big(C_i(x_j) \neq y_j\big)\right]$ {Calculate the weighted error.}
8: **if** $\epsilon_i > 0.5$ **then**
9: $\mathbf{w} = \{w_j = 1/N \mid j = 1, 2, \ldots, N\}$. {Reset the weights for all N examples.}
10: Go back to Step 4.
11: **end if**
12: $\alpha_i = \frac{1}{2} \ln \frac{1-\epsilon_i}{\epsilon_i}$.
13: Update the weight of each example according to Equation 4.103.
14: **end for**
15: $C^*(\mathbf{x}) = \underset{y}{\operatorname{argmax}} \sum_{j=1}^{T} \alpha_j \delta(C_j(\mathbf{x}) = y)\big)$.

Σχήμα 3.10: Ψευδοκώδικας Αλγορίθμου Adaboost [15]

3.6.6 Αλγόριθμος των Κ-μέσων

Ο αλγόριθμος των Κ-μέσων (K-means) είναι ένας από τους κλασσικότερους αλγόριθμους συσταδοποίησης και ανήκει στην διαιρετική συσταδοποίηση. Η ιδέα πίσω από την υλοποίηση του είναι αρκετά απλή. Αρχικά καθορίζεται ο αριθμός των συστάδων, έστω k. Σε κάθε συστάδα ανατίθεται ένα κέντρο (centroid) που συνήθως είναι το μέσον της συστάδας με βάση τους τύπους των αποστάσεων που αναφέρθηκαν παραπάνω. Ο αλγόριθμος λειτουργεί επαναληπτικά επαναυπολογίζοντας κάθε φορά τα νέα κέντρα μέχρι οι συστάδες να μην αλλάζουν άλλο [15]. Στο σχήμα 3.11 παρουσιάζονται στιγμιότυπα της δημιουργίας των συστάδων σε κάθε επανάληψη του αλγορίθμου.



Σχήμα 3.11: Δημιουργία των συστάδων με τον K-means [15]

Στο σχήμα 3.12 παρουσιάζεται ο ψευδοκώδικας του αλγορίθμου.

Algorithm 1.4: k-Means Learner

Function k-means ()
Initialize k prototypes $(\boldsymbol{w_1} \ldots, \boldsymbol{w_k})$ so that the weighted distance between the clusters becomes $\boldsymbol{w_j} = \boldsymbol{i_l}, \boldsymbol{j} \in \{\mathbf{1}, \ldots, \boldsymbol{k}\}, \boldsymbol{l} \in \{\mathbf{1}, \ldots, \boldsymbol{n}\}$
Associate each cluster $\boldsymbol{C_j}$ with the prototype weight $\boldsymbol{w_j}$
Repeat
for each input vector $\boldsymbol{i_l}, \boldsymbol{l} \in \{\mathbf{1}, \ldots, \boldsymbol{n}\}$
do
Assign $\boldsymbol{i_l}$ to cluster $\boldsymbol{C_{j_*}}$ with the nearest $\boldsymbol{w_{j_*}}$
for each cluster $\boldsymbol{C_{j_*}}: \boldsymbol{j} \in \{\mathbf{1}, \ldots, \boldsymbol{k}\}$, **do**;
Update the prototype $\boldsymbol{w_j}$ to be centroid of the sample observations in the current $\boldsymbol{C_{j_*}}$; $\boldsymbol{w_j} = \sum_{i_{l \epsilon c_j}} \boldsymbol{i_l} / \left| \boldsymbol{C_j} \right|$

$$\boldsymbol{E} = \sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{k}} \sum_{\boldsymbol{i_l} \epsilon \boldsymbol{C_j}} \left| \boldsymbol{i_l} - \boldsymbol{w_j} \right|^{\mathbf{2}}$$

until E becomes constant or does not change significantly.

Σχήμα 3.12: Ψευδοκώδικας του αλγορίθμου K-μέσων [16]

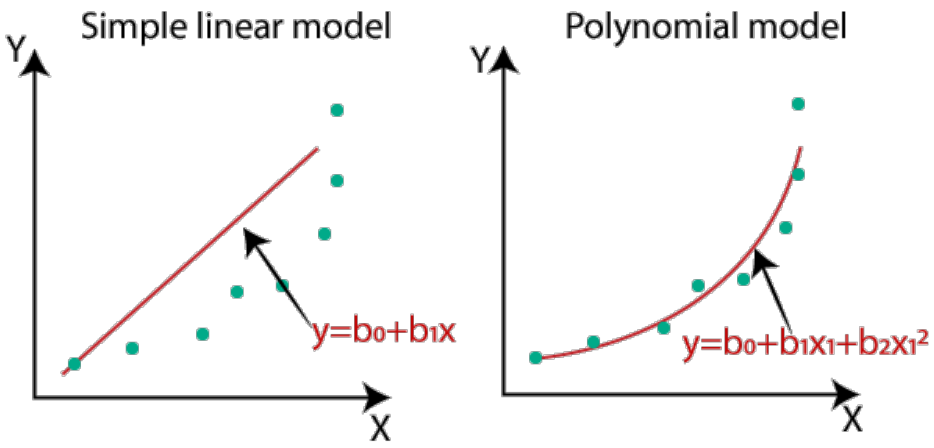
3.6.7 Γραμμική Παλινδρόμηση

Ο αλγόριθμος της Γραμμικής Παλινδρόμησης (Linear Regression) [32] είναι ο πιο απλός αλγόριθμος παλινδρόμησης, ο οποίος μοντελοποιεί την σχέση μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών προσαρμόζοντας μια γραμμική συνάρτηση (ευθεία) στα δεδομένα. Μια ευθεία γραμμικής παλινδρόμησης έχει την μορφή $Y = a + bX$ όπου η Y είναι μεταβλητή-στόχος και η είναι η ανεξάρτητη μεταβλητή (predictor). Η κλίση της συνάρτησης

βρίσκεται με μεθόδους αριθμητικής ανάλυσης, όπως η μέθοδος των ελαχίστων τετραγώνων (Least Squares) ή της απότομης καθόδου (Gradient Descent).

3.6.8 Πολυωνυμική Παλινδρόμηση

Η Πολυωνυμική Παλινδρόμηση (Polynomial Regression) [33] χρησιμοποιεί αντί για γραμμική συνάρτηση, ένα πολυώνυμο δεύτερου ή μεγαλύτερου βαθμού για να προσεγγίσει τα δεδομένα. Η πολυωνυμική παλινδρόμηση δεν απαιτεί γραμμική συσχέτιση μεταξύ των εξαρτημένων και των ανεξάρτητων μεταβλητών, συνεπώς προσεγγίζει καλύτερα τα μη γραμμικά δεδομένα. Η επιλογή του βαθμού του πολυωνύμου γίνεται επαναληπτικά, αυξάνοντας κάθε φορά τον βαθμό του και ελέγχοντας αν επιτυγχάνεται αντίστοιχη μείωση του σφάλματος (RMSE ή MAE). Για την εύρεση των συντελεστών του πολυωνύμου χρησιμοποιούνται αντίστοιχοι μέθοδοι αριθμητικής ανάλυσης με την γραμμική παλινδρόμηση. Στο



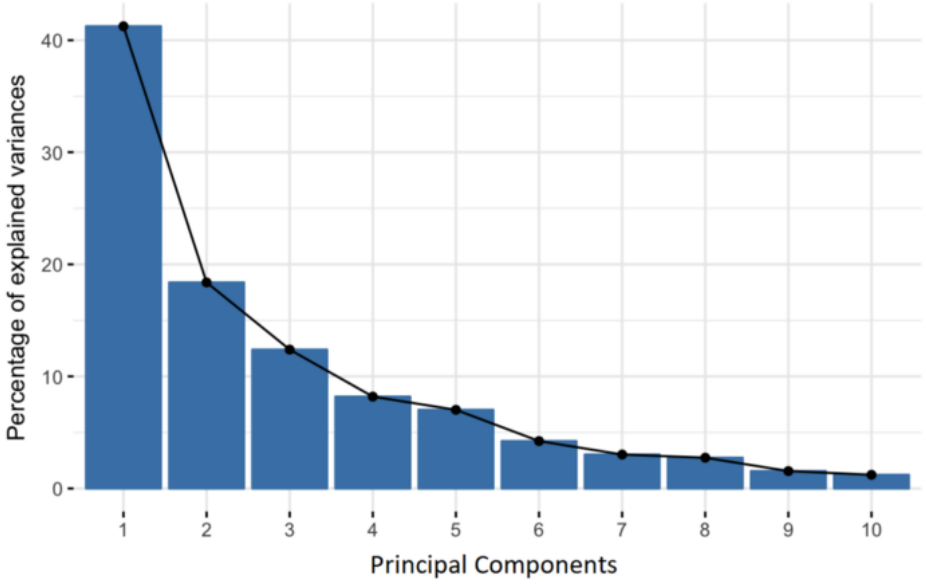
Σχήμα 3.13: Γραμμική και πολυωνυμική παλινδρόμηση [33]

σχήμα 3.13 φαίνεται πως η πολυωνυμική παλινδρόμηση με πολυώνυμο δευτέρου βαθμού προσεγγίζει καλύτερα τα δεδομένα από την γραμμική.

3.6.9 Ανάλυση Κύριων Συνιστωσών

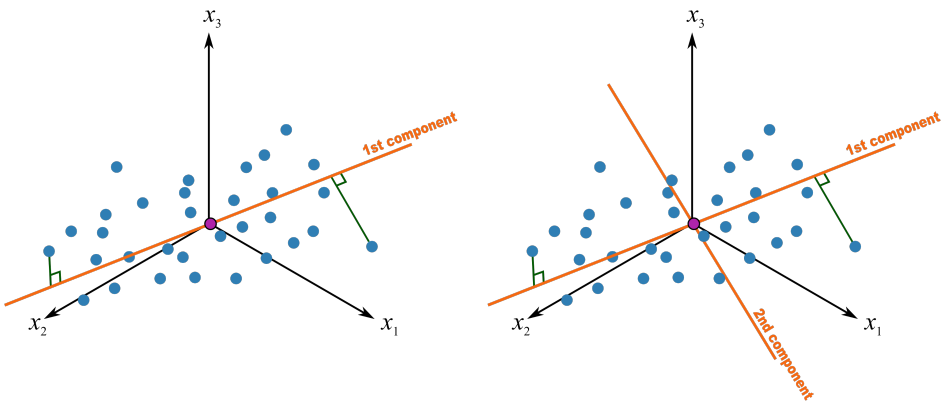
Η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA) είναι μια μέθοδος για την μείωση των διαστάσεων του συνόλου δεδομένων, ενώ παράλληλα στοχεύει στο να διατηρηθεί μεγάλο κομμάτι της πληροφορίας του αρχικού συνόλου μεταβλητών [34]. Πιο συγκεκριμένα, η ΑΚΣ προσπαθεί να εντοπίσει συνιστώσες οι οποίες αποτυπώνουν την μέγιστη διασπορά των δεδομένων, όπως στην εικόνα 3.14. Οι Κύριες Συνιστώσες είναι νέες

μεταβλητές η οποίες προκύπτουν από συνδυασμό των αρχικών μεταβλητών του dataset. Η ιδέα είναι ότι αν υπάρχουν δέκα μεταβλητές, από την ΑΚΣ θα προκύψουν 10 κύριες συνιστώσες, όμως η πρώτη ΚΣ θα συγκεντρώνει την περισσότερη πληροφορία, η δεύτερη ΚΣ την αμέσως περισσότερη κοκ.



Σχήμα 3.14: Κύριες Συνιστώσες [34]

Οι ΚΣ κατασκευάζονται με τρόπο τέτοιο ώστε να αποτυπώνουν την μεγαλύτερη δυνατή διασπορά (variance). Η πρώτη ΚΣ θα περιέχει την μεγαλύτερη διασπορά. Η δεύτερη ΚΣ θα είναι κάθετη ως προς την πρώτη και υπό αυτόν τον περιορισμό θα λάβει το μεγαλύτερο μέρος της εναπομείνουσας διασποράς και ούτω καθεξής, όπως απεικονίζεται στο σχήμα 3.15. Ο αλγόριθμος βασίζεται στις έννοιες του πίνακα διασποράς (variance matrix) και συνδιασποράς (covariance matrix) και στις ιδιοτιμές/ιδιοδιανύσματά τους.



Σχήμα 3.15: Ορθογωνικότητα μεταξύ των ΚΣ [34]

3.7 Το επιδημολογικό μοντέλο SIR

Το μοντέλο SIR [35] [4] είναι ενα τμηματικό (compartmental) μοντέλο πρόβλεψης, που χρησιμοποείται ευρέως για την μοντελοποίηση και πρόβλεψη επιδημιών. Έχει μια σειρά παραλλαγών (SEIR, SEIRD, MSEIR κτλ.) στις οποίες προστίθενται επιπλέον τμήματα (compartments) στο μοντέλο. Η απλούστερη μορφή του, η S-I-R αποτελείται από τα εξής τμήματα:

- S - Susceptible: Ο υγιής πληθυσμός που είναι δυνατόν να μολυνθεί.
- I - Infected: Το τμήμα του πληθυσμού που έχει ήδη μολυνθεί.
- R - Recovered: Το τμήμα του πληθυσμού που έχει ανακάμψει από την λοίμωξη και δεν ξαναμολύνεται.

Τα S, I, R είναι μεταβλητές συναρτήσει του χρόνου, για παράδειγμα αν ο χρόνος μετριέται σε ημέρες τότε $I(t)$ είναι ο αριθμός των ανθρώπων που έχουν μολυνθεί την ημέρα t. Οι παράμετροι του μοντέλου είναι οι εξής:

- b - beta: Ο αριθμός των ατόμων που κάποιος φορέας του ιού μολύνει ανα μέρα.

$$b = infection\_probability * contacts\_per\_day$$

- D - Days: Ο αριθμός των ημερών που κάποιος φορέας μπορεί να μεταδίδει τον ιό.
- R_0 - Ο δείκτης αναπαραγωγής του ιού (reproduction number). Αναφέρεται στο πόσους ανθρώπους μπορει να μολύνει ένας φορέας συνολικά.

$$R_0 = b * D$$

- g - gamma: Ο ρυθμός ανακάμψεων ανά ημέρα

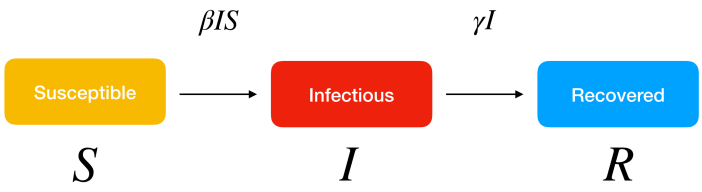
$$g = 1/D$$

- N - Ο συνολικός πληθυσμός

Οι 3 βασικές μεταβλητές συνδέονται μεταξύ τους με το εξής σύστημα διαφορικών εξισώσεων (σχ. 3.16) ενώ στο σχήμα 3.17 παρουσιάζεται η μετάβαση από το ένα τμήμα στο άλλο:

$$\frac{dS}{dt} = -\beta \cdot I \cdot \frac{S}{N}$$
$$\frac{dI}{dt} = \beta \cdot I \cdot \frac{S}{N} - \gamma \cdot I$$
$$\frac{dR}{dt} = \gamma \cdot I$$

Σχήμα 3.16: Σύστημα ΣΔΕ που περιγράφει την σχέση μεταξύ των μεταβλητών του SIR



Σχήμα 3.17: Μεταβάσεις μεταξύ των τμημάτων του SIR μοντέλου

Τέτοιου τύπου μοντέλα είναι εξαιρετικά ευαίσθητα στις αρχικές συνθήκες. Για να μπορέσουν να περιγράψουν με σχετική ακρίβεια τις συνθήκες του πραγματικού κόσμου, χρειάζεται η προσθήκη επιπλέον τμημάτων, για το τμήμα του πληθυσμού που πεθαίνει (D - dead), για το τα άτομα που ναι μεν είναι φορείς αλλά δεν μεταδίδουν τον ιό (E - Exposed), για τους ασθενείς σε κρίσιμη κατάσταση (C - Critical) κτλ. Όσο προστίθενται επιπλέον τμήματα στο μοντέλο τόσο μεγαλώνει η πολυπλοκότητα του. Βασικό ζήτημα είναι επίσης να μοντελοποιηθεί η επίδραση που έχουν τα περιοριστικά μέτρα στον δείκτη αναπαραγωγής, ο οποίος πρέπει να μειώνεται ή να αυξάνεται ανάλογως.

Κεφάλαιο 4

Εξερεύνηση των δεδομένων

4.1 Τα δεδομένα

Για αυτό το κομμάτι της εργασίας χρησιμοποιήθηκαν τα ακόλουθα σύνολα δεδομένων (datasets):

1. `covid_19_data.csv` Αθροιστική καταγραφή των κρουσμάτων, των θανάτων και των ιάσεων σε κάθε χώρα ανα ημέρα [36].

2. `country_info.csv` Δημογραφικές και ιατρικές πληροφορίες για κάθε χώρα (πληθυσμός, ΑΕΠ, δαπάνες για την υγεία, μέση θερμοκρασία κτλ) [37].

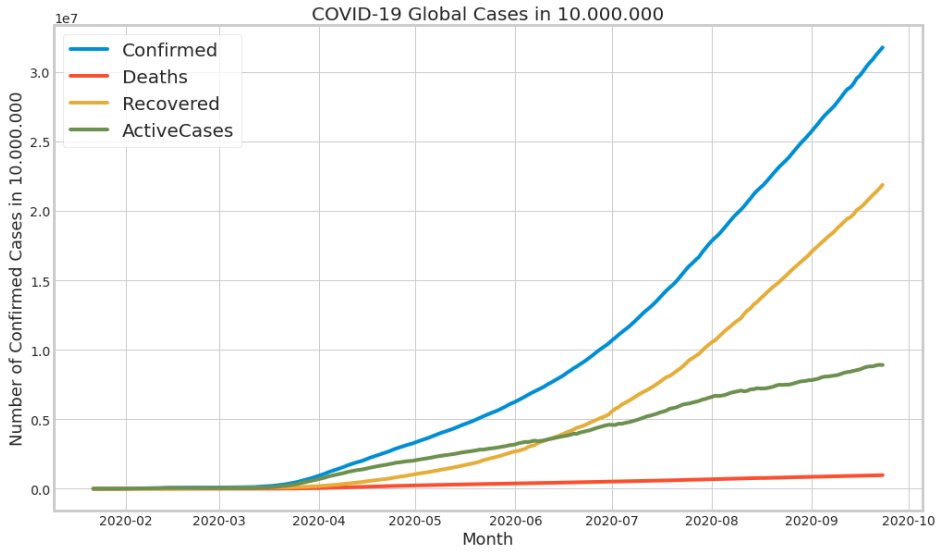
4.2 Εισαγωγή

Σε αυτό το εισαγωγικό τμήμα της εργασίας έγινε προσπάθεια να εξερευνηθούν τα επιδημιολογικά δεδομένα και να εξαχθούν κάποια συμπεράσματα που αφορούν την πορεία και εξάπλωση του COVID-19, με την βοήθεια διαγραμμάτων και γραφικών. Πιο συγκεκριμένη ανάλυση έγινε για τις χώρες της Ευρώπης. Στην συνέχεια εφαρμόστηκαν μέθοδοι συσταδοποίησης και συγκεκριμένα ο αλγόριθμος των K-μέσων (K-means), σε μια προσπάθεια ομαδοποίησης των χωρών ανάλογα με την πορεία των θανάτων στην εξέλιξη της πανδημίας. Με στόχο να αποκωδικοποιηθεί το ποιοι παράγοντες σχετίζονται με την διάδοση και την θνησιμότητα του ιού, τα δεδομένα ενοποιήθηκαν με ενα ακόμα σύνολο δεδομένων, το οποίο περιέχει σημαντικές δημογραφικές και ιατρικές πληροφορίες (μέση ηλικία, ΑΕΠ, δαπάνες για την υγεία, νοσοκομειακά κρεβάτια, αριθμώς καπνιστών κτλ) για κάθε χώρα, με στόχο

την εύρεση συσχετίσεων μεταξύ των μεταβλητών αυτών και των κρουσμάτων/θανάτων του Covid-19. Δόθηκε ιδιαίτερο βάρος στην οπτικοποίηση (visualization) των αποτελεσμάτων με διαγράμματα διαφόρων ειδών και απεικόνιση σε χάρτες, ώστε να τα αποτελέσματα να είναι όσο το δυνατόν πιο ξεκάθαρα στον αναγνώστη.

4.3 Η κατάσταση παγκοσμίως στα τέλη Σεπτεμβρίου 2020

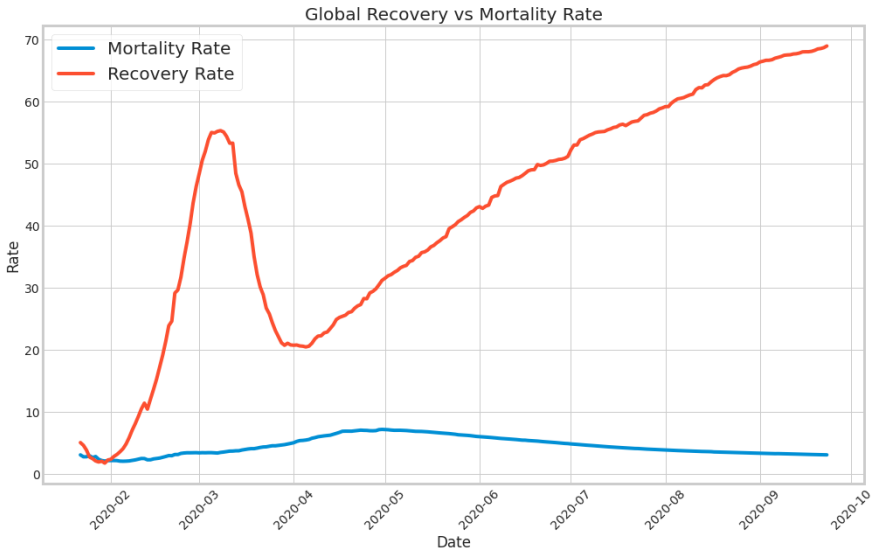
Το σύνολο δεδομένων αναφέρεται σε ημερομηνίες από τις 22/1/2020 έως τις 23/9/2020. Αποτυπώνει δηλαδή πλήρως την εξέλιξη του πρώτου κύματος της επιδημίας του Covid-19, την ύφεση μετά την λήψη περιοριστικών μέτρων από τις διάφορες χώρες (κλείσιμο σχολείων, εστίασης, lockdown) και τις απαρχές του δεύτερου και ισχυρότερου κύματος που βρίσκεται σε πλήρη εξέλιξη κατά τον χρόνο συγγραφής της εργασίας. Στο Σχήμα 4.1, παρουσιάζεται η εξέλιξη της πανδημίας από τον Ιανουάριο έως τον Σεπτέμβριο του 2020 σε επίπεδο κρουσμάτων, θανάτων και ανακάμψεων, ενώ στο Σχήμα 4.2, παρουσιάζεται η γραφική παράσταση του ρυθμού της θνησιμότητας και των ιάσεων από την αρχή της πανδημίας.



Σχήμα 4.1: Η εξέλιξη της πανδημίας ώς τα τέλη Σεπτεμβρίου 2020

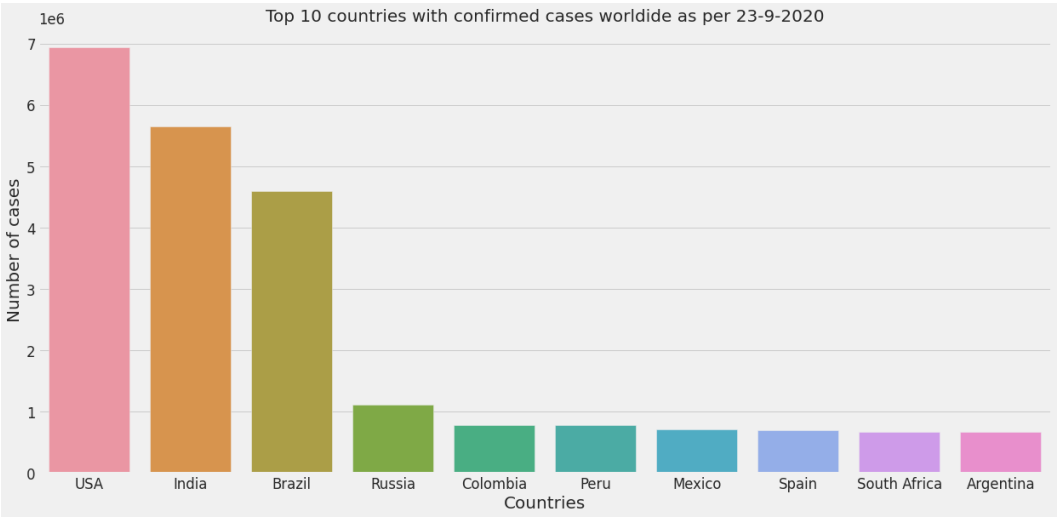
Η επιδημιολογική κατάσταση στις 23/09/2020 στον κόσμο

Ημερ/νια	Κρούσματα	Θάνατοι	Ιάσεις	Θνησιμότητα
2020-09-23	31.779.835	975.104	21.890.442	3,068310

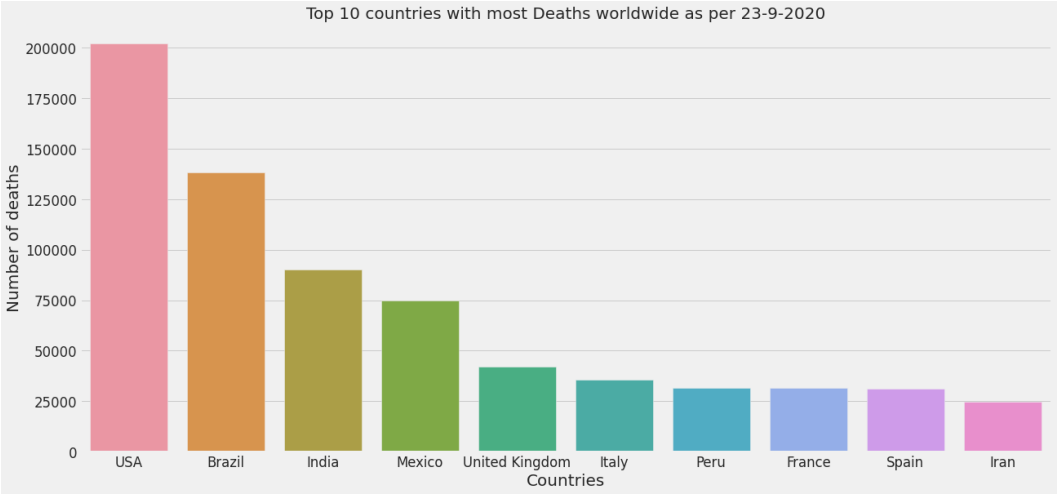


Σχήμα 4.2: Ρυθμός θνησιμότητας και ιάσεων

Εξετάζοντας περαιτέρω τα δεδομένα σε επίπεδο χώρας, στο σχήμα 4.3 παρουσιάζονται με φθίνουσα σειρά οι δέκα χώρες με τα περισσότερα κρούσματα παγκοσμίως ενώ στο σχήμα 4.4 η κατάταξη γίνεται με βάση τον αριθμό των θανάτων ανα χώρα. Παρατηρούμε ότι στην πρώτη θέση βρίσκονται οι ΗΠΑ τόσο όσον αφορά τα κρούσματα όσο και στο επίπεδο των θανάτων και ακολουθούν χώρες όπως η Βραζιλία, η Ινδία και η Ρωσία. Αξιοσημείωτο είναι ότι χώρες όπως το Μεξικό αλλά και οι Ευρωπαϊκές χώρες που χτύπηθηκαν ιδαίτερα σκληρά στο πρώτο κύμα της πανδημίας (Ισπανία, Αγγλία, Ιταλία) βρίσκονται αρκετά ψηλά στο γράφημα με τους θανάτους ενώ δεν βρίσκονται στην πρώτη δεκάδα των κρουσμάτων. Η θνησιμότητα σε αυτές τις χώρες είναι αρκετά υψηλότερη του μέσου όρου.

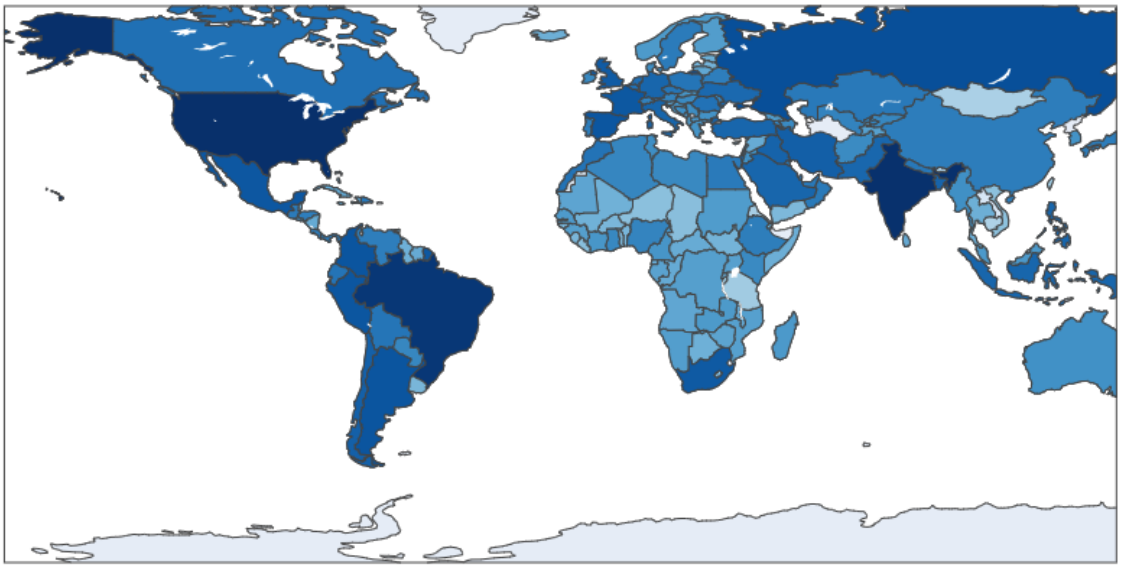


Σχήμα 4.3: Οι δέκα χώρες με τα περισσότερα κρούσματα

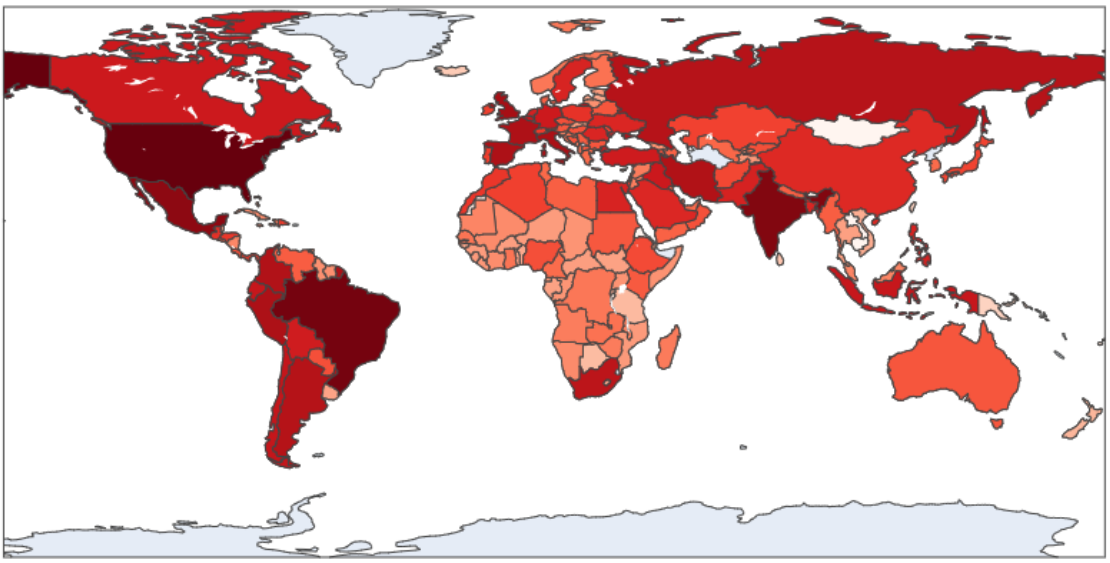


Σχήμα 4.4: Οι δέκα χώρες με τους περισσότερους θανάτους

Μια πολυ καλή οπτική περιγραφή της κατάστασης προσφέρουν οι Χάρτες Θερμότητας (Heat Maps) όπου με σκουρότερα χρώματα αναπαρίσταται ο μεγαλύτερος αριθμός κρουσμάτων και θανάτων. Τα Σχήματα 4.5 και 4.6 περιέχουν τέτοιους χάρτες σε παγκόσμιο επίπεδο με μπλέ χρώμα για τα κρούσματα και κόκκινο για τους θανάτους αντίστοιχα. Παρατηρούμε ότι στα τέλη του Σεπτεμβρίου, ο ιός έχει απλωθεί σε όλο τον κόσμο. Λιγότερο επηρεασμένη ήπειρος φαίνεται να είναι η Αφρική, ενώ περισσότερο επηρεάστηκαν η Ευρώπη και η Αμερική (Βόρεια και Νότια).



Σχήμα 4.5: Χάρτης θερμότητας που απεικονίζει την πυκνότητα των κρουσμάτων



Σχήμα 4.6: Χάρτης θερμότητας που απεικονίζει την πυκνότητα των θανάτων

4.4 Ομαδοποίηση χωρών βάσει της εξέλιξης της πανδημίας

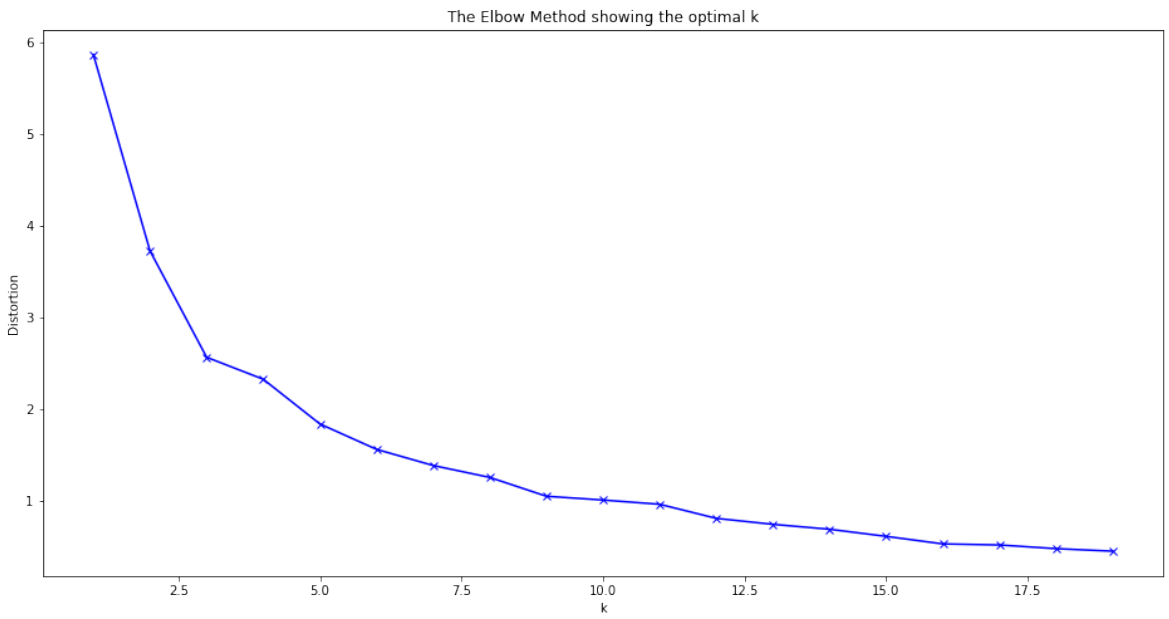
Στην προσπάθεια να παρατηρήσουμε την εξέλιξη της πανδημίας, είναι χρήσιμο να προσπαθήσουμε να ομαδοποίησουμε χώρες στις οποίες η εξάπλωση του ιού είχε παρόμοια χαρακτηριστικά. Η υλοποίηση βασιστηκε στο άρθρο [38].

Ως γνώρισμα για την ομαδοποίηση - συσταδοποίση (clustering) των χωρών χρησιμοποιήθηκαν οι θάνατοι/100.000 κατοίκους. Αυτό έγινε για τους εξής δύο λόγους:

1. Τα κρούσματα ως μέτρο είναι επιρρεπή στον θόρυβο (πχ εξάρτηση από αριθμό τεστ) ενώ οι θάνατοι όχι.

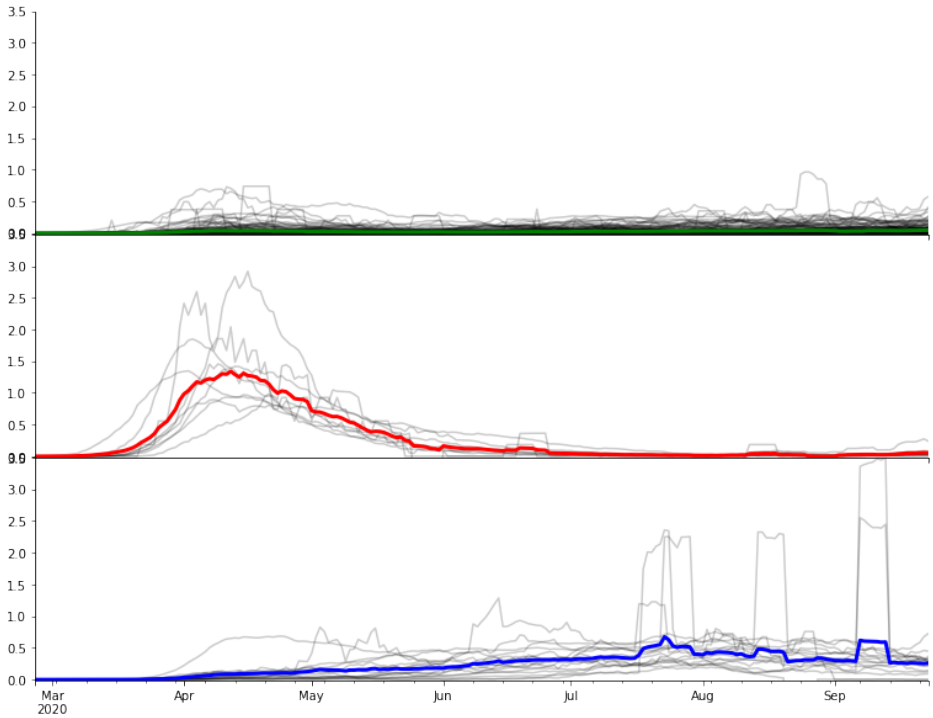
2. Η αναγωγή σε νούμερο ανα 100.000 κατοίκους, φέρνει τα δεδομένα στην ίδια κλίμακα και απεμπλέκει τον αριθμό των θανάτων από τον πληθυσμό κάθε χώρας.

Ο αλγόριθμος που χρησιμοποίηθηκε ήταν ο αλγόριθμος των K-μέσων με αριθμό συστάδων $K = 3$. Ο αριθμός των συστάδων προέκυψε από την μέθοδο του αγκώνα (elbow rule) μετά από δοκιμές για K από 1 εως 20. Το αποτέλεσμα της μεθόδου φαίνεται στο σχήμα 4.7. Παρατηρούμε ότι στις 3 συστάδες ο ρυθμός μείωσης του σφάλματος σταθεροποιείται σημαντικά.



Σχήμα 4.7: Η μέθοδος του αγκώνα για την εύρεση του βέλτιστου αριθμού συστάδων

Ως δεδομένα για την εκπαίδευση του K-means χρησιμοποιήθηκε ο κινητός μέσος (moving average) των θανάτων, για αριθμό ημερών = 7 για κάθε χώρα. Τα αποτελέσματα του αλγορίθμου παρουσιάζονται στο σχήμα 4.8.



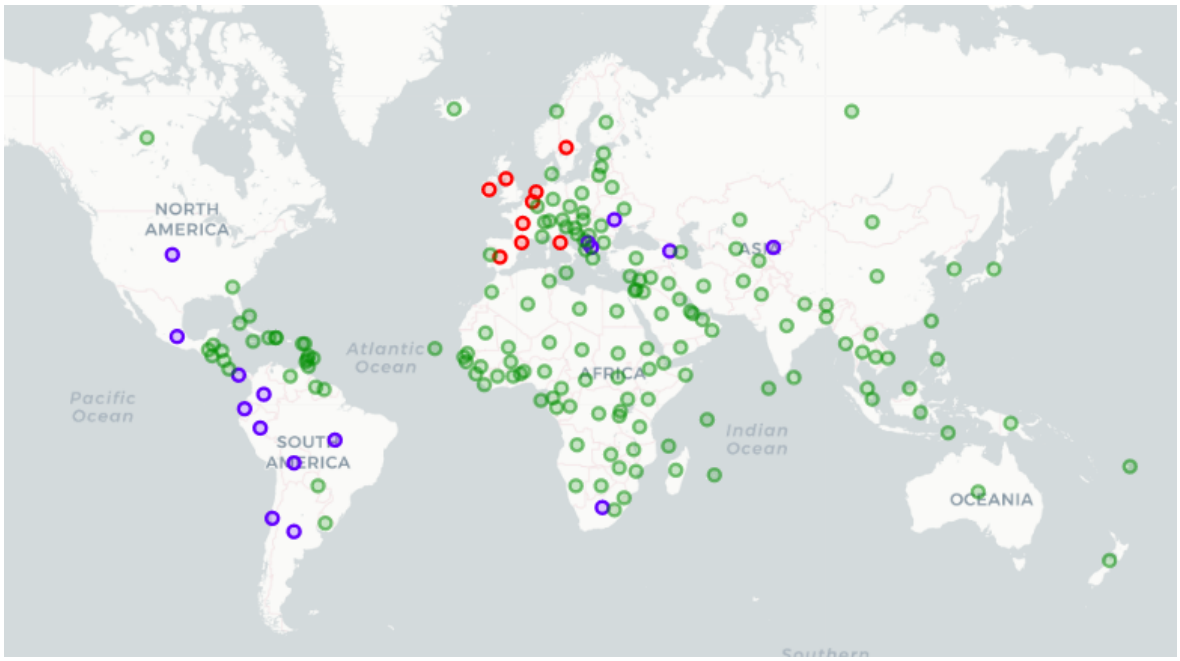
Σχήμα 4.8: Οι τρεις συστάδες που προέκυψαν από τον αλγόριθμο K-means

Η πράσινη συστάδα αριθμεί 158 χώρες, η μπλέ 16 και η κόκκινη 9 χώρες. Παρατηρώντας

τα διαγράμματα βρίσκουμε τα εξής μοτίβα όσον αφορά την συμπεριφορά του αριθμού των θανάτων στις διάφορες χώρες:

- Πράσινη συστάδα: Χώρες στις οποίες ο αριθμός των θανάτων δεν ανέβηκε ποτέ δραματικά κατά τη διάρκεια του πρώτου κύματος και παρέμεινε χαμηλός έκτοτε. Αυτή η συστάδα περιλαμβάνει τις περισσότερες χώρες.
- Κόκκινη συστάδα: Χώρες που στο πρώτο κύμα σημείωσαν πολύ μεγάλο αριθμό θανάτων, αλλά κατάφεραν να "επιπεδώσουν την καμπύλη" και ο αριθμός θανάτων να πέσει σε μικρούς αριθμούς.
- Μπλε συστάδα: Χώρες που ο αριθμός των θανάτων αυξάνει σταθερά από την αρχή της επιδημίας.

Για καλύτερη ερμηνεία των αποτελεσμάτων του αλγορίθμου το Σχήμα 4.9 παρουσιάζει τα αποτελέσματα του αποτυπωμένα σε χάρτη.



Σχήμα 4.9: Αποτύπωση της συσταδοποίησης σε χάρτη

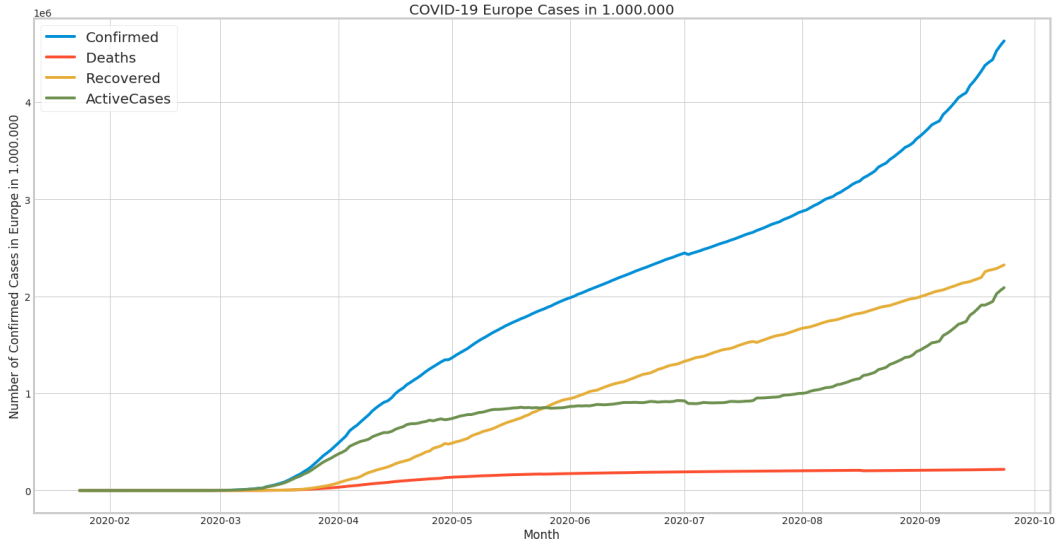
Οι χώρες που αποτελούν την κόκκινη συστάδα, είναι οι Ευρωπαικές χώρες που χτυπήθηκαν ισχυρά στο πρώτο κύμα (Ηνωμένο Βασίλειο, Ισπανία, Ιταλία, Σουηδία, Γαλλία κτλ) ενώ στην μπλε συστάδα ανήκουν κατά βάση χώρες της Αμερικής (ΗΠΑ και χώρες της Ν. Αμερικής) οι οποίες παρουσιάζουν σταθερά ανοδική πορεία όσον αφορά τους θανάτους απο Covid-19.

4.5 Εστιάζοντας στην Ευρώπη

4.5.1 Η πορεία της πανδημίας στην Ευρώπη

Η Ευρώπη ήταν η πρώτη ήπειρος η οποία δοκιμάστηκε από το πρώτο κύμα της πανδημίας. Στην παρούσα εργασία επιλέχθηκε να γίνει ειδική ανάλυση της εξάπλωσης του Covid-19 στην Ευρωπαϊκή ήπειρο. Προστέθηκε μια επιπλέον μεταβλητή `continent` στα δεδομένα όπου σε κάθε χώρα αντιστοιχήθηκε η ήπειρός της. Έπειτα επιλέχθηκε το κομμάτι του συνόλου δεδομένων που αναφέρεται στην Ευρώπη. Στόχος είναι να μελετηθεί η πορεία του ιού τόσο στο σύνολο της ηπείρου όσο και στο επίπεδο επιμέρους χωρών - αυτών που παρουσίασαν τους μεγαλύτερους αριθμούς κρουσμάτων.

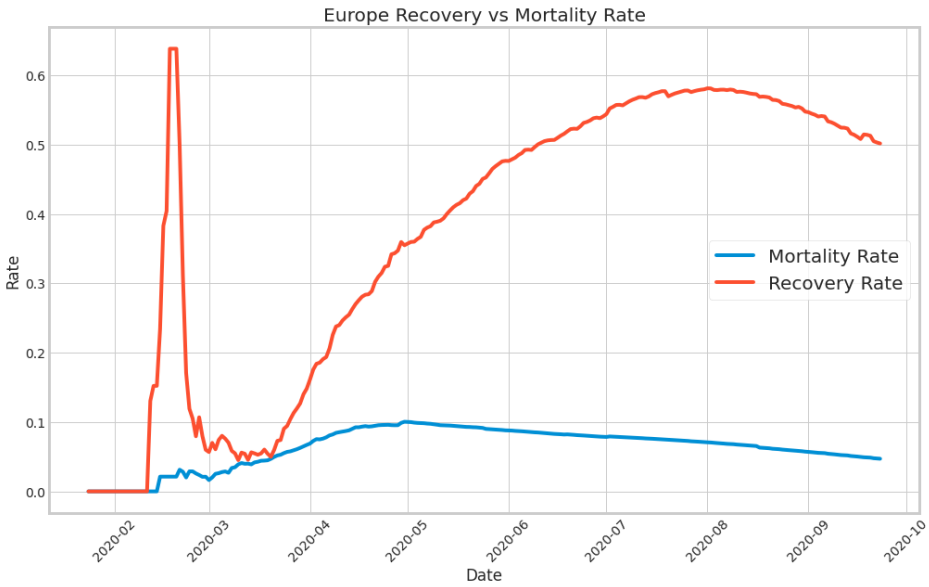
Στο σχήμα 4.10 βλέπουμε την πορεία της πανδημίας στην Ευρώπη, ενώ στο σχήμα 4.11 τον ρυθμό της θνησιμότητας και των ανακάμψεων.



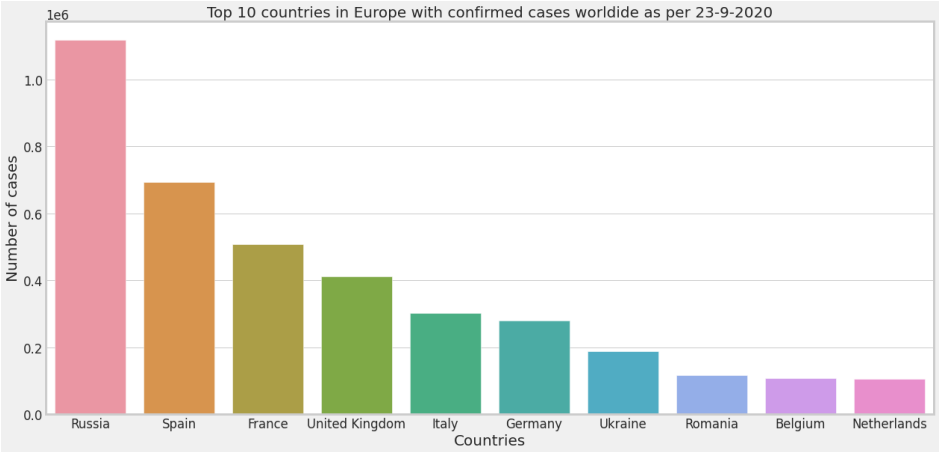
Σχήμα 4.10: Η εξέλιξη της πανδημίας στην Ευρώπη

Διακρίνεται αρκετά καθαρά η έξαρση του πρώτου κύματος την περίοδο Απρίλιου-Μαΐου, η σχετική κάμψη την περίοδο του καλοκαιριού και η απότομη αύξηση που παρουσίασαν τα κρούσματα από το τέλος του καλοκαιριού και έπειτα που οδήγησε στο δεύτερο κύμα της επιδημίας. Η δε θνησιμότητα κορυφώθηκε στο τέλος του Μαίου για να ακολουθήσει μια σταθερή πτωτική πορεία μέχρι το τέλος των ημερομηνιών των δεδομένων.

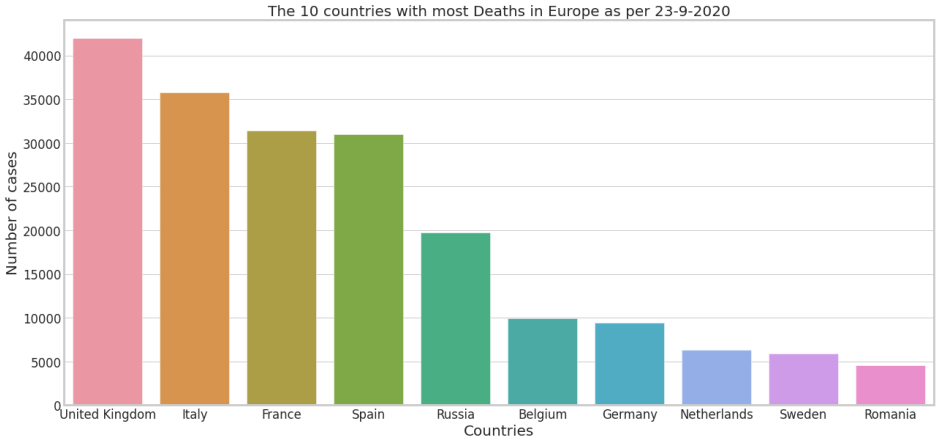
Στη συνέχεια, στα διαγράμματα 4.12 και 4.13 παρουσιάζονται οι 10 χώρες με τα περισσότερα κρούσματα και θανατους, αντίστοιχα.



Σχήμα 4.11: Θνησιμότητα και ιάσεις στην Ευρώπη



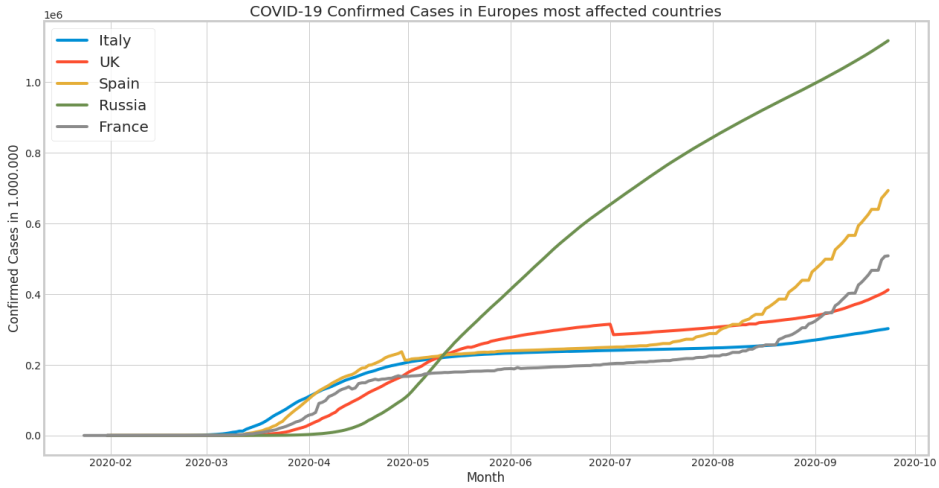
Σχήμα 4.12: Οι δέκα χώρες μετα περισσότερα κρούσματα στην Ευρώπη



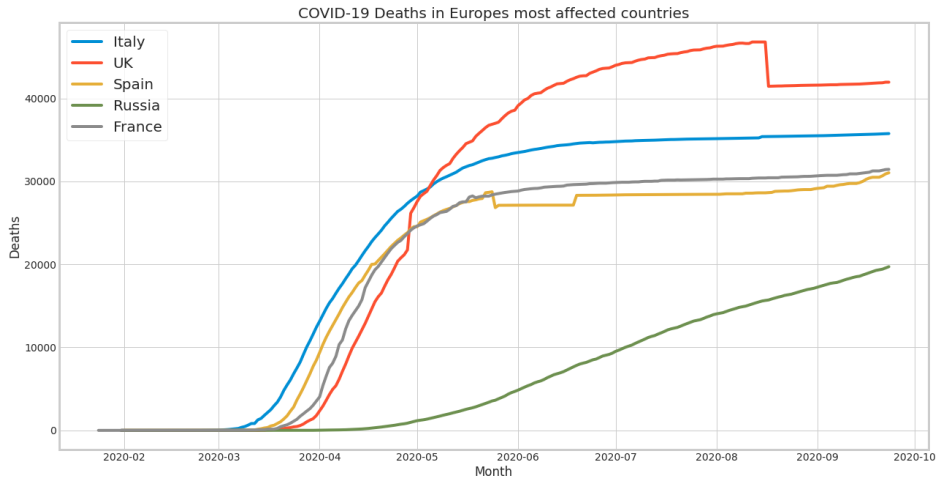
Σχήμα 4.13: Οι δέκα χώρες με τους περισσότερους θάνατους στην Ευρώπη

Ενδιαφέρον παρουσιάζει η περίπτωση της Σουηδίας η οποία δεν βρίσκεται στην πρώτη δεκάδα των κρουσμάτων, αλλά καταλαμβάνει θέση στην δεκάδα των θανάτων. Αυτό πιθανότατα να οφείλεται στην διαφορετική προσέγγιση της χώρας στο ζήτημα των περιοριστικών μέτρων κατά την διάρκεια του πρώτου κύματος.

Οι πέντε χώρες όπου παρατηρείται ο μεγαλύτερος αριθμός κρουσμάτων είναι: Ρωσία, Ισπανία, Γαλλία, Ηνωμένο Βασίλειο και Ιταλία. Στα διαγράμματα 4.14, 4.15 και 4.16 περιγράφονται τα κρούσματα, οι θάνατοι και ο ρυθμός θνησιμότητας των πέντε αυτών χωρών.

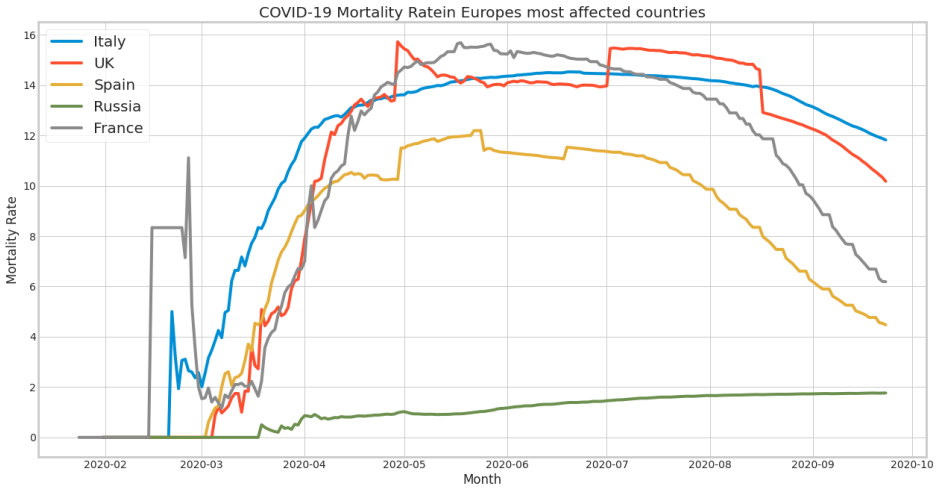


Σχήμα 4.14: Η εξέλιξη των κρουσμάτων στις 5 πιο επηρεασμένες χώρες της Ευρώπης



Σχήμα 4.15: Η εξέλιξη των θανάτων στις 5 πιο επηρεασμένες χώρες της Ευρώπης

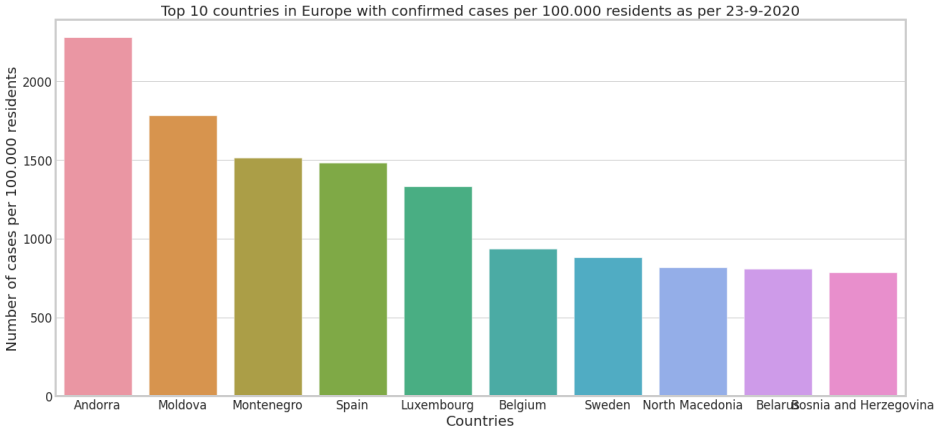
Παρατηρούμε ότι η μεγάλη διασπορά του ιού στην Ρωσία ξεκίνησε περίπου έναν μήνα μετά τις υπόλοιπες τέσσερις χώρες του γραφήματος. Ενδιαφέρον παρουσιάζει το γεγονός ότι η Ρωσία δείχνει μια αυτόνομη συμπεριφορά - οι υπόλοιπες χώρες "επιπέδωσαν" την καμπύλη



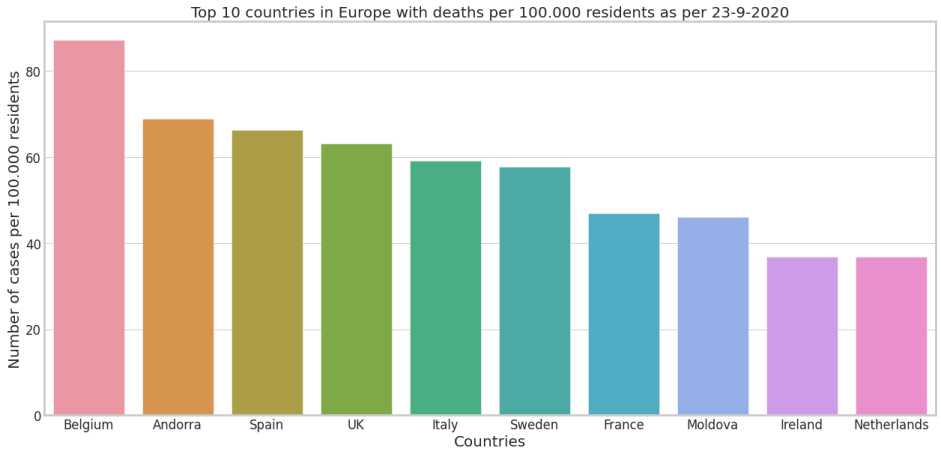
Σχήμα 4.16: Ο ρυθμός θνησιμότητας στις 5 πιο επηρεασμένες χώρες της Ευρώπης

τους, ενώ στην Ρωσία ο ρυθμός αύξησης των κρουσμάτων δεν μειώθηκε αλλά συνεχίζει να διατηρεί ανοδική πορεία. Μια εξήγηση για αυτό θα ήταν το μέγεθος της χώρας, καθώς είναι τόσο μεγάλη όσο σχέδον όλη η Ευρώπη. Παρόλη την σταθερή αύξηση των κρουσμάτων στο έδαφος της η Ρωσία διατηρεί σημαντικά χαμηλότερα την θνησιμότητα συκριτικά με τις άλλες τέσσερις χώρες που εξετάζουμε. Όσον αφορά τις υπόλοιπες χώρες, διακρίνεται ξεκάθαρα η "επιπέδωση της καμπύλης" από τον Μαϊο και ύστερα, καθώς και η αρχή του δεύτερου κύματος της πανδημίας μετά το τέλος του καλοκαιριού.

Προκειμένου να εξεταστούν τα κρούσματα/θάνατοι στην ίδια κλίμακα για κάθε χώρα (και να απεμπλακεί ο πληθυσμός της) πρέπει να αναγάγουμε την μεταβλητή που μας ενδιαφέρει σε νούμερο ανα κάποιον αριθμό κατοίκων. Στα σχήματα 4.17 και 4.18 παρουσιάζεται η αναγωγή κρουσμάτων και θανάτων ανά 100.000 κατοίκους.



Σχήμα 4.17: Κρούσματα ανα 100.000 κατοίκους



Σχήμα 4.18: Θάνατοι ανα 100.000 κατοίκους

Παρατηρούμε ότι στο διάγραμμα των κρουσμάτων καταλαμβάνουν τις πρώτες θέσεις πολύ μικρές χώρες όπως η Ανδόρρα, η Μολδαβία, το Λουξεμβούργο κτλ. Ο πολύ μικρός πληθυσμός αυτών των χωρών ίσως να λειτουργεί παραπλανητικά στον υπολογισμό. Αντίθετα στο διάγραμμα των θανάτων ανα 100.000 κατοίκους ξαναβλέπουμε την πεντάδα χωρών που χαρακτηρίσαμε ως πιο επηρεασμένες από την πανδημία στα προηγούμενα διαγράμματα. Χαρακτηριστικό είναι ότι το Βέλγιο είναι πρώτο στην αναγωγή των θανάτων.

4.5.2 Εμπλουτισμός των δεδομένων με περισσότερες πληροφορίες

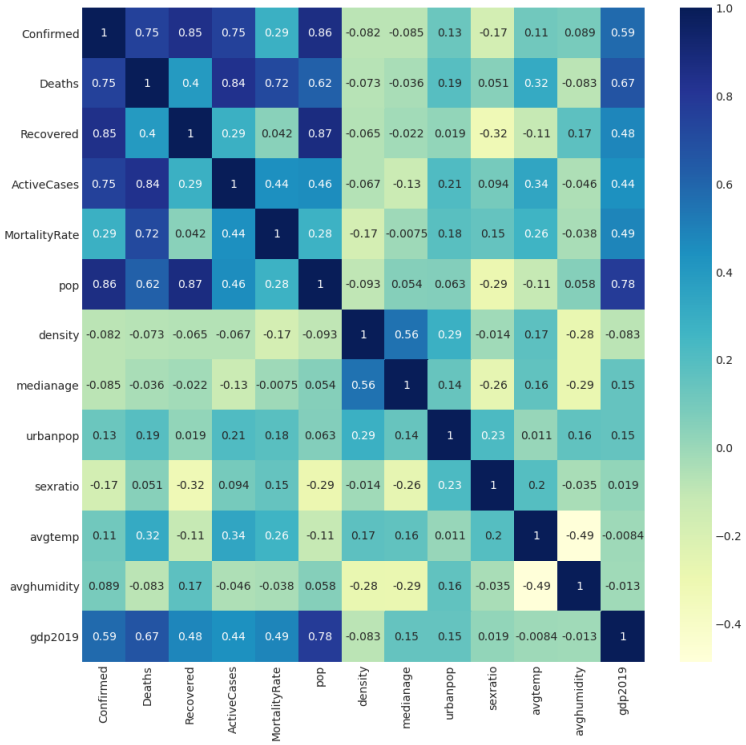
Στην προσπάθεια για λυθεί ο γρίφος γιατί κάποιες χώρες είχαν πολύ περισσότερα κρούσματα/θανάτους από άλλες, ενώθηκε το παρόν σύνολο δεδομένων με το `country_info.csv` το οποίο περιέχει δημογραφικές πληροφορίες για κάθε χώρα. Στόχος είναι η εύρεση συσχετίσεων (correlations) μεταξύ των κρουσμάτων-θανάτων και άλλων μεταβλητών που αφορούν την κάθε χώρα. Η συσχέτιση δεν σημαίνει απαραίτητα και αιτιότητα. Το μόνο ασφαλές συμπέρασμα που μπορούμε να εξάγουμε είναι οτι οι μεταβλητές που συσχετίζονται θετικά αυξάνουν ταυτόχρονα και οι μεταβλητές που συσχετίζονται αρνητικά το αντίστροφο [39].

Για να προκύψουν σχετικά μικροί σε μέγεθος και εύκολα ερμηνεύσιμοι χάρτες συσχέτισεων (correlation maps), από τις 60 μεταβλητές του νέου συνόλου δεδομένων, αρχικά χρησιμοποιήθηκαν οι παρακάτω οκτώ μεταβλητές οι οποίες έχουν περισσότερο δημογραφικό χαρακτήρα:

country_info.csv - Δημογραφικές μεταβλητές

Μεταβλητή	Περιγραφή
pop	Πληθυσμός σε εκατομμύρια
density	Αριθμός ανθρώπων που ζουν ανά m^2
median_age	Μέση ηλικία της χώρας
urban_pop	Ποσοστό που ζει σε πόλεις
sexratio	Λόγος ανδρών/γυναικών
avgtemp	Μέση Θερμοκρασία
avghumidity	Μέση υγρασία
gdp2019	ΑΕΠ για το 2019 σε εκατομμύρια δολλάρια

Στην συνέχεια ενώθηκαν τα δύο σύνολα δεδομένων, σε ένα ενιαίο το οποίο περιέχει, εκτός από τις πληροφορίες για τον Covid-19 για κάθε χώρα (κρούσματα,θάνατοι,θνησιμότητα), και τις παραπάνω μεταβλητές. Στο Σχήμα 4.19 παρουσιάζεται ο χάρτης συσχέτισης όλων των μεταβλητών του συνόλου.



Σχήμα 4.19: Χάρτης συσχετίσεων για τις επιπλέον δημογραφικές μεταβλητές

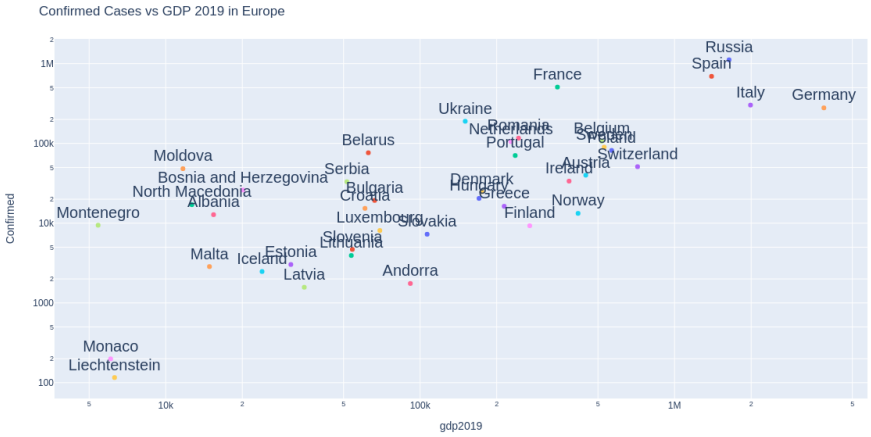
Μας ενδιαφέρουν οι συσχετίσεις που αφορούν τις σχετιζόμενες με τον Covid-19 μεταβλητές (Confirmed, Deaths, Mortality Rate) με τις υπόλοιπες μεταβλητές του συνόλου. Πα-

ρατηρούνται οι εξής συσχετίσεις με βάση τον συντελεστή Pearson (Pearson coefficient) ενώ θα αγνοηθούν προφανείς συσχετίσεις όπως Πληθυσμός - Κρούσματα κτλ:

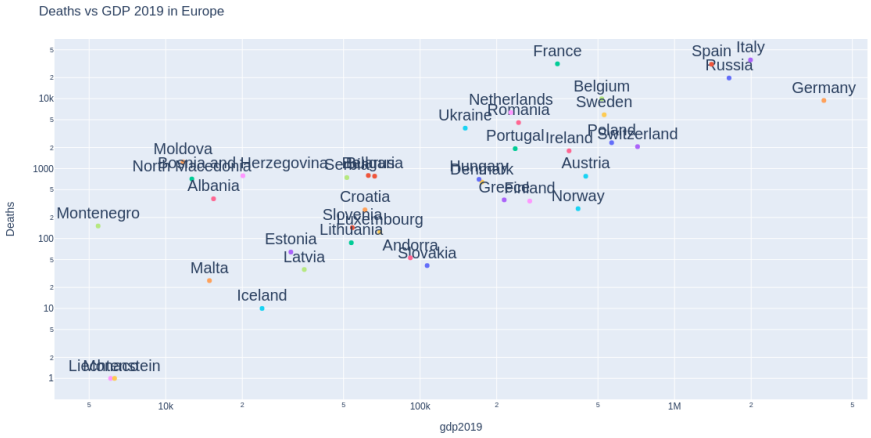
- avgtemp - Deaths : 0.32
- gdp2019 - Confirmed : 0,67
- gdp2019 - Deaths : 0,59

Οι πιο ενδιαφέρουσες συσχετίσεις αφορούν το ΑΕΠ κάθε χώρας σε σχέση με τα κρούσματα και τους θανάτους. Αυτό μπορεί να έχει πολλαπλές εξηγήσεις. Το ΑΕΠ είναι άμεσα συσχετισμένο με τον πληθυσμό της χώρας, οπότε μια πρώτη λογική εξήγηση είναι ότι όσο μεγαλύτερος είναι ο πληθυσμός της χώρας, τόσο αυξάνονται και τα κρούσματα/θάνατοι σε απόλυτους αριθμούς. Επίσης οι πλουσιότερες χώρες πιθανόν να μπορούσαν να διεξάγουν περισσότερα μοριακά τεστ ανίχνευσης του ιού από τις φτωχότερες, ειδικά στην πρώτη φάση της πανδημίας όπου τα τεστ ήταν δυσεύρετα. Όμως αυτό δεν εξηγεί την συσχέτιση του ΑΕΠ με τον αριθμό των θανάτων ο οποίος παραμένει ανεξάρτητος από τον αριθμό των τεστ που διεξήχθησαν. Μια άλλη πιθανή εξήγηση είναι ότι οι χώρες με το μεγαλύτερο ΑΕΠ βασίζονται περισσότερο στην βαριά βιομηχανία η οποία δεν σταμάτησε να λειτουργεί ενώ είναι γνωστό ότι οι μαζικοί χώροι εργασίας αποτελούν εστίες υπερμετάδοσης του ιού, ιδαίτερα όταν δεν τηρούνται ευλαβικά τα μέτρα προστασίας. Θετική συσχέτιση υπάρχει επίσης μεταξύ της μέσης θερμοκρασίας της χώρας και του αριθμού των θανάτων. Στα επόμενα διαγράμματα, παρουσιάζονται γραφήματα διασποράς (scatter plots) που απεικονίζουν την θέση της κάθε χώρας σε σχέση με τις μεταβλητές που ερευνούμε.

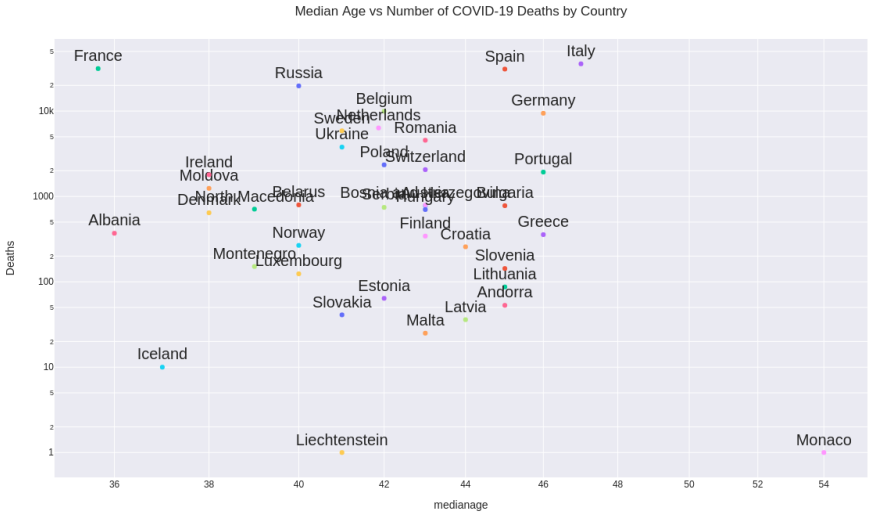
Παρατηρούμε όντως ότι οι χώρες στην πρώτη πεντάδα του ΑΕΠ είναι ψηλά σε κρούσματα (σχ. 4.20) και θανάτους (σχ 4.21), παρατήρηση που συνάδει με το σχήμα 4.12 που απεικόνιζε τις 10 χώρες με τον υψηλότερο αριθμό κρουσμάτων. Σχεδόν μηδενική συσχέτιση έδειξαν οι μεταβλητές της μέσης ηλικίας ανα χώρα με τα κρούσματα/θανάτους. Στο Σχήμα 4.22 παρουσιάζεται το διάγραμμα διασποράς που αφορά την μέση ηλικία ανα χώρα και την θνησιμότητα. Ξεχωρίζει η περίπτωση της Ιταλίας με μέση ηλικία τα 47 έτη και την υψηλότερη θνησιμότητα στην ήπειρο της τάξης του 11%. Μία από τις εξηγήσεις που έδωσαν οι επιστήμονες για το φαινόμενο της υψηλής θνησιμότητας στην χώρα ήταν ο γερασμένος πληθυσμός της. Από μόνη της αυτή η εξήγηση δεν αρκεί, βλέποντας ότι χώρες όπως η Γερμανία, η Πορτογαλία και η Ελλάδα με μέση ηλικία πληθυσμού τα 46 έτη παρουσιάζουν αρκετά χαμηλότερη θνησιμότητα.



Σχήμα 4.20: Διάγραμμα διασποράς κρουσμάτων και ΑΕΠ στην Ευρώπη



Σχήμα 4.21: Διάγραμμα διασποράς θανάτων και ΑΕΠ στην Ευρώπη



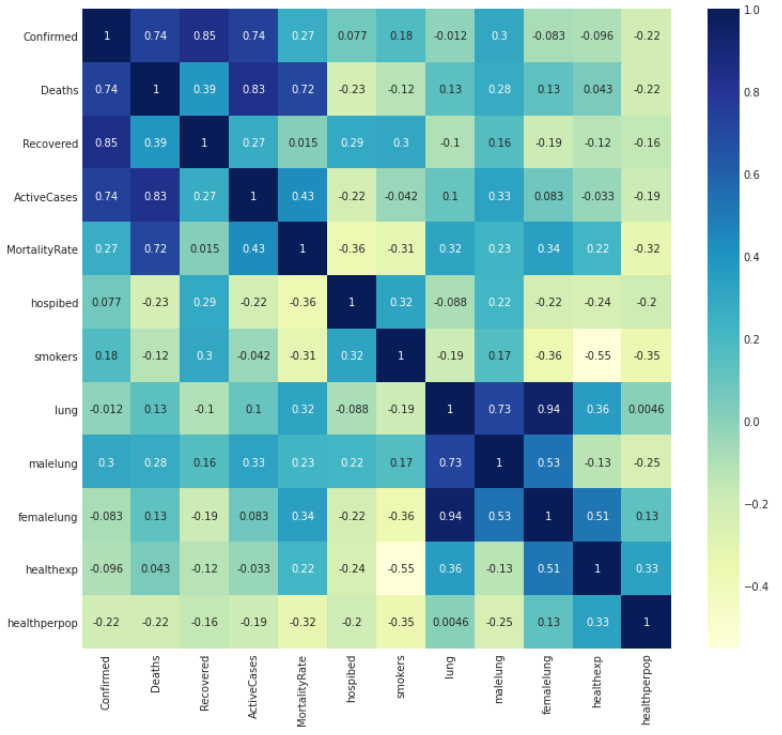
Σχήμα 4.22: Διάγραμμα διασποράς μέσης ηλικίας και θανάτων στην Ευρώπη

Σε μια δεύτερη απόπειρα να ανακαλυφθούν ενδιαφέρουσες συσχετίσεις, προστίθενται στα δεδομένα μεταβλητές που αφορούν το σύστημα υγείας της κάθε χώρας (κρεββάτια νοσοκομείων, δαπάνες που αφορούν την υγεία) αλλά και ποσοστό καπνιστών, ποσοστό ανθρώπων με πνευμονολογικά προβλήματα ανα φύλλο κτλ.

country_info.csv - Ιατρικές μεταβλητές

variable	description
hospibed	Αριθμός νοσοκομειακών κλινων ανά 1000 κατοίκους
healthexp	Δαπάνες για την υγεία σε εκατομμύρια δολλάρια
healthperpop	Δαπάνες για την υγεία ανα κάτοικο
smokers	Ποσοστό καπνιστών στον γενικό πληθυσμό
lung	Θνησιμότητα λόγων πνευμονολογικών παθήσεων ανα 100.000 κατοίκους
malelung	Θνησιμότητα ανδρών λόγων πνευμονολογικών παθήσεων ανα 100.000 κατοίκους
femalelung	Θνησιμότητα γυναικών λόγων πνευμονολογικών παθήσεων ανα 100.000 κατοίκους

Παρατίθεται ο χάρτης συσχετίσεων στο Σχήμα 4.23

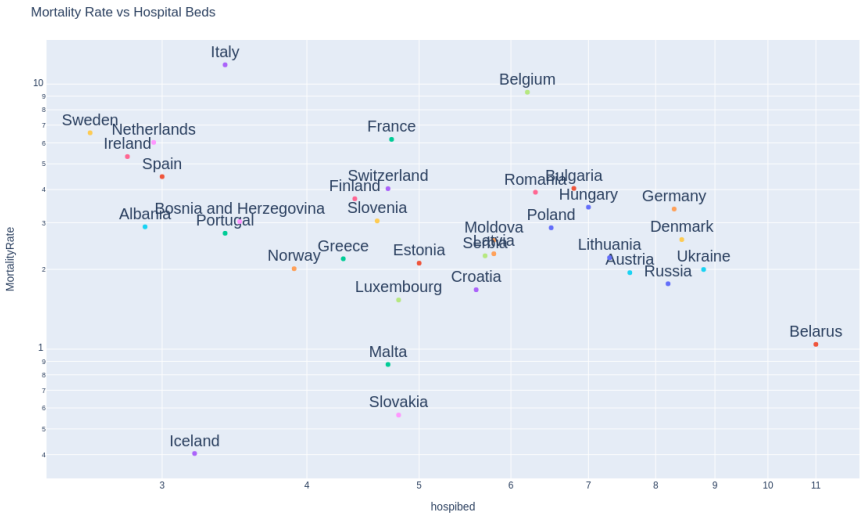


Σχήμα 4.23: Χάρτης συσχετίσεων για τις μεταβλητές ιατρικού χαρακτήρα

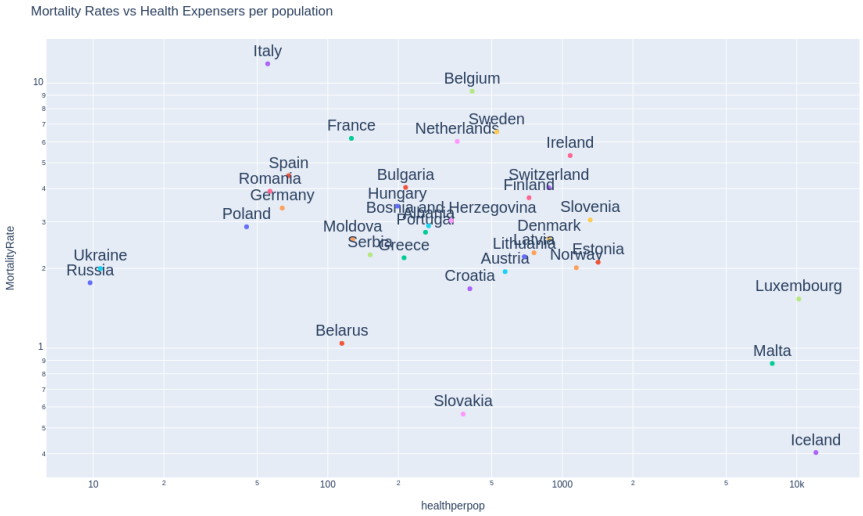
Παρατηρούνται οι εξής συσχετίσεις στον χάρτη:

- hospibed - MortalityRate : -0.36
- malelung - MortalityRate : 0,34
- femalelung - MortalityRate : 0,23
- smokers - MortalityRate : -0.31
- healthexpperpop - MortalityRate : -0.32

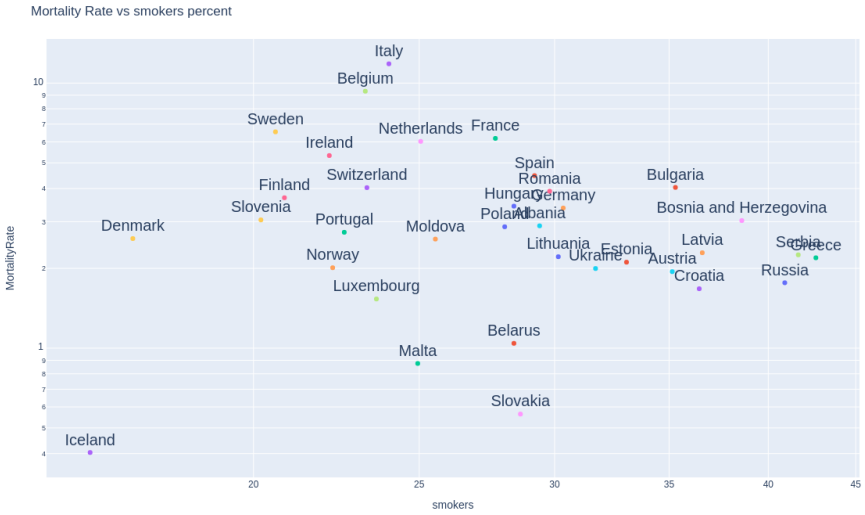
Διακρίνουμε αρνητική συσχέτιση μεταξύ θνησιμότητας και νοσοκομειακών κλινών και δαπανών για την υγεία ανα κάτοικο. Αυτό επιβεβαιώνει την υπόθεση ότι κρίσιμος παράγοντας στην επιβίωση των ασθενών με Covid-19 είναι η κατάσταση του συστήματος υγείας και η απρόσκοπτη πρόσβαση όλου του πληθυσμού στην περίθαλψη. Θετική συσχέτιση υπάρχει μεταξύ παθήσεων των πνευμόνων και θνησιμότητας. Οι ασθενείς με παθήσεις των πνευμόνων (ΧΑΠ, άσθμα κτλ.) θεωρούνται ευπαθείς ομάδες. Μια ενδιαφέρουσα συσχέτιση ειναι η αρνητική συσχέτιση ποσοστού καπνιστών με την θνησιμότητα. Θα εξετάσουμε το διάγραμμα διασποράς αυτών των δύο μεταβλητών (καπνιστές,θνησιμότητα) για να προσπαθήσουμε να εξηγήσουμε το αποτέλεσμα. Ακολουθούν ξανά τα διαγράμματα διασποράς για τα νοσοκομειακά κρεβάτια (σχ. 4.24), τις δαπάνες για την υγεία ανα κατοικό (σχ. 4.25) και το ποσοστό καπνιστών σε σχέση με την θνησιμότητα (σχ. 4.26).



Σχήμα 4.24: Διάγραμμα διασποράς θνησιμότητας και νοσοκομειακών κλινών ανα 1000 κατοίκους



Σχήμα 4.25: Διάγραμμα διασποράς θνησιμότητας και δαπανών για την υγεία.



Σχήμα 4.26: Διάγραμμα διασποράς θνησιμότητας και καπνιστών

Πράγματι, παρατηρούμε ότι χώρες με μεγάλη θνησιμότητα (Ιταλία, Ηνωμένο Βασίλειο) βρίσκονται στις χαμηλότερες θέσεις όσον αφορά τις νοσοκομειακές κλίνες ανα χίλιους κάτοικους. Όσον αφορά την αρνητική συσχέτιση μεταξύ καπνιστών και θνησιμότητας από το διάγραμμα διασποράς παρατηρούμε ότι τα υψηλότερα ποσοστά καπνιστών σημειώνονται στις Βαλκανικές Χώρες (Ελλάδα, Σερβία, Κροατία, Βοσνία κτλ) οι οποίες επλήγησαν λιγότερο από το πρώτο κύμα του Covid-19.

Κεφάλαιο 5

Προβλέψεις

5.1 Εισαγωγή

Το παρόν κεφάλαιο ασχολείται με τον τομέα των προβλέψεων της πορείας του Covid-19 στο μέλλον, δηλαδή σε ημερομηνίες πέραν αυτών που περιέχονται στα δεδομένα. Αυτό έχει εξαιρετικά μεγάλη σημασία έτσι ώστε να μπορεί να προβλεφθεί η πορεία της επιδημίας. Τα μοντέλα προβλέψεων παίζουν εξαιρετικά σημαντικό ρόλο στις αποφάσεις των χωρών για την λήψη ή την άρση περιοριστικών μέτρων ή για την εκτίμηση του αν το σύστημα υγείας θα αντέξει ή θα καταρρεύσει σε μια ενδεχόμενη αύξηση των κρουσμάτων. Στο συγκεκριμένο κεφάλαιο θα προσπαθήσουμε με την βοήθεια αλγορίθμων μηχανικής μάθησης/εξόρυξης δεδομένων να "προβλέψουμε" τον αριθμό των κρουσμάτων παγκόσμια αλλά και ανα χώρα.

Ο αριθμός των κρουσμάτων/θανάτων αποτελεί μια συνεχή μεταβλητή, οπότε η πρόβλεψη της εξέλιξής της ανάγεται σε ένα πρόβλημα παλινδρόμησης (regression). Στην ενότητα 5.3 θα εφαρμόσουμε μια σειρά αλγορίθμων - μοντέλων παλινδρόμησης και θα προσπαθήσουμε να προβλέψουμε τον αριθμό των κρουσμάτων τόσο σε παγκόσμιο επίπεδο όσο και σε επιμέρους χώρες. Οι Οι αλγόριθμοι που θα εφαρμοστούν είναι οι εξής: Γραμμική Παλινδρόμηση (Linear Regression), Πολυωνυμική Παλινδρόμηση (Polynomial Regression) και Παλινδρόμηση Διανυσματικής Υποστήριξης (Support Vector Machine (SVM) Regression). Έπειτα θα αξιολογηθούν συνολικά οι επιδόσεις του κάθε μοντελου με μέτρο την Ρίζα του Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error - RMSE).

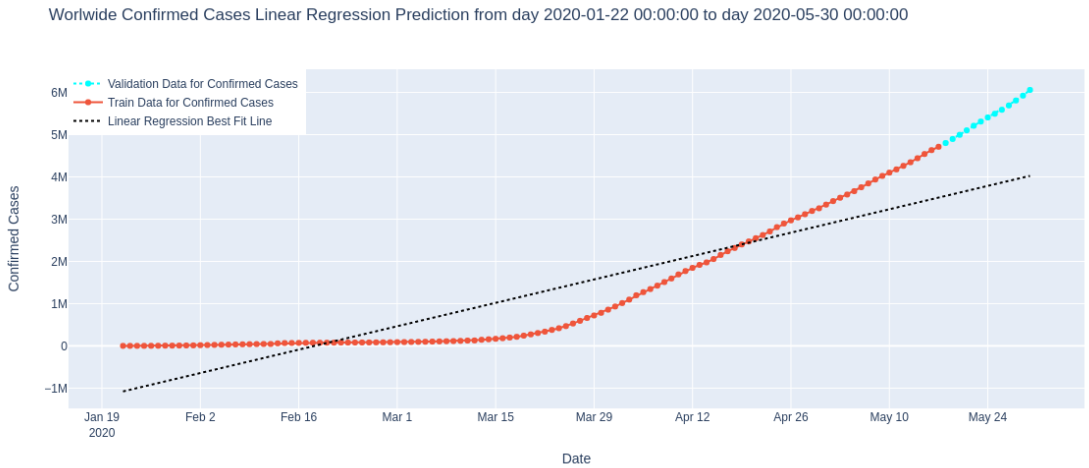
5.2 Δεδομένα

Για αυτό το κεφάλαιο χρησιμοποιήθηκε ξανά το σύνολο δεδομένων `covid_data_19.csv` [36] που ήταν και το βασικό σύνολο δεδομένων του προηγούμενου κεφαλαίου. Ως σύνολο εκπαίδευσης χρησιμοποιήθηκε το 90% (22/1 - 29/8) του αρχικού, ενώ το υπόλοιπο 10% (30/8-23/9) χρησιμοποιήθηκε ως σύνολο ελέγχου. Ο λόγος που δεν έγινε ο κλασσικός χωρισμός 80-20 είναι για να περιέχει το σύνολο εκπαίδευσης την αύξηση που παρουσίασαν τα κρούσματα στα τέλη του καλοκαιριού - αρχές φθινοπώρου.

5.3 Αλγόριθμοι Παλινδρόμησης

5.3.1 Γραμμική Παλινδρόμηση (Linear Regression)

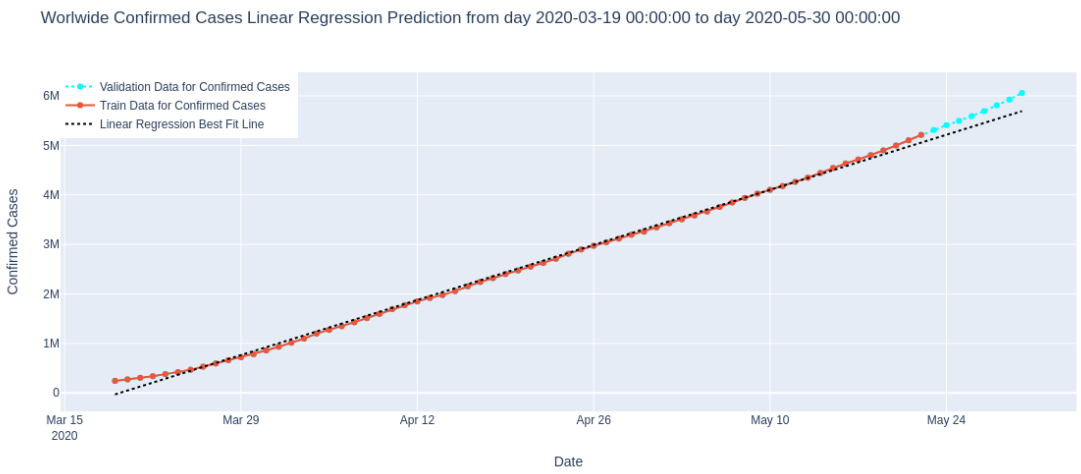
Ως πρώτη προσέγγιση χρησιμοποιείται ο αλγόριθμος της γραμμικής παλινδρόμησης (Linear Regression). Στην γραμμική παλινδρόμηση το μοντέλο θα προσπαθήσει να προσεγγίσει (fit) τα δεδομένα με μια γραμμική συνάρτηση - ευθεία. Στο Σχήμα 5.1 απεικονίζεται η προσπάθεια του μοντέλου να προσεγγίσει τα δεδομένα του συνόλου εκπαίδευσης και να προβλέψει την περαιτέρω εξέλιξη.



Σχήμα 5.1: Πρόβλεψη με χρήση του αλγορίθμου της γραμμικής παλινδρόμησης

Παρατηρούμε ότι το μοντέλο αποτυγχάνει πλήρως να προσεγγίσει τα δεδομένα, πράγμα λογικό εφόσον η γραφική παράσταση των κρουσμάτων δείχνει ότι δεν είναι μια γραμμική συνάρτηση αλλά μια καμπύλη που είναι αδύνατον να προσεγγιστεί με ακρίβεια από μια ευθεία. Στο τέλος των δεδομένων (23/9) τα πραγματικά επιβεβαιωμένα κρούσματα ήταν 31.77

εκατομμύρια ενώ το μοντέλο της γραμμικής παλινδρόμησης προβλέπει 21.44 εκατομύρια κρούσματα. Το σφάλμα RMSE που χρησιμοποείται και ως μέτρο απόδοσης του κάθε μοντέλου είναι: 840.055. Στο Σχήμα 5.2 το εύρος των ημερομηνιών μειώθηκε στο διάστημα από 15/3 μέχρι 30/5, στην περίοδο του πρώτου κύματος του κορονοϊού και το μοντέλο επανεκπαιδεύτηκε στο μικρότερο dataset.



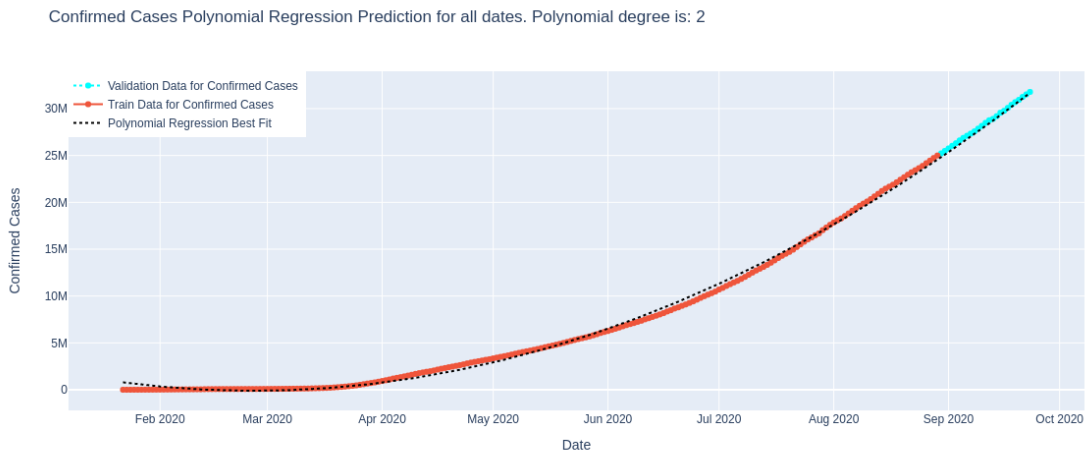
Σχήμα 5.2: Μοντέλο γραμμικής παλινδρόμησης στο διάστημα 15-3 έως 30-5

Βλέπουμε ότι η γραφική παράσταση των κρουσμάτων έχει αλλάξει σχεδόν σε μια ευθεία γραμμή με σταθερή κλίση. Σε αυτήν την περίπτωση το μοντέλο του γραμμικού παλινδρομητή αποδίδει αρκετά καλύτερα. Η πρόβλεψή του για τις 30/5 είναι 5.69 εκατομύρια κρούσματα, ενώ τα πραγματικά είναι 6.05 εκ. Το σφάλμα είναι 255.517 στην περίπτωση του μικρότερου dataset.

5.3.2 Πολυωνυμική παλινδρόμηση (Polynomial Regression)

Η πολυωνυμική παλινδρόμηση είναι μια ειδική περίπτωση της γραμμικής, όπου η εξίσωση που προσπαθεί να προσεγγίσει τα δεδομένα δεν είναι μια γραμμική συνάρτηση αλλά ένα πολυώνυμο βαθμού n. Στο συγκεκριμένο παράδειγμα, κατά την εκπαίδευση του μοντέλου δοκιμάστηκαν πολυωνυμικοί βαθμοί από το 1 έως το 20 και υπολογίστηκε το RMSE για κάθε βαθμό. Το μικρότερο σφάλμα προκύπτει για βαθμό πολυωνύμου 4 και έχει τιμή 302.994. Παρατηρούμε την αισθητή μείωση του σφάλματος συγκριτικά με το γραμμικό μοντέλο. Επίσης η πρόβλεψη της πολυωνυμικής παλινδρόμησης για την τελευταία ημερομηνία των δεδομένων είναι 31,67 εκ. κρούσματα ενώ ο πραγματικός αριθμός είναι 31,77 εκατομμύρια, δηλαδή διαφορά 100.000 κρουσμάτων. Παρακάτω, στο σχήμα 5.3 παρουσιάζεται η

εφαρμογή του μοντέλου της πολυωνυμικής παλινδρόμησης στα δεδομένα για όλες τις ημερομηνίες του συνόλου δεδομένων.

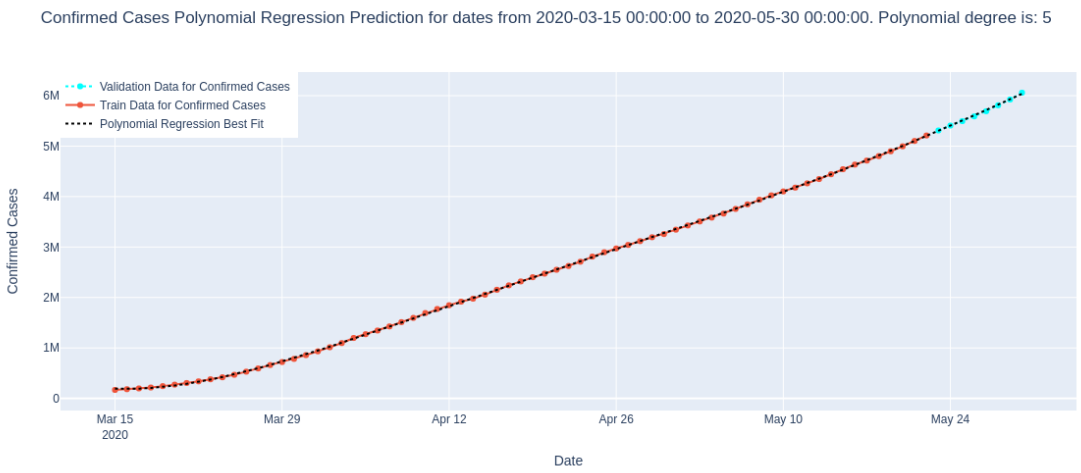


Σχήμα 5.3: Πολυωνυμική παλινδρόμηση

Παρατηρούμε πόσο καλύτερα προσεγγίζει το σύνολο εκπαίδευσης η πολυωνυμική καμπύλη, όπως επίσης και την ακρίβεια που επιτυγχάνεται στο σύνολο ελέγχου. Η καμπύλη της πρόβλεψης "ακολουθεί" την καμπύλη των πραγματικών κρουσμάτων. Μάλιστα αν "σπάσουμε" τα δεδομένα σε μικρότερα κομμάτια, η ακρίβεια που επιτυγχάνει η πολυωνυμική παρεμβολή βελτιώνεται ακόμα περισσότερο.

Στο Σχήμα 5.4 σπάμε ξανά τα δεδομένα στο κομμάτι από 15/3 έως 30/5 και εφαρμόζουμε το μοντέλο της πολυωνυμικής παλινδρόμησης. Βλέπουμε ότι η ακρίβεια προσέγγισης των δεδομένων αυξάνεται, και η πολυωνυμική παλινδρόμηση προβλέπει πολύ καλά την εξέλιξη των κρουσμάτων. Το σφάλμα RMSE μειώνεται σε 15.588 ενώ για την συγκεκριμένη περίπτωση επιλέγεται πολυώνυμου 5ου βαθμού. Συγκεκριμένα στις 30/5 το μοντέλο προβλέπει 6,0367 εκατομμύρια κρούσματα ενώ τα πραγματικά είναι 6,059 εκατομμύρια, διαφορά δηλαδή 22.300 κρουσμάτων (πολύ λιγότερο από τον αριθμό κρουσμάτων ανα ημέρα σε παγκόσμια κλίμακα).

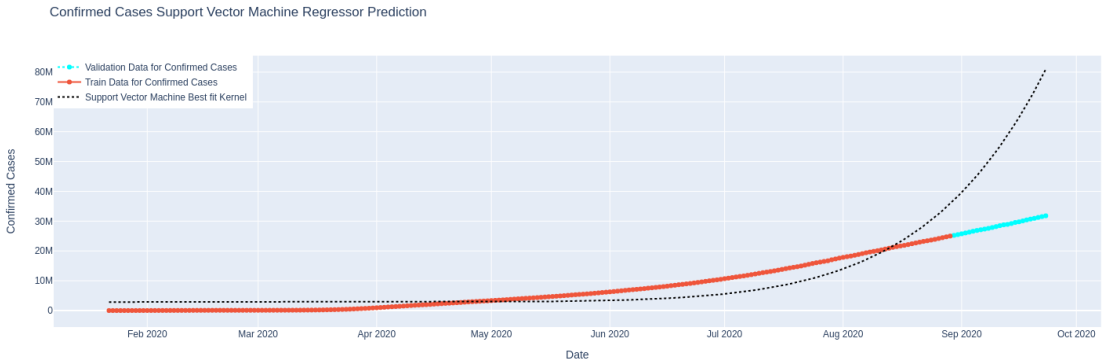
Συμπεραίνουμε ότι το μοντέλο της πολυωνυμικής παλινδρόμησης είναι καταλληλότερο για να περιγράψει τα δεδομένα μας από το απλό μοντέλο της γραμμικής. Επίσης μπορεί να χρησιμοποιηθεί ακριβώς λόγω της αυξημένης ακρίβειας για να προβλέψει (με ένα ανεκτό περιθώριο σφάλματος) την μελλοντική εξέλιξη των κρουσμάτων.



Σχήμα 5.4: Μοντέλο πολυωνυμικής παλινδρόμησης στο διάστημα 15-3 έως 30-5

5.3.3 Μηχανές Διανυσματικής Υποστήριξης - Support Vector Machines (SVM)

Ο αλγόριθμος SVM χρησιμοποιείται συνήθως ως αλγόριθμος ταξινόμησης, αλλά μπορεί να χρησιμοποιηθεί και σε προβλήματα παλινδρόμησης. Στο Σχήμα 5.5 παρουσιάζεται η εφαρμογή του αλγορίθμου. Παρατηρούμε ότι οι προβλέψεις του SVM παλινδρομητή αποκλίνουν δραματικά απο την πραγματική εξέλιξη των κρουσμάτων. Στην αξιολόγηση των αποτελεσμάτων το σφάλμα RMSE είναι 30.215.665 ενώ ο αλγόριθμος προβλέπει 80,962 εκατομμύρια κρούσματα με τα πραγματικά να είναι 31.77 εκατομμύρια. Η απόκλισή του είναι πολύ μεγάλη και ο αλγοριθμός αδυνατεί να προσεγγίσει σωστά τα δεδομένα άρα δεν μπορεί να χρησιμοποιηθεί για πραγματικές προβλέψεις.



Σχήμα 5.5: Μοντέλο Μηχανών Διανυσματικής Υποστήριξης

5.3.4 Μελλοντικές προβλέψεις και συνολική αξιολόγηση των μοντέλων

Σε αυτήν την ενότητα θα συνοψιστούν τα αποτελέσματα και των τριών μοντέλων που εφαρμόστηκαν στα δεδομένα. Επίσης θα εφαρμοστούν ξανά οι αλγορίθμοι ώστε να επεκτείνουμε τις προβλέψεις τους σε ημερομηνίες πέραν της 23/09 και συγκεκριμένα μέχρι την ημερομήνια της 1/11/2020. Στον πίνακα που ακολουθεί βλέπουμε τα συγκεντρωτικά αποτελέσματα των τριών αλγορίθμων. Το μοντέλο που υλοποιήθηκε με χρήση του αλγορίθμου της πολυωνυμικής παρεμβολής καταγράφει το μικρότερο σφάλμα RMSE και την μεγαλύτερη ακρίβεια στην πρόβλεψη των κρουσμάτων στην τελευταία ημερομηνία των δεδομένων.

Αξιολόγηση αλγορίθμων

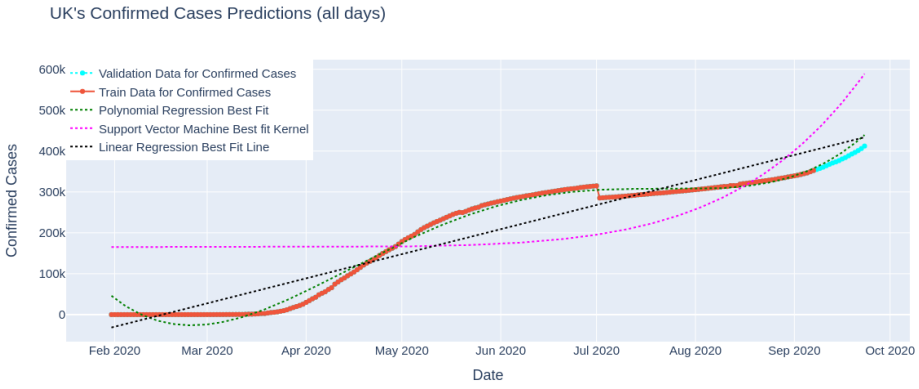
Αλγόριθμος	Σφάλμα RMSE	Πρόβλεψη(23/9)	Κρούσματα(23/9)
Γραμμική Παλινδρόμηση	8.400.055	21.445.260	31.779.840
Πολυωνυμική Παλινδρόμηση	302.994	31.674.290	31.779.840
SVM Παλινδρόμηση	30.215.665	80.962.730	31.779.840

Στο Σχήμα 5.6 παρουσιάζεται η πρόβλεψη των μοντέλων που υλοποιήθηκαν για το διάστημα 24/9 έως 1/11, πέρα απο την τελευταία ημερομηνία των δεδομένων. Ο αλγόριθμος πολυωνυμικής παλινδρόμησης προβλέπει για την 1/11 σύνολο κρουσμάτων 44,235 εκατομμύρια κρούσματα. Η ιστοσελίδα Worldometer δηλώνει για την συγκεκριμένη ημερομηνία 43,8 εκατομμύρια κρούσματα. Παρατηρούμε λοιπόν ότι ακόμα και για διάστημα μεγαλύτερο του ενός μήνα, ο αλγόριθμος διατηρεί την αξιοπιστία του.

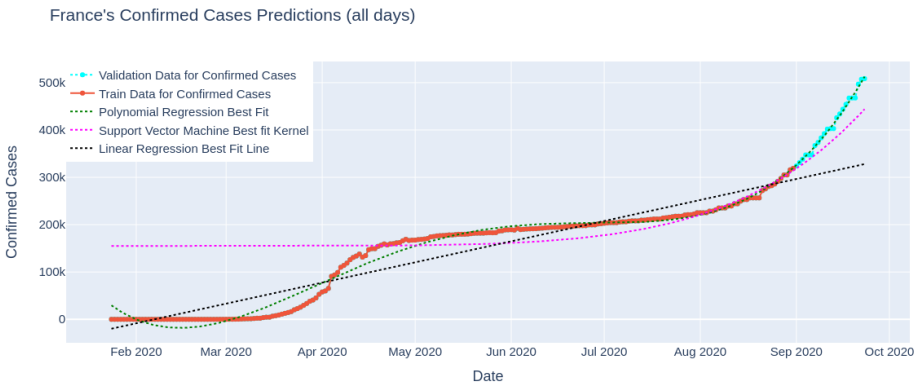
Σχήμα 5.6: Πρόβλεψη και των τριών αλγορίθμων μέχρι την 1/11

5.3.5 Εφαρμογή σε επίπεδο χώρας

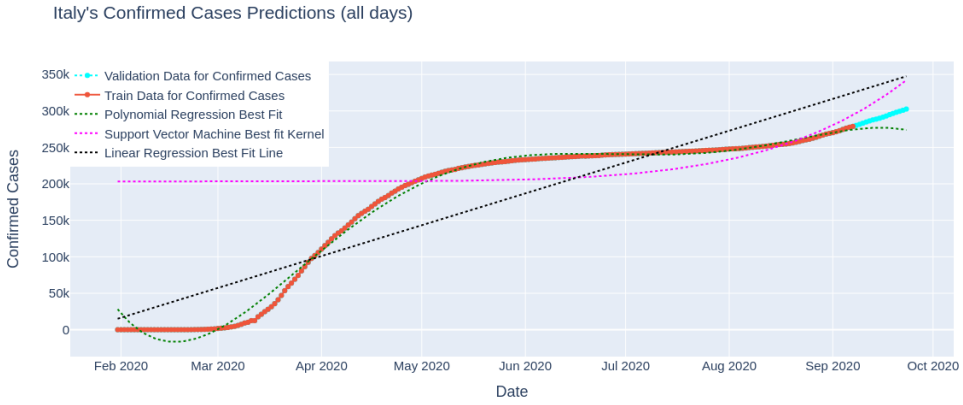
Σε αυτήν την υποενότητα θα εφαρμοστούν οι παραπάνω αλγόριθμοι προβλέψης σε επίπεδο χώρας και συγκεκριμένα στις πέντε χώρες με τα περισσότερα κρούσματα στην Ευρώπη (Μεγάλη Βρετανία, Γαλλία, Ρωσία, Ιταλία και Ισπανία). Στα παρακάτω σχήματα φαίνονται τα αποτελέσματα που είχε κάθε αλγόριθμος σε κάθε χώρα. Παρατηρούμε οτι ο αλγόριθμος της πολωνυμικής παλινδρόμησης (πράσινη διακεκομένη γραμμή) σε γενικές γραμμές προσεγγίζει και πάλι καλύτερα τα δεδομένα από τους υπόλοιπους και είναι ακριβέστερος στις προβλέψεις που αφορούν το σύνολο ελέγχου. Εξαίρεση αποτελεί η περίπτωση της Ιταλίας (Σχήμα 5.9) όπου βλέπουμε ότι ο αλγόριθμος της πολυωνιμικής παρεμβολής στο κομμάτι του συνόλου ελέγχου αντί να συνεχίσει την ανοδική πορεία αρχίζει και φθίνει προβλέποντας μειωμένο αριθμό κρουσμάτων. Αυτό είναι τυπικό παράδειγμα υπερπροσαρμογής(overfitting) που ενώ στα δεδομένα εκπαίδευσης το μοντέλο συμπεριφέρεται σωστά, αποτυγχάνει να κάνει το ίδιο σε άγνωστα δεδομένα.



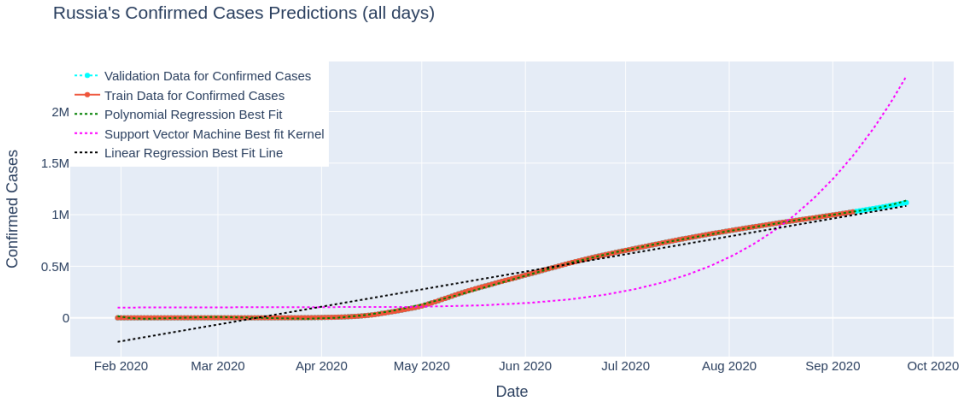
Σχήμα 5.7: Εφαρμογή και των τριών αλγορίθμων στο Ηνωμένο Βασίλειο



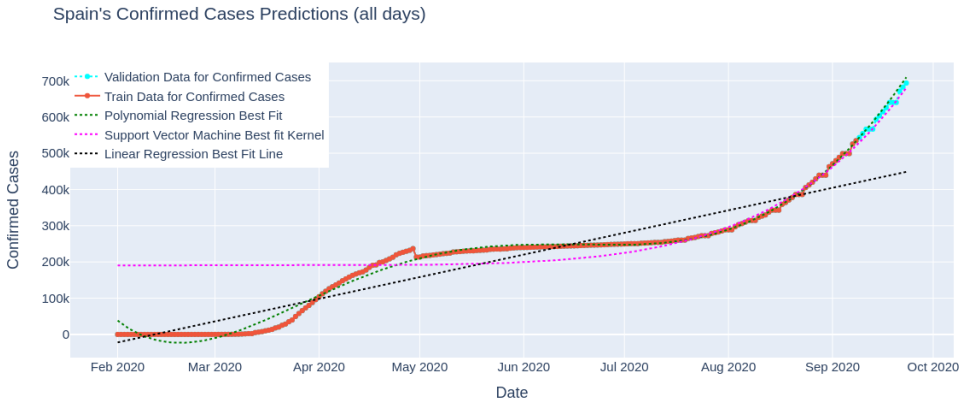
Σχήμα 5.8: Εφαρμογή και των τριών αλγορίθμων στην Γαλλία



Σχήμα 5.9: Εφαρμογή και των τριών αλγορίθμων στην Ιταλία



Σχήμα 5.10: Εφαρμογή και των τριών αλγορίθμων στην Ρωσία



Σχήμα 5.11: Εφαρμογή και των τριών αλγορίθμων στην Ισπανία

5.3.6 Συμπεράσματα

Στο κεφάλαιο αυτό εφαρμόστηκαν αλγόριθμοι παλινδρόμησης και μοντέλα χρονοσειρών με στόχο να προβλεφθούν τα κρούσματα του Covid-19 σε μελλοντικό χρόνο. Ο αλγόριθμος της πολυωνυμικής παρεμβολής έδειξε αρκετά καλή ακρίβεια στην πρόβλεψη, τόσο σε παγκόσμιο επίπεδο όσο και σε επίπεδο χώρας. Οι προβλέψεις του, ακόμα και σε βάθος ενός μήνα, προσεγγίζουν αρκετά καλά τα πραγματικά δεδομένα και μπορούν να χρησιμοποιηθούν από τους επιστήμονες για να έχουν μια εικόνα του μέλλοντος, ώστε να να εκτιμήσουν τα μέτρα που θα εφαρμοστούν κτλ. Το έλειμμα που παρουσιάζουν οι συγκεκριμένοι αλγόριθμοι είναι το εξής: δεν μπορούν να εκτιμήσουν το πότε η καμπύλη θα καμθφεί ή θα αρχίσει να αυξάνεται. Βασιζόμενοι μόνο στο παρελθόν και μη εκτιμώντας ως παραμέτρους την επίδραση που έχουν τα περιοριστικά μέτρα στην ανακοπή της εξάπλωσης του ιού, στα συγκεκριμένα παραδείγματα που είδαμε η εκτίμηση των μοντέλων είναι οτι τα κρούσματα θα συνεχίσουν να αυξάνονται επ' αόριστον. Γνωρίζουυε όμως ότι από τα τέλη Οκτωβρίου αρκετές Ευρωπαϊκές χώρες, εφάρμοσαν μέτρα αντίστοιχα με αυτά του πρώτου κύματος (lockdown, αναστολή πτήσεων κτλ) και σε κάποιες από αυτές η αύξηση των κρουσμάτων έχει ήδη αρχίσει να μειώνεται. Συμπερασματικά, οι αλγόριθμοι αυτοί μπορούν να μας δώσουν (σχετικά) αξιόπιστα αποτελέσματα για βραχυπρόθεσμες προβλέψεις αλλά αδυνατούν να προβλέψουν το σημείο που θα αρχίσουν τα κρούσματα να αυξάνονται ή να μειώνονται επηρεασμένα από εξωγενείς παράγοντες.

Κεφάλαιο 6

Πρόγνωση Ασθενούς

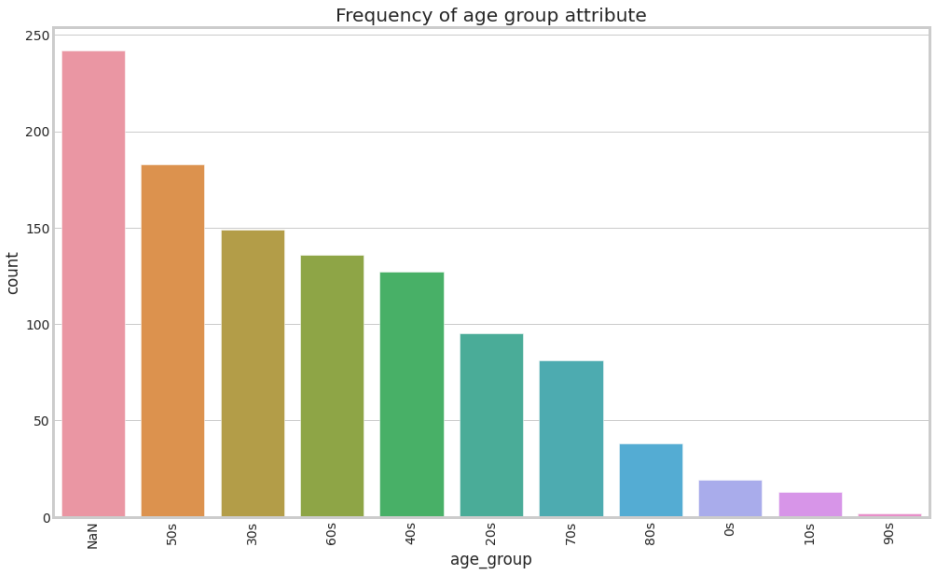
6.1 Εισαγωγή

Στο κεφάλαιο 6 θα μελετήσουμε την χρησιμότητα των αλγορίθμων κατηγοριοποίησης (classification) σε δεδομένα που αφορούν ασθενείς του κορονοϊού. Θα προσπαθήσουμε να προβλέψουμε αν ο ασθενής είναι θετικός ή όχι στον Covid-19, βασισμένοι σε δεδομένα που περιέχουν κλινικές πληροφορίες ασθενών. Επίσης θα γίνει προσπάθεια να να προβλέψουμε αν οι ασθενείς θα επιβιώσουν ή θα πεθάνουν βασισμένοι σε δεδομένα που αφορούν την ηλικία τους, τα συμπτώματά τους, το φύλο τους κτλ.

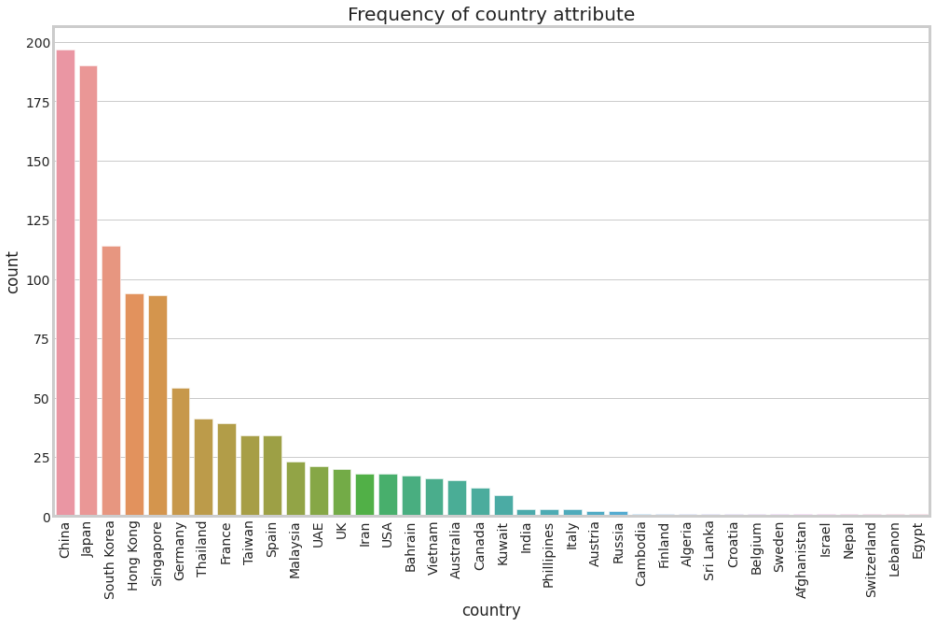
6.2 Πρόγνωση αποτελέσματος ασθενή

Το παρόν υποκεφάλαιο ασχολείται με την πρόβλεψη της εξέλιξης της υγείας των ασθενών που έχουν προσβληθεί από Covid-19 και συγκεκριμένα αν θα ξεπεράσουν την ασθένεια ή αν θα καταλήξουν από αυτήν. Τα δεδομένα που χρησιμοποιήθηκαν είναι τα ίδια με αυτά που χρησιμοποίησαν οι συγγραφείς του άρθρου [40] και αποτελεί μια επεξεργασμένη έκδοση του συνόλου δεδομένων που είναι δημοσιευμένο στο αποθετήριο Kaggle με τίτλο "Novel Corona Virus 2019 Dataset". Οι πηγές του είναι από τον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ - WHO) και το πανεπιστήμιο John Hopkins. Τα δεδομένα του αφορούν ασθενείς του πρώτου κύματος.

Εξερευνώντας τα δεδομένα, στα παρακάτω σχήματα βλέπουμε την ηλικιακή κατανομή των δεδομένων (σχ. 6.1), τις χώρες από τις οποίες προήλθαν οι ασθενείς (σχ. 6.2) καθώς και την συχνότητα των συμπωμάτων που παρουσίασαν (σχ. 6.3). Παρατηρώντας τα διαγράμ-



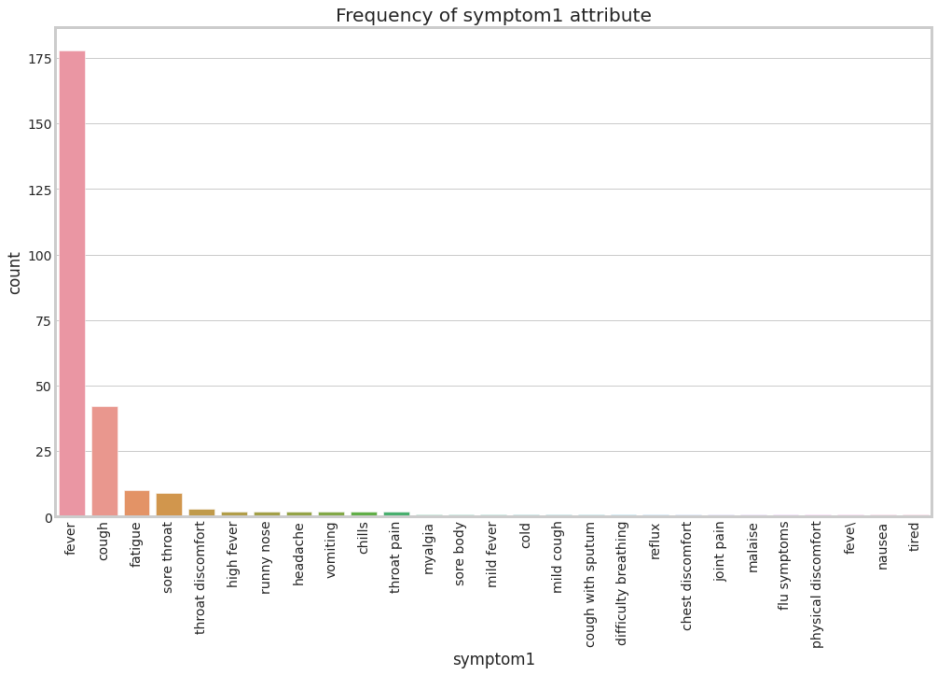
Σχήμα 6.1: Ηλικιακή κατανομή των ασθενών



Σχήμα 6.2: Συχνότερες χώρες καταγωγής των ασθενών

ματα βλέπουμε ότι οι περισσότεροι ασθενείς βρίσκονται στην ηλικιακή ομάδα των 50, ακολουθούμενοι από την ηλικιακή ομάδα των 30, με βασικές χώρες προέλευσης την Κίνα, την Ιαπωνία και τη Ν. Κορέα, ενώ τα συχνότερα συμπτώματα είναι ο πυρετός ακολουθούμενος από βήχα,κόπωση και πονόλαιμο.

Τα δεδομένα επεξεργάστηκαν περαιτέρω για να γίνουν κατάλληλα για τα μοντέλα που θα εφαρμοστούν. Συγκεκριμένα, όλα τα μη-αριθμητικά χαρακτηριστικά μετατράπηκαν σε αριθμητικά με την τεχνική του Label Encoding, όπου σε κάθε διακριτό χαρακτηριστικό ανα-



Σχήμα 6.3: Συχνότητα συμπτωμάτων των ασθενών

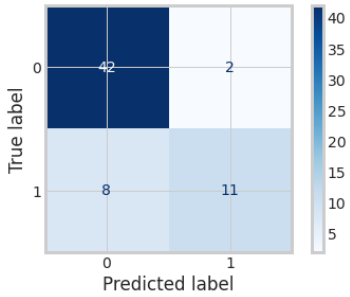
τίθεται ένας αριθμός. Επίσης, δημιουργήθηκε μια νέα μεταβλητή, η `diff_sym_hos`, που δηλώνει τον αριθμό των ημερών που περάσαν από την ημέρα εμφάνισης των συμπτωμάτων του ασθενούς μέχρι την ημέρα που προσήλθε στο νοσοκομείο. Τέλος, κρατήθηκαν μόνο οι εγγραφές του συνόλου δεδομένων που είχαν γνωστή την μεταβλητή - στόχο `death`, δηλαδή ήταν γνωστή η έκβαση της νοσηλείας του ασθενή. Το τελικό dataset έχει μέγεθος 207 γραμμές και 13 γνωρίσματα (207x13).

Το σύνολο δεδομένων χωρίστηκε σε συνολα εκπαίδευσης και ελέγχου με αναλογία 70-30 και εφαρμόστηκαν οι εξής αλγόριθμοι κατηγοριοποίησης: Δέντρα Απόφασης (Decision Trees), Ταξινομητής Μηχανών Διανυσματικής Υποστήριξης (Support Vector Machine - SVM), Ταξινομητής Κοντινότερου Γείτονα (K Nearest Neighboor Classifier - KNN) καθώς και οι μετα-αλγόριθμοι (ensemble) Τυχαίου Δάσους (Random Forest) και Adaboost.

Για κάθε αλγόριθμο που χρησιμοποιήθηκε, σχεδιάστηκε ο πίνακας σύγχυσης (confusion matrix) και βάσει αυτού εξήχθησαν ως μέτρα απόδοσης τα εξής: accuracy, recall , precision και F1 score. Επειδή το dataset δεν έχει ισορροποημένο αριθμό κλάσεων (οι ασθενείς που επιβιώνουν είναι πολύ περισσότεροι από αυτούς που πεθαίνουν), ως μέτρο σύγκρισης μεταξύ των αλγορίθμων επιλέχθηκε το F1 score που είναι το σταθμισμα του recall και του precision.

Δέντρα Απόφασης (Δ.Α.)

Από την εφαρμογή του αλγορίθμου των Δέντρων Απόφασης προκύπτει ο εξής πίνακας σύγχυσης (σχ. 6.4):



Σχήμα 6.4: Πίνακας Σύγχυσης για τα Δέντρα Απόφασης

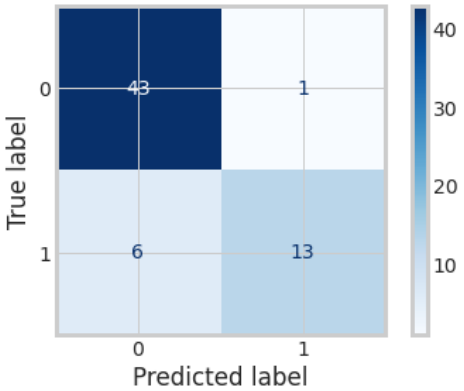
Στην περίπτωση του αλγορίθμου των Δέντρων Απόφασης, προβλέφθηκαν σωστά 42 ασθενείς που επιβίωσαν ενώ προβλέφθηκαν λάθος δύο ασθενείς. Στους ασθενείς που πέθαναν, 11 προβλέφθηκαν σωστά και οκτώ λάθος. Προκύπτουν τα ακόλουθα μέτρα απόδοσης του αλγορίθμου.

Μετρα Απόδοσης : Δ.Α.

Recall	0.77
Precision	0.84
F1 Score	0.79
Accuracy	0.84

Μηχανές Διανυσματικής Υποστήριξης (SVM)

Από την εφαρμογή του αλγορίθμου SVM προκύπτει ο εξής πίνακας σύγχυσης(σχ. 6.5):



Σχήμα 6.5: Πίνακας σύγχυσης για τον αλγόριθμο SVM

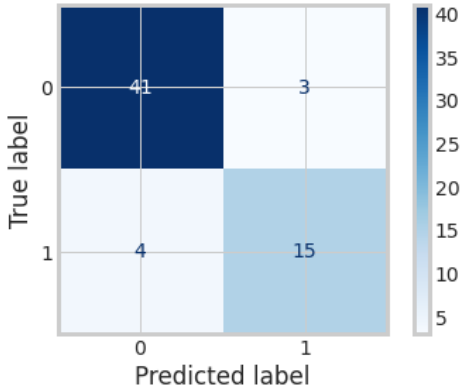
Ο αλγόριθμος SVM ταξινόμησε σωστά 43 ασθενείς που επιβίωσαν και λάθος έναν. Στους ασθενείς που πέθαναν ταξινόμησε σωστά 13 περιπτώσεις και λάθος έξι. Τα μέτρα απόδοσής του έχουν ως εξής:

Μετρα Απόδοσης: SVM

Recall	0.83
Precision	0.90
F1 Score	0.86
Accuracy	0.89

Κ Κοντινότεροι Γείτονες (KNN)

Από την εφαρμογή του αλγορίθμου KNN προκύπτει ο εξής πίνακας σύγχυσης (σχ. 6.6):



Σχήμα 6.6: Πίνακας σύγχυσης για τον αλγόριθμο KNN

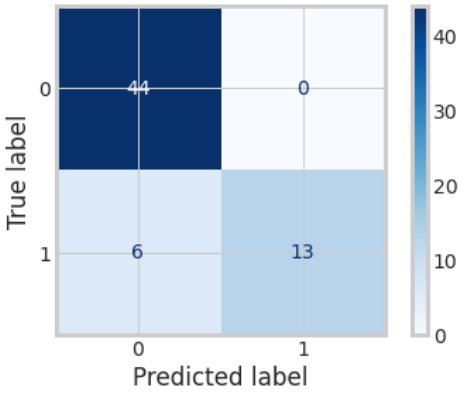
Ο αλγόριθμος KNN ταξινομεί σωστά 41 ασθενείς που επιβίωσαν και τρεις λάθος, ενώ στους ασθενείς που πέθαναν ταξινομεί 15 σωστά και τέσσερις λάθος. Τα μέτρα απόδοσής του έχουν ως εξής:

Μέτρα αποδοσης: KNN

Recall	0.86
Precision	0.87
F1 Score	0.87
Accuracy	0.89

Τυχαίο Δάσος (Random Forest)

Από την εφαρμογή του αλγορίθμου του Τυχαίου Δάσους προκύπτει ο ακόλουθος πίνακας σύγχυσης (σχ. 6.7):



Σχήμα 6.7: Πίνακας σύγχυσης Τυχαίου Δάσους

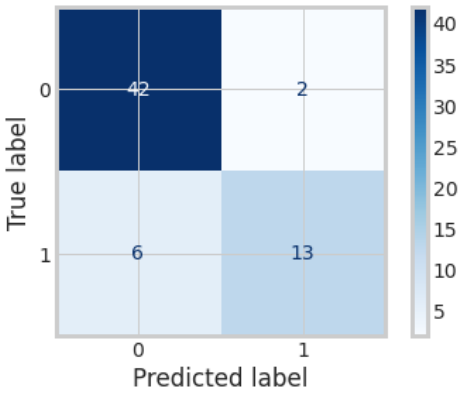
Ο αλγόριθμος πετυχαίνει σωστά την πρόβλεψη για όλους τους ασθενείς που επιβίωσαν (44) ενώ στην κλάση των ασθενών που πέθαναν προβλέπει 13 σωστά και έξι λάθος. Τα μέτρα απόδοσης του φαίνονται παρακάτω:

Μέτρα Απόδοσης: Τ.Δ.

Recall	0.84
Precision	0.94
F1 Score	0.87
Accuracy	0.90

Adaboost (σε Δέντρα Απόφασης)

Από την κατηγοριοποίηση με τον Adaboost προέκυψε ο ακόλουθος πίνακα σύγχυσης (σχ. 6.8):



Σχήμα 6.8: Πίνακας σύγχυσης Adaboost

Ο αλγόριθμος Adaboost ταξινομεί σωστά 42 ασθενείς που επιβίωσαν και 2 λάθος, ενώ

στους ασθενείς που πέθαναν ταξινομεί 13 σωστά και έξι λάθος. Τα μέτρα απόδοσης του φαίνονται παρακάτω:

Μέτρα Απόδοσης: Adaboost

Recall	0.82
Precision	0.87
F1 Score	0.84
Accuracy	0.87

6.2.1 Σύγκριση όλων των αλγορίθμων

Παίρνοντας υπόψιν όλα τα μέτρα που παρατέθηκαν παραπάνω, με βασικό μέτρο σύγκρισης το F1-score, την καλύτερη απόδοση την έχει ο αλγόριθμος Τυχαίου Δάσους με F1-score **0.87** και πετυχαίνοντας μεγαλύτερα νούμερα στα υπόλοιπα μέτρα (accuracy, recall, precision). Όλοι οι αλγόριθμοι κινήθηκαν σε κοντινά επίπεδα με εξαίρεση τον αλγόριθμο του Δέντρου Απόφασης όπου η απόδοση του ήταν αισθητά χαμηλότερη από τους υπόλοιπους.

Συγκεντρωτικά F1 Scores

Δέντρα Απόφασης	0.79
SVM	0.86
KNN	0.87
Τυχαίο Δάσος	0.87
Adaboost	0.84

Εφόσον επιλέχθηκε ο αλγόριθμος Τυχαίου Δάσους ως ο καταλληλότερος από τους αλγορίθμους που δοκιμάστηκαν, μπορούμε να προσπαθήσουμε να ρυθμίσουμε τις υπερπαραμέτρους του ώστε να βελτιώσουμε περισσότερο την απόδοσή του. Με την βοήθεια της βιβλιοθήκης SVgrid δοκιμάστηκαν διάφοροι συνδυασμοί παραμέτρων με στόχο την μεγιστοποίηση του F1-score στα δεδομένα εκπαίδευσης. Ο συνδυασμός παραμέτρων που επέλεξε η αυτόματη βελτιστοποίηση είναι ο εξής:

- class_weight: 1: 2
- max_depth: 10
- max_features: 0.6
- n_estimators: 100

Αυτός ο συνδυασμός παραμέτρων πέτυχε F1-score 92.82 στο σύνολο εκπαίδευσης. Παρόλη την αύξηση όμως του μέτρου F1 στα δεδομένα εκπαίδευσης, τα αποτελέσματα του τρεξίματος στο σύνολο ελέγχου παρέμειναν ακριβώς ίδια με αυτά πριν την βελτιστοποίηση.

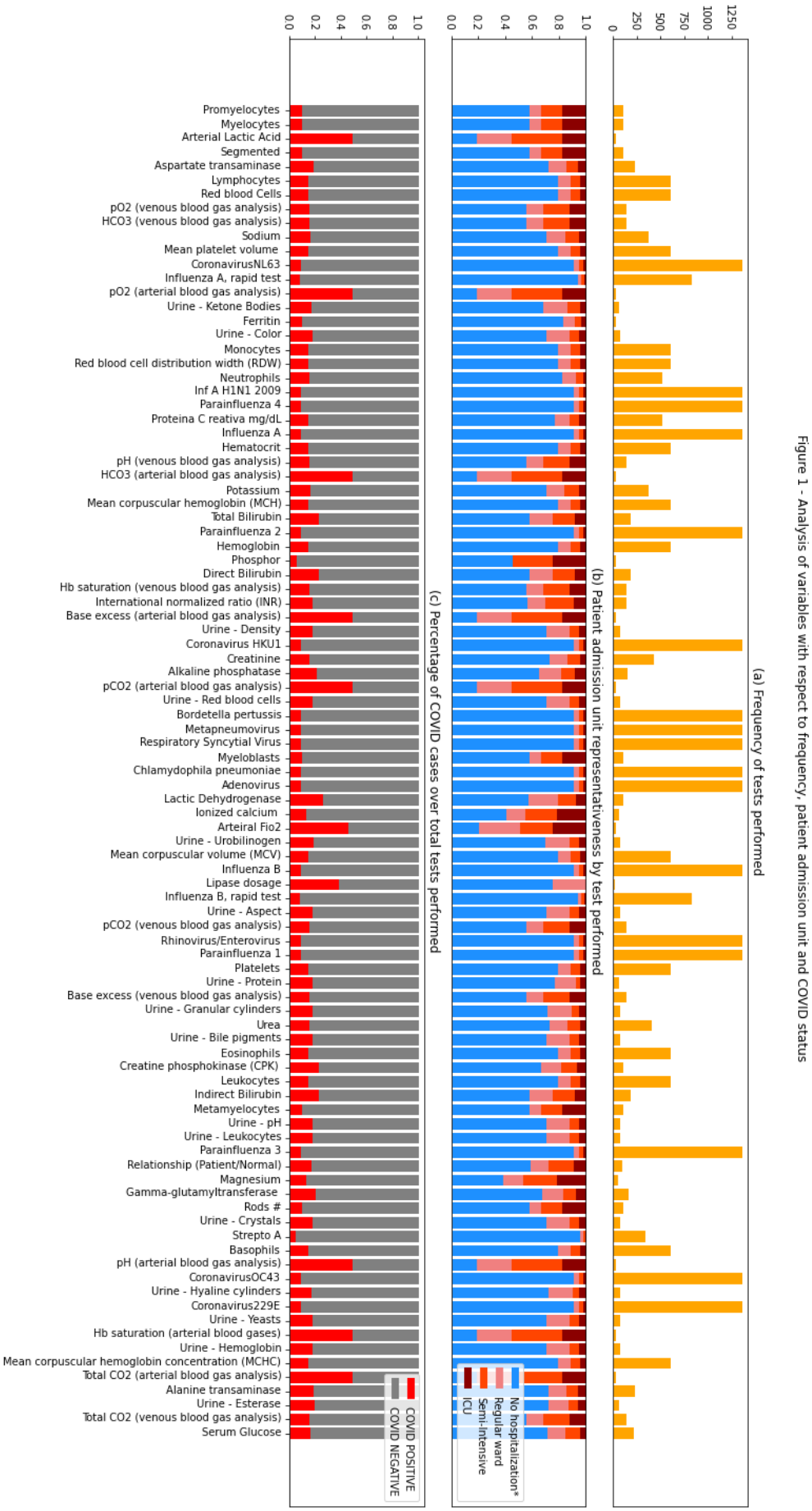
6.3 Πρόβλεψη θετικότητας ασθενούς βάσει κλινικών εξετάσεων

Σε αυτήν την υποενότητα θα γίνει προσπάθεια να προβλεφθεί αν ο ασθενής είναι θετικός ή όχι στον Covid-19 με βάση τις ιατρικές εξετάσεις που διενεργούν τα νοσοκομεία όταν ο ασθενής εισάγεται με συμπτώματα της νόσου. Τα δεδομένα που χρησιμοποιήθηκαν περιέχει ανώνυμα δεδομένα ασθενών από το νοσοκομείο "Israelita Albert Einstein" στο Σάο Πάολο της Βραζιλίας [41]. Περιέχει πριν την επεξεργασία 5664 εγγραφές και 111 μεταβλητές που περιέχουν το αποτέλεσμα του μοριακού (PCR) τέστ Covid-19 και μια σειρά άλλων εξετάσεων (αιματολογικές, ούρων κτλ) καθώς και το επίπεδο νοσηλείας του ασθενούς (κανονική κλίνη, εντατική, ημι-εντατική).

Προεπεξεργασία Δεδομένων

Κατά την προεπεξεργσία των δεδομένων διαπιστώθηκε ότι το 88,1% του συνόλου δεδομένων είναι κενές τιμές (NaN). Επίσης ενα μεγάλο κομμάτι των τιμών έχει κατηγορικές τιμές τύπου συμβολοσειράς (string). Όλα αυτά πρέπει να μετατραπούν σε αριθμητικά δεδομένα (numerical data) για να μπορέσουν να εφαρμοστούν οι αλγόριθμοι. Αυτό επιτυγχάνεται εφαρμόζοντας Label Encoding, και πλέον τα δεδομένα περιέχουν μόνο αριθμητικές τιμές. Έπειτα βρίσκουμε τις εξετάσεις εκείνες οι οποίες δεν διενεργήθηκαν σε θετικούς ασθενείς και τις διαγράφουμε από το dataset σε μια προσπάθεια να μειώσουμε τις διαστάσεις του. Μετά την αφαίρεση αυτών την μεταβλητών, το πλήθος των μεταβλητών μειώνεται σε 95.

Στο σχήμα 6.9 βλέπουμε την συχνότητα κάθε εξέτασης, την σχέση κάθε εξέτασης με την πιθανότητα να νοσηλευθεί ο ασθενής σε ΜΕΘ καθώς και την σχέση κάθε εξέτασης με το αποτέλεσμα του ασθενούς στο PCR τεστ (θετικό ή αρνητικό). Κάποια πρόχειρα συμπεράσματα που μπορούν να εξαχθούν από το διάγραμμα είναι ότι οι εξετάσεις που πραγματοποιούνται με μικρότερη συχνότητα (σπανιότερες) σχετίζονται με υψηλότερη πιθανότητα ο ασθενής να είναι θετικός ή/και να νοσηλευτεί σε ΜΕΘ.

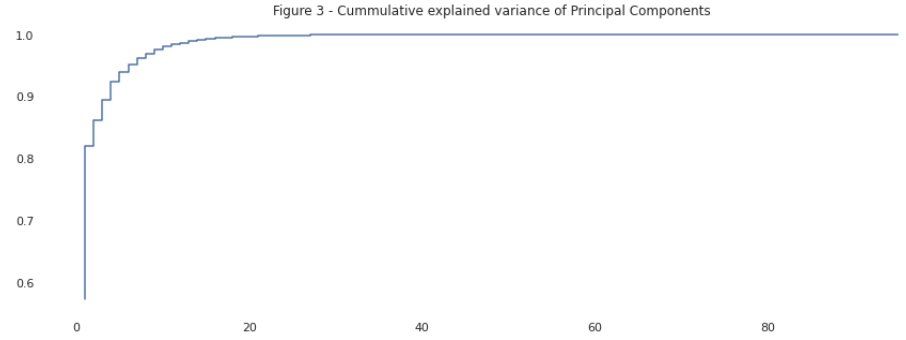


Σχήμα 6.9: Συχνότητα εξετάσεων και συσχέτιση με το αποτέλεσμα των τεστ

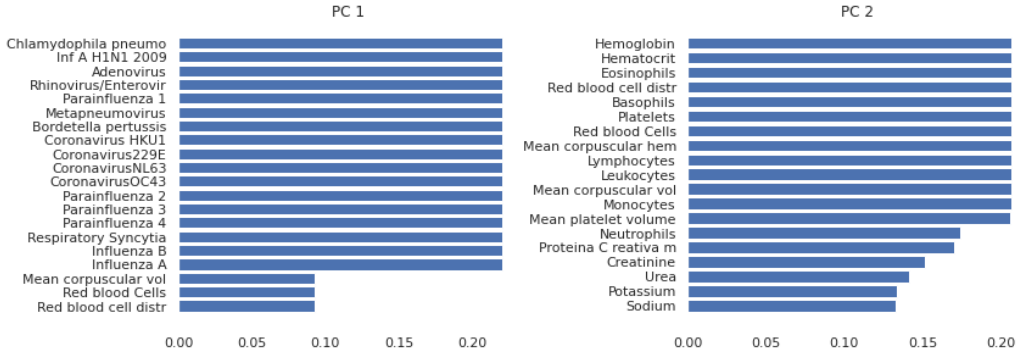
6.3.1 Ανάλυση Κύριων Συνιστωσών - Principal Components Analysis (PCA)

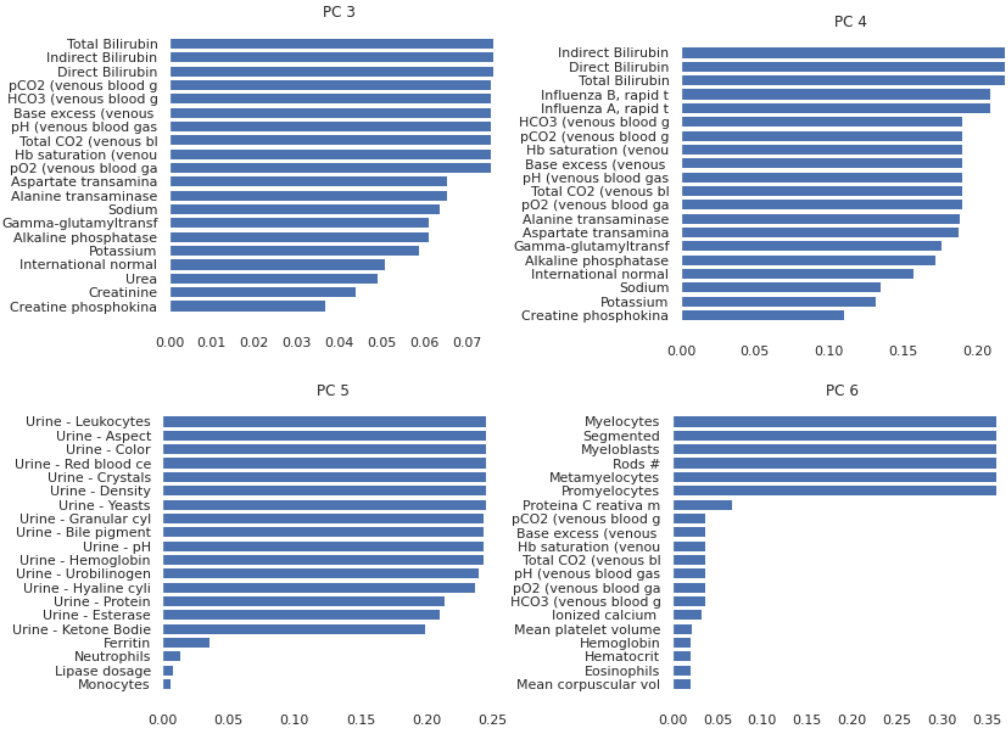
Το πρόβλημα των μεγάλων διαστάσεων του συνόλου δεδομένων καθώς και των πολλών κενών τιμών συνεχίζει να υφίσταται. Αν επιλέξουμε απλά να διαγράψουμε από τα δεδομένα τις γραμμές που περιέχουν κενές τιμές, το σύνολο θα μικρύνει υπερβολικά. Οπότε θα δοκιμαστεί η μέθοδος της Ανάλυσης Κύριων Συνιστωσών (ΑΚΣ - PCA) για εύρεση των κύριων συνιστωσών στο σύνολο δεδομένων. Στην ουσία με την Ανάλυση Κύριων Συνιστωσών θα προσπαθήσουμε να βρούμε ποιες εξετάσεις συνήθως γίνονται μαζί (coocurrence).

Το σχήμα 6.10 δείχνει την αύξηση της πληροφορίας (explained variance) σε σχέση με τις Κύριες Συνιστώσες (ΚΣ). Παρατηρούμε ότι μετά την έκτη ΚΣ η πληροφορία αυξάνει με πολύ αργό ρυθμό. Άρα θεωρούμε ότι οι έξι ΚΣ είναι αντιπροσωπευτικό δείγμα για τα δεδομένα. Στο Σχήμα 6.11 απεικονίζονται τα σημαντικότερα γνωρίσματα (features) για κάθε μία από τις έξι πρώτες Κύριες Συνιστώσες.



Σχήμα 6.10: Σχέση μεταξύ πληροφορίας και κύριων συνιστωσών





Σχήμα 6.11: Οι έξι Κύριες Συνιστώσες του συνόλου δεδομένων

Μια πρώτη ανάγνωση του σχήματος 6.11 μας δίνει την παρακάτω ερμηνεία όσον αφορά την κάθε ΚΣ.

- PC1 : Αποτελέσματα τεστ για γρίπη (influenza) και άλλους κορονοϊούς και γενικά ιούς του αναπνευστικού συστήματος.
- PC2 : Αιματολογικές εξετάσεις (λευκά και ερυθρά αιμοσφαίρια)
- PC3 : Αέρια αίματος και εξετάσεις που αφορούν το συκώτι
- PC4 : Εξετάσεις που αφορούν τα νεφρά
- PC5 : Εξετάσεις ούρων
- PC6 : Εξετάσεις που αφορούν μυελό των οστών

Με βάση την ΑΚΣ και εφόσον δεν μπορούμε να υποκαταστήσουμε τις τιμές που λείπουν για παράδειγμα με την μέση τιμή ή με μηδέν επειδή κάτι τέτοιο θα αλλοίωνε εξαιρετικά τα αποτελέσματα, θα προσπαθήσουμε να εξετάσουμε ποιες ΚΣ μας δίνουν ένα ικανό τμήμα του συνόλου δεδομένων με το οποίο μπορούμε να εργαστούμε. Η ΚΣ 2 που περιέχει αιματολογικές εξετάσεις, όταν διαγράψουμε τις κενές τιμές καταλήγει σε ένα σύνολο με 420 γραμμές

και 19 μεταβλητές. Από την ΚΣ 1 που περιέχει τα τεστ ιών του αναπνευστικού προκύπτει ένα σύνολο δεδομένων με 1352 γραμμές και 20 μεταβλητές. Τα υπόλοιπα σετ εξετάσεων που προέκυψαν από την ΑΚΣ μας δίνουν πολύ μικρά σύνολα δεδομένων όταν διαγράψουμε τις κενές τιμές. Άρα θα προχωρήσουμε στην κατηγοριοποίηση βασίζόμενοι στις αιματολογικές εξετάσεις (ΚΣ 2) και στα τεστ που αφορούν τους ιούς του αναπνευστικού συστήματος (ΚΣ 1).

6.3.2 Κατηγοριοποίηση

Σε αυτήν την ενότητα θα εφαρμοστούν μοντέλα κατηγοριοποίησης, ξεχωριστά σε κάθε μια απο τις δύο ομάδες εξετάσεων που ξεχωρίσαμε στην προηγούμενη ενότητα. Θα εφαρμόσουμε τους αλγορίθμους των Δεντρων Απόφασης, Κ-κοντινότερους Γείτονες (KNN), Μηχανές Διανυσματικής Υποστήριξης (SVM) καθώς και μετα-αλγόριθμους, τους Random Forest, και Adaboost σε Δέντρα Απόφασης. Σαν μέτρο απόδοσης θα χρησιμοποιήσουμε το F1 score μιας και αυτό το σύνολο δεδομένων είναι δυσανάλογο στις κλάσεις του, με τους περισσότερους ασθενείς να είναι αρνητικοί στον ιό.

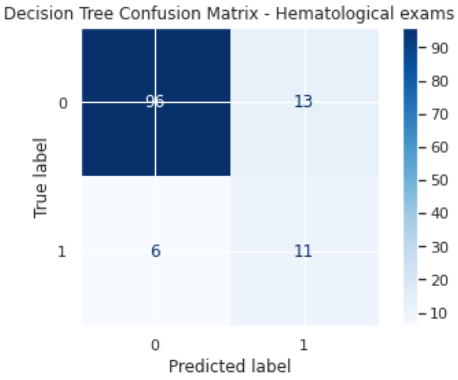
Εξαιρετικά μεγάλη σημασία έχει και το μέτρο του Recall το οποίο στην προκειμένη περίπτωση απαντά στο ερώτημα "από τους ασθενείς που είναι πραγματικά θετικοί, πόσους αναγνωρίσαμε σωστά;". Προφανώς έχει μεγαλύτερο "κόστος" να κατηγοριοποίησουμε εναν θετικό ασθενή ως αρνητικό από το αντίστροφο δεδομένης της διασποράς του ιού από τους φορείς και την πιθανή επιδείνωση της υγείας του αν δεν λάβει ιατρική περίθαλψη. Σε όλους τους αλγόρθμους έχει εφαρμοστεί βελτιστοποίηση των υπερ-παραμέτρων τους με χρήση της βιβλιοθήκης SVgrid με στόχο την μεγιστοποίηση του F1 score.

6.3.3 Αιματολογικές εξετάσεις

Εδώ τα δεδομένα χωρίστηκαν σε ποσοστό 70-30 (σύνολα εκπαίδευσης και ελέγχου). Το σύνολο εκπαίδευσης περιέχει τις αιματολογικές εξετάσεις που ανήκουν στην ΚΣ 2 καθώς και την ηλικία του ασθενούς. Το σύνολο ελέγχου περιέχει και το αποτέλεσμα του PCR test (θετικό/αρνητικό). Για να ορίσουμε μια βάση στα αποτελέσματα μας, δημιουργήσαμε εναν εικονικό κατηγοριοποιητή (dummy classifier) ο οποίος κατηγοριοποιεί όλες τις εγγραφές στην συχνότερη κλάση, δηλαδή στους αρνητικούς ασθενείς. Αυτός ο κατηγοριοποιητής κατέγραψε F1 score 0.46. Άρα θέλουμε τα μοντέλα μας να έχουν μεγαλύτερο F1 από 0.46 τουλάχιστον.

Δέντρα Απόφασης (Δ.Α.)

Στο σχήμα 6.12 βλέπουμε τον πίνακα σύγχυσης για τον αλγόριθμο των Δέντρων Απόφασης. Στους αρνητικούς ασθενείς ο αλγόριθμος προέβλεψε 96 σωστά και 13 λάθος ενώ στους θετικούς, 11 κατηγοριοποιήθηκαν σωστά και έξι λάθος.



Σχήμα 6.12: Πίνακας σύγχυσης για τον αλγόριθμο των Δέντρων Απόφασης

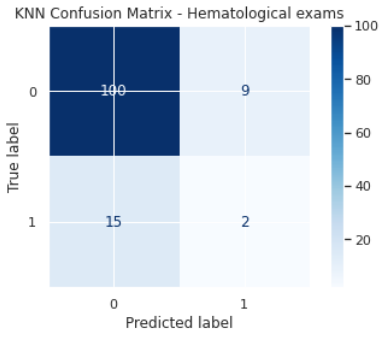
Τα μέτρα απόδοσής του έχουν ως εξής:

Μέτρα αποδοσης: Δ.Α.

Recall	0.76
Precision	0.70
F1 Score	0.72
Accuracy	0.85

Κ κοντινότεροι γείτονες (KNN)

Στο σχήμα 6.13 βλέπουμε τον πίνακα σύγχυσης για τον αλγόριθμο των Κ-κοντινότερων γειτόνων. Στους αρνητικούς ασθενείς ο αλγόριθμος προέβλεψε 100 σωστά και εννιά λάθος ενώ στους θετικούς, δύο κατηγοριοποιήθηκαν σωστά και 15 λάθος.



Σχήμα 6.13: Πίνακας σύγχυσης για τον αλγόριθμο KNN

Τα μέτρα απόδοσής του έχουν ως εξής:

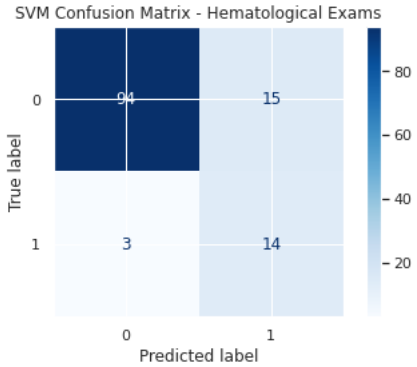
Μέτρα αποδοσης: KNN

Recall	0.52
Precision	0.53
F1 Score	0.52
Accuracy	0.81

Παρατηρούμε ότι ο KNN μετα βίας ξεπερνά το κατώφλι του εικονικού ταξινομητή όσον αφορά το μέτρο F1 score.

Μηχανές Διανυσματικής Υποστήριξης (SVM)

Στο σχήμα 6.14 βλέπουμε τον πίνακα σύγχισης για τον αλγόριθμο SVM. Στους αρνητικούς ασθενείς ο αλγόριθμος προέβλεψε 94 σωστά και 15 λάθος ενώ στους θετικούς, 14 κατηγοριοποιήθηκαν σωστά και τρείς λάθος.



Σχήμα 6.14: Πίνακας σύγχυσης για τον αλγόριθμο SVM

Τα μέτρα απόδοσής του έχουν ως εξής:

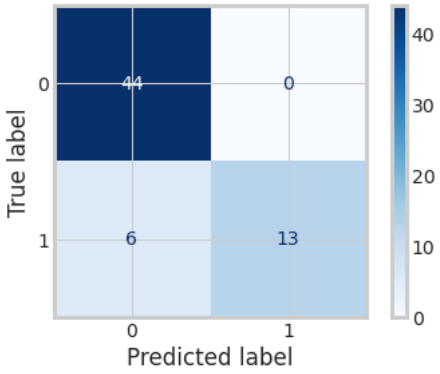
Μέτρα αποδοσης: SVM

Recall	0.84
Precision	0.73
F1 Score	0.76
Accuracy	0.86

Τυχαίο Δάσος (Τ.Δ.)

Στο σχήμα 6.15 βλέπουμε τον πίνακα σύγχισης για τον αλγόριθμο Τυχαίου Δάσους. Στους αρνητικούς ασθενείς ο αλγόριθμος προέβλεψε 94 σωστά και 15 λάθος ενώ στους θε-

τικούς, 12 κατηγοριοποιήθηκαν σωστά και πέντε λάθος.



Σχήμα 6.15: Πίνακας σύγχυσης για τον αλγόριθμο Τυχαίου Δάσους

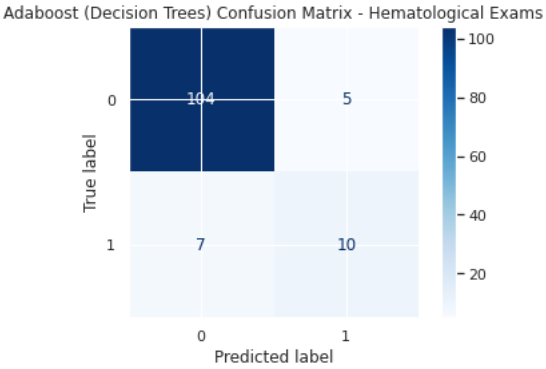
Τα μέτρα απόδοσής του έχουν ως εξής:

Μέτρα αποδοσης: Τ.Δ.

Recall	0.78
Precision	0.70
F1 Score	0.72
Accuracy	0.84

Adaboost (σε Δέντρα Απόφασης)

Στο σχήμα 6.16 βλέπουμε τον πίνακα σύγχυσης για τον αλγόριθμο Adaboost. Στους αρνητικούς ασθενείς ο αλγόριθμος προέβλεψε 94 σωστά και 15 λάθος ενώ στους θετικούς, 12 κατηγοριοποιήθηκαν σωστά και πέντε λάθος.



Σχήμα 6.16: Πίνακας σύγχυσης για τον αλγόριθμο Adaboost

Τα μέτρα απόδοσής του έχουν ως εξής:

Μέτρα αποδοσης: Adaboost

Recall	0.78
Precision	0.80
F1 Score	0.79
Accuracy	0.90

Συγκεντρωτική Σύγκριση

Στην προηγούμενη υποενότητα είδαμε τον πίνακα σύγχισης για κάθε αλγόριθμο και μελετήσαμε την απόδοσή του ως προς τα μέτρα: F1, accuracy, precision, recall. Ως βασικό μέτρο για την σύγκριση των αλγορίθμων θα χρησιμοποιηθεί το F1 score για τους λόγους που εξηγήθηκαν στην αρχή της ενότητας. Επίσης θα ληφθεί υποψιν σοβαρά η απόδοση στο μέτρο Recall, γιατί το "κόστος" του να χαρακτηριστεί ενας φορέας του ιού αρνητικός είναι μεγάλο. Πρακτικά μας ενδιαφέρει να μην "χάσουμε" θετικούς ασθενείς στην κατηγοριοποίηση γιατί αυτό θα σημαίνει περαιτέρω διάδοση του ιού.

Συγκεντρωτικά μέτρα απόδοσης

Αλγόριθμος	F1 score	Recall
Δέντρα Απόφασης	0.72	0.76
KNN	0.52	0.51
SVM	0.76	0.84
Τυχαίο Δάσος	0.72	0.78
Adaboost	0.78	0.77

Με μια επισκόπηση του συγκεντρωτικού πίνακα βλέπουμε ότι το μεγαλύτερο F1 score επέτεχε ο αλγόριθμος Adaboost σε Δέντρα Απόφασης, ακολουθούμενος από τον SVM. Λαμβάνοντας υπόψιν όμως το πολύ καλύτερο Recall score του SVM ίσως θα μπορούσαμε να πούμε ότι ο SVM είναι καταλληλότερος για την συγκεκριμένη κατηγοριοποίηση αν μας ενδιαφέρει το να κατηγοριοποίησουμε σωστά όσους περισσότερους θετικούς ασθενείς γίνεται. Σε γενικές γραμμές όλοι οι αλγόριθμοι πέτυχαν αρκετά μεγαλύτερες τιμές από το κατώφλι που τέθηκε (0.46) με εξαίρεση τον KNN του οποίου η απόδοση ήταν αρκετά χαμηλή. Επίσης μπορούμε να συμπεράνουμε ότι η κατηγοριοποίηση των ασθενών με βάση τις αιματολογικές εξετάσεις τους είναι σχετικά εφικτή.

Εξετάσεις για άλλους ιούς του αναπνευστικού

Το επόμενο σύνολο εξετάσεων που μας υπέδειξε η ΑΚΣ και είναι αρκετά μεγάλο σε μέγεθος ώστε να είναι εφικτό να δουλέψουμε με αυτό, είναι αυτό που περιέχει τις εξετάσεις για τους διάφορους ιούς του αναπνευστικού συστήματος (γρίπη, άλλοι κορονοϊοί κτλ). Ξανά δημιουργήθηκε ενας εικονικός ταξινομητής και ορίσθηκε ως κατώφλι η τιμή 0.48 για το F1 score. Εφαρμόστηκαν οι ίδιοι αλγόριθμοι με αυτούς της προηγούμενης ενότητας. Κανένας αλγόριθμος δεν κατάφερε να πετύχει αισθητά υψηλότερη τιμή από το κατώφλι του εικονικού κατηγοριοποιητή. Παρακάτω παρατίθεται ο πίνακας με τα F1 scores των αλγορίθμων για αυτήν την κατηγορία εξετάσεων.

F1 Score - Ιοί του Αναπνευστικού

Αλγόριθμος	F1 score
Δέντρα Απόφασης	0.48
KNN	0.48
SVM	0.59
Τυχαίο Δάσος	0.58
Adaboost	0.50

Κρίνεται ότι δεν μπορούμε να βασιστούμε σε αυτήν την κατηγορία εξετάσεων για να κατηγοριοποιήσουμε τους ασθενείς.

Κεφάλαιο 7

Εξόρυξη κειμένου σε Ειδησεογραφικά Άρθρα και Tweets

7.1 Εισαγωγή

Το συγκεκριμένο κεφάλαιο στοχεύει να αναλύσει μια σειρά άρθρων από ειδησεογραφικά πρακτορεία από όλο τον κόσμο καθώς και έναν αριθμό από tweets, σε μια προσπάθεια να ανακαλύψει πόσο και με ποιον τρόπο επηρέασε την κοινή γνώμη η πανδημία του Covid-19. Έπειτα εστιάζει στην δημιουργία ενός ανιχνευτή ψευδών ειδήσεων (fake news classifier) με σκοπό να διακρίνονται αυτόματα οι ψευδείς ειδήσεις από τις πραγματικές, κάνοντας χρήση αλγορίθμων εξόρυξης κειμένου. Η εξόρυξη κειμένου (text mining) αποτελεί υποκατηγορία της εξόρυξης δεδομένων, όπου τα δεδομένα είναι σε μορφή φυσικής γλώσσας (κειμένου) και όχι αριθμητικά. Σε αυτό το κεφάλαιο εξετάστηκαν τα εξής σύνολα δεδομένων:

- Δύο σύνολα δεδομένων που περιέχουν τίτλους και περιλήψεις άρθρων από γνωστούς ειδησεογραφικούς ιστότοπους όπως: BBC, New York Times, Al Jazeera, RT, The Guardian κτλ. Το πρώτο σύνολο συλλέχθηκε την περίοδο της Άνοιξη του 2020 και συκεκριμένα τις ημερομηνίες 11/3 έως 30/4. Το δεύτερο σύνολο δεδομένων συλλέχθηκε στην περίοδο του δεύτερου κύματος, το φθινόπωρο του 2020, μεταξύ 28/10 και 8/12. Όλα τα άρθρα είναι στην Αγγλική γλώσσα. Τα άρθρα συλλέχθηκαν μέσω web crawler που γράφτηκε σε Python από τα RSS feeds των ιστοσελίδων με χρήση της βιβλιοθλήκης feedparser.

- Σύνολο δεδομένων που περιέχει 1000 ψευδείς και πραγματικές ειδήσεις που αφορούν

τον Covid-19. Περιέχεται πέρα από τον τίτλο και το κείμενο του άρθρου, η πηγή καθώς και το αν το άρθρο είναι ψευδές ή πραγματικό. [42].

- Σύνολο δεδομένων με tweets που δημοσιεύθηκε στην ιστοσελίδα Kaggle. Περιλαμβάνει Tweets με το hashtag #covid19 που ανακτήθηκαν το καλοκαίρι του 2020 στο διάστημα 25/7 με 29/8 και περιέχει περίπου 180.000 tweets στην Αγγλική γλώσσα. Περιλαμβάνει επίσης πληροφορίες όπως τον αριθμό των followers του χρήστη που έγραψε το tweet, πότε δημιουργήθηκε ο λογαριασμός του, εαν το tweet είναι πρωτότυπο ή retweet κτλ [43].

Η ανάλυση περιλαμβάνει συσταδοποίηση (clustering) με τον αλγόριθμο K-means, για την ανάδειξη θεματικών (topic labeling) στο κάθε dataset καθώς και συναισθηματική ανάλυση (sentiment analysis), για την αντιστοίχηση κάθε άρθρου/tweet με κάποιο συναίσθημα βάσει των λέξεων που το αποτελούν.

7.2 Προεπεξεργασία κειμένου

Στόχος της προεπεξεργασίας είναι να "καθαριστεί" όσο το δυνατόν περισσότερο το κείμενο και να αφαιρεθούν λέξεις οι οποίες δεν προσφέρουν πληροφορία στο κείμενο (stopwords). Η προσέγγιση που ακολουθήθηκε είναι κοινή σε όλα τα δεδομένα και περιλαμβάνει τα παρακάτω βήματα:

1. "Καθάρισμα" του κειμένου από άχρηστες πληροφορίες όπως html tags, links, usernames κτλ με χρήση κανονικών εκφράσεων (regular expressions).

2. Μια σειρά βημάτων Επεξεργασίας Φυσικής Γλώσσας ώστε να μειωθεί το μέγεθος του κειμένου (αφαίρεση μικρών λέξεων μικρότερων από 3 χαρακτήρες, αφαίρεση stopwords, διαχωρισμός σε λεκτικές μονάδες (tokenization), λημματοποίηση, αποκοπή καταλήξεων). Στόχος είναι να μειωθεί ο αριθμός των λέξεων κάθε εγγράφου χωρίς να χαθεί χρήσιμη πληροφορία.

3. Διανυσματοποίηση του κειμένου με βάση το Term Frequency - Inverse Document Frequence (TFIDF Vectorization), με στόχο την κατάταξη των λέξεων του κάθε κειμένου ανάλογα με την σημασία/βαρος τους.

7.3 Ειδησεογραφικά Άρθρα

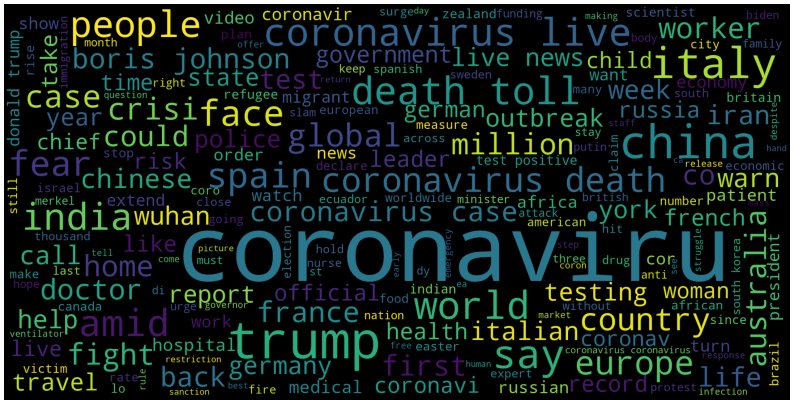
Σε κάθε μία από τις δυο χρονικές περιόδους που συλλέχθηκαν τα αρθρα, ανακτήθηκε το σύνολο των άρθρων που η κάθε ιστοσελίδα προσέφερε μέσω της ροής RSS. Κατα την επεξεργασία και μετά την αφαίρεση των διπλότυπων άρθρων, την περίοδο της Άνοιξης ανακτήθηκαν συνολικά 4.337 άθρθα από τις διάφορες πηγές. Θέλοντας να κρατήσουμε μόνο τα άρθρα σχετικά με τον Covid-19 έγινε αναζήτηση στον τίτλο κάθε άρθρου για λέξεις-κλειδιά όπως “corona”, “virus”, “covid”, “COVID”, “Covid19”, “lockdown”, “measures” κτλ και κρατήθηκαν μόνο τα σχετικά άρθρα. Σε σύνολο 4.337 άρθρων, 2.732 ήταν σχετικά με τον Covid-19, πάνω από τα μισά. Την περιόδο του Φθινοπώρου συλλέχθηκαν συνολικά 5.919 άρθρα ενώ σχετικά με τον κορονοϊό βρέθηκαν τα 1.869.

7.3.1 Άρθρα από την Άνοιξη του 2020

Μετά την προεπεξεργασία του κειμένου προκύπτουν οι εξής μεταβλητές:

1. `text`: η ενοποίηση της περίληψης και του τίτλου του άρθρου.
2. `text_tokenized`: το κείμενο χωρισμένο σε λεκτικές μονάδες (tokens) μετά την αφαίρεση stopwords, μικρών λέξεων κτλ.
3. `text_lemmatized`: οι λεκτικές μονάδες μετά την λημματοποίηση (lemmatization)
4. `text_stemmed`: οι λεκτικές μονάδες μετά την αποκοπή καταλήξεων (stemming).
5. `final_string`: το κείμενο μετά την προεπεξεργασία έτοιμο για διανυσματοποίηση.

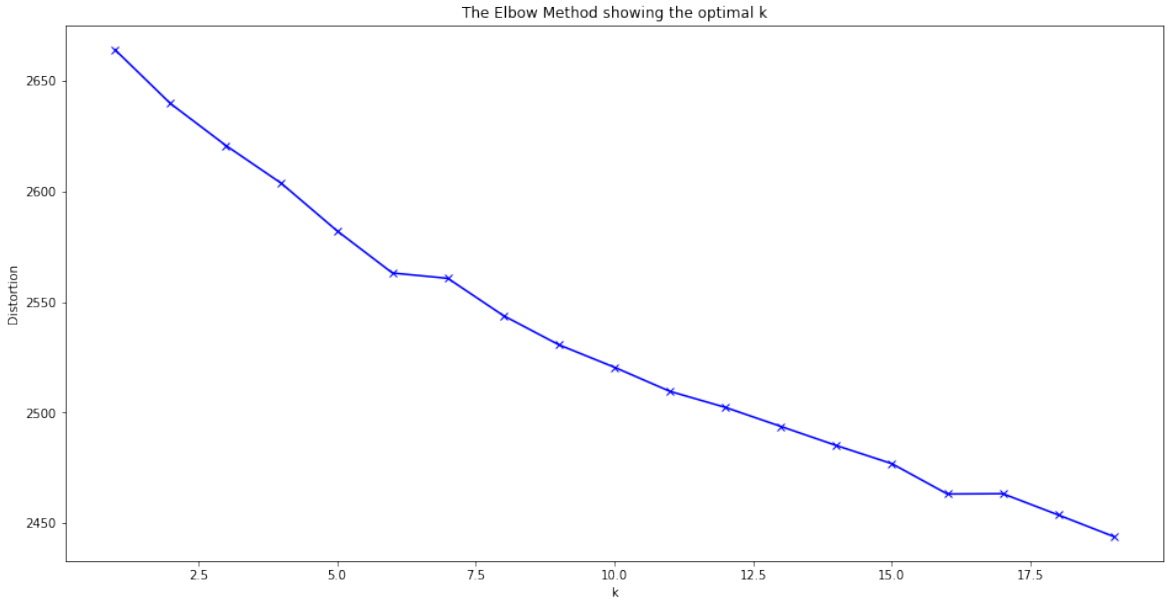
Στο σχήμα 7.1 βλέπουμε το WordCloud για τις 200 συχνότερες λέξεις του dataset.



Σχήμα 7.1: Οι 200 συχνότερες λέξεις (Ειδησεογραφικά Άρθρα - Άνοιξη 2020)

Συσταδοποίηση - Clustering

Για να προχωρήσουμε στην συσταδοποίηση των άρθρων πρέπει να οριστεί από πριν ο αριθμός των συστάδων. Εφαρμόζοντας την "μέθοδο του αγκώνα" (elbow rule), μετά από διαδοχικές επαναλήψεις του αλγορίθμου, για αριθμό συστάδων από 0 εως 20, βλέπουμε κάθε φορά την μείωση του άθροισματος των τετραγώνων των αποστάσεων (sum of squares) εντός των συστάδων (inertia). Στο σχήμα 7.2 παρατηρούμε την γραφική παράσταση των αποστάσεων σε σχέση με τον αριθμό των συστάδων. Διακρίνουμε ότι περίπου στις 7 συστάδες υπάρχει μια σταθεροποίηση του σφάλματος, οπότε επιλεγέται αριθμός συστάδων k = 7.



Σχήμα 7.2: Η μέθοδος του αγκώνα για τα ειδησεογραφικά άρθρα (Άνοιξη 2020)

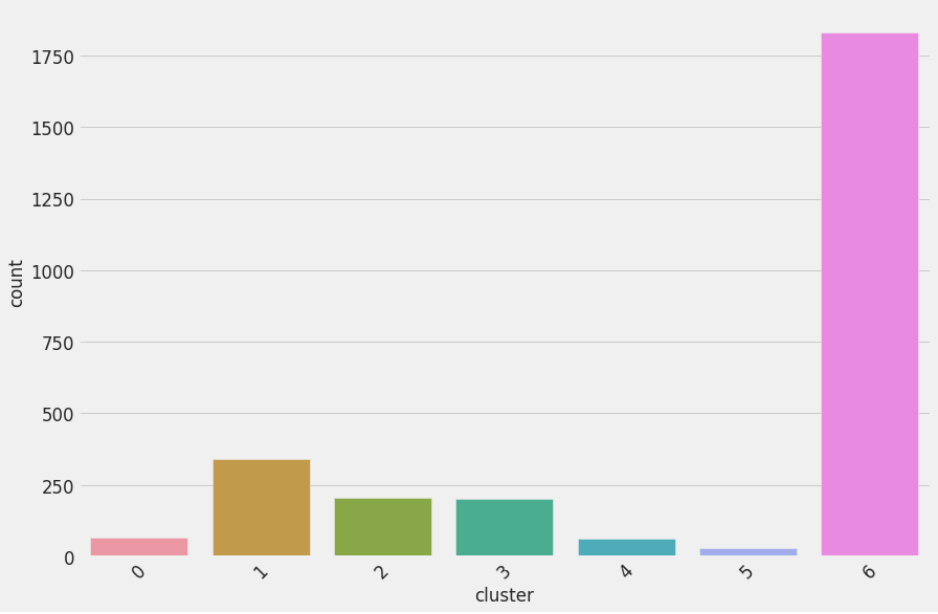
Μετά την εφαρμογή του K-means για 7 συστάδες, στον παρακάτω πίνακα παρουσιάζονται οι 10 σημαντικότερες λέξεις των άρθρων που αποτελούν την κάθε συστάδα. Στόχος είναι να μπορέσουμε να καταλάβουμε την θεματική της κάθε συστάδας ώστε να ορίσουμε το σύνολο των θεματικών που απασχόλησαν τον κόσμο για το συγκεκριμένο χρονικό διάστημα.

Οι 10 ισχυρότερες/σημαντικότερες λέξεις κάθε συστάδας

- **Cluster 0:** mask, face mask, face, wearing, wear, protective, public, official, say, protect
- **Cluster 1:** lockdown, coronavirus lockdown, people, india, country, measure, covid, home, lockdown,police, week

- **Cluster 2:** death, toll, death toll, case, number, coronavirus death, coronavirus case, italy, country, live
- **Cluster 3:** trump, president, donald trump, donald, pandemic, president trump, funding, state, china, president donald
- **Cluster 4:** johnson, boris, boris johnson, minister, prime minister, prime, intensive care, ıntensive, hospital, care
- **Cluster 5:** stock, stock market, market, business news, market business, news, coronavirus, business, coronavirus outbreak, news, outbreak
- **Cluster 6:** pandemic, virus, people, world, country, view, health, spread,outbreak, government

Στο σχήμα 7.3 βλέπουμε την κατανομή των συστάδων που προέκυψαν από τον K-means. Η τελευταία συστάδα έχει αισθητά μεγαλύτερο μέγεθος από τις υπόλοιπες περιλαμβάνοντας περίπου 1750 άρθρα. Οι ισχυρότερες λέξεις της συστάδας, μας οδηγούν στο συμπέρασμα πως περιλαμβάνει γενικά άρθρα για την εξάπλωση του κορονοϊού. Έχοντας κατηγοριοποιήσει πλέον τα άρθρα σε συστάδες, μπορούμε να δούμε 5 τυχαίους τίτλους άρθρων απο κάθε συστάδα. Αυτό θα μας βοηθήσει περαιτέρω στο να περιγράψουμε την θεματική κάθε συστάδας.



Σχήμα 7.3: Μέγεθος συστάδων

5 τίτλοι άρθρων από κάθε συστάδα

- **Cluster 0:** Police In Latin America Are Using The Coronavirus To Abuse Their Power | Who's Making Hong Kong's Ubiquitous Face Masks? Prisoners, Among Others | Can a face mask protect me from coronavirus? , Vulnerable prisoners 'exploited' to make coronavirus masks and hand gels Covid-19 myths busted | Leading by example: Indian deities don masks to 'spread awareness' about coronavirus (PHOTO)

- **Cluster 1:** WHO warns Europe still in 'the eye of the storm' of Covid-19, as Denmark & Austria lift lockdown measures | Lockdown extended, but tests lagging: Where does India stand after 21 days of shutdown? | Chinese to hit the road again on first national holiday since lockdown | 'Beyond insane'? Video of cops going after Australian 'rooftop drinkers' violating Covid-19 lockdown pushes people's buttons | How are the UK's territories dealing with the coronavirus?

- **Cluster 2:** Spain endures another record day for Covid-19 deaths but rate of new infections drops significantly | China reports no new coronavirus deaths in ten days: Live updates | The Coronavirus Death Toll Is Rising At Different Rates In Different Countries. These Charts Help Explain Why. | Italy records highest single toll from coronavirus: Live updates | Coronavirus Live Updates: Job Losses in America Soar, Part of Global Economic Collapse

- **Cluster 3:** Fauci confirms reports Trump rebuffed social distancing advice – video | 'There may be retaliation': Trump says India may face US wrath if PM Modi fails to overturn export ban on Covid-19 | Coronavirus pandemic: US suspends travel from Europe for 30 days | The pandemic is producing a 'Trump bump' in the polls – but it may not last | Trump Attacks W.H.O. and Ousts Watchdog for Pandemic Fund

- **Cluster 4:** Face Coverings And Drug Treatments Will Help The UK Manage Coronavirus Until A Vaccine Is Found | Boris Johnson: second Covid-19 peak will be disaster if lockdown lifted too early – video | Inside an NHS coronavirus intensive care unit on the frontline – video | Boris Johnson Has Been Admitted To Hospital For Tests After Having Coronavirus For 10 Day | Boris Johnson Aide Attended Secretive U.K. Coronavirus Panel

- **Cluster 5:** Stock Market Live Updates and Tracker | Stock Market Live Tracker During

Coronavirus Pandemic | The U.S. Sought Passenger Data, but Airlines Said No | U.S. Oil Prices Plunge Into Negative Territory: Live Markets Updates | Consumer Survey Shows Continued Concern

- **Cluster 6:** Australia Is Banning All Visitors To Immigration Detention Centres Because Of The Coronavirus | US launches airstrikes in Iraq in retaliation for rocket attack that killed three | Coronavirus: Health workers around the world on fears and fighting virus | S Korea reports no new domestic coronavirus cases: Live updates | European hospitals 'running out' of essential ICU meds for Covid-19, group warns

Από τα παραπανω μπορούν να εξαχθούν οι εξής θεματικές για κάθε συστάδα:

Θεματικές Ενότητες

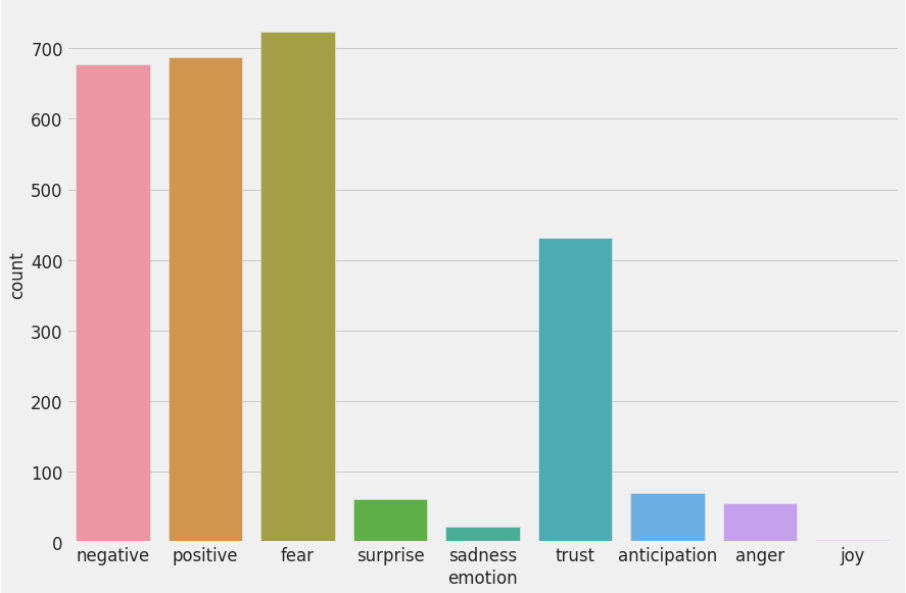
- **Cluster 0:** προστατευτικές μάσκες για τον Covid-19
- **Cluster 1:** lockdown και περιοριστικά μέτρα
- **Cluster 2:** θάνατοι από τον κορονοϊό
- **Cluster 3:** ειδήσεις που αφορούν τον πρόεδρο των ΗΠΑ
- **Cluster 4:** ειδήσεις που αφορούν τον πρωθυπουργό της Μεγάλης Βρετανίας
- **Cluster 5:** οικονομικές ειδήσεις
- **Cluster 6:** γενική κατηγορία άρθρων για τον ιό

Ανάλυση Συναισθημάτων

Έχοντας πλεον χωρίσει τα άρθρα σε συστάδες - θεματικές ενότητες, το επόμενο βήμα της ανάλυσης είναι η ανάλυση συναισθημάταν/εξόρυξη γνώμης (sentiment analysis/opinion mining) για κάθε άρθρο. Στόχος είναι να αντιστοιχηθεί κάθε κείμενο με ένα ή περισσότερα συναισθήματα (emotions) ώστε να μπορούμε να εξάγουμε συνολικα συμπεράσματα για το θέμα που μελετάμε. Υπάρχουν πολλές μέθοδοι για να διεξαχθεί η ανάλυση συναισθηματων. Στην συγκεκριμένη εργασία επιλέχθηκε το λεξικό NRC Lexicon της Αγγλικής γλώσσας στο οποίο κάθε λέξη έχει αντιστοιχηθεί με ένα ή περισσότερα συναισθήματα. Το λεξικό έχει υλοποιηθεί και ως βιβλιοθήκη της Python με τίτλο `nrclex`. Τα οκτώ διαφορετικά συναισθήματα που καλύπτει είναι: fear, anger, anticipation, trust, surprise, positive, negative, sadness,

disgust, joy δηλαδή: φόβος, θυμός, προσμονή, εμπιστοσύνη, έκπληξη, θετικό , αρνητικό, αποστροφή, χαρά. Στην ανάλυση που έγινε, για κάθε άρθρο κρατήθηκε το ισχυρότερο συναίσθημα που όρισε το λεξικό.

Στο σχήμα 7.4 βλέπουμε την συνολική εικόνα για όλα τα άρθρα. Η χρωματική παλέτα (αντιστοιχία χρώματος-συναισθήματος) θα παραμείνει σταθερή σε όλα τα διαγράμματα που ακολουθούν. Παρατηρούμε ότι το συναίσθημα του φόβου (fear) είναι κυριάρχο και χαρακτηρίζει πάνω από 700 άρθρα, ακολουθούμενο από τα αρνητικά και θετικά συναισθήματα. Μια πρώτη γενική παρατήρηση είναι ότι αθροιστικά, τα αρνητικά συναισθήματα ξεπερνούν τα θετικά, πράγμα λογικό εφόσον το γενικό θέμα είναι ένας νέος, άγνωστος, αρκετά επικίνδυνος ιός.



Σχήμα 7.4: Ανάλυση συναισθημάτων για όλα τα άρθρα

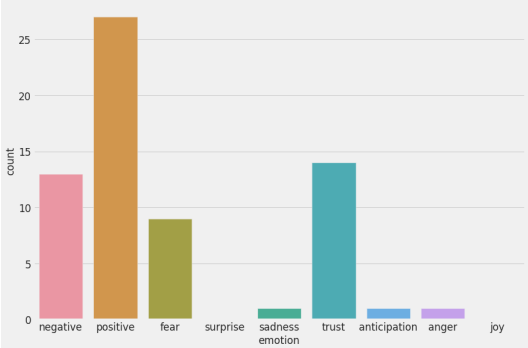
Στα σχήματα παρακάτω εξετάζεται η ανάλυση συναισθημάτων για κάθε συστάδα ξεχωριστά. Παρατηρούμε τα εξής:

- Η συστάδα 0 που αναφέρεται στις μάσκες χαρακτηρίζεται κυρίως από θετικά συναισθήματα (positive,trust).
- Η πρώτη συστάδα που αναφέρεται στο lockdown και τα περιοριστικά μέτρα έχει βασικά χαρακτηριστικά τον φόβο, αλλά και αρκετά άρθρα που χαρακτηρίζονται ως θετικά.
- Η δεύτερη συστάδα που αποτελείται από άρθρα που αναφέρονται σε θανάτους και κρούσματα κυριαρχείται εξ' ολοκλήρου από αρνητικά συναισθήματα (fear, negative).

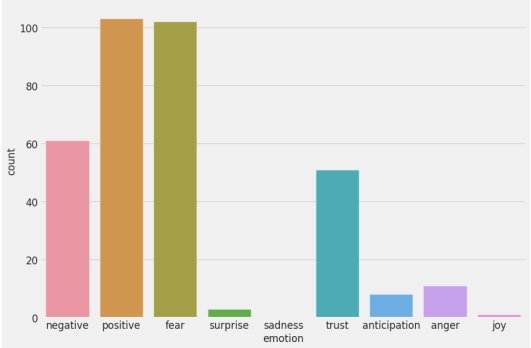
- Η τρίτη και τέταρτη συστάδα που αναφέρονται στον Πρόεδρο των ΗΠΑ και τον πρωθυπουργό της Μ. Βρετανίας αντίστοιχα έχουν μικτά συναισθήματα, τόσο αρνητικά όσο και θετικά. Σημειώνεται ότι ο Boris Johnson νόσησε βαριά από κορονοϊό κατά την περίοδο του πρώτου κύματος.

- Η πέμπτη συστάδα που αποτελείται από οικονομικού περιεχομένου άρθρα έχει ως βασικό χαρακτηριστικό τον φόβο πιθανότατα λόγω της σύνδεσης της υγιειονομικής κρίσης με μια επερχόμενη οικονομική.

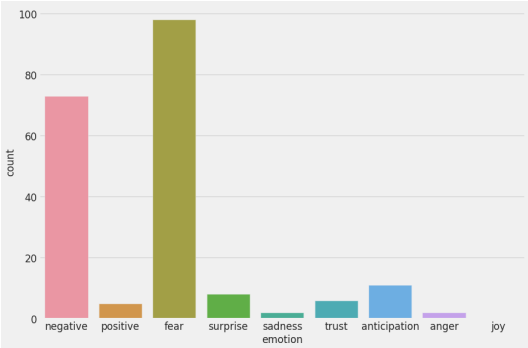
- Η έκτη και μεγαλύτερη συστάδα χαρακτηρίζεται τόσο από θετικά όσο και από αρνητικά συναισθήματα.



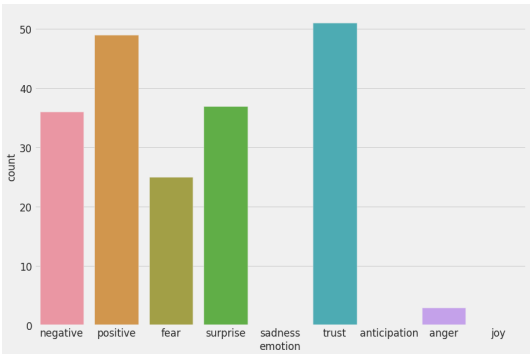
Σχήμα 7.5: Cluster 0 - Μάσκες



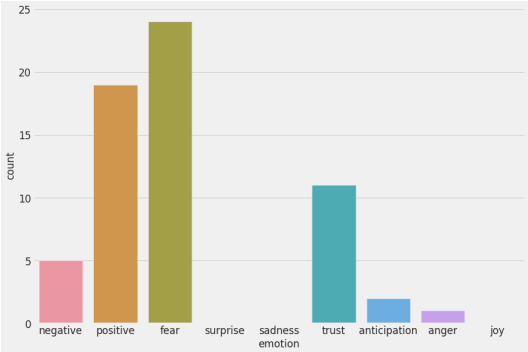
Σχήμα 7.6: Cluster 1 - Lockdown



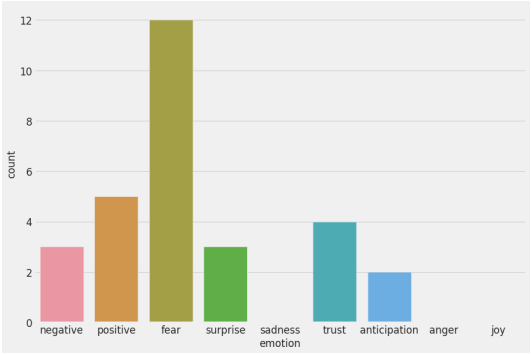
Σχήμα 7.7: Cluster 3 - Θάνατοι και κρούσματα



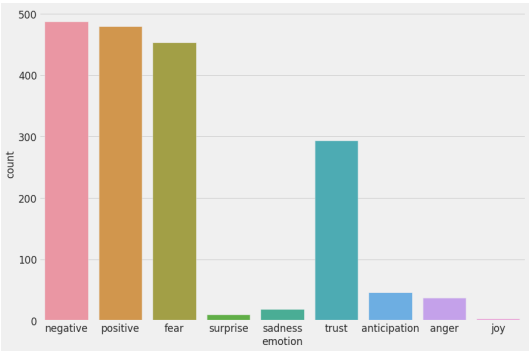
Σχήμα 7.8: Cluster 4 - Donald Trump



Σχήμα 7.9: Cluster 3 - Boris Johnson



Σχήμα 7.10: Cluster 5 - Οικονομικά Άρθρα



Σχήμα 7.11: Cluster 6 - Γενικά άρθρα για τον ιο

Η έκτη συστάδα, όπως αναφέρθηκε παραπάνω, είναι ιδιαίτερα μεγάλη, οι ισχυρότερες λέξεις της αρκετά γενικές και είναι δύσκολο να αντιστοιχηθεί σε κάποια συγκεκριμένη θεματική. Εφαρμόζοντας εκ νέου συσταδοποίηση στην συγκεκεκριμένη συστάδα για επτά νέες υπο - συστάδες ο αλγόριθμος k-means μας δίνει τα παρακάτω αποτελέσματα:

Ανάλυση της έκτης συστάδας

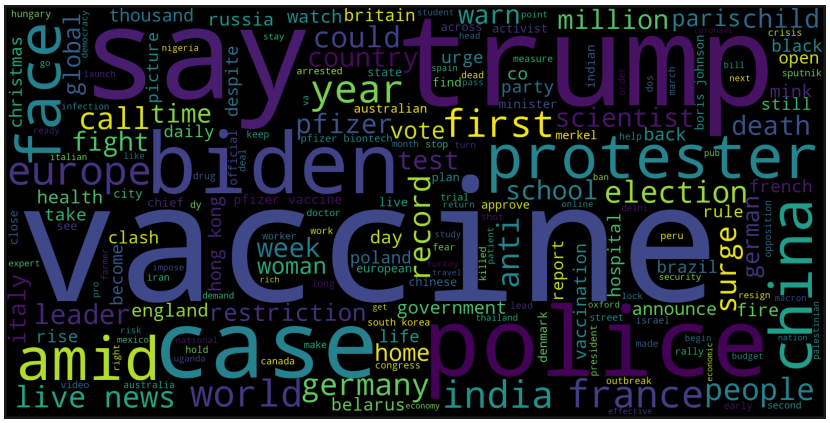
- **Cluster 6.0:** virus, country, people, spread, case, health, outbreak, said, test,say
- **Cluster 6.1:** view, people, news, government, week, said, chinese, told, australia, going
- **Cluster 6.2:** pandemic, coronavirus pandemic, covid pandemic, people, government, country, global, year, could, said
- **Cluster 6.3:** amid, doctor, patient, amid coronavirus, amid covid, crisis, medical, hospital, call, fear

- **Cluster 6.4:** italy, home, stay home, stay, people, italian, spain, minister, care, country
- **Cluster 6.5:** world, organization, world health, health, health organization, around, around world, pandemic, country, global,
- **Cluster 6.6:** china, wuhan, outbreak, chinese, city, pandemic, case, virus, coronavirus outbreak, people

Ενδιαφέρον παρουσιάζουν οι συστάδες 6.4, που αναφέρεται στην Ιταλία και την Ισπανία, δύο από τις πιο σκληρά χτυπημένες χώρες της Ευρώπης στο πρώτο κύμα, η συστάδα 6.5, που αναφέρεται στον Παγκόσμιο Οργανισμό Υγείας και 6.6 που αναφέρεται στην Κίνα και στην πόλη Wuhan από οπου και ξεκίνησε ο Covid-19.

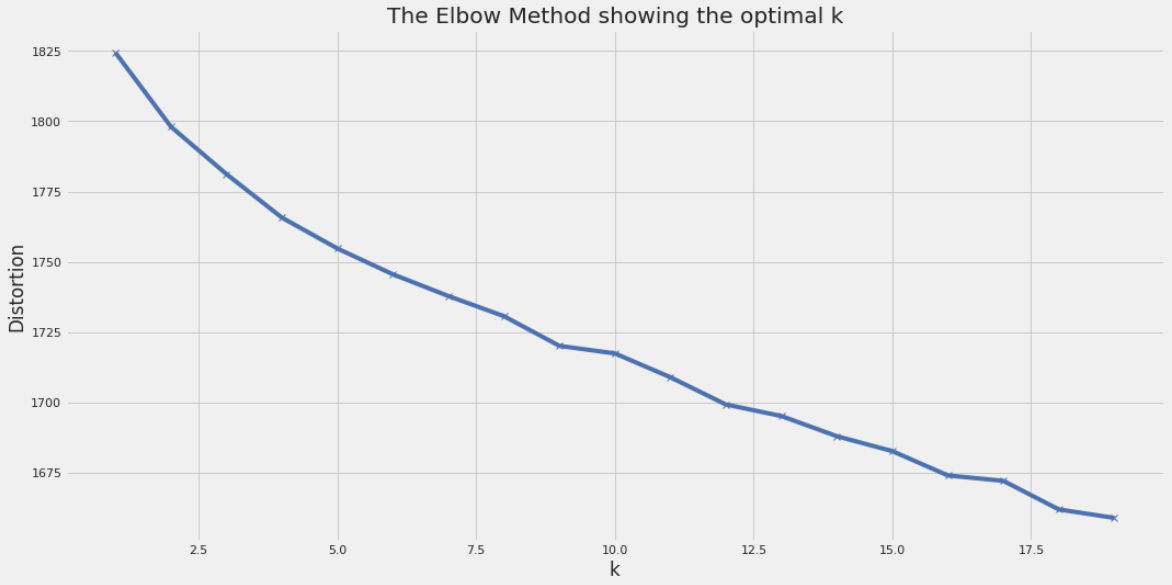
7.3.2 Ειδησεογραφικά άρθρα από το Φθινόπωρο του 2020

Το Φθινόπωρο του 2020, συλλέχθηκαν συνολικά 5.919 τίτλοι και περιλήψεις άρθρων από τις ίδιες πηγές με την πρώτη περίοδο. Με βάση τις ίδιες λέξεις-κλειδιά βρέθηκαν 1.869 άρθρα σχετικά με τον κορονοϊό, συγκριτικά λιγότερα άρθρα σε σχέση με την άνοιξη του 2020, παρότι ο συνολικός αριθμός άρθρων που συλλέχθηκε ήταν μεγαλύτερος. Στο σχήμα 7.13 βλεπουμε το WordCloud για τις συχνότερες λέξεις των συγκεκριμένων άρθρων. Παρατηρούμε με μεγάλο μέγεθος την λέξη “vaccine” (εμβόλιο), τις λέξεις “trump” και “biden”, καθώς και τις λέξεις “police” και “protester”. Τον Νοέμβριο του 2020 διεξήχθησαν οι Αμερικάνικες Εκλογές. Επίσης το φθινόπωρο του 2020 βρέθηκαν σε τελικό στάδιο έγκρισης διάφορα εμβόλια κατά του κορονοϊού. Παράλληλα έγιναν επίσης σε όλο τον κόσμο αρκετές διαδηλώσεις και συγκρούσεις (κάποιες ενάντια στα μέτρα που επεβλήθησαν λόγω του Covid-19).



Σχήμα 7.12: Οι 200 συχνότερες λέξεις (Ειδησεογραφικά Άρθρα - Φθινόπωρο 2020)

Η διαδικασία προεπεξεργασίας του κειμένου κάθε άρθρου είναι η ίδια με την προηγούμενη. Η μέθοδος του αγκώνα για τον K-means έδειξε σταθεροποίηση του σφάλματος στις εννέα συστάδες οπότε επιλέχθηκε αυτός ο αριθμός ως παράμετρος του αλγορίθμου.



Σχήμα 7.13: Μέθοδος του αγκώνα - Ειδησεογραφικά Άρθρα (Φθινόπωρο 2020)

Οι 10 ισχυρότερες/σημαντικότερες λέξεις κάθε συστάδας

- **Cluster 0:** china, government, measure, world, mink, minister, health, outbreak, country, spread
- **Cluster 1:** pandemic, year, coronavirus pandemic, child, people, time, home, world, amid, city
- **Cluster 2:** virus, mask, case, today, face, wearing, infected, record, american, test
- **Cluster 3:** lockdown, case, second, restriction, record, country, wave death, christmas, germany
- **Cluster 4:** trump, biden, election, president, donald, donald trump, elect, president elect, republican, voter
- **Cluster 5:** pfizer, vaccine, biontech, pfizer biontech, covid vaccine, first, approval, coronavirus vaccine, emergency, pfizer covid
- **Cluster 6:** positive, test, tested, tested positive, positive covid, testing, test positive, positive coronavirus, virus, trump

- **Cluster 7:** police, protester, protest, lockdown, anti, clash, anti lockdown, video, arrest, german

- **Cluster 8:** vaccine, covid vaccine, coronavirus vaccine, trial, sputnik, astrazeneca, russia, oxford, said, country

5 τίτλοι άρθρων από κάθε συστάδα

- **Cluster 0:** NASA discovers WATER spread out across Moon's sunlit surface | Coronavirus: Calls in Denmark to dig up millions of dead mink | Coronavirus: Germany hails couch potatoes in new videos | Jerrold Post: The man who analysed the minds of world leaders | Denmark finds 214 people with mink-related coronavirus

- **Cluster 1:** Philip Green is the Scrooge who haunts millions of garment workers | How to reinvent cities for the post-pandemic world | Cities can lead a green revolution after Covid. In Barcelona, we're showing how | It's A US Territory Where The Coronavirus Never Arrived — But Some Residents Can't Get Home | As Pandemic Threatens Britain's Mental Health, These 'Fishermen' Fight Back

- **Cluster 2:** Key test: South Koreans sit university exam amid COVID-19 surge | Is this the beginning of an mRNA vaccine revolution? | T-cell Covid immunity 'present in adults six months after first infection' | Fighting the virus, the University of Michigan battens down the hatches. | Biden calls for masks as some Americans face new restrictions.

- **Cluster 3:** Coronavirus: Germany restricts social life in 'lockdown light' | Dr Anthony Fauci says US in 'very difficult situation' as Covid infections continue to grow | A Takeout Order At KFC At 1:30 A.M. Raised Suspicions. It Led To A 26,000 Coronavirus Fine. | Leaders at a loss as coronavirus catches up with central Europe | Covid-19 in the US: Is this coronavirus wave the worst yet?

- **Cluster 4:** Donald Trump suggests 2024 presidential bid: 'I'll see you in four years' | 'If I lost, I'd be a very gracious loser': Trump pushes false fraud claims in Georgia | Biden plans to urge all Americans to wear masks for 100 days after inauguration | US Covid response in chaos as controversial Trump pandemic adviser Atlas resigns |

Biden unveils national security and foreign policy team as Republicans urge Trump to concede

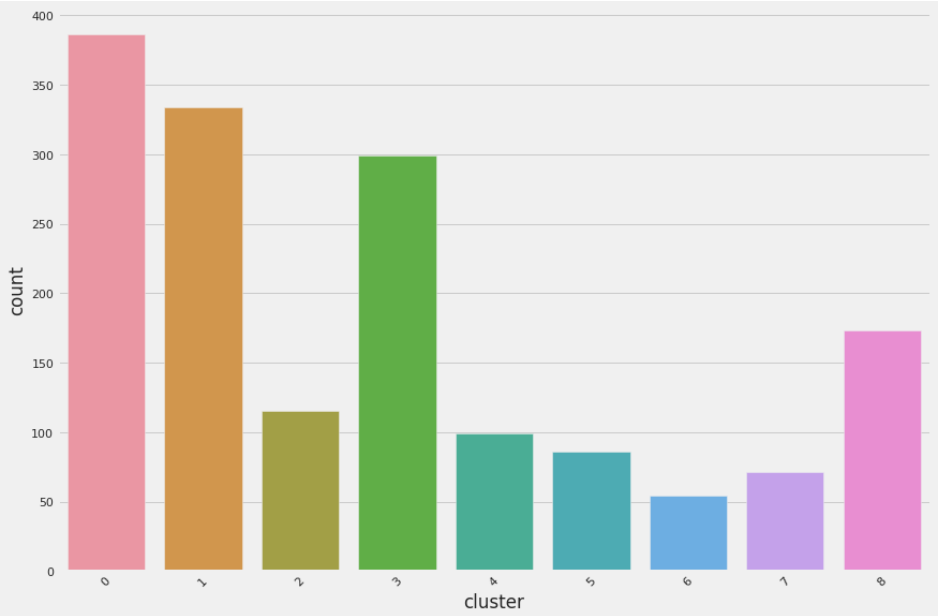
- **Cluster 5:** Coronavirus live news: Pfizer and Moderna file for EU approval of vaccines | Dow surges 1,600 points on Pfizer, BioNTech vaccine trial results | 'V-Day': UK rolls out vaccine, 90-year-old woman first in line | Pfizer seeks US FDA approval for COVID-19 vaccine's emergency use | Dow opens higher on more upbeat COVID-19 vaccine news

- **Cluster 6:** COVID-19 scare on cruise ship a false alarm, Singapore says | Boris Johnson Self-Isolates After Possible Exposure to Covid-19 | N.Y.C. says 3 percent of its coronavirus tests reveal infections. Why does the state disagree? | Anger in North Dakota after governor asks Covid-positive health workers to keep working | Slovakia carries out Covid mass testing of two-thirds of population

- **Cluster 7:** Black Lives Matter: Activists demand #EndSARS protesters' release | More than 60 arrested in anti-lockdown protests in London | Covid-19: Thailand's food hawkers sell at protests to stay afloat | Protesters flee as Greek police fire tear gas on anniversary of 1973 student revolt | Violent clashes in Barcelona as protesters hurl missiles at police after Catalonia closes borders over Covid

- **Cluster 8:** Which countries and hackers are targeting Covid vaccine developers? | The Covid vaccine results are great news, but it's not all over yet | Coronavirus live news: Italy reports 630 Covid-linked deaths in a day | The UK Is Now The First Western Country To Approve A Coronavirus Vaccine | 'The scientists have done it': Boris Johnson hails Covid vaccine

Στο σχήμα 7.14 βλέπουμε την κατανομή των άρθρων ανα συστάδα. Παρατηρούμε μια πιο ομοιόμορφη κατανομή σε σχέση με τα προηγούμενα άρθρα, καθώς δεν υπάρχει κάποια συστάδα με πολύ μεγαλύτερο αριθμό άρθρων από τις υπόλοιπες. Οι συστάδες με αριθμό 0, 1, 3 και 8 έχουν μεγαλύτερο αριθμό άρθρων από τις υπόλοιπες.



Σχήμα 7.14: Μέγεθος συστάδων

Από τις σημαντικότερες λέξεις κάθε συστάδας μπορούμε να εξάγουμε τις εξής θεματικές ενότητες:

Θεματικές Ενότητες

- **Cluster 0:** -
- **Cluster 1:** -
- **Cluster 2:** προστατευτικές μάσκες και κρούσματα
- **Cluster 3:** lockdown & περιορισμοί. Δεύτερο κύμα στην Ευρώπη.
- **Cluster 4:** Αμερικάνικες Προεδρικές Εκλογές
- **Cluster 5:** εμβόλιο (Pfizer)
- **Cluster 6:** θετικά τεστ
- **Cluster 7:** διαδηλώσεις και συγκρούσεις (anti-lockdown protests)
- **Cluster 8:** εμβόλιο (Ρώσικο - Sputnik και Αγγλικό - Astrazeneca)

Οι πρώτες δύο συστάδες (0,1) κρίνοντας από τις ισχυρότερες λέξεις τους, δεν φαίνεται να έχουν ξεκάθαρο θέμα. Η συστάδα 0 φαίνεται να αναφέρεται στην Κίνα, στην Κίνα, στα

μέτρα πρόληψης του ιού, καθώς στο θέμα που προέκυψε στην Δανία με τα μινκ (γουνοφόρα ζώα) στα οποία μεταδόθηκε ο κορονοίός. Η συστάδα με αριθμό 1 φαίνεται να αναφέρεται στην πανδημία αλλά με αρκετές λέξεις που δεν αποκωδικοποιούνται εύκολα όπως "child", "people", "country" κτλ. Οι συστάδες αυτές δεν παρουσιάζουν ισχυρή συνοχή σε σχέση με τις υπόλοιπες.

Ανάλυση Συναισθημάτων

Στο σχήμα 7.15 βλέπουμε την ανάλυση συναισθημάτων για όλα τα άρθρα. Παρατηρούμε ότι το κυρίαρχο συναίσθημα είναι το θετικό (positive) ενώ ακολουθούν αρκετά ψηλά τα αρνητικά (negative) και ο φόβος (fear). Στα επόμενα σχήματα παρουσιάζεται η ανάλυση συναισθημάτων για κάθε συστάδα που δημιούργησε ο K-means. Παρατηρούμε τα εξής:

- Στις συστάδες 0 και 1, που δεν καταφέραμε να προσδιορίσουμε ακριβώς την θεματική τους, βλέπουμε ότι τα αρνητικά και θετικά συναισθήματα σχεδόν ισοβαθμούν. Οι συστάδες αυτές δεν έχουν ξεκάθαρο συναίσθημα που να τις χαρακτηρίζει.

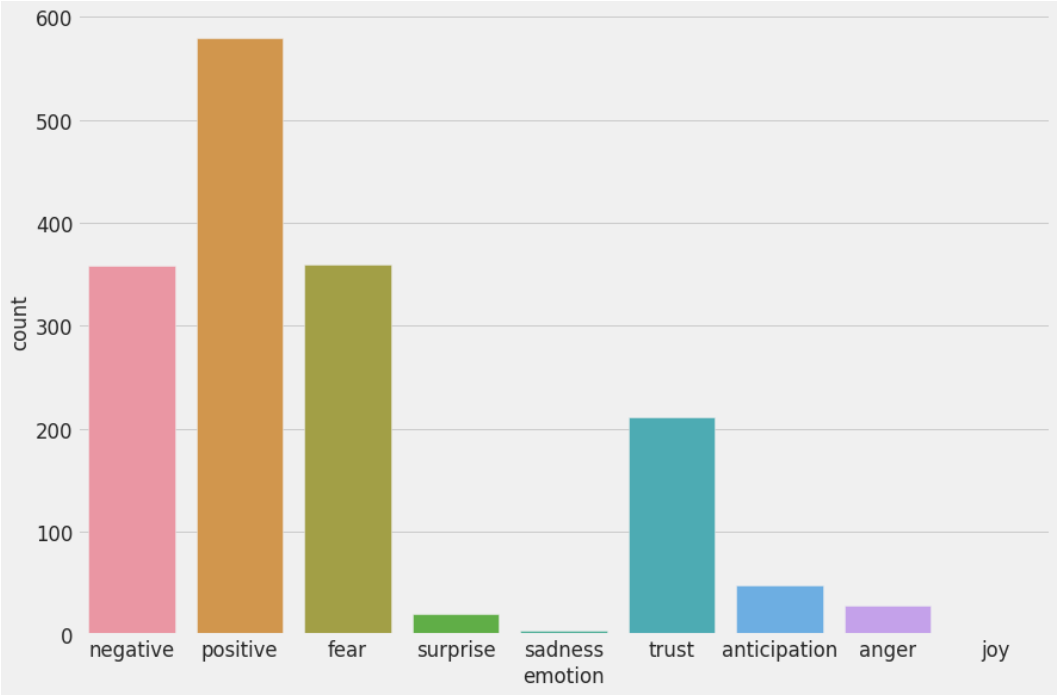
- Η δεύτερη συστάδα που αφορά τις προστατευτικές μάσκες αλλά και τα κρούσματα, ενώ χαρακτηρίζεται από το αρνητικό συναίσθημα έχει ισχυρή παρουσία θετικών συναισθημάτων.

- Η τρίτη συστάδα που αναφέρεται στο Lockdown κυριαρχείται από το συναίσθημα του φόβου.

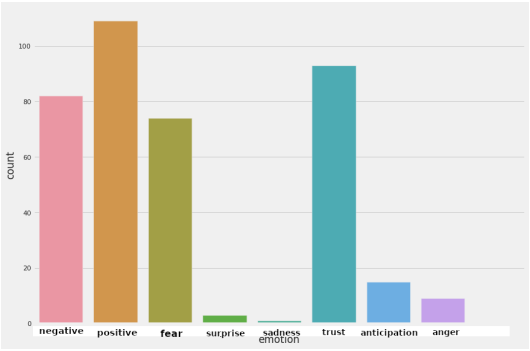
- Η τέταρτη και η έκτη συστάδα έχουν ανάμικτα συναισθήματα.

- Η έβδομη συστάδα που αναφέρεται στις διαδηλώσεις ενάντι στο lockdown, κυριαρχείται από το συνάισθημα του φόβου. Είναι η μοναδική συστάδα που έχει σχετική παρουσία το συναίσθημα της οργής (anger).

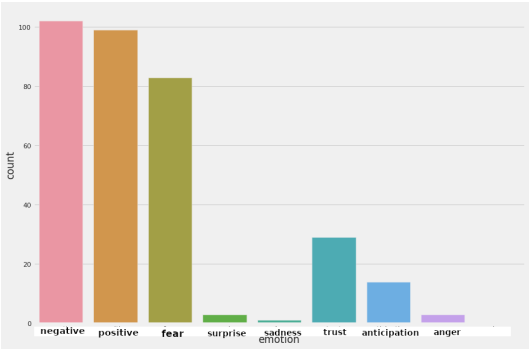
- Η πέμπτη και η όγδοη συστάδα που αναφέρονται στα εμβόλια κυριαρχούνται απόλυτα από θετικά συναισθήματα.



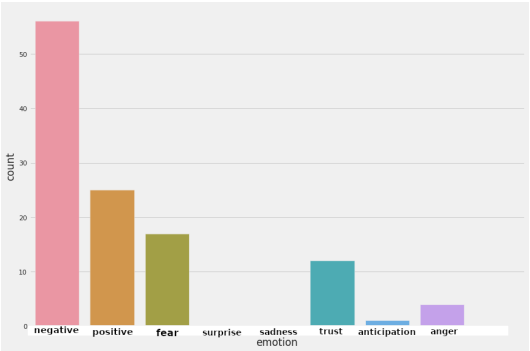
Σχήμα 7.15: Ανάλυση συναισθημάτων (όλα τα άρθρα)



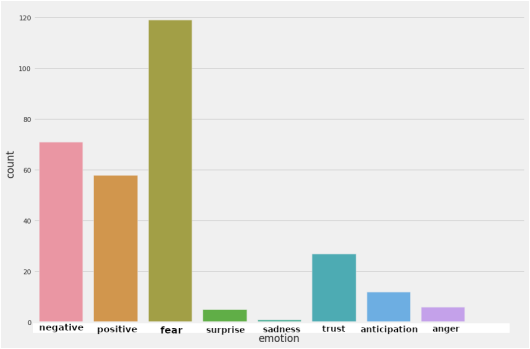
Σχήμα 7.16: Cluster 0 -



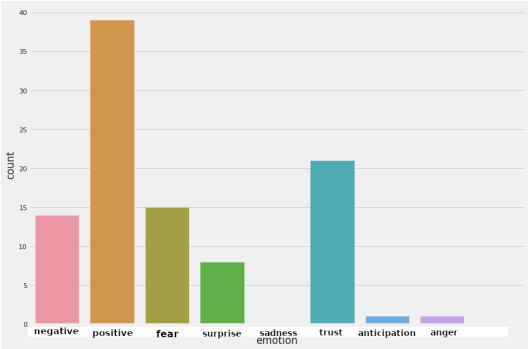
Σχήμα 7.17: Cluster 1 -



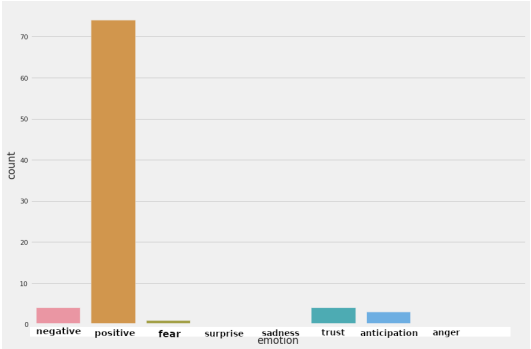
Σχήμα 7.18: Cluster 2 - Προστατευτικές μάσκες



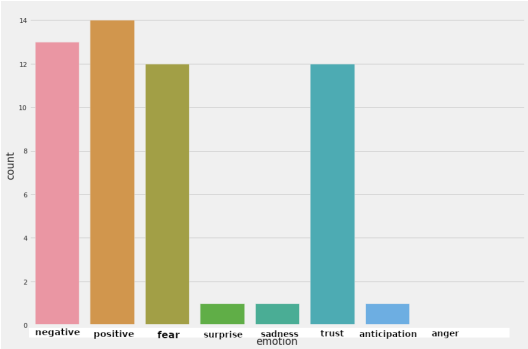
Σχήμα 7.19: Cluster 3 - Lockdown



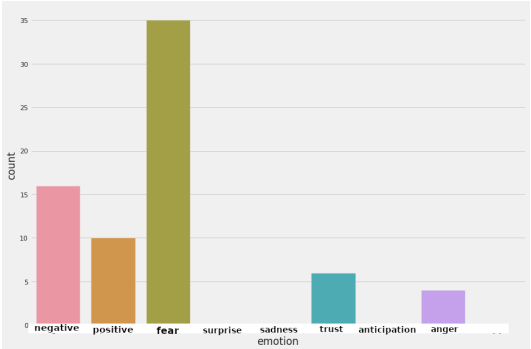
Σχήμα 7.20: Cluster 4 - Αμερικάνικες Εκλογές



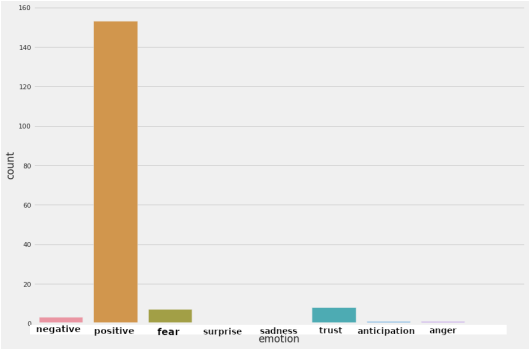
Σχήμα 7.21: Cluster 5 - Εμβόλιο (pfizer)



Σχήμα 7.22: Cluster 6 - Θετικά τεστ



Σχήμα 7.23: Cluster 7 - Διαδηλώσεις



Σχήμα 7.24: Cluster 8 - Εμβόλιο (Sputnik & Astrazeneca)

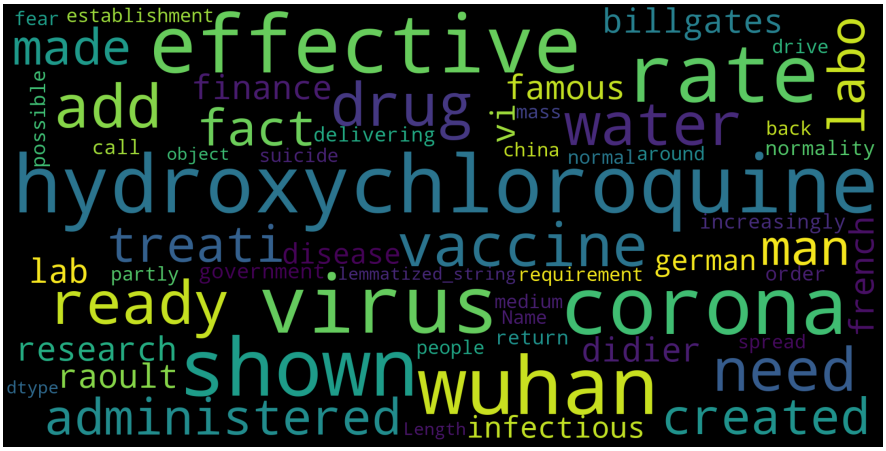
7.4 Ανιχνευτής Ψευδών Ειδήσεων

Σε αυτήν την ενότητα θα εκπαιδεύσουμε διάφορα μοντέλα κατηγοριοποίησης από αυτά που χρησιμοποιήθηκαν στο αντίστοιχο κεφάλαιο, σε ένα σύνολο δεδομένων το οποίο περιέχει πραγματικές και ψευδείς ειδήσεις που αναφέρονται στον Covid-19. Το σύνολο δεδομένων που χρησιμοποιήθηκε περιέχει 585 πραγματικές ειδήσεις και 574 ψευδείς (τίτλους και κείμενο). Οι κυριότεροι ιστότοποι από όπου προέρχονται οι ψευδείς ειδήσεις είναι: naturalnews.com, web.archive.org, orthomolecular.org, facebook.com κτλ., ενώ οι πραγματικές ειδήσεις προέρχονται από τα health.harvard.edu, nytimes.com, globalhealthnow.org, who.int, cdc.gov. κτλ Ενδεικτικά παρατίθενται ένα παράδειγμα ψευδούς και ένα παράδειγμα μιας πραγματικής είδησης:

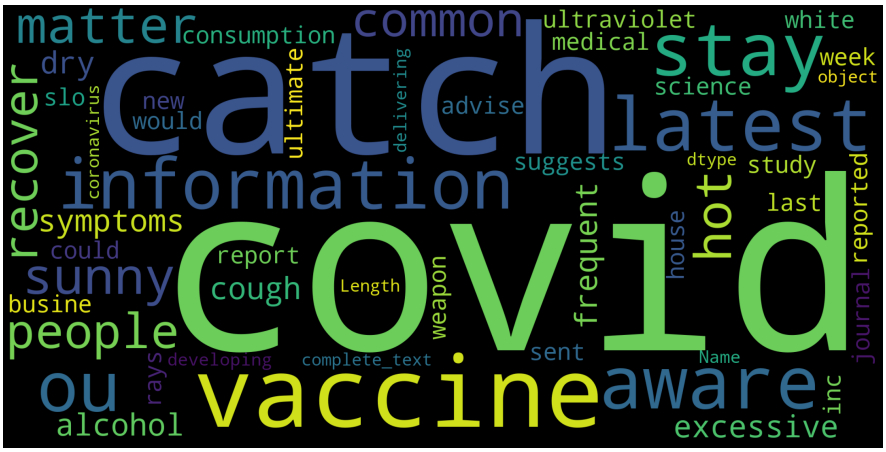
Fake News: "You just need to add water, and the drugs and vaccines are ready to be administered. There are two parts to the kit: one holds pellets containing the chemical machinery that synthesises the end product, and the other holds pellets containing instructions that tell the drug which compound to create. [...]"

Real News: "Stay aware of the latest information on the COVID-19 outbreak, available on the WHO website and through your national and local public health authority. Most people who become infected experience mild illness and recover[...]"

Παρακάτω βλέπουμε τα WordClouds απο κάθε κλάση των δεδομένων (Fake σχ. 7.25 και True 7.26).



Σχήμα 7.25: Ψευδείς Ειδήσεις - συχνότερες λέξεις



Σχήμα 7.26: Πραγματικές ειδήσεις - συχνότερες λέξεις

Η διαδικασία που ακολουθήθηκε στην προεπεξεργασία του κειμένου ήταν παρόμοια με αυτήν που ακολουθήθηκε στα ειδησεογραφικά άρθρα και περιλαμβάνει:

- Καθάρισμα κειμένου από stop-words και μικρές λέξεις (λιγότεροι απο δύο χαρακτήρες)
- Καθάρισμα από αριθμούς και σημεία στίξης
- Λημματοποίηση
- Αποκοπή καταλήξεων (stemming)
- Διανυσματοποίηση κάθε κειμένου μέσω TFIDF vectorization

Τα δεδομένα χωρίστηκαν σε σύνολο εκπαίδευσης και ελέγχου σε αναλογία 80/20, ενώ χρησιμοποιήθηκαν οι αλγόριθμοι των Δέντρων Απόφασης, Κ-κοντινότεροι γείτονες (KNN), Μηχανές Διανυσματικής Υποστήριξης (SVM), Τυχαίο Δάσος και Adaboost. Ως μέτρο της απόδοσης κάθε αλγορίθμου χρησιμοποιήθηκε η ακρίβεια (accuracy) δεδομένου ότι και οι κλάσεις στα δεδομένα ήταν ισομοιρασμένες. Μετά από δοκιμές τόσο στο απλό κείμενο χωρίς προπεξεργασία όσο και στο προεπεξεργασμένο κείμενο, η μεγαλύτερα ακρίβεια προέκυψε από το κείμενο που είχε υποστεί αποκοπή καταλήξεων (stemming). Τα αποτελέσματα κάθε αλγορίθμου παρουσιάζονται στον παρακάτω πίνακα:

Απόδοση Αλγορίθμων

Αλγόριθμος	Accuracy
Δέντρα Απόφασης	0.75
KNN	0.85
SVM	0.94
Τυχαίο Δάσος	0.90
Adaboost	0.80

Φαίνεται ότι ο αλγόριθμος SVM αποδίδει καλύτερα από τους τρεις αλγορίθμους που δοκιμάστηκαν. Δοκιμάζοντας να ρυθμίσουμε αυτόματα τις υπερ-παραμέτρους του μέσω της βιβλιοθήκης SVgrid, με στόχο την μεγιστοποίηση του accuracy, η ακρίβεια βελτιώθηκε και έφτασε στο 0.95. Μετά την επιλογή του SVM ως τον αλγόριθμο με την μεγαλύτερη ακρίβεια από τους τρείς, δοκιμάστηκαν κάποια άρθρα από το διαδίκτυο που δεν περιλαμβάνονται στο dataset.

Άρθρο1 - Fake News (naturalnews.com) *Big Tech COVID-19 censorship is endangering the population We constantly hear that we're living in the Information Age, but the truth is that we don't enjoy quite as much unfettered access to knowledge via the internet as we'd like to believe. We have Big Tech to thank for that. As more and more people depend on Google, Amazon, Facebook and other platforms, their power only grows. And after all they've done in the past few years to silence those who speak out about vaccines, it should come as no surprise that they're up to the same old tricks when it comes to the coronavirus pandemic and its response. [...]*

Άρθρο2 - Real News (cnn.com) *As vaccines trickle across the US, more Americans are now hospitalized with Covid-19 than ever before After a day of celebration and heartache, Americans face a harsh reality with the Covid-19 crisis. A record 110,549 Covid-19 patients were hospitalized Monday, according to the Covid Tracking Project. That will inevitably lead to more deaths as Christmas and New Year's Day get closer. And while more doses of the Pfizer/BioNTech vaccine get sent across the country this week, there won't be enough for everyone for months. [...]*

Ο αλγόριθμος SVM κατηγοριοποιεί σωστά το πρώτο άρθρο ως Fake news και το δεύτερο ως πραγματική είδηση.

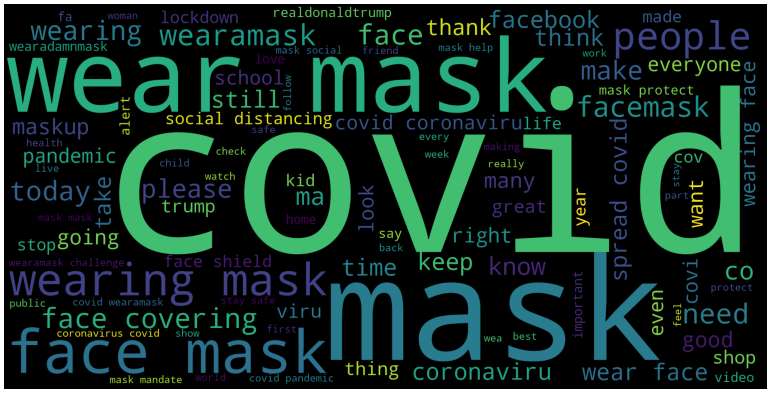
7.5 Ανάλυση Tweets

Το σύνολο δεδομένων, περιέχει 178.663 μοναδικά tweets μετά την αφαίρεση των διπλότυπων (retweets). Αντί να αναλυθεί ολόκληρο, χωρίστηκε σε μικρότερα σύνολα με συγκεκριμένο θέμα το καθένα, με θεματολογία που προέκυψε και στην ανάλυση των ειδησεογραφικών άρθρων, όπως πχ οι μάσκες ή οι διαδηλώσεις. Στόχος είναι να να γίνει ανάλυση συναισθημάτων σε αυτές τις θεματικές ενότητες.

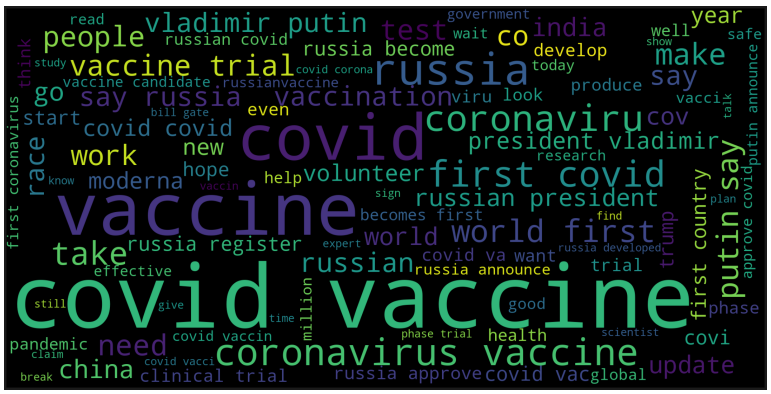
Τα μικρότερα σύνολα δεδομένων είχαν θέμα και μέγεθος:

- Μάσκες (10.034 tweets)
- Εμβόλια (4.636 tweets)
- Διαδηλώσεις (4.682 tweets)
- Άρθρα-απάτη (hoaxes) (1.903 tweets)

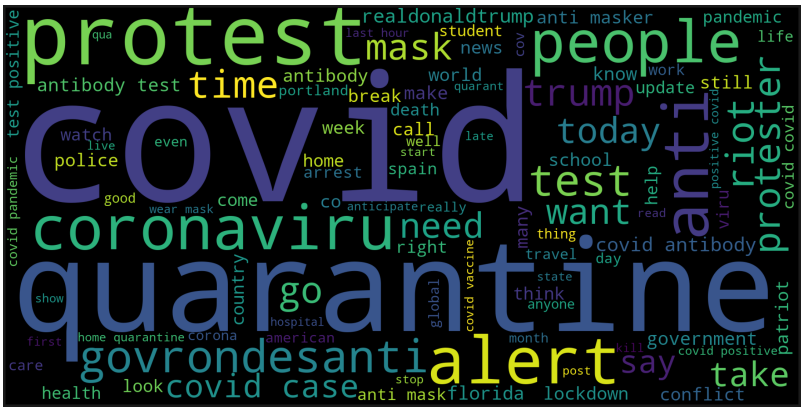
Στα σχήματα παρακάτω παρουσιάζονται τα WordClouds κάθε συνόλου δεδομένων.



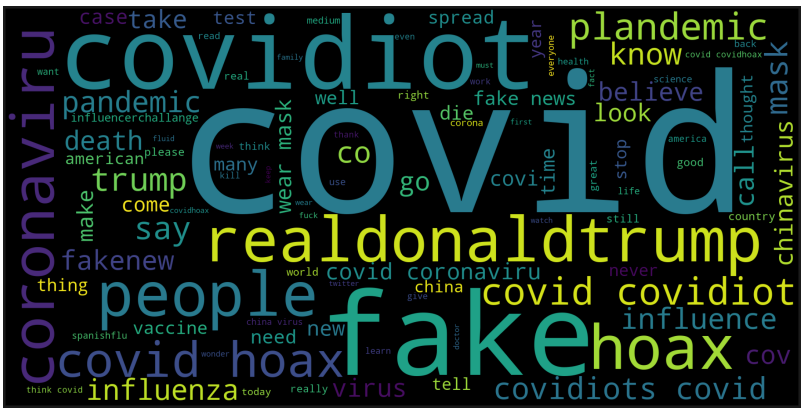
Σχήμα 7.27: Tweets σχετικά με μάσκες



Σχήμα 7.28: Tweets σχετικά με εμβόλια



Σχήμα 7.29: Tweets σχετικά με διαδηλώσεις



Σχήμα 7.30: Tweets σχετικά με hoaxes

Παρατηρούμε ότι το WordCloud που αφορά το εμβόλιο κυριαρχείται από λέξεις που αφορούν την Ρωσσία. Αυτό συμβαίνει γιατί την περίοδο που έγινε η συλλογή των tweets (τέλη καλοκαιριού του 2020) η Ρωσία ήταν η πρώτη χώρα που είχε ανακοινώσει την ύπαρξη εμβολίου, του Sputnik-5. Τα υπόλοιπα εμβόλια ήταν σε φάση δοκιμών.

Στο σύνολο δεδομένων που αφορά τις διαδηλώσεις/αναταραχές με μεγαλύτερη γραμματοσειρά παρουσιάζεται η λέξη “quarantine”. Ενδιαφέρον παρουσιάζει και το “realdonaldtrump” που είναι το handler του προέδρου των ΗΠΑ Donald Trump στο twitter. Ο πρόεδρος των ΗΠΑ παρουσιάζεται και στο WordCloud που αφορά τα hoaxes, μαζί με λέξεις όπως “plandemic” και “covidiot” που είναι hashtags που χρησιμοποιούνται συχνά από συνομωσιολόγους/αρνητές του ιού.

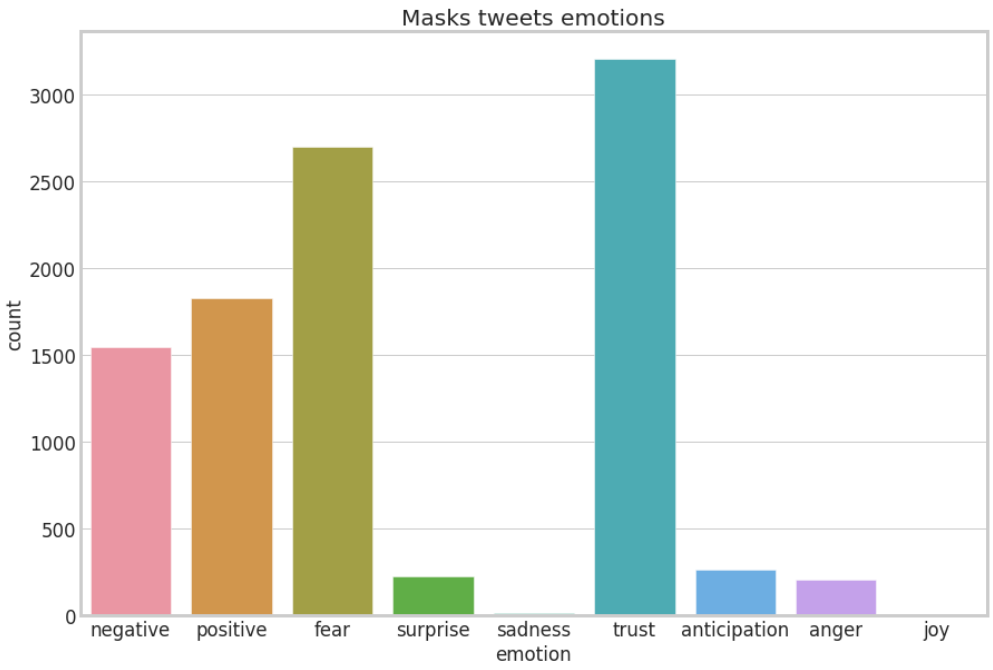
Το WordCloud που αφορά τις μάσκες περιέχει λέξεις που αφορούν την χρήση μάσκας και γενικά φαίνεται να προτρέπει στην χρήση μάσκας. Θα εξετάσουμε κάθε υπο-σύνολο δεδομένων ξεχωριστά, με στόχο να βρεθούν τα κυρίαρχα συναισθήματα που το χαρακτηρίζουν.

Αξίζει να αναφερθεί επίσης, ότι έγινε συσταδοποίηση σε κάθε ξεχωριστό σύνολο δεδομένων η οποία όμως δεν απέδωσε ιδιαίτερα. Σε αντίθεση με την συσταδοποίηση που έγινε στα ειδησεογραφικά άρθρα, οι σημαντικοτέρες λέξεις κάθε συστάδας δεν βοήθησαν στο να διακριθούν ξεκάθαρες θεματικές ενότητες, για αυτό και τα αποτελέσματα δεν παρατίθενται αναλυτικά.

7.5.1 Μάσκες

Το σύνολο δεδομένων προέκυψε από tweets του αρχικού, τα οποία περιέχουν στο κείμενό τους μια από τις ακόλουθες λέξεις-κλειδιά: “mask”, “facemask”, “masks”, “mask”, “face-cover”, “face”, “face cover”, “face mask”, “protective mask”.

Ανάλυση Συναισθημάτων

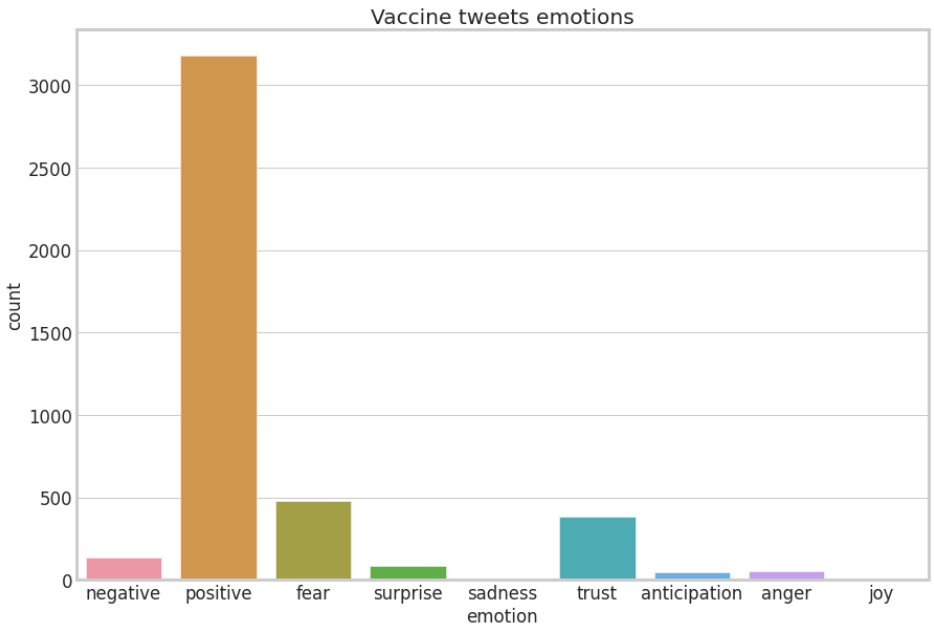


Σχήμα 7.31: Ανάλυση συναισθημάτων στα tweets σχετικά με μάσκες

7.5.2 Εμβόλια

Λέξεις-κλειδιά: “vaccine”, “vaccines”, “pfizer”, “moderna”, “astrazeneca”, “oxford”, “sputnik”, “vaccination”.

Ανάλυση Συναισθημάτων

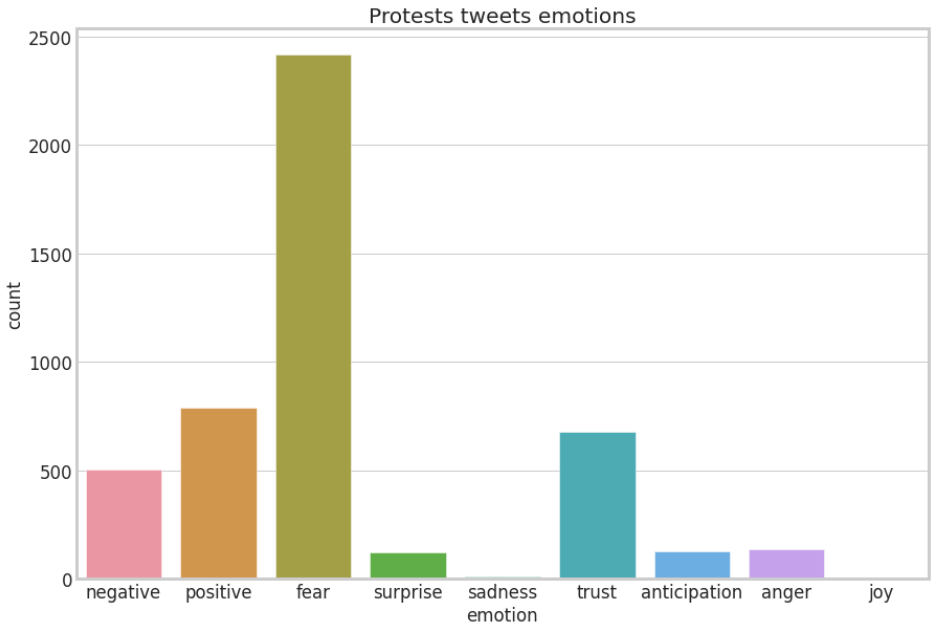


Σχήμα 7.32: Ανάλυση συναισθημάτων στα tweets σχετικά με εμβόλια

7.5.3 Διαδηλώσεις

Λέξεις-κλειδιά: “protest”, “protester”, “anti”, “clash”, “demonstration”, “conflict”, “riot”

Ανάλυση Συναισθημάτων

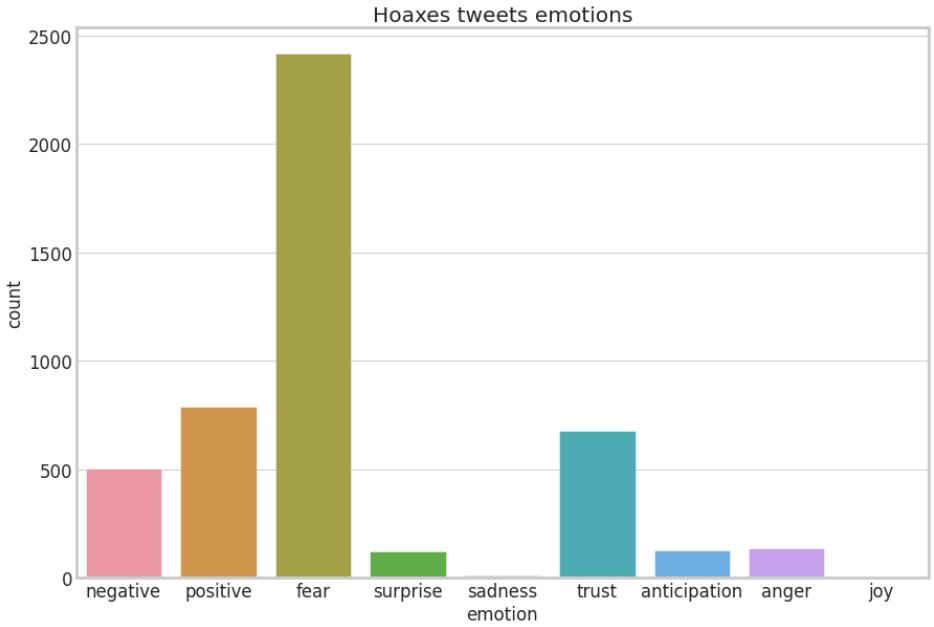


Σχήμα 7.33: Ανάλυση συναισθημάτων στα tweets σχετικά με διαδηλώσεις

7.5.4 Hoaxes

Λέξεις-Κλειδιά: “hoax””, “plandemic”, “china virus”, “CoronaJihad”, “Chinesevirus”, “bioweapon”, “chinavirus”, “fake”, “fakevirus”

Ανάλυση Συναισθημάτων



Σχήμα 7.34: Ανάλυση συναισθημάτων στα tweets σχετικά με hoaxes

7.5.5 Συμπεράσματα

Η τεχνική της συσταδοποίησης στα συγκεκριμένα tweets δεν βοήθησε στο να ξεδιαλύνει περισσότερο το ερώτημα “τι συζητούν οι πολίτες στο twitter όταν αναφέρονται στις μάσκες/εμβόλια κοκ”. Παρ’ όλα αυτά μπορούμε να εξάγουμε κάποια συμπεράσματα από την ανάλυση των tweets. Όσον αφορά τις μάσκες,συμπεραίνουμε ότι τα περισσότερα tweets προτρέπουν σε χρήση μάσκας ως προστατευτικό μέσο για τον Covid-19. Από την ανάλυση συναισθημάτων παρατηρούμε ότι το κυρίαρχο συναίσθημα στα tweets του υπο-συνόλου δεδομένων είναι η εμπιστοσύνη (trust) ακολουθούμενη από τον φόβο (fear). Ψηλά βρίσκονται επίσης και τα θετικα και αρνητικά συναισθήματα. Συμπεραίνουμε ότι -στο συγκεκριμένο σύνολο δεδομένων τουλάχιστον - τα tweets που αναφέρονται στην χρήση μάσκας έχουν κατά κανόνα θετικό συναίσθημα, με την μάσκα να αποτελεί εργαλείο στην αντιμετώπιση του

Covid-19. Βλέπουμε επίσης στο WordCloud (σχ. 7.27) να ξεχωρίζουν λέξεις όπως "wearamask". Κάτι τέτοιο συνάδει με τα αποτελέσματα που προέκυψαν απο την ανάλυση των ειδησεογραφικών άρθρων.

Σχετικά με τα εμβόλια, βλέπουμε ότι τα περισσότερα tweets σχετίζονται με την Ρωσία και το εμβόλιό της. Από την ανάλυση συναισθηματων προκύπτει ότι σχεδόν όλο το σύνολο δεδομένων είναι χαρακτηρισμένο από το θετικό (positive) συναίσθημα. Σε πολύ μικρότερα μεγέθη υπάρχουν ο φόβος και η εμπιστοσύνη. Συμπεραίνουμε ότι τα tweets του συκεκριμένου συνόλου είναι θετικά προς το εμβόλιο.

Στα σύνολα δεδομένων που αναφέρονται στις διαδηλώσεις και τα hoaxes, στην ανάλυση συναισθημάτων κυριαρχεί ο φόβος και στα δύο σύνολα δεδομένων. Από το WordCloud και τις συχνότερες λέξεις μπορούμε να συμπεράνουμε ότι οι τα περισσότερα tweets που αναφέρονται σε διαδηλώσεις αφορούν διαδηλώσεις ενάντια στα περιοριστικά μέτρα-καραντίνα.

Κεφάλαιο 8

Τεχνικές λεπτομέρειες

Σε αυτό το κεφάλαιο θα γίνει μια συνοπτική περιγραφή των εργαλείων που χρησιμοποιήθηκαν για την υλοποίηση του πειραματικού (προγραμματιστικού) μέρους της διπλωματικής εργασίας. Ο κώδικας της εργασίας γράφτηκε σε γλώσσα Python χρησιμοποιοώντας το online προγραμματιστικό περιβάλλον (IDE) Google Colab.

8.1 Η γλώσσα προγραμματισμού Python

Η γλώσσα προγραμματισμού Python είναι μια από τις πιο διαδεδομένες και ισχυρές γλώσσες προγραμματισμού σήμερα. Η Python είναι μια αντικειμενοστραφής γλώσσα προγραμματισμού, όμως δεν επιβάλλει στον προγραμματιστή το αντικειμενοστραφές μοντέλο. Τα προγράμματα στην Python μπορούν να γραφτούν και με τις αρχές του δομημένου προγραμματισμού. Στον τομέα της εξόρυξης δεδομένων έχει το πλεονέκτημα να διαθέτει έναν μεγάλο αριθμό βιβλιοθηκών τόσο για χειρισμό δεδομένων και οπτικοποίηση (pandas, matplotlib, seaborn κ.α.) όσο και για για υλοποίηση αλγορίθμων μηχανικής μάθησης (sklearn, tensorflow κ.α.) καθώς και επεξεργασίας φυσικής γλώσσας (nltk, textblob). Παράλληλα οι δυνατοτήτες της Python δεν περιορίζονται μόνο στον τομέα της Επιστήμης Δεδομένων, δίνοντας την δυνατότητα οι κώδικες που γράφτηκαν να μετασχηματιστούν σε πλήρως λειτουργικές εφαρμογές. Παρακάτω αναφέρονται μερικές από τις βιβλιοθήκες που χρησιμοποιήθηκαν:

- `pandas`: Η βιβλιοθήκη της python για χειρισμό και ανάλυση δεδομένων. Βασίζεται στην βιβλιοθήκη `numpy` η οποία ειναι η βασική βιβλιοθήκη της python για χειρισμό πινάκων και επιστημονικούς υπολογισμούς. Αναγνωρίζει και διαχειρίζεται αρχεία csv,

json, xls κ.α. Τα κύρια εργαλεία της είναι τα Dataframes για δεδομένα σε μορφή πίνακα και Series για δεδομένα σε μορφή χρονοσειρών. Διαθέτει πολλά εργαλεία για καθάρισμα και προεπεξεργασία δεδομένων.

- `sklearn`: Η βασική βιβλιοθήκη μηχανικής μάθησης της γλώσσας Python. Περιλαμβάνει υλοποιήσεις πολλών αλγορίθμων κατηγοριοποίησης, παλινδρόμησης και συσταδοποίησης τους οποίους ο προγραμματιστής μπορεί να χρησιμοποίησει ως έτοιμες συναρτήσεις στον κώδικά του. Εκτός από τους αλγόριθμους μηχανικής μάθησης η βιβλιοθήκη περιλαμβάνει μεθόδους όπως η ΑΚΣ για μείωση διαστάσεων, μεθόδους μετατροπής κατηγορικών μεταβλητών σε αριθμητικές κ.α.

- `nltk`: Βιβλιοθήκη της Python για επεξεργασία φυσικής γλώσσας. Παρέχει στον προγραμματιστή πληθώρα μεθόδων προεπεξεργασίας κειμένου (λημματοποίηση, stemming, αναγνώριση μέρος του λόγου, διανυσματοποίηση κ.α.)

- `nrclex`: Η συγκεκριμένη βιβλιοθήκη περιέχει περίπου 18.000 λέξεις αντιστοιχισμένες η καθεμία με ένα συναίσθημα (φόβος, λύπη, χαρά, οργή κτλ). Μπορεί να αναλύσει προτάσεις και ολόκληρα κείμενα και να τα συσχετίσει με ένα ή περισσότερα συναισθήματα.

8.2 Google Colab

Το Google Colaboratory (Colab) είναι ένα online προγραμματιστικό περιβάλλον για την γλώσσα Python που παρέχει η Google και είναι δωρεάν. Το περιβάλλον του είναι παρόμοιο με το Jupyter Notebook, και αποτελείται από κελιά τα οποίο μπορούν να περιέχουν είτε εκτελέσιμο κώδικα είτε κείμενο σε μορφή Markdown. Κάποια από τα πλεονοκτήματα του είναι τα εξής:

- Η χρήση του απαιτεί έναν απλό λογαριασμό Gmail και σύνδεση στο διαδίκτυο και είναι δωρεάν.

- Οι κώδικες εκτελούνται σε εικονική μηχανή (virtual machine) και όχι στο μηχάνημα του προγραμματιστή. Έτσι ακόμα και ιδιαίτερα απαιτητικές εφαρμογές μπορούν να εκτελεστούν και σε παλιότερα μηχανήματα.

- Διαθέτει προεγκατεστημένες σχεδόν όλες τις βιβλιοθήκες που απαιτούνται για προγράμματα που αφορούν μηχανική μάθηση/εξόρυξη δεδομένων. Όσες δεν είναι ήδη εγκατεστημένες, προστίθενται με μεγάλη ευκολία.
- Πολύ εύκολη διαμοίραση των πηγαίων κωδίκων με άλλους προγραμματιστές και συλλογική εργασία πάνω σε projects.

Κεφάλαιο 9

Επίλογος

Το τελευταίο κεφάλαιο αυτής της διπλωματικής εργασίας συνοψίζει την μελέτη και τα αποτελέσματα που προέκυψαν κατά την διάρκεια της εκπόνησής της. Επίσης αναφέρονται και μελλοντικές επεκτάσεις που μπορούν να πραγματοποιηθούν βασισμένες στην παρούσα εργασία.

9.1 Σύνοψη και συμπεράσματα

Στόχος αυτής της διπλωματικής εργασίας ήταν να παρουσιάσει στον αναγνώστη μια ολοκληρωμένη μελέτη που αφορά τον νέο κορονοϊό Covid-19 εξετάζοντας πτυχές επιδημιολογικών, ιατρικών και κοινωνιολογικών δεδομένων κάνοντας χρήση αλγορίθμων εξόρυξης δεδομένων, προκειμένου να εξαχθούν χρήσιμα συμπεράσματα από τα δεδομένα. Ως εργαλείο για την υλοποίηση του πειραματικού (προγραμματιστικού) μέρους της διπλωματικής εργασίας χρησιμοποιηθήκε η γλώσσα προγραμματισμού Python. Η χρήση της Python προκρίθηκε λόγω της πληθώρας βιβλιοθηκών που διαθέτει για χειρισμό δεδομένων και εφαρμογή αλγόριθμων μηχανικής μάθησης/εξόρυξης δεδομένων, αλλά και επεξεργασίας φυσικής γλώσσας (pandas, sklearn, nltk κ.α.).

Αρχικά στο Κεφάλαιο 4 έγινε μια παρουσίαση των επιδημιολογικών δεδομένων από την αρχή της πανδημίας μέχρι τα τέλη Σεπτεμβρίου του 2020. Παρουσιάστηκαν διαγράμματα που δείχνουν την εξέλιξη της πανδημίας τόσο παγκόσμια όσο και στην Ευρωπαϊκή ήπειρο, ενώ μελετήθηκαν συσχετίσεις δημογραφικών και ιατρικών μεταβλητών με τον αριθμό των κρουσμάτων, των θανάτων και της θνησιμότητας. Ταυτόχρονα έγινε μια προσπάθεια ερμηνείας των αποτελεσμάτων αυτών.

Έπειτα, στο επόμενο Κεφάλαιο 5, μελετήθηκε η πιθανή συνεισφορά των αλγορίθμων παλινδρόμησης στην πρόβλεψη της μελλοντικής πορείας των κρουσμάτων. Υλοποιήθηκαν τρία μοντέλα παλινδρόμησης (Γραμμική, Πολυωνυμική και SVM Παλινδρομητής) και βρέθηκε ότι καλύτερη απόδοση, βάσει του σφάλματος RMSE, είχε ο αλγόριθμος της πολυωνυμικής παλινδρόμησης. Σε αυτό το σημείο σχολιάστηκε η αδυναμία των αλγορίθμων παλινδρόμησης να προβλέψουν μεταβολές στην πορεία των κρουσμάτων λόγω της επίδρασης εξωγενών παραγόντων (π.χ. εφαρμογή περιοριστικών μέτρων).

Στην συνέχεια, στο Κεφάλαιο 6 εφαρμόστηκαν μοντέλα κατηγοριοποίησης σε δύο σύνολα δεδομένων που περιείχαν πληροφορίες ασθενών. Σκοπός ήταν να προβλεφθεί η εξέλιξη της υγείας του ασθενούς καθώς και αν ο ασθενής είναι θετικός στον Covid-19. Στο πρώτο σύνολο δεδομένων, με βάση το F1 score ως μέτρο απόδοσης, βρέθηκε ότι ο αλγόριθμος Τυχαίου Δάσους αποδίδει καλύτερα από τους υπόλοιπους, με F1 score 0,87. Στην περίπτωση του δεύτερου συνόλου συνόλου, στην προσπάθεια για να μειωθεί το μέγεθος του σύνολου δεδομένων, χρησιμοποιήθηκε η Ανάλυση Κύριων Συνιστωσών. Οι δύο Κύριες Συνιστώστες που προέκυψαν με μέγεθος ικανό ώστε να χρησιμοποιηθούν για πειράματα, αποτελούνταν από εξετάσεις για ιούς του αναπνευστικού και αιματολογικές εξετάσεις. Στις αιματολογικές εξετάσεις καλύτερη απόδοση με βάση το F1 score πέτυχε ο αλγόριθμος Adaboost. Την υψηλότερη απόδοση με βάση το Recall την επέτυχε ο αλγόριθμος των Μηχανών Διανυσματικής Υποστήριξης (SVM). Στην ΚΣ που αποτελούνταν από εξετάσεις για άλλους ιους του αναπνευστικού, διαπιστώθηκε ότι κανένας αλγόριθμος δεν επέτυχε απόδοση αισθητά καλύτερη από 0.50.

Τέλος, στο Κεφάλαιο 7, χρησιμοποιήθηκαν αλγόριθμοι εξόρυξης κειμένου για να αναλυθούν ειδησεογραφικά άρθρα και tweets σχετικά με τον Covid-19. Το κεφάλαιο αυτό αποσκοπεί στο να μελετήσει πως ο Covid-19 επηρέασε τον δημόσιο διάλογο σε ΜΜΕ και μέσα κοινωνικής δικτύωσης. Εφαρμόστηκε συσταδοποίηση στα κείμενα με σκοπό την ανάδειξη θεματικών σε κάθε χρονική περίοδο. Την περίοδο της Άνοιξης του 2020 βρέθηκαν ως θεματικές τα περιοριστικά μέτρα για τον ιό, ειδήσεις σχετικές με την οικονομία, τα μέτρα προστασίας (μάσκες), τα κρούσματα και οι θάνατοι κτλ. Το Φθινόπωρο, κυρίαρχες θεματικές ήταν τα εμβόλια, οι Αμερικάνικες εκλογές καθώς και οι διαδηλώσεις που έλαβαν χώρα ενάντια στα περιοριστικα μέτρα. Ξαναμεφανίζονται επίσης συστάδες σχετικές με τις μάσκες και τον αριθμό των κρουσμάτων/θανάτων. Έπειτα πραγματοποιήθηκε ανάλυση συναισθημάτων. Στην πρώτη χρονική περίοδο, κυρίαρχο αίσθημα ήταν ο φόβος ενώ στην δεύτερη

κυριαρχούν τα θετικά συναισθήματα. Χαρακτηριστικό είναι ότι οι συστάδες που σχετίζονται με τα εμβόλια και τις μάσκες κυριαρχούνται από θετικά συναισθήματα. Οι συστάδες που αναφέρονται σε κρούσματα και θανάτους χαρακτηρίζονται από το συναίσθημα του φόβου.

Στην ανάλυση που πραγματοποιήθηκε στα tweets βρέθηκε ότι τα συναισθήματα συμφωνούν γενικά με αυτά που βρέθηκαν στα ειδησεογραφικά άρθρα. Η συσταδοποίηση δεν μπόρεσε να προσφέρει επιπλέον γνώση στον χωρισμό των tweets σε θεματικές ενότητες.

Τέλος, υλοποιήθηκε ένας κατηγοριοποιητής ψευδών ειδήσεων που δέχεται ως είσοδο άρθρα και αποφαίνεται αν αυτά είναι πραγματικές ή ψευδείς ειδήσεις. Εκπαιδεύτηκαν μοντέλα κατηγοριοποίησης σε ένα σύνολο δεδομένων με ψευδείς και πραγματικές ειδήσες όπου την καλύτερη απόδοση την είχε ο αλγόριθμος SVM με ακρίβεια 0.95.

9.2 Μελλοντικές επεκτάσεις

Ως μελλοντικές επεκτάσεις αυτής της εργασίας θα μπορούσαν να αναφερθούν τα παρακάτω:

- Εφαρμογή μοντέλων χρονοσειρών για προβλέψεις. Η ικανότητα αυτών των αλγορίθμων να προσαρμόζονται στην εποχικότητα (seasonality) των δεδομένων, πιθανόν να τους καθιστά καταλληλότερους για προβλέψεις σε σχέση με τους κλασικούς αλγόριθμους παλινδρόμησης.

- Τα δεδομένα με κλινικές πληροφορίες ασθενών που ήταν διαθέσιμα κατά την περιόδο της εκπόνησης της εργασίας ήταν περιορισμένα και ελλιπή. Πληρέστερα και λιγότερο θορυβώδη δεδομένα θα μπορούσαν να οδηγήσουν σε ακριβέστερες κατηγοριοποιήσεις ασθενών. Ένα πληρέστερο σύστημα πρόγνωσης θετικότητας/αποτελέσματος ασθενούς θα μπορούσε να περιλαμβάνει χρήση απεικονιστικών εξετάσεων και ανάλυση τους με χρήση νευρωνικών δικτύων, σε συνδυασμό με την κατηγοριοποίηση με βάση τις αιματολογικές εξετάσεις.

- Στο κομμάτι της εξόρυξης κειμένου για την ανάδειξη θεματικών (topic labeling), ως εναλλακτικές του αλγορίθμου των K-μέσων θα μπορούσαν να χρησιμοποιηθούν ο αλγόριθμος Latent Dirichlet Allocation (LDA) και ο αλγόριθμος Latent Semantics Analyis (LSA).

- Οι κώδικες που γράφτηκαν για την διπλωματική εργασία θα μπορούσαν, με χρήση του κατάλληλου framework της γλώσσας Python (πχ Django), να αποτελέσουν την βάση για μια web εφαρμογή που θα παρουσιάζει στον χρήστη σε πραγματικό χρόνο επιδημολογικά δεδομένα, προβλέψεις κτλ

Βιβλιογραφία

[1] D. Reddy, "A review on data mining from past to the future," *International Journal of Computer Applications*, vol. 975, p. 8887, 2011, doi: https://doi.org/10.5120/1961-2623.

[2] "Where did covid come from?" https://www.nature.com/articles/d41586-020-03165-9, Ημερομηνία πρόσβασης: 14-01-2021.

[3] "Coronavirus," https://www.who.int/health-topics/coronavirus#tab=tab_1, Ημερομηνία πρόσβασης: 14-01-2021.

[4] I. Cooper, A. Mondal, and C. G. Antonopoulos, "A sir model assumption for the spread of covid-19 in different communities," *Chaos, Solitons & Fractals*, vol. 139, p. 110057, 2020, doi: https://doi.org/10.1016/j.chaos.2020.110057.

[5] S. F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. R. Varkonyi-Koczy, U. Reuter, T. Rabczuk, and P. M. Atkinson, "Covid-19 outbreak prediction with machine learning," *Algorithms*, vol. 13, p. 249, 2020, doi: https://doi.org/10.3390/a13100249.

[6] P. Wang, X. Zheng, J. Li, and B. Zhu, "Prediction of epidemic trends in covid-19 with logistic model and machine learning technics," *Chaos, Solitons & Fractals*, vol. 139, p. 110058, 2020, doi: https://doi.org/10.1016/j.chaos.2020.110058.

[7] V. Chaurasia and S. Pal, "Covid-19 pandemic: Application of machine learning time series analysis for prediction of human future," *Research on Biomedical Engineering*, 2020, doi: https://doi.org/10.1007/s42600-020-00105-4.

[8] L. Muhammad, M. M. Islam, U. S. Sharif, and S. I. Ayon, "Predictive data mining models for novel coronavirus (covid-19) infected patients recovery," *SN Computer Science*, vol. 1, p. 206, 2020, doi: https://doi.org/10.1007/s42979-020-00216-w.

[9] A. A. Osi, H. G. Dikko, M. Abdu, A. Ibrahim, L. A. Isma'il, H. Sarki, U. Muhammad, A. A. Suleiman, S. S. Sani, and M. Z. Ringim, "A classification approach for predicting covid-19 patient survival outcome with machine learning techniques," *medRxiv*, 2020, doi: https://doi.org/10.1101/2020.08.02.20129767.

[10] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf, and M. M. U. Din, "Machine learning based approaches for detecting covid-19 using clinical text data," *International Journal of Information Technology*, vol. 12, no. 3, pp. 731–739, 2020, doi: https://doi.org/10.1007/s41870-020-00495-9.

[11] A. Z. Khuzani, M. Heidari, and S. A. Shariati, "Covid-classifier: An automated machine learning model to assist in the diagnosis of covid-19 infection in chest x-ray images," *medRxiv*, 2020, doi: https://doi.org/10.1101/2020.05.09.20096560.

[12] J. Xue, J. Chen, R. Hu, C. Chen, C. Zheng, Y. Su, and T. Zhu, "Twitter discussions and emotions about the covid-19 pandemic: Machine learning approach," *Journal of medical Internet research*, vol. 22, no. 11, p. e20550, 2020, doi: https://doi.org/10.2196/20550.

[13] Y. Zhang, H. Lyu, Y. Liu, X. Zhang, Y. Wang, and J. Luo, "Monitoring depression trend on twitter during the covid-19 pandemic," *arXiv preprint arXiv:2007.00228*, 2020, doi: https://doi.org/10.2196/preprints.26769.

[14] P. Patwa, S. Sharma, S. PYKL, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, "Fighting an infodemic: Covid-19 fake news dataset," *arXiv preprint arXiv:2011.03327*, 2020, oNLINE: https://arxiv.org/abs/2011.03327.

[15] P. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*, 2nd ed. USA: Pearson Education, 2016.

[16] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," pp. 3–21, 2020, doi: https://doi.org/10.1007/978-3-030-22475-2_1.

[17] F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Machine learning*, vol. 42, no. 3, pp. 203–231, 2001, doi: https://doi.org/10.1023/A:1007601015854.

[18] M. Hossin and M. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, p. 1, 2015, doi: https://doi.org/10.5121/ijdkp.2015.5201.

[19] "What is text mining, text analytics and natural language processing?" https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing, Ημερομηνία πρόσβασης: 10-01-2021.

[20] M. U. Maheswari and J. Sathiaseelan, "Text mining: Survey on techniques and applications," *Int. J. Sci. Res.*, vol. 6, no. 6, pp. 45–56, 2017, doi: https://doi.org/10.4304/jetwi.1.1.60-76.

[21] "All you need to know about text preprocessing for nlp and machine learning," https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html, Ημερομηνία πρόσβασης: 11-01-2021.

[22] "Natural language processing: Text data vectorization," https://medium.com/@paritosh_30025/natural-language-processing-text-data-vectorization-af2520529cf7, Ημερομηνία πρόσβασης: 11-01-2021.

[23] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986, oNLINE: https://link.springer.com/article/10.1007/BF00116251.

[24] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Occam's razor," *Information processing letters*, vol. 24, no. 6, pp. 377–380, 1987, doi: https://doi.org/10.1016/0020-0190(87)90114-1.

[25] "A friendly introduction to support vector machines," https://www.kdnuggets.com/2019/09/friendly-introduction-support-vector-machines.html, Ημερομηνία πρόσβασης: 9-01-2021.

[26] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006, doi: https://doi.org/10.1038/nbt1206-1565.

[27] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009, doi: https://doi.org/10.4249/scholarpedia.1883.

[28] A. S. Arefin, C. Riveros, R. Berretta, and P. Moscato, "Gpu-fs-k nn: A software tool for fast and scalable k nn computation using gpus," *PloS one*, vol. 7, no. 8, p. e44000, 2012, doi: https://doi.org/10.1371/journal.pone.0044000.

[29] "Random forest explained," https://towardsdatascience.com/random-forest-explained-7eae084f3ebe, Ημερομηνία πρόσβασης: 10-01-2021.

[30] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: https://doi.org/10.1023/A:1010933404324.

[31] "Boosting and adaboost clearly explained," https://towardsdatascience.com/boosting-and-adaboost-clearly-explained-856e21152d3e, Ημερομηνία πρόσβασης: 12-01-2021.

[32] "Linear regression explained," https://towardsdatascience.com/linear-regression-explained-d0a1068accb9, Ημερομηνία πρόσβασης: 12-01-2021.

[33] "Understanding polynomial regression," https://medium.com/analytics-vidhya/understanding-polynomial-regression-5ac25b970e18, Ημερομηνία πρόσβασης: 12-01-2021.

[34] "A step-by-step explanation of principal component analysis," https://builtin.com/data-science/step-step-explanation-principal-component-analysis, Ημερομηνία πρόσβασης: 11-01-2021.

[35] "Infectious disease modelling: Understanding the models that are used to model coronavirus," https://towardsdatascience.com/infectious-disease-modelling-part-i-understanding-sir-28d60e29fdfc, Ημερομηνία πρόσβασης: 14-01-2021.

[36] "Kaggle dataset: Novel corona virus 2019 dataset," https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset/version/143, Ημερομηνία πρόσβασης: 14-01-2021.

[37] "Kaggle dataset: countryinfo," https://www.kaggle.com/koryto/countryinfo?select=covid19countryinfo.csv, Ημερομηνία πρόσβασης: 14-01-2021.

[38] "Covid-19: What is hidden behind the official numbers?" https://towardsdatascience.com/which-countries-are-affected-the-most-by-covid-19-4d4570852e31, Ημερομηνία πρόσβασης: 12-01-2021.

[39] "Correlation vs. causation," https://www.jmp.com/en_au/statistics-knowledge-portal/what-is-correlation/correlation-vs-causation.html, Ημερομηνία πρόσβασης: 13-01-2021.

[40] C. Iwendi, A. K. Bashir, A. Peshkar, R. Sujatha, J. M. Chatterjee, S. Pasupuleti, R. Mishra, S. Pillai, and O. Jo, "Covid-19 patient health prediction using boosted random forest algorithm," *Frontiers in Public Health*, vol. 8, p. 357, 2020, doi: https://doi.org/10.3389/fpubh.2020.00357.

[41] "Kaggle dataset: Covid-19 - clinical data to assess diagnosis," https://www.kaggle.com/S%C3%ADrio-Libanes/covid19, Ημερομηνία πρόσβασης: 15-01-2021.

[42] "Covid-19 fake news dataset," https://github.com/susanli2016/NLP-with-Python/blob/master/data/corona_fake.csv, Ημερομηνία πρόσβασης: 17-01-2021.

[43] "Kaggle dataset: Covid19 tweets," https://www.kaggle.com/gpreda/covid19-tweets, Ημερομηνία πρόσβασης: 17-01-2021.

Παράρτημα

Κώδικες και Δεδομένα

Οι κώδικες (Jupyter notebooks) που χρησιμοποιήθηκαν στο προγραμματιαστικό μέρος της εργασίας, βρίσκονται διαθέσιμοι εδώ, ενώ τα αντίστοιχα σύνολα δεδομένων βρίσκονται εδώ. Όλοι οι υπερσύνδεσμοι οδηγούν στο αποθετήριο Github.