ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ
ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ
ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΝΑΠΤΥΞΗ ΤΕΧΝΙΚΩΝ ΠΡΟΒΛΕΨΗΣ ΓΙΑ ΤΗΝ ΕΙΣΑΓΩΓΗ ΑΣΘΕΝΩΝ ΣΕ ΜΟΝΑΔΕΣ ΥΓΕΙΑΣ ΜΕ ΤΗΝ ΧΡΗΣΗ BIG DATA

ΣΟΦΙΑ ΣΑΡΑΦ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ
ΔΗΜΗΤΡΙΟΣ ΑΣΚΟΥΝΗΣ
ΚΑΘΗΓΗΤΗΣ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΕΘΝΙΚΟΥ ΜΕΤΣΟΒΙΟΥ ΠΟΛΥΤΕΧΝΕΙΟΥ

ΔΕΚΕΜΒΡΙΟΣ 2022

Περιεχόμενα

Περίληψη

Ο όγκος των δεδομένων στον χώρο της υγείας συνεχώς αυξάνεται . Τεράστιες ποσότητες δεδομένων, όπως τα ιατρικά στοιχεία, οι εγγραφές ασθενών, οι απεικονιστικές εξετάσεις, οι συμπεριφορές και οι απαιτήσεις των ασθενών ακόμα και τα οικονομικά στοιχεία που σχετίζονται με την περίθαλψη και την παροχή υπηρεσιών υγείας συλλέγονται καθημερινά στα πληροφοριακά συστήματα των δομών υγείας και πρόνοιας. Η επεξεργασία αυτών των δεδομένων απαιτεί ευφυείς λύσεις για λήψη έγκυρων και αξιόπιστων πληροφοριών, αποφάσεων και προβλέψεων. Τα προγνωστικά μοντέλα αναλύουν τα υπάρχοντα δεδομένα με στόχο ασφαλείς προβλέψεις για μελλοντικά γεγονότα και τάσεις. Η χρήση αυτών των μοντέλων από μονάδες υγείας βοηθά στην βελτίωση του κόστους και της ποιότητας των παρεχόμενων υπηρεσιών αλλά και στον καθορισμό των βέλτιστων κλινικών πρακτικών. Στόχος της συγκεκριμένης διπλωματικής εργασίας είναι η συλλογή και ανάλυση δεδομένων που αφορούν εισαγωγές ασθενών μονάδων υγείας με σκοπό την εξαγωγή χρήσιμων συμπερασμάτων. Η προγνωστική ανάλυση είναι ένας τομέας όπου με την χρήση διάφορων στατιστικών μεθόδων σε έναν πολύ μεγάλο αριθμό δεδομένων ασθενών ανιχνεύονται μοτίβα και συσχετίσεις που θα ήταν πολύ δύσκολο για ένα μεμονωμένο άτομο να συνδυάσει όλα αυτά τα δεδομένα. Με βάση ειδικούς αλγόριθμους που συσχετίζουν την κατάσταση του ασθενή με πολλές παρόμοιες περιπτώσεις που είναι περασμένες σε μία τεράστια αποθήκη δεδομένων μπορεί να βοηθηθεί πάρα πολύ στην διάγνωση και στην αντίστοιχη θεραπεία που θα δοθεί στον ασθενή.

Λέξεις-κλειδιά: Χρονοσειρές, Python, Τεχνικές Προβλέψεων, Μηχανική Μάθηση

Abstract

The volume of data in the field of health is constantly increasing. Huge amounts of data such as medical records, patient records, imaging diagnostics, patient behaviors and demands and even financial data related to health care and services are collected daily in the information systems of health and welfare structures. Processing of the data requires intelligent solutions to obtain valid and reliable information, decisions and forecasts. Predictive models are analyzing existing data for secure forecasts of future events and trends. Use of these models by health units helps to improve the cost and quality of services provided but also to determine best clinical practices. The aim of this thesis is the collection and analysis of data related to the admission of patient health units in order to draw useful conclusions. Prognostic analysis that is using various statistical methods in a very large number of patient data patterns and correlations, is recognized to be very difficult for an individual to combine all this data. Based on special algorithms that correlate the patient's condition with many similar cases that are passed in a huge data warehouse, prognostic analysis can be very helpful in the diagnosis and the corresponding treatment that will be given to the patient.

Key words: Time-Series, Python, Forecasting Methods, Machine Learning

Ευχαριστίες

Η ολοκλήρωση του παρόντος Μεταπτυχιακού Προγράμματος Σπουδών αποτελεί το αποτέλεσμα όχι ατομικής, αλλά ομαδικής εργασίας. Οφείλω λοιπόν ένα θερμό ευχαριστώ στους συναδέλφους του μεταπτυχιακού που συνεργαστήκαμε και μελετήσαμε μαζί καθ' όλη την διάρκεια του προγράμματος. Ευχαριστώ από καρδιάς τον επιβλέποντα Καθηγητή μου κ. Ασκούνη Δημήτριο για τη στήριξη και τις συμβουλές του κατά τη συγγραφή της διπλωματικής εργασίας.

1 Εισαγωγή

1.1 Ορισμός Big Data

Αν και τα Big Data είναι μια δημοφιλής λέξη-trend τόσο στον ακαδημαϊκό χώρο όσο και στον επιχειρηματικό κλάδο, το νόημά της εξακολουθεί να καλύπτεται από πολλές εννοιολογικές ασάφειες. Ο όρος χρησιμοποιείται για να περιγράψει ένα ευρύ φάσμα εννοιών: από την τεχνολογική ικανότητα αποθήκευσης, συγκέντρωσης και επεξεργασίας δεδομένων, έως την πολιτιστική αλλαγή που εισβάλλει διάχυτα στις επιχειρήσεις και την κοινωνία, συνδυαστικά με την υπερφόρτωση πληροφοριών. Η έλλειψη ενός επίσημου ορισμού οδήγησε την έρευνα να εξελιχθεί σε πολλαπλές διαδρομές. Επιπλέον, η υπάρχουσα ασάφεια μεταξύ ερευνητών και επαγγελματιών υπονομεύει την αποτελεσματική ανάπτυξη του θέματος. [1]

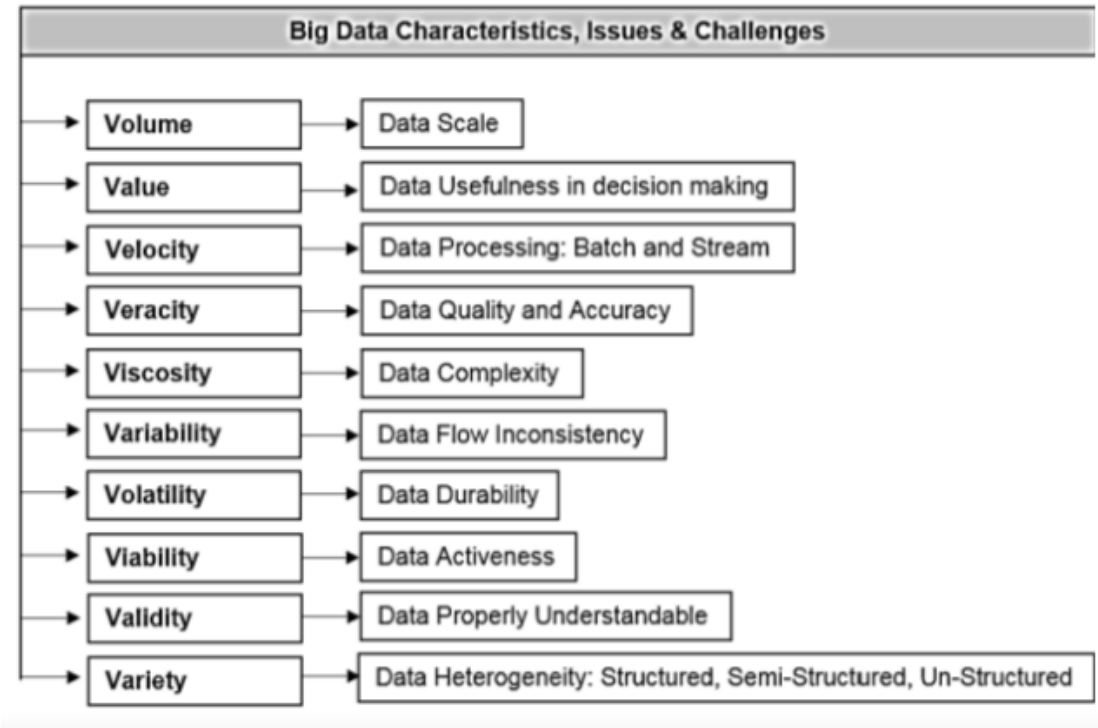
Ο σημερινός διεθνής πληθυσμός ξεπερνά τα 7,2 δισεκατομμύρια και πάνω από 2 δισεκατομμύρια από αυτούς τους ανθρώπους είναι συνδεδεμένοι στο Διαδίκτυο. Επιπλέον, 5 δισεκατομμύρια άτομα χρησιμοποιούν διάφορες κινητές συσκευές. Ως αποτέλεσμα αυτής της τεχνολογικής επανάστασης, αυτά τα εκατομμύρια άνθρωποι παράγουν τεράστιες ποσότητες δεδομένων μέσω της αυξημένης χρήσης τέτοιων συσκευών. Συγκεκριμένα, οι απομακρυσμένοι αισθητήρες παράγουν συνεχώς πολλά ετερογενή δεδομένα που είναι είτε δομημένα είτε μη. Αυτά τα δεδομένα είναι γνωστά ως Big Data. Τα μεγάλα δεδομένα χαρακτηρίζονται από τρεις πτυχές: (α) τα δεδομένα είναι πολυάριθμα, (β) τα δεδομένα δεν μπορούν να κατηγοριοποιηθούν σε κανονικές σχεσιακές βάσεις δεδομένων και (γ) τα δεδομένα παράγονται, συλλέγονται και επεξεργάζονται πολύ γρήγορα. Τα Big Data είναι πολλά υποσχόμενα ως προς την επιχειρηματική τους εφαρμογή και αυξάνονται ραγδαία ως τμήμα του κλάδου της πληροφορικής. Έχουν δημιουργήσει σημαντικό ενδιαφέρον σε διάφορους τομείς, συμπεριλαμβανομένων των κλάδων της κατασκευής μηχανημάτων, της υγειονομικής περίθαλψης, των τραπεζικών συναλλαγών, των μέσων κοινωνικής δικτύωσης και της δορυφορικής απεικόνισης. Παραδοσιακά, τα δεδομένα αποθηκεύονται σε μια εξαιρετικά δομημένη μορφή για τη μεγιστοποίηση του ενημερωτικού τους περιεχομένου. Ωστόσο, οι όγκοι των δεδομένων καθορίζονται τόσο από μη δομημένα όσο και από ημι-δομημένα δεδομένα. Επομένως, η επεξεργασία από άκρο σε άκρο μπορεί να παρεμποδιστεί από τη μετάφραση μεταξύ δομημένων δεδομένων σε

σχεσιακά συστήματα διαχείρισης βάσεων δεδομένων και μη δομημένων δεδομένων για ανάλυση. Ο εκπληκτικός ρυθμός αύξησης του όγκου των συλλεγόμενων δεδομένων δημιουργεί πολυάριθμα κρίσιμα ζητήματα και προκλήσεις, όπως η ταχεία ανάπτυξη δεδομένων, η ταχύτητα μεταφοράς και τα θέματα ασφάλειας. Ωστόσο, οι εξελίξεις στις τεχνολογίες αποθήκευσης και εξόρυξης δεδομένων επιτρέπουν τη διατήρηση αυτών των αυξημένων ποσοτήτων δεδομένων. Οι μελλοντικές ερευνητικές κατευθύνσεις σε αυτόν τον τομέα καθορίζονται από ευκαιρίες και αρκετά ανοιχτά ζητήματα στην κυριαρχία των Big Data. [2]

1.2 Χαρακτηριστικά των Big Data

Τα Μεγάλα Δεδομένα ορίζονται συνήθως με όρους των 3V, μια ονομασία που αναπτύχθηκε αρχικά από την Gartner Doug Laney το 2001: Volume (Όγκος), Velocity (Ταχύτητα), Variety (Ποικιλία). Ο όγκος υποδεικνύει τη ποσότητα δεδομένων, η ταχύτητα υποδεικνύει την υψηλή ταχύτητα επεξεργασίας δεδομένων (για παράδειγμα 500 TB δεδομένων ανά ημέρα μπορούν να θεωρηθούν μεγάλα δεδομένα) και η ποικιλία υποδηλώνει την πολυπλοκότητα των δεδομένων (τρεις βασικές κατηγορίες δεδομένων: δομημένα, ημι-δομημένα και αδόμητα).

Στην πορεία στα χαρακτηριστικά των Big Data προστέθηκαν επιπλέον φτάνοντας στα 5Vs με την προσθήκη των Veracity (Αληθοφάνεια) και Value (Αξία) στα υπάρχοντα 3Vs. Η αληθοφάνεια (Veracity) αναφέρεται σε πηγές που επηρεάζουν την ακρίβεια, όπως ασυνέπειες, έλλειψη δεδομένων, αμφισημίες, απάτη, επανάληψη, καθυστέρηση. Η αξία (Value) για μεγάλα δεδομένα υποδηλώνει την έννοια ότι εάν συγκεκριμένα δεδομένα παρέχουν ή όχι σημαντική αξία, η οποία σχετίζεται με την ανάλυση Μεγάλων Δεδομένων. Τα Big Data πρόσφατα ορίζονται και σε όρους 10Vs. Τα πέντε επιπλέον χαρακτηριστικά είναι : Validity (Εγκυρότητα), Variability (Μεταβλητότητα), Viscosity (Ιξώδες-Συνθετότητα), Viability (Βιωσιμότητα), and Volatility (Μεταβλητότητα). Αν και αυτά τα 10V ανήκουν στα χαρακτηριστικά των Big Data, είναι γνωστά ως τα 10 «Μεγάλες προκλήσεις» για τα Big Data. Στην παρακάτω εικόνα φαίνονται τα 10Vs χαρακτηριστικά των μεγάλων δεδομένων. [3]



1. *Χαρακτηριστικά Μεγάλων Δεδομένων (Πηγή: M. A. H. S. G. B. A. A. A. S. S. Nawsher Khan, «The 10 Vs, Issues and Challenges of Big Data,» σε Proceedings of the 2018 International Conference on Big Data and Education, 2018 [4])*

1.3 Data Analytics

Ο όρος «Data Analytics είναι ένας όρος που χρησιμοποιείται για να περιγράψει διάφορους στόχους και τεχνικές επεξεργασίας ενός συνόλου δεδομένων.

Υπάρχουν τρεις τύποι Data Analytics:

- Descriptive Analysis-Περιγραφική ανάλυση: είναι μια διαδικασία για τη σύνοψη του υπό διερεύνηση δεδομένων. Μπορεί να χρησιμοποιηθεί για τη δημιουργία τυπικών αναφορών που μπορεί να είναι χρήσιμες για την αντιμετώπιση ερωτήσεων όπως «Τι συνέβη; Ποιο είναι το πρόβλημα? Τι ενέργειες χρειάζονται;»
- Predictive Analysis- Προγνωστική Ανάλυση: τα περιγραφικά analytics, δυστυχώς δεν λένε τίποτα για το μέλλον, αυτός είναι ο λόγος που χρειάζονται predictive analytics. Η προγνωστική ανάλυση χρησιμοποιεί στατιστικά μοντέλα των ιστορικών συνόλων δεδομένων για να προβλέψει το μέλλον. Τα προγνωστικά αναλυτικά στοιχεία είναι χρήσιμα για να απαντηθούν ερωτήσεις όπως «Γιατί συμβαίνει αυτό; Τι θα συμβεί μετά?". Η προγνωστική ικανότητα εξαρτάται από την καλή προσαρμογή του στατιστικού μοντέλου.

- Prescriptive Analysis- Καθοδηγητική ανάλυση: είναι το είδος των αναλυτικών στοιχείων που βοηθούν στη χρήστη διαφορετικών σεναρίων του μοντέλου δεδομένων (προσομοίωση πολλαπλών μεταβλητών, ανίχνευση κρυφών σχέσεων μεταξύ διαφορετικών μεταβλητών). Είναι χρήσιμη για να απαντηθούν ερωτήματα όπως «Τι θα συμβεί αν χρησιμοποιηθεί το συγκεκριμένο σενάριο πόρων», «Ποιο είναι το βέλτιστο σενάριο;». Τα prescriptive δεδομένα χρησιμοποιούνται γενικότερα σε προβλήματα βελτιστοποίησης με χρήση εξελιγμένων αλγορίθμων για την βέλτιστη λύση (π.χ. κλινικές δοκιμές). [5]

2 Big Data στον τομέα της Υγείας

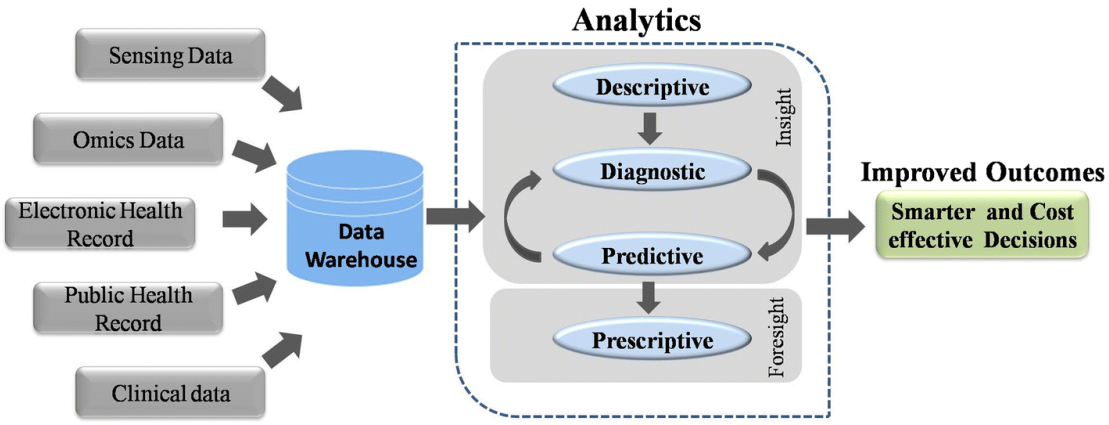
Το ταχέως αναπτυσσόμενο πεδίο της ανάλυσης των μεγάλων δεδομένων έχει ήδη αρχίσει να παίζει έναν κεντρικό ρόλο στην εξέλιξη των πρακτικών αλλά και της έρευνας στον χώρο της υγείας. Παρέχει εργαλεία που αφορούν τη συγκέντρωση, τη διαχείριση, την ανάλυση και την αφομοίωση μεγάλων όγκων από ανόμοια, δομημένα και μη δομημένα δεδομένα που αναπαράγονται από τα τρέχοντα συστήματα υγειονομικής περίθαλψης. Η ανάλυση μεγάλων δεδομένων έχει πρόσφατα εφαρμοστεί για να βοηθήσει τη διαδικασία παροχής φροντίδας και εξερεύνησης ασθενειών. Ωστόσο, το ποσοστό υιοθέτησης και ο ρυθμός ανάπτυξη της έρευνας σε αυτόν τον χώρο εξακολουθούν να παρεμποδίζονται από ορισμένα θεμελιώδη προβλήματα που υπάρχουν στο ζήτημα των μεγάλων δεδομένων. Η υγειονομική περίθαλψη είναι ένα χαρακτηριστικό παράδειγμα του πώς τα τρία Vs των δεδομένων, η ταχύτητα (ταχύτητα παραγωγής δεδομένων), η ποικιλία και ο όγκος, είναι μια έμφυτη πτυχή των δεδομένων που παράγει. Αυτά τα δεδομένα διαδίδονται μεταξύ πολλών συστημάτων υγειονομικής περίθαλψης, ασφαλιστών υγείας, ερευνητών, κυβερνητικών φορέων κ.λπ. Επιπλέον, κάθε ένα από αυτά τα αποθετήρια δεδομένων είναι αποσιωπημένο και εγγενώς ανίκανο να παρέχει μια πλατφόρμα για παγκόσμια διαφάνεια δεδομένων. Για να προσθέσουμε στα τρία V, η ακρίβεια των δεδομένων υγειονομικής περίθαλψης είναι επίσης κρίσιμη για την ουσιαστική χρήση τους για την ανάπτυξη έρευνας. Παρά την πολυπλοκότητα των δεδομένων του χώρου της υγειονομικής περίθαλψης, υπάρχει δυνατότητα του οφέλους από την ανάπτυξη και την εφαρμογή λύσεων μεγάλων δεδομένων σε αυτό τον χώρο. Έπειτα από δεκαετίες τεχνολογικής υστέρησης, ο τομέας της ιατρικής έχει αρχίσει να εγκλιματίζεται στη σημερινή εποχή με την προσφορά των ψηφιακών δεδομένων. Οι νέες τεχνολογίες κάνουν δυνατή τη συλλογή τεράστιων ποσοτήτων πληροφοριών για τον κάθε μεμονωμένο ασθενή σε μεγάλη χρονική κλίμακα. Ωστόσο, παρόλη την εμφάνιση των ιατρικών ηλεκτρονικών ειδών, τα δεδομένα που συγκεντρώθηκαν από τους ασθενείς παρέμειναν σε μεγάλο βαθμό μη επαρκώς αξιοποιημένα και ως εκ τούτου κατέληξαν χαμένα. Η κατανόηση και η πρόβλεψη ασθενειών είναι ζητήματα που απαιτούν μια συγκεντρωτική προσέγγιση όπου δομημένα και αδόμητα δεδομένα που προέρχονται από μια μυριάδα κλινικών και μη κλινικών μέσων που θα μπορούσαν να χρησιμοποιούνται για μια πιο ολοκληρωμένη προοπτική των καταστάσεων της νόσου. Οι ερευνητές μελετούν αυτή τη σύνθετη φύση των δεδομένων υγειονομικής περίθαλψης τόσο σχετικά τα χαρακτηριστικά των ίδιων

των δεδομένων όσο και την ταξινόμηση των αναλυτικών στοιχείων που μπορούν να πραγματοποιηθούν με νόημα σε αυτά. [6]

2.1 Big data analytics στην υγειονομική περίθαλψη

Ο όγκος δεδομένων υγείας αναμένεται να αυξηθεί δραματικά τα επόμενα χρόνια. Επιπλέον, τα μοντέλα αποζημίωσης για την υγειονομική περίθαλψη αλλάζουν. Η ουσιαστική χρήση και η πληρωμή επί της απόδοσης αναδεικνύονται ως οι νέοι κρίσιμοι παράγοντες στο σημερινό περιβάλλον υγειονομικής περίθαλψης. Αν και το κέρδος δεν είναι και δεν πρέπει να είναι πρωταρχικό κίνητρο στον χώρο της υγείας, είναι ζωτικής σημασίας για τους οργανισμούς υγειονομικής περίθαλψης να αποκτήσουν τα διαθέσιμα εργαλεία, υποδομές και τεχνικές για να αξιοποιήσουν αποτελεσματικά τα μεγάλα δεδομένα, διαφορετικά κινδυνεύουν να χάσουν δυνητικά πολλά εκατομμύρια εσόδων και κερδών.

Οι υπάρχουσες τεχνικές ανάλυσης μπορούν να εφαρμοστούν στον τεράστιο όγκο των υπαρχόντων δεδομένων υγείας και ιατρικών δεδομένων που σχετίζονται με τον ασθενή για να επιτευχθεί μια βαθύτερη κατανόηση των αποτελεσμάτων, τα οποία στη συνέχεια μπορούν να εφαρμοστούν στο επίπεδο της περίθαλψης. Στην ιδανική περίπτωση στο μέλλον, τα ατομικά δεδομένα και τα δεδομένα πληθυσμού θα ενημερώνουν κάθε γιατρό και τον ασθενή του κατά τη διαδικασία λήψης αποφάσεων και θα βοηθούν στον καθορισμό της καταλληλότερης θεραπευτικής επιλογής για τον συγκεκριμένο ασθενή. Οι αποθήκες δεδομένων αποθηκεύουν τεράστιες ποσότητες δεδομένων που παράγονται από διάφορες πηγές. Αυτά τα δεδομένα υποβάλλονται σε επεξεργασία χρησιμοποιώντας αναλυτικούς αγωγούς για την απόκτηση εξυπνότερων και προσιτών επιλογών υγειονομικής περίθαλψης. [7]



2. *Διαδικασία ανάλυσης δεδομένων υγείας (Πηγή: S. K. S. M. S. &. S. K. Sabyasachi Dash, «Big data in healthcare: management, analysis and future prospects,» Journal of Big Data, 2019 [8])*

2.2 Πλεονεκτήματα στην υγειονομική περίθαλψη

Με την εξέλιξη της ψηφιοποίησης, τον συνδυασμό και την αποτελεσματική χρήση των μεγάλων δεδομένων, οι οργανισμοί υγειονομικής περίθαλψης (από τα ιατρεία των μεμονωμένων ιδιωτών ιατρών έως και τα μεγάλα νοσοκομειακά δίκτυα και τους οργανισμούς περίθαλψης) θα εκμεταλλευτούν σημαντικά οφέλη. Τα πιο πιθανά οφέλη περιλαμβάνουν την ανίχνευση ασθενειών σε αρχικά στάδια που έτσι μπορούν να αντιμετωπιστούν πιο εύκολα και αποτελεσματικά, την διαχείριση συγκεκριμένων ατομικών και πληθυσμιακών προβλημάτων υγείας αλλά και τον ενδεχόμενο εντοπισμός απάτης στον τομέα της υγειονομικής περίθαλψης πιο άμεσα και αποτελεσματικά. Κάποιες εξελίξεις ή αποτελέσματα μπορεί να προβλεφθούν ή/και να εκτιμηθούν με βάση τις τεράστιες διαθέσιμες ποσότητες ιστορικών δεδομένων, όπως είναι η πρόβλεψη της διάρκειας παραμονής του ασθενή (LOS- length of stay). Προβλέψεις και αποτελέσματα όπως ασθενείς που πιθανότατα δεν θα ωφεληθούν από μια χειρουργική επέμβαση, τυχόν ιατρικές επιπλοκές, ασθενείς που διατρέχουν κίνδυνο για τέτοιες ιατρικές επιπλοκές, ασθενείς που διατρέχουν κίνδυνο για σήψη ή κάποια άλλη νοσοκομειακή ασθένεια, η ίδια η εξέλιξη της ασθένειας, οι αιτιώδεις παράγοντες της νόσου και της προόδου της νόσου και άλλα παρόμοια αποτελέσματα εκτιμάται ότι μπορούν να παραχθούν από την ανάλυση των μεγάλων δεδομένων και να επιτρέψουν και την εξοικονόμηση τεράστιων χρηματικών ποσών στην υγειονομική περίθαλψη. Οι βασικές κατηγορίες που τα μεγάλα δεδομένα θα μπορούσαν να βοηθήσουν στη μείωση της σπατάλης και της αναποτελεσματικότητας είναι:

- Οι κλινικές επεμβάσεις: Στα πλαίσια συγκριτικής έρευνας αποτελεσματικότητας για τον προσδιορισμό πιο σχετικών κλινικά καθώς και οικονομικά αποδοτικών τρόπων διάγνωσης και θεραπείας ασθενειών.
- Η έρευνα και ανάπτυξη: 1) μέσω μοντελοποίησης προβλέψεων για τη μείωση της φθοράς και την χρήση πιο απλών και στοχευμένων φαρμάκων και συσκευών. 2) μέσω στατιστικών εργαλείων και αλγορίθμων για τη βελτίωση του σχεδιασμού των κλινικών δοκιμών σε ασθενείς για καλύτερη αντιστοίχιση των θεραπειών με μεμονωμένους ασθενείς, μειώνοντας έτσι τις αποτυχίες των δοκιμών και επιταχύνοντας νέες θεραπείες στην αγορά. και 3) μέσω της ανάλυσης κλινικών δοκιμών καθώς και αρχείων των ασθενών για τον εντοπισμό επακόλουθων ενδείξεων και την ανακάλυψη άμεσα τυχόν ανεπιθύμητων ενεργειών πριν τα προϊόντα φτάσουν στην αγορά και κατ' επέκταση σε ασθενείς.
- Η δημόσια υγεία: 1) μέσω της ανάλυσης προτύπων ασθενειών και τη παρακολούθηση των εστιών ασθενειών και της μετάδοσης τους για τη βελτίωση της επιτήρησης της δημόσιας υγείας και της ταχύτητας απόκρισης 2) μέσω ταχύτερης ανάπτυξης εμβολίων με μεγαλύτερη αποτελεσματικότητα και 3) μέσω μετατροπής του μεγάλου όγκου δεδομένων σε πληροφορίες που μπορούν να χρησιμοποιηθούν για τον εντοπισμό αναγκών, την παροχή υπηρεσιών, την πρόβλεψη και την πρόληψη κρίσεων, ειδικά προς όφελος του γενικότερου πληθυσμού.
- Η ιατρική βάσει τεκμηρίωσης: Ο συνδυασμός και η ανάλυσης μιας ποικιλίας δομημένων και μη δομημένων δεδομένων, οικονομικών και λειτουργικών, κλινικών και γονιδιωματικών δεδομένων θα μπορέσουν να ταιριάξουν τις θεραπείες με τα αποτελέσματα, και να προβλέψουν τα άτομα σε κίνδυνο για νόσο ή επανεισδοχή και να προσφερθεί αποτελεσματικότερη φροντίδα.
- Οι γονιδιωματικές αναλύσεις ως μέρος της τακτικής διαδικασίας λήψης αποφάσεων σχετικά με την ιατρική περίθαλψη και το αυξανόμενο ιατρικό αρχείο των ασθενών.
- Η ανάλυση απάτης πριν από την εκδίκαση: Η ανάλυση των αιτημάτων αξιώσεων με σκοπό τη μείωση της απάτης, της σπατάλης και της κατάχρησης στο υγειονομικό τομέα.
- Οι Συσκευές και η απομακρυσμένη παρακολούθηση: μέσω της λήψης και ανάλυσης σε πραγματικό χρόνο μεγάλου όγκου δεδομένων γρήγορης κίνησης από

συσκευές στο νοσοκομείο αλλά και στο σπίτι, ως παρακολούθηση ασφάλειας και τυχόν ανεπιθύμητων ενεργειών.

2.3 Πλαίσιο μεθοδολογίας ανάλυσης και επεξεργασίας των μεγάλων δεδομένων

Το εννοιολογικό πλαίσιο για ένα έργο ανάλυσης μεγάλων δεδομένων στον χώρο της υγειονομικής περίθαλψης είναι παρόμοιο με αυτό ενός παραδοσιακού έργου πληροφορικής ή ανάλυσης στον χώρο της υγείας. Η βασική διαφορά έγκειται στον τρόπο εκτέλεσης της επεξεργασίας. Σε ένα κανονικό έργο ανάλυσης υγείας, αυτή μπορεί να εκτελεστεί με ένα εργαλείο επιχειρηματικής ευφυΐας εγκατεστημένο σε ένα αυτόνομο σύστημα. Επειδή τα μεγάλα δεδομένα είναι εξ ορισμού μεγάλου όγκου, η επεξεργασία τους αναλύεται και εκτελείται σε πολλαπλούς κόμβους. Η έννοια της κατανεμημένης επεξεργασίας υπάρχει εδώ και δεκαετίες. Αυτό όμως που είναι σχετικά καινούργιο είναι η χρήση της στην ανάλυση πολύ μεγάλων συνόλων δεδομένων καθώς οι πάροχοι υγειονομικής περίθαλψης αρχίζουν να αξιοποιούν τα μεγάλα αποθετήρια δεδομένων τους για να αποκτήσουν τις πληροφορίες για τη λήψη καλύτερα ενημερωμένων και αποτελεσματικών αποφάσεων σχετικά με τον τομέα της υγείας. Επιπλέον, πλατφόρμες ανοιχτού κώδικα όπως το Hadoop/MapReduce, που είναι διαθέσιμες στο cloud, έχουν ενθαρρύνει την εφαρμογή αναλυτικών στοιχείων μεγάλων δεδομένων στην υγειονομική περίθαλψη.

Ενώ οι αλγόριθμοι και τα μοντέλα είναι παρόμοια, οι διεπαφές του χρήστη των παραδοσιακών εργαλείων ανάλυσης και εκείνων που χρησιμοποιούνται για μεγάλα δεδομένα είναι τελείως διαφορετικές. Τα παραδοσιακά εργαλεία ανάλυσης υγείας έχουν γίνει πολύ φιλικά και διαφανή. Τα εργαλεία ανάλυσης μεγάλων δεδομένων, από την άλλη, είναι εξαιρετικά πολύπλοκα, απαιτούν ένταση προγραμματισμού και εφαρμογή ποικίλων δεξιοτήτων. Έχουν εμφανιστεί, με ad hoc τρόπο ως επί το πλείστων, ως εργαλεία και πλατφόρμες ανάπτυξης ανοιχτού κώδικα, με αποτέλεσμα να μην διαθέτουν την υποστήριξη και τη φιλικότητα προς τον χρήστη που διαθέτουν τα ιδιόκτητα εργαλεία που βασίζονται στον προμηθευτή. Η πολυπλοκότητα όμως ξεκινά από τα ίδια τα δεδομένα.

Τα μεγάλα δεδομένα της υγειονομικής περίθαλψης μπορούν να προέρχονται από εσωτερικές πηγές (ηλεκτρονικά αρχεία, συστήματα υποστήριξης κλινικών αποφάσεων κ.λπ.) καθώς και εξωτερικές (κρατικές υπηρεσίες, εργαστήρια, φαρμακεία, ασφαλιστικές εταιρείες κ.λπ.), συχνά σε πολλαπλές μορφές (επίπεδα αρχεία, .csv, σχεσιακούς πίνακες, απλά κείμενα κ.λπ.) και σε πολλαπλές τοποθεσίες (γεωγραφικές τοποθεσίες καθώς και σε ιστότοπους διαφορετικών παρόχων υγειονομικής περίθαλψης) σε πολλές παλαιού τύπου και άλλες εφαρμογές (εφαρμογές επεξεργασίας συναλλαγών, βάσεις δεδομένων κ.λπ.). Οι πηγές και οι τύποι δεδομένων περιλαμβάνουν:

1. Δεδομένα ιστού και μέσων κοινωνικής δικτύωσης: Δεδομένα αλληλεπίδρασης από εφαρμογές όπως είναι το Facebook, το Twitter, το LinkedIn, τα ιστολόγια και οτιδήποτε αντίστοιχο. Μπορούν επίσης να περιλαμβάνονται επίσης δεδομένα ιστότοπων υγείας, εφαρμογών smartphone κ.λπ.
2. Δεδομένα από μηχανή σε μηχανή: μετρήσεις από απομακρυσμένους αισθητήρες, μετρητές και άλλες αντίστοιχες συσκευές.
3. Μεγάλα δεδομένα συναλλαγών: αξιώσεις που αφορούν την υγειονομική περίθαλψη και άλλα αρχεία τιμολόγησης διατίθενται όλο και περισσότερο σε ημι-δομημένες και μη δομημένες μορφές δεδομένων.
4. Βιομετρικά δεδομένα: δακτυλικά αποτυπώματα, σαρώσεις αμφιβληστροειδούς, ακτινογραφίες και άλλες ιατρικές απεικονίσεις, η αρτηριακή πίεση, οι μετρήσεις παλμών και παλμικής οξυμετρίας και άλλα παρόμοια είδη δεδομένων.
5. Δεδομένα που δημιουργούνται από τον ίδιο τον άνθρωπο: μη δομημένα και ημι-δομημένα δεδομένα όπως σημειώσεις γιατρών, email και έντυπα έγγραφα.

Για τους σκοπούς της ανάλυσης μεγάλων δεδομένων, αυτά τα δεδομένα πρέπει αρχικά να συγκεντρωθούν. Τα δεδομένα βρίσκονται σε «ακατέργαστη» κατάσταση και πρέπει να υποστούν επεξεργασία ή μετασχηματισμό, οπότε υπάρχουν στην συνέχεια πολλές επιλογές. Μια αρχιτεκτονική προσέγγιση προσανατολισμένη στις υπηρεσίες σε συνδυασμό με υπηρεσίες Web (ενδιάμεσο λογισμικό) είναι μια τέτοια επιλογή. Τα δεδομένα παραμένουν ακατέργαστα και οι υπηρεσίες χρησιμοποιούνται για την κλήση, την ανάκτηση και την επεξεργασία των δεδομένων. Μια άλλη επιλογή είναι η αποθήκευση δεδομένων όπου τα δεδομένα από διάφορες πηγές συγκεντρώνονται και είναι έτοιμα για επεξεργασία. Μέσω των βημάτων εξαγωγής, μετασχηματισμού και φόρτωσης, τα δεδομένα από διάφορες πηγές «καθαρίζονται» και προετοιμάζονται.

Ανάλογα με το αν τα δεδομένα είναι δομημένα ή μη, μπορούν να εισαχθούν διάφορες μορφές δεδομένων στην πλατφόρμα ανάλυσης μεγάλων δεδομένων.
Στο επόμενο στάδιο του πλαισίου μεθοδολογίας, λαμβάνονται αρκετές αποφάσεις σχετικά με την προσέγγιση της εισαγωγής δεδομένων, τον κατανεμημένο σχεδιασμό, την επιλογή εργαλείων και τα επιλεγμένα μοντέλα ανάλυσης. Με βάση τομείς όπως η στατιστική, η επιστήμη των υπολογιστών, τα εφαρμοσμένα μαθηματικά και τα οικονομικά, μια μεγάλη ποικιλία τεχνικών και τεχνολογιών έχει αναπτυχθεί και έχει προσαρμοστεί αντίστοιχα για συγκέντρωση, χειρισμό, ανάλυση και οπτικοποίηση των μεγάλων δεδομένων στον τομέα της υγειονομικής περίθαλψης.

Η πιο γνωστή πλατφόρμα για την ανάλυση των μεγάλων δεδομένων είναι η πλατφόρμα επεξεργασίας κατανεμημένων δεδομένων ανοιχτού κώδικα Hadoop (πλατφόρμα Apache) που αρχικά είχε αναπτυχθεί για λειτουργίες ρουτίνας όπως η συγκέντρωση ευρετηρίων αναζήτησης ιστού. Ανήκει στην κατηγορία τεχνολογιών «NoSQL» και διαθέτει τη δυνατότητα επεξεργασίας εξαιρετικά μεγάλων ποσοτήτων δεδομένων, κυρίως μέσω της κατανομής των διαμελισμένων συνόλων δεδομένων σε πολλούς διακομιστές (κόμβους), καθένας από τους οποίους επιλύει διαφορετικά μέρη του μεγαλύτερου προβλήματος και στη συνέχεια ενσωματώνονται για το τελικό αποτέλεσμα. Το Hadoop εξυπηρετεί τους διπλούς ρόλους του οργανωτή δεδομένων και του εργαλείου ανάλυσης. Προσφέρει επίσης πολλές δυνατότητες για να επιτρέψει στις επιχειρήσεις να αξιοποιήσουν τα δεδομένα, πράγμα που μέχρι τώρα ήταν δύσκολο να διαχειριστούν και να αναλύσουν. Πιο συγκεκριμένα, το Hadoop καθιστά δυνατή την επεξεργασία εξαιρετικά μεγάλου όγκου δεδομένων με διάφορες δομές ή και με καθόλου δομή. Παρόλα αυτά, το Hadoop μπορεί να είναι δύσκολο να εγκατασταθεί, καθώς και να ρυθμιστεί και να διαχειριστεί, και άτομα με δεξιότητες Hadoop δεν βρίσκονται εύκολα. Επιπλέον, για τους παραπάνω λόγους, φαίνεται ότι οι οργανισμοί δεν είναι αρκετά προετοιμασμένοι να αποδεχτούν πλήρως το Hadoop.
Ενώ τα διαθέσιμα πλαίσια και τα εργαλεία είναι ως επί το πλείστων ανοιχτού κώδικα και περιστρέφονται γύρω από το Hadoop και τις αντίστοιχες πλατφόρμες, υπάρχουν πολλά θέματα που πρέπει να ληφθούν υπόψη από τους προγραμματιστές και τους χρήστες των αναλυτικών στοιχείων μεγάλων δεδομένων στον τομέα της υγείας. Ενώ το κόστος ανάπτυξης μπορεί να είναι χαμηλότερο, καθώς αυτά τα εργαλεία είναι ανοιχτού κώδικα και δωρεάν, τα μειονεκτήματα τους είναι η έλλειψη τεχνικής υποστήριξης και η ελάχιστη ασφάλεια που προσφέρουν. Στον κλάδο της υγειονομικής

περίθαλψης, αυτά είναι, φυσικά πολύ σημαντικά μειονεκτήματα λόγω της φύσης των δεδομένων, και ως εκ τούτου πρέπει να αντιμετωπιστούν. Επιπλέον, αυτές οι πλατφόρμες/εργαλεία απαιτούν πολύ προγραμματισμό και δεξιότητες που μπορεί να μην διαθέτει ένας τυπικός τελικός χρήστης στην υγειονομική περίθαλψη. Ακόμη, λαμβάνοντας υπόψη τη πολύ πρόσφατη εμφάνιση των αναλυτικών στοιχείων μεγάλων δεδομένων στην υγειονομική περίθαλψη, σημαντικά ζητήματα διακυβέρνησης, όπως η ιδιοκτησία, το απόρρητο, η ασφάλεια και τα πρότυπα, δεν έχουν ακόμη αντιμετωπιστεί. [9]

Μέσω λοιπόν μιας σειράς επαναλήψεων και αναλύσεων what-if, αποκτάται η γνώση από την ανάλυση μεγάλων δεδομένων και θα μπορέσουν να ληφθούν τεκμηριωμένες αποφάσεις για τον τομέα της υγείας. Τα μοντέλα και τα ευρήματά τους ελέγχονται και επικυρώνονται και παρουσιάζονται στα ενδιαφερόμενα μέρη για να προχωρήσουν σε δράση. Η υλοποίηση είναι μια σταδιακή προσέγγιση με ενσωματωμένους σταθμούς ανάδρασης σε κάθε βήμα για την ελαχιστοποίηση του κινδύνου αποτυχίας.
Απαιτούνται προηγμένοι αλγόριθμοι για την εφαρμογή προσεγγίσεων ML και AI για ανάλυση μεγάλων δεδομένων σε συμπλέγματα υπολογιστών. Μια γλώσσα προγραμματισμού κατάλληλη για εργασία σε μεγάλα δεδομένα (Python, R ή άλλες γλώσσες που αναλύονται παρακάτω στην παρούσα εργασία) θα μπορούσε να χρησιμοποιηθεί για τη σύνταξη τέτοιων αλγορίθμων ή λογισμικού. Επομένως, η καλή γνώση της πληροφορικής είναι απαιτούμενη για τον χειρισμό των μεγάλων δεδομένων από τη έρευνα στον χώρο της υγείας.

3 Ανάπτυξη τεχνικών πρόβλεψης

3.1 Ανάλυση Χρονοσειρών

3.1.1 Εισαγωγή στις χρονοσειρές

Σε πολλούς επιστημονικούς κλάδους παρατηρείται η προσπάθεια που γίνεται με διάφορες μελέτες/έρευνες ώστε να δημιουργηθούν τα μοντέλα που θα εκφράζουν την εξάρτηση ενός μεγέθους σε σχέση με άλλα μεγέθη. Τα μοντέλα αυτά αναπτύσσονται με βάση την όποια υπάρχουσα γνώση έχουν ήδη οι ερευνητές για το πρόβλημα που αφορά η μελέτη και αυτά τα μοντέλα ονομάζονται μοντέλα βασικών αρχών. Παρόλα αυτά πολλές φορές ισχύει το γεγονός οι ερευνητές να μην γνωρίζουν αυτές τις σχέσεις για την ανάπτυξη των μοντέλων βασικών αρχών ή να μην θέλουν να δημιουργήσουν αυτά τα μοντέλα και έτσι η μελέτη των φαινόμενων που ενδιαφέρουν καταλήγει να γίνεται με βάση την παρατήρηση των σχετικών μεγεθών. Αυτού του είδους τα μοντέλα ονομάζονται εμπειρικά ή παραγόμενα από τα υπάρχοντα δεδομένα (Empirical or Data Driven Models). Τα μοντέλα που όπως έχει σημειωθεί και παραπάνω, δεν προϋποθέτουν γνώση του προβλήματος, αλλά στηρίζονται στα υπάρχοντα δεδομένα και πιο συγκεκριμένα σε υπάρχοντα δεδομένα με χρονική διάταξη είναι οι χρονοσειρές. Πιο συγκεκριμένα, οι χρονοσειρές αφορούν προβλήματα στα οποία ένα μέγεθος αλλάζει τιμές στον χρόνο και γίνεται μελέτη της εξάρτησης του μεγέθους αυτού σε κάποια χρονική στιγμή t, σε σχέση με τις τιμές του ίδιου μεγέθους σε προηγούμενες χρονικές στιγμές t-1, t-2,...Με αυτόν τον τρόπο εξετάζεται η εξέλιξη της διαδικασίας που παράγει αυτό το παρατηρούμενο μέγεθος. Με τον όρο χρονοσειρά ή χρονολογική σειρά ορίζεται μια ακολουθία xt : t = 0, 1, 2, ..., όπου κάθε xt εκφράζει την κατά την χρονική στιγμή t κατάσταση ενός συστήματος που εξελίσσεται στο χρόνο κατά ένα τυχαίο τρόπο. Οποιοδήποτε αντίστοιχο φαινόμενο που εξελίσσεται στον χρόνο θεωρείται μια στοχαστική διαδικασία, και έτσι καταλήγουμε στο συμπέρασμα ότι οι χρονοσειρές αποτελούν στοχαστικές διαδικασίες [10]. Στις χρονοσειρές οι παρατηρήσεις της υπό έλεγχο μεταβλητής εμφανίζουν προκαθορισμένα και ίσα χρονικά διαστήματα που σημαίνει ότι χαρακτηρίζονται από σταθερό χρονικό βήμα ή διαφορετικά σταθερή δειγματοληψία. Παρόλα αυτά σε κάποιες περιπτώσεις ο χρόνος δειγματοληψίας δεν είναι σταθερός και σε τέτοιες περιπτώσεις χρειάζεται ειδική επεξεργασία της χρονοσειράς για να γίνει η ανάλυση της. Οι χρονοσειρές έχουν εφαρμογή σε διάφορους επιστημονικούς τομείς, όπως τις Οικονομικές Επιστήμες, την

Ιατρική, την Κοινωνιολογία κ.α. και αυτό διότι η ανάλυση των χρονοσειρών επιτρέπει την πρόβλεψη των τιμών του μεγέθους που γίνεται η παρατήρηση. Οι χρονοσειρές καταγράφουν τις παρελθοντικές τιμές ενός μεγέθους με σκοπό την εκτίμηση των μελλοντικών τιμών. Οι χρονολογικές σειρές δεν χρησιμοποιούνται αποκλειστικά και μόνο ως ένα εργαλείο πρόγνωσης και πρόβλεψης, αλλά είναι ένα εργαλείο σίγουρα χρήσιμο για την επεξήγηση και την κατανόηση της συμπεριφοράς του ίδιου του φαινομένου, αφού γίνεται καταγραφή όλης της ιστορίας του. [11]

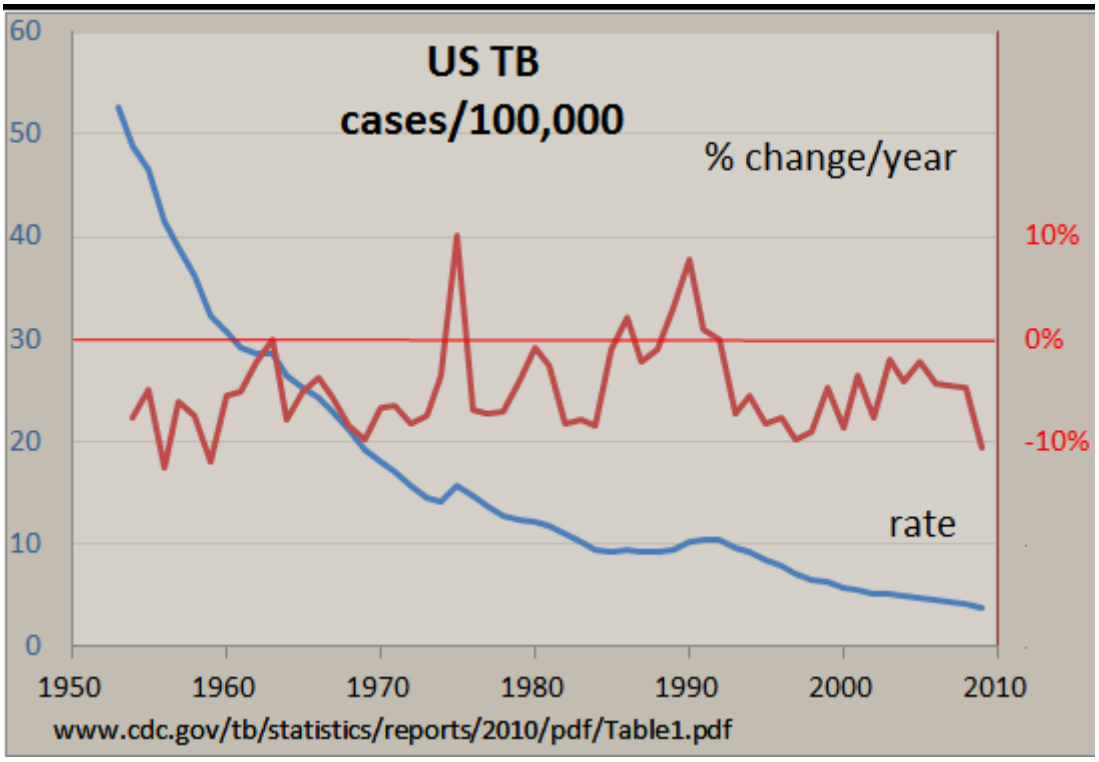
Η μελέτη των χρονοσειρών γίνεται με την ανάλυση των χρονοσειρών. Η ανάλυση των χρονοσειρών έχει ως στόχο τα παρακάτω σημεία:

1. Την αναζήτηση του μαθηματικού μοντέλου που θα αναλύει και θα περιγράφει το υπό ανάλυση φαινόμενο.

2. Την αναζήτηση του κατάλληλου μοντέλου (όχι απαραίτητα και του αληθινού μοντέλου), το οποίο θα παράγει τις καλύτερες προβλέψεις που ενδιαφέρουν τον μελετητή.

Στην παρούσα διπλωματική εργασία γίνεται αναφορά αυτής της αναζήτησης των κατάλληλων μοντέλων με στόχο την παραγωγή προβλέψεων σε διάφορα σύνολα δεδομένων. Τα μοντέλα προβλέψεων των χρονοσειρών χωρίζονται σε γραμμικά ή μη-γραμμικά ανάλογα με το αν η συνάρτηση η οποία περιγράφεται από το κάθε μοντέλο είναι γραμμική ή όχι. Επιπλέον, γίνεται διαχωρισμός τους σε υποκειμενικές/ποιοτικές (subjective/qualitive) και σε αντικειμενικές/ποσοτικές (objective/quantitive). Στα υποκειμενικά μοντέλα οι προβλέψεις εφαρμόζονται από έμπειρους αναλυτές, οι οποίοι χρησιμοποιούν κυρίως την προσωπική τους κρίση, ενώ στα αντικειμενικά μοντέλα, οι προβλέψεις στηρίζονται αποκλειστικά και μόνο σε κάποιο μαθηματικό μοντέλο. Ο βασικός στόχος της ανάλυσης αυτής είναι η ανεύρεση των βασικών χαρακτηριστικών της χρονοσειράς που εξετάζεται και η περιγραφή της εσωτερικής δομής της.

Σε αρχικό στάδιο η ανάλυση μίας χρονοσειράς περιλαμβάνει την γραφική απεικόνιση των τιμών της σε συνάρτηση με το χρόνο. Με αυτό τον τρόπο, τα βασικά χαρακτηριστικά της χρονοσειράς, η τάση, η κυκλικότητα, η εποχικότητα και οι ακραίες τιμές, αποτυπώνονται ως γραφικά μοτίβα. Η αναγνώριση αυτών των βασικών χαρακτηριστικών της χρονοσειράς καθορίζει και το είδος της ανάλυσης που θα πρέπει να ακολουθηθεί με την επιλογή και του πλέον κατάλληλου μοντέλου.



3. *Η επίπτωση της φυματίωσης ΗΠΑ 1953-2009 (Πηγή: «https://www.wikiwand.com/el/Χρονολογικές_Σειρές,» [12])*

Τέλος στόχος της ανάλυσης των χρονοσειρών αποτελεί όπως έχει αναφερθεί ξανά, η πρόβλεψη των μελλοντικών τιμών της χρονοσειράς και ο προσδιορισμός της αβεβαιότητας αυτών των προβλέψεων που προκύπτουν. Ο κύριος στόχος αυτής της πρόβλεψης είναι όσο το δυνατό η μεγαλύτερη ακρίβεια των προβλεπόμενων τιμών συγκριτικά με τις πραγματικές μελλοντικές τιμές έτσι ώστε να βοηθήσει στη σωστή και έγκαιρη λήψη αποφάσεων με βάση τις προβλέψεις αυτές.

Βασικά Χαρακτηριστικά και Κατηγορίες Χρονοσειρών

Η χρονολογική σειρά αναλύεται στα επιμέρους χαρακτηριστικά της. Με βάση αυτά τα επιμέρους χαρακτηριστικά, οι χρονοσειρές κατηγοριοποιούνται και αντίστοιχα. Για να γίνει μελέτη της χρονοσειράς πρέπει να δημιουργηθεί και το αντίστοιχο γράφημα των τιμών της στο πεδίο του χρόνου. Αυτό είναι το πρώτο βήμα της ανάλυσης μιας χρονοσειράς που παρατηρούνται τα βασικά χαρακτηριστικά της εύκολα, χαρακτηριστικά όπως η τάση, η κυκλικότητα, η εποχικότητα και οι ακραίες τιμές.

Οι χρονοσειρές διαχωρίζονται σε μονοδιάστατες και πολυδιάστατες χρονοσειρές με γνώμονα με το πλήθος των μεγεθών που καταγράφονται. Οι μονοδιάστατες

χρονοσειρές καταγράφουν τιμές ενός μεγέθους και αποτελούνται από μία ακολουθία τιμών της ίδιας μεταβλητής στο πέρασμα του χρόνου. Όταν μια χρονοσειρά περιλαμβάνει παραπάνω από μία μεταβλητές, τότε μιλάμε για μια πολυδιάστατη χρονοσειρά. Στις πολυδιάστατες χρονοσειρές υπάρχει η δυνατότητα ταυτόχρονης παρατήρησης πολλών μεγεθών για το ίδιο σύστημα. Σε αυτή την περίπτωση, είναι συχνό οι μεταβλητές να αλληλοσυσχετίζονται με την πάροδο του χρόνου. Αν μία μεταβλητή Χ είναι χρήσιμη για την πρόγνωση μελλοντικών τιμών της μεταβλητής Υ τότε η πολυδιάστατη μεταβλητή είναι ομογενής (homogeneous), διαφορετικά είναι ετερογενής (heterogeneous). Στις ομογενείς πολυδιάστατες χρονοσειρές, με την οποιαδήποτε αλλαγή που μπορεί να προκληθεί σε ένα στοιχείο των παρατηρήσεων της μίας μεταβλητής θα προκύψει η ίδια αλλαγή στις παρατηρήσεις των άλλων μεταβλητών που σχετίζονται με το φαινόμενο που παρατηρούμε. Στην περίπτωση των πολυδιάστατων χρονοσειρών είναι σημαντικό να αναφερθεί και η έννοια των διανυσματικών χρονοσειρών που περιέχει ως συστατικά μονοδιάστατες χρονοσειρές.

Στάσιμες Χρονοσειρές (Stationary Time Series) είναι οι χρονοσειρές στις οποίες τα στατιστικά μέτρα τους, όπως η μέση τιμή, η διασπορά, η μικτή ροπή 2ας τάξης, δηλαδή η συνδιακύμανση μένουν αναλλοίωτα στο χρόνο. Όταν όλα τα στατιστικά μέτρα μένουν αναλλοίωτα στο χρόνο τότε μιλάμε για μια αυστηρή στασιμότητα Η μη-στασιμότητα αποτελεί ένα σημαντικό πρόβλημα στην ανάλυση χρονοσειρών και πολύ περισσότερο όταν γίνεται προσπάθεια παραγωγής προβλέψεων μέσω της ανάλυσης αυτής. Ως χαρακτηριστικά παρουσίας μη στασιμότητας καταγράφονται κυρίως η ύπαρξη τάσης, εποχικότητας, κυκλικότητας και ακραίων τιμών.

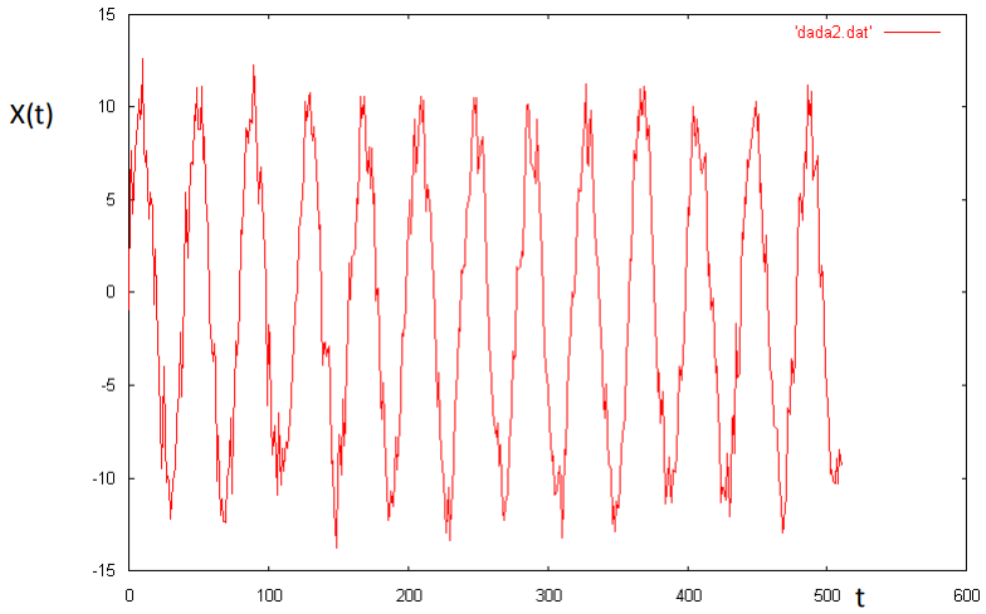
Τάση (Trend)

Η μακροχρόνια ομαλή κεντρική κίνηση που ακολουθείται από τη χρονολογική σειρά κατά τη διάρκεια μιας ολόκληρης χρονικής περιόδου ονομάζεται τάση (trend). Η τάση μπορεί να είναι ανοδική, καθοδική ή σύνθετη. Η τάση θεωρείται ανύπαρκτη εάν η κεντρική ομαλή κίνηση της χρονοσειράς ακολουθεί μια νοητή ευθεία παράλληλη με τον άξονα του χρόνου. Η τάση μπορεί να αναπαρασταθεί είτε ως απλή γραμμική συνάρτηση με το χρόνο ή και ως πολυωνυμική συνάρτηση του χρόνου ή εκθετική. Γενικότερα όταν η τάση της χρονοσειράς περιγράφεται από κάποια γνωστή ή κάποια εκτιμώμενη συνάρτηση του χρόνου, τότε ονομάζεται καθοριστική η τάση της (deterministic trend). Αν όμως η τάση δεν μπορεί να περιγραφεί από κάποια γνωστή

(παραμετρική) συνάρτηση του χρόνου, παρουσιάζει δηλαδή αργές μεταβολές σε σχέση με το χρόνο αλλά όχι με κάποιο καθοριστικό τρόπο, τότε αυτή η τάση ονομάζεται στοχαστική (stochastic trend). [13]

Εποχικότητα (Seasonality)

Οι χρονοσειρές που τα δεδομένα τους εμφανίζονται κατ' επανάληψη με τον ίδιο περίπου τρόπο σε συγκεκριμένα χρονικά διαστήματα, τότε αυτές οι χρονοσειρές παρουσιάζουν εποχικότητα. Ως εποχικότητα καλείται η περιοδική διακύμανση η οποία έχει σταθερό και μικρότερο ή ίσο μήκος του ενός έτους. Όταν η εποχική διακύμανση εμφανίζεται με τρόπο συστηματικό, τότε είναι ένα χαρακτηριστικό που εύκολα μπορεί να αναγνωριστεί οπτικά, να μετρηθεί και να απομονωθεί ώστε να μην επηρεάζει τα δεδομένα που μελετούνται. Η νέα χρονοσειρά που προκύπτει με αυτό τον τρόπο της απομόνωσης ονομάζεται αποεποχικοποιημένη χρονοσειρά. [14]



4. Παράδειγμα τάσης και εποχικότητας (Πηγή: «ΥΔΡΟΛΟΓΙΚΗ ΠΡΟΣΟΜΟΙΩΣΗ ΚΑΙ ΠΡΟΒΛΕΨΗ,» Τμήμα Πολιτικών Μηχανικών Πανεπιστήμιο Θεσσαλίας [15])

Κυκλικότητα (Cyclic)

Η χρονοσειρά μπορεί να εμφανίζει κυκλικότητα στη συμπεριφορά των τιμών της κατά την παρατήρηση των διακυμάνσεων με ανοδικές και καθοδικές φάσεις που επαναλαμβάνονται διαδοχικά γύρω από τη γραμμή τάσης. Η κυκλική συμπεριφορά ορίζεται από δύο κάτω σημεία καμπής (trough) και ένα άνω σημείο καμπής (peak) το

οποίο παρεμβάλλεται μεταξύ τους. Οι κυκλικές μεταβολές δεν επαναλαμβάνονται σε κανονικά χρονικά διαστήματα με αποτέλεσμα η συχνότητα μεταξύ των δύο "peak" ή των δύο "trough" να μην είναι σταθερή, και έτσι να μην υπάρχει καθορισμένη, με σταθερό μήκος, περίοδος. Η κυκλικότητα αυτή ως χαρακτηριστικό της χρονοσειράς παρουσιάζεται γραφικά ως μεταβολή που κυμαίνεται από μία χαμηλή στάθμη σε μία πιο υψηλή και παρατηρείται ολοκλήρωση του κύκλου αυτού σε χρονικό διάστημα μεγαλύτερου του ενός έτους, συνήθως σε διάστημα της πενταετίας ή της δεκαετίας. Η κυκλική κίνηση δεν ακολουθεί κανονικό μοντέλο και κινείται απρόβλεπτα, γι' αυτό και στην πράξη οι κυκλικές αυξομειώσεις είναι δύσκολο να αντιμετωπιστούν. Η κυκλικότητα συμβαίνει κυρίως σε χρονοσειρές της Οικονομικής Επιστήμης, όπως για παράδειγμα το Ακαθάριστο Εθνικό Προϊόν, λόγω των ανόδων και των υφέσεων που εμφανίζουν οι οικονομίες. [16]

Ιδιάζοντα Σημεία (Outliers)

Τα ιδιάζοντα σημεία (outliers) είναι οι παρατηρήσεις που είναι απομονωμένες, δηλαδή αφορούν ακραίες τιμές που εμφανίζονται σε ένα γράφημα κάποιας χρονοσειράς ως αλλαγές απότομες ως προς το πρότυπο της συμπεριφοράς της. Τα ιδιάζοντα σημεία χαρακτηρίζονται ως μη προβλέψιμα και η επίδρασή που έχουν στην χρονοσειρά έχει μικρή χρονική διάρκεια. Η ερμηνεία αυτών των παρατηρήσεων πρέπει να γίνεται με πολύ προσοχή γιατί απαιτεί θεωρητική γνώση, κριτική ικανότητα και κοινή λογική. Ένα ιδιάζων σημείο μπορεί να αντιπροσωπεύει μια παρατήρηση ασυνήθιστη που έχει προκληθεί από κάποιο απρόβλεπτο γεγονός ή κάποιο σφάλμα του συστήματος καταγραφής παρατηρήσεων. Ένα τέτοιο παράδειγμα είναι το γεγονός μιας απεργίας μπορεί να προκαλέσει μια μεγάλη πτώση της παραγωγής ενός εργοστασίου. [17]

Συσχέτιση Χρονοσειράς

Με τον όρο συσχέτιση αναφερόμαστε στη σχέση μεταξύ δύο τυχαίων μεταβλητών. Πιο συγκεκριμένα, η συσχέτιση μίας χρονοσειράς έχει να κάνει με την ύπαρξη εξάρτησης μεταξύ μίας τιμής της χρονοσειράς μια συγκεκριμένη χρονική στιγμή και μίας άλλης τιμής με χρονική υστέρηση h (lag= h). Αυτό σημαίνει ότι η μεταβολή μίας τιμής οφείλεται στη συμπεριφορά της προηγούμενης από αυτή τιμής αν $h = 1$ ή της h - υστέρησής της .

Λευκός Θόρυβος

Το βασικό δομικό στοιχείο των χρονοσειρών είναι ο Λευκός Θόρυβος (White Noise). Θεωρώντας τα διαδοχικά στοιχεία της χρονοσειράς ως τυχαίες μεταβλητές, τότε αυτές αποτελούν ανεξάρτητες τυχαίες μεταβλητές με ίδια κατανομή (independent and identically distributed, iid) με δεδομένο ότι οι τυχαίες μεταβλητές για $\tau > 1$ έχουν την ίδια κατανομή και είναι ανεξάρτητες μεταξύ τους. Μια iid χρονοσειρά είναι εντελώς τυχαία και δεν περιέχει αυτοσυσχετίσεις (γραμμικές ή μη-γραμμικές), δηλαδή δεν υπάρχουν συσχετίσεις μεταξύ των τυχαίων μεταβλητών της χρονοσειράς. Μια iid χρονοσειρά ονομάζεται και λευκός θόρυβος (white noise). Αν επιπρόσθετα αυτές οι τυχαίες μεταβλητές της χρονοσειράς λευκού θορύβου ακολουθούν κανονική (Γκαουσιανή) κατανομή, τότε η χρονοσειρά ονομάζεται Γκαουσιανός λευκός θόρυβος (Gaussian white noise).

3.1.2 Πρόβλεψη

Ο τελικός και βασικός στόχος της ανάλυσης μίας χρονοσειράς που μελετάται και στην παρούσα διπλωματική εργασία, είναι η πραγματοποίηση των προβλέψεων (Forecasting) που αφορούν τις μελλοντικές τιμές της χρονοσειράς, καθώς και ο προσδιορισμός της αβεβαιότητας αυτών των προβλέψεων που προκύπτουν. Η ανάλυση μιας χρονολογικής σειράς δεδομένων μιας μεταβλητής έχει ως στόχο την πρόβλεψη αυτή των μελλοντικών τιμών της αντίστοιχης μεταβλητής, με βάση, κυρίως τις παρελθούσες τιμές της ίδιας της μεταβλητής που αναλύεται. Έτσι γίνεται η υπόθεση ότι η συμπεριφορά του παρελθόντος της χρονολογικής σειράς θα είναι όμοια στο παρόν και στο μέλλον χωρίς ιδιαίτερες διαφοροποιήσεις. Αυτή η υπόθεση αποτελεί το βασικό μειονέκτημα της προβλεπτικής ικανότητας των χρονολογικών σειρών γιατί άλλοτε μπορεί να ισχύει και άλλοτε όχι. Η χρήση ερμηνευτικών μεταβλητών στις χρονολογικές σειρές λύνει το πρόβλημα αυτό σε κάποιο επίπεδο, καθώς υπάρχουν παράγοντες που επιδρούν στη διαμόρφωση των τιμών μιας χρονολογικής σειράς από την χρονική στιγμή κατασκευής του υποδείγματος αλλά υπάρχουν και κάποιοι άλλοι οι οποίοι πιθανόν να μην έχουν συμπεριληφθεί στο υπόδειγμα. Ως αποτέλεσμα αυτού, η βασικότερη υπόθεση που μπορεί να γίνει με την ανάλυση των χρονολογικών σειρών είναι ότι το πρότυπο συμπεριφοράς της χρονοσειράς θα είναι όμοιο στο μέλλον. Αυτό σημαίνει ότι αν οι εξωτερικοί παράγοντες που διαμορφώνουν σημαντικά τις τιμές μιας

χρονολογικής σειράς παραμένουν σταθεροί, τότε η χρονολογική σειρά αυτή δεν θα παρουσιάζει έντονες διαφοροποιήσεις στις μελλοντικές τιμές της και η πρόβλεψη των μελλοντικών τιμών θα είναι αρκετά ικανοποιητική ως προς την βεβαιότητα τους.

Άλλο ένα ζήτημα που πρέπει να ληφθεί υπόψιν κατά τις προβλέψεις με βάση την ανάλυση των χρονολογικών σειρών είναι αυτό του χρονικού μήκους των προβλεπόμενων τιμών, το πόσο δηλαδή απέχει χρονικά η ταυτοποίηση του υποδείγματος από τις προβλεπόμενες τιμές που παράγονται από το υπόδειγμα αυτό. Ένα χαρακτηριστικό των χρονολογικών σειρών είναι ότι οι προβλέψεις αυτών είναι ακριβείς για σύντομο χρονικό διάστημα στο μέλλον. Στην πραγματικότητα, οι προβλεπόμενες μελλοντικές τιμές μιας χρονολογικής σειράς είναι αποτέλεσμα της σύνθεσης προβλέψεων των κύριων συνιστωσών αυτής, δηλαδή των χαρακτηριστικών της όπως της τάσης, της εποχικότητας, της κυκλικότητας και της άρρυθμης μεταβολής. Οι τρεις πρώτες κύριες συνιστώσες μπορούν να εντοπιστούν ως τμήματα του συνολικού υποδείγματος, επομένως μπορούν και να μοντελοποιηθούν αντίστοιχα με αποτέλεσμα να δημιουργούν προβλέψεις για το μέλλον. Στην αντίθετη πλευρά από τις τρεις αυτές βασικές συνιστώσες, η άρρυθμη μεταβολή είναι αυτή που δεν μπορεί να προβλεφθεί. Οπότε αν η συνεισφορά της άρρυθμης συνιστώσας είναι υψηλή στο τελικό υπόδειγμα της χρονολογικής σειράς, τότε μειώνεται η προβλεπτική ικανότητα του υποδείγματος που χρησιμοποιείται. [18]

3.2 Ανάπτυξη Εφαρμογών Ανάλυσης Δεδομένων

3.2.1 Γλώσσες και τεχνολογίες

Οι πιο δημοφιλείς γλώσσες ανάπτυξης εφαρμογών για την ανάλυση δεδομένων είναι οι R, η Python, καθώς και οι Java και C++ που ξεχωρίζουν λόγω της ταχύτητας τους και αυτό τις κάνει αρκετά δημοφιλείς σε σύγκριση με άλλες γλώσσες προγραμματισμού. Παρακάτω αναφέρονται τα βασικά χαρακτηριστικά των συγκεκριμένων γλωσσών προγραμματισμού και πως χρησιμοποιούνται για την ανάλυση δεδομένων.

Η R είναι γλώσσα και περιβάλλον για την ανάλυση δεδομένων, την ανάπτυξη στατιστικών μοντέλων και την δημιουργία γραφικών παραστάσεων. Αναπτύχθηκε το 1995 από στατιστικολόγους ως γλώσσα ανοιχτού κώδικα, εναλλακτική των ακριβών

σουιτών στατιστικού λογισμικού όπως ήταν το SAS και το Matlab. Η R αρχικά χρησιμοποιήθηκε για ακαδημαϊκούς σκοπούς και για έρευνα, αλλά πλέον θεωρείται μια από τις πιο αναπτυσσόμενες γλώσσες στον χώρο των επιχειρήσεων. Καθώς η επιστήμη των δεδομένων είναι πλέον ζωτικής σημασίας για την πλειονότητα των επιχειρήσεων, η δημοτικότητα της R έχει εκτιναχτεί στα ύψη. Οργανισμοί και μεγάλες επιχειρήσεις όπως η Google, Facebook, και Microsoft έχουν στραφεί προς την R για την ανάλυση και οπτικοποίηση των δεδομένων τους, καθώς και για την δημιουργία σχετικών εκθέσεων. Σε αντίθεση με τις αντικειμενοστραφείς γλώσσες προγραμματισμού όπως είναι η Java και η Python, η R είναι μια διαδικαστική γλώσσα, που βασίζεται σε μια σειρά βημάτων βάση ρουτίνας για να εκτελεστεί μια εργασία προγραμματισμού, υπάρχει δηλαδή σταθερή σειρά ενεργειών που πρέπει να πραγματοποιηθούν. Η βασική διαφορά είναι ότι η R χρησιμοποιεί διαδικασίες για να επεξεργαστεί τα δεδομένα. Το πλεονέκτημα των διαδικασιών που εκτελεί η R είναι ότι προσφέρει ορατότητα σε πολύπλοκες εργασίες με πολλές εξαρτήσεις, η οποία μπορεί να είναι πολύ σημαντική σε πολλές εφαρμογές της ανάλυσης δεδομένων. Το αρνητικό είναι ότι αυτός ο τρόπος συνήθως απαιτεί περισσότερες γραμμές κώδικα να γραφούν και να εκτελεστούν από ότι θα χρειάζονταν οι αντικειμενοστρεφείς γλώσσες προγραμματισμού. Κάποια επιπλέον πλεονεκτήματα της R είναι η τεράστια κοινότητα που υπάρχει και παρέχει την υποστήριξη της μέσω λίστας ηλεκτρονικού ταχυδρομείου, σχετικών οδηγιών χρήσης και ομάδας στο stack overflow. Υπάρχει επίσης το CRAN, μια βιβλιοθήκη με πακέτα που συνεισφέρουν οι χρήστες της R. Τα πακέτα αυτά κάνουν πιο εύκολο το να έχει κανείς πρόσβαση στις τελευταίες τεχνικές που αναπτύσσονται στην R και για τις λειτουργίες της χωρίς να χρειάζεται να αναπτυχθούν όλα από το μηδέν.

Η επόμενη γλώσσα που αξίζει την παρατήρηση μας, η Python δημιουργήθηκε από τον Guido Van Rossem το 1991 δίνοντας έμφαση στην παραγωγικότητα και την αναγνωσιμότητα του κώδικα Σε αρχικό στάδιο, χρησιμοποιήθηκε ως εισαγωγή στους υπολογιστές και τον προγραμματισμό, στην πορεία όμως η Python χρησιμοποιήθηκε κυρίως από προγραμματιστές που ασχολούνται με ανάλυση δεδομένων ή θέλουν να εφαρμόσουν στατιστικές μεθόδους. Είναι μια πολύ ευέλικτη γλώσσα εστιάζοντας στην αναγνωσιμότητα και στην απλότητα της. Το θετικό χαρακτηριστικό της Python είναι ότι μπορούν να εκφραστούν έννοιες με πολύ λιγότερες γραμμές κώδικά από γίνεται στην Java ή στην C++. Όπως ισχύει στην R, έτσι και η Python έχει βιβλιοθήκες. Το

PyPi αποτελείται από βιβλιοθήκες της Python που της επιτρέπει να εκτελεί ένα ευρύ φάσμα εντολών. Πιο συγκεκριμένα, οι βιβλιοθήκες NumPy και matplotlib επιτρέπουν στην Python να εκτελέσει πολλές από τις λειτουργίες της Matlab όπως είναι η ανάλυση και η οπτικοποίηση δεδομένων. Στις βιβλιοθήκες NumPy και matplotlib είναι ενσωματωμένη και η βιβλιοθήκη scikit-learn. Η βιβλιοθήκη αυτή είναι ένα πολύ απλό και αποτελεσματικό εργαλείο για την ανάλυση και την εξόρυξη δεδομένων,δηλαδή στην ουσία είναι η βιβλιοθήκη μηχανικής μάθησης της γλώσσας Python. Με την χρήση της βιβλιοθήκης scikit-learn προσφέρεται η δυνατότητα να εφαρμοστούν τεχνικές μηχανικής μάθησης όπως η ταξινόμηση, η παλινδρόμηση, η ομαδοποίηση, η μείωση διάστασης, η επιλογή μοντέλου και η προ επεξεργασία. Όπως συμβαίνει με την R έτσι και η Python αντίστοιχα έχει μεγάλη κοινότητα, απλά οι πληροφορίες είναι λίγο ποιο διάσπαρτες διότι είναι και μια γλώσσα γενικού σκοπού. Παρ' όλα αυτά, η Python αναπτύσσεται ραγδαία στον τομέα της ανάλυσης δεδομένων και είναι η γλώσσα που χρησιμοποιήθηκε και κατά την πειραματική διαδικασία της παρούσας εργασίας στην συνέχεια.

Η Java είναι μια γλώσσα που είναι ισχυρή, φορητή, και επεκτάσιμη, χαρακτηριστικά που την καθιστούν ιδανική για τη δημιουργία εφαρμογών των επιχειρήσεων και για να έχει και την αντίστοιχη υποστήριξη ως προς την ανάπτυξη της. Η Java περιλαμβάνει πολλά εργαλεία όπως το Java Runtime Environment, Java plug-ins, και την εικονική μηχανή της Java (JVM). Αυτά τα εργαλεία κάνουν πολύ απλό τον προγραμματισμό και την χρήση εντολών με Java καθώς και την υποστήριξης σε κάθε επίπεδο, δίνοντας στους προγραμματιστές ότι χρειάζονται για την κατασκευή Web συστημάτων και εφαρμογών Java. Η ταχύτητα της είναι αυτή που βοήθησε την Java να ξεπεράσει άλλες γλώσσες και να γίνει πιο κατάλληλη για εφαρμογές μεγάλης κλίμακας. Αυτός ήταν και ένας λόγος που έκανε το Twitter να αλλάξει την μηχανή αναζήτησης του σε Java από Ruby on Rails που ήταν πριν. Ένα ακόμη βασικό στοιχείο της Java είναι ότι είναι πολύ κοντά στο να είναι 100% αντικειμενοστραφής. Έχοντας αυτό δεδομένο, έχει όλα τα οφέλη του αντικειμενοστραφούς προγραμματισμού και παράλληλα διατηρεί τα χαρακτηριστικά της ευελιξίας και επεκτασιμότητας. Ως μια από τις πιο ευρέως γνωστές γλώσσες προγραμματισμού, είναι εύκολο για τις επιχειρήσεις να βρουν τους αντίστοιχους προγραμματιστές ώστε να υποστηρίξουν την χρήση της στα έργα τους. Λόγω της τεράστιας κοινότητας της Java υπάρχει αρκετή και άριστη υποστήριξη.

Στο διαδίκτυο υπάρχουν πολλές συγκρίσεις σχετικά με την δημοτικότητα των δύο βασικών γλωσσών της R και της Python. Παρόλο που οι αριθμοί αυτοί παρουσιάζουν μια εικόνα για το πώς αυτές οι δύο γλώσσες εξελίσσονται στον τομέα της επιστήμης των υπολογιστών, είναι και πάλι αρκετά δύσκολο να συγκριθούν η μια με την άλλη. Ο βασικός λόγος της δυσκολίας αυτής είναι ότι την R την βρίσκουμε μόνο σε περιβάλλον επιστήμης των δεδομένων (data science), σε αντίθεση με την Python που είναι μια γλώσσα πιο ευρύ σκοπού και χρησιμοποιείται και σε άλλους τομείς όπως για παράδειγμα την ανάπτυξη ιστοσελίδων. Έτσι η κατάταξη αναγκαστικά ωθείται υπέρ της Python. Αν παρατηρηθούν οι έρευνες όμως που αναφέρονται σε γλώσσες προγραμματισμού που χρησιμοποιούνται μόνο για την ανάλυση δεδομένων τότε η R είναι αναμφισβήτητα ο νικητής. Παρατηρείται επίσης ότι υπάρχει και μια ομάδα χρηστών που χρησιμοποιεί και τις δύο γλώσσες όπου και όταν χρειάζεται ανάλογα με το έργο κάθε φορά.

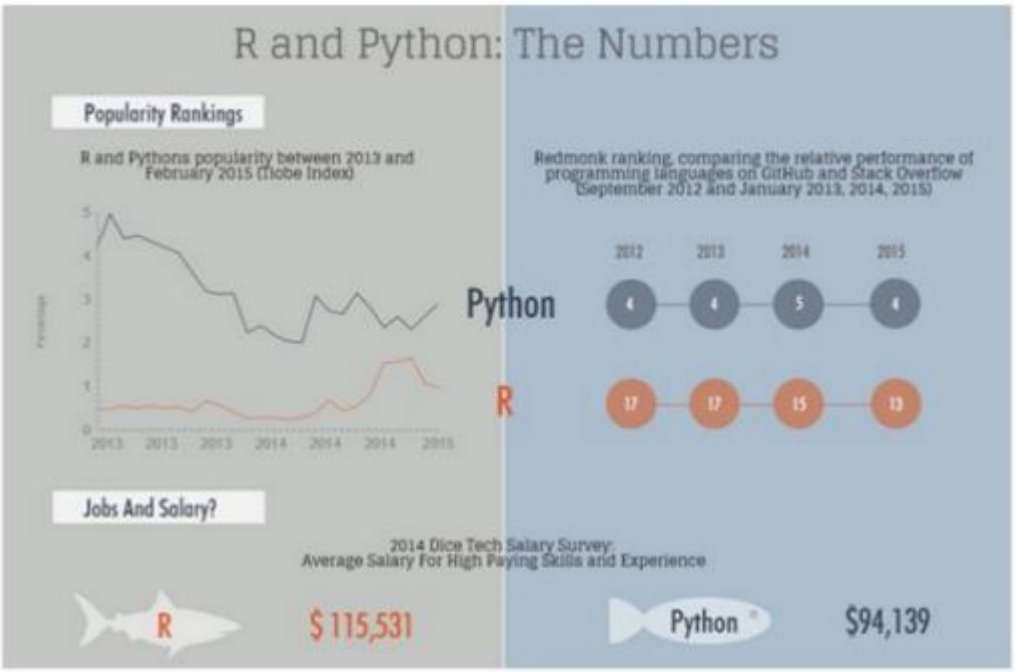
Αν υπάρχει ανάγκη στατιστικής ανάλυσης σε βάθος, τότε η R είναι η ιδανική γλώσσα αφού αναπτύχθηκε από στατιστικολόγους γι' αυτό το σκοπό και είναι ιδανική για δεδομένα που λαμβάνονται από αισθητήρες συστημάτων Internet of Things ώστε να επεξεργαστεί οικονομικά μοντέλα. Υποστηρίζεται πολύ καλά από το CRAN το οποίο περιέχει χιλιάδες πακέτα που επιτρέπουν να εκτελούνται οι πιο περίπλοκες εργασίες ανάλυσης και οπτικοποίησης δεδομένων. Η R παράγει υψηλής ποιότητας διαγράμματα και παραστάσεις, με έμφαση στην εύκολη παραγωγή. Επιτρέπεται επίσης η μετατροπή των εικόνων σε διαδραστικές Web εφαρμογές.

Παρόλα τα παραπάνω θετικά της, η R είναι σημαντικά πιο αργή από την Python ή την Java και επειδή η R είναι ιδανική μόνο για περιβάλλοντα της επιστήμης των δεδομένων, σε κάποιες εφαρμογές δεν είναι λειτουργική. Για τους χρήστες με υπόβαθρο στα μαθηματικά και τη στατιστική η σύνταξη της R είναι σχετικά απλή. Αντίθετα για αυτούς με εμπειρία στο προγραμματισμό η προσέγγιση της R θα φανεί παράλογη και δύσκολη.

Η ευελιξία της Python την κάνει μια πολύ δημοφιλή επιλογή για τους προγραμματιστές που εφαρμόζουν στατιστικές μεθόδους ή ασχολούνται με την ανάλυση δεδομένων, όπως επίσης για τους επιστήμονες των δεδομένων των οποίων τα καθήκοντα πρέπει να ενσωματωθούν σε εφαρμογές Web ή σε περιβάλλοντα παραγωγής. Έτσι η Python είναι ένα ενιαίο εργαλείο για τη διαχείριση ολόκληρης της ροής των δεδομένων. Ο

συνδυασμός των εξειδικευμένων βιβλιοθηκών μηχανικής μάθησης της Python (όπως είναι οι scikit-learn, PyBrain, και TensorFlow) καθώς και η ευελιξία γενικού σκοπού που έχει αναφερθεί παραπάνω, καθιστούν την Python κατάλληλη για την ανάπτυξη μοντέλων πρόβλεψης που συνδέονται απευθείας με συστήματα παραγωγής. Παρόλο που η κοινότητα της Python αναπτύσσεται συνεχώς, εξακολουθούν να υπάρχουν πακέτα της R που δεν μπορείς να τα βρεις στην Python. Αν ζητούνται ιδιαίτερες δυνατότητες με εξειδικευμένα πακέτα, τότε η R είναι καλύτερη επιλογή.

Η ταχύτητα της Java από την άλλη, την κάνει ιδανική για την δημιουργία συστημάτων μεγάλης κλίμακας. Ενώ η Python είναι πολύ πιο γρήγορη από ότι είναι η R, η Java παρέχει ακόμα μεγαλύτερη απόδοση και από τις δύο. Η ταχύτητα και η επεκτασιμότητα της είναι οι βασικές αιτίες που το Twitter, το Facebook και το LinkedIn βασίζονται σε αυτήν. Η εικονική μηχανή της Java (JVM) είναι ένα μεγάλο περιβάλλον για την ανάπτυξη προσαρμοσμένων εργαλείων σε πολύ σύντομο χρονικό διάστημα. Η γλώσσα προγραμματισμού Scala τρέχει σε JVM και είναι δημοφιλής στους επιστήμονες των δεδομένων για το συνδυασμό του αντικειμενοστραφούς και λειτουργικού προγραμματισμού. Παρόλα αυτά, ανάμεσα σε αυτές τις τρεις γλώσσες, η Java είναι σίγουρα η λιγότερο κατάλληλη για την ανάλυση των δεδομένων αφού έχει σημαντική έλλειψη πακέτων σε αυτόν τον τομέα και σίγουρα η R και η Python είναι οι καλύτερες επιλογές. [19]



5. *Σύγκριση R & Python σε αριθμούς(Πηγή: «https://www.datacamp.com/tutorial/r-or-python-for-data-analysis,» 1 2020 [20])*

3.3 Μηχανική μάθηση

3.3.1 Στάδια μηχανικής μάθησης

Η επιστήμη της μηχανικής μάθησης μπορεί να διακριθεί στα παρακάτω βασικά στάδια:

1. Συλλογή δεδομένων (data gathering): Απαιτείται να υπάρχει ένα σύνολο δεδομένων προς επεξεργασία και προς την εξαγωγή συμπερασμάτων.

2. Προ επεξεργασία ή προετοιμασία δεδομένων (data pre-processing): Τα δεδομένα στις περισσότερες περιπτώσεις βρίσκονται σε μια μορφή που δεν εξυπηρετεί την τροφοδότηση τους σε κάποιο μοντέλο μηχανικής μάθησης, συνεπώς πρέπει να ακολουθηθεί η κατάλληλη επεξεργασία με συγκεκριμένα εργαλεία ώστε να έρθουν στην κατάλληλη μορφή η οποία να μπορεί να χρησιμοποιηθεί από το μοντέλο μηχανικής μάθησης. Ένα χαρακτηριστικό εργαλείο προ επεξεργασίας δεδομένων της γλώσσας Python είναι η βιβλιοθήκη preprocessing για να εξυπηρετήσει αυτό το σκοπό. Παραδείγματα προ επεξεργασίας είναι η τακτοποίηση των δεδομένων σε μορφή διακριτών χαρακτηριστικών με αποδεκτή μορφή (π.χ. με την χρήση βιβλιοθηκών Pandas και Numpy της Python), η μετατροπή του χρόνου σε ένα συνεπές σύστημα μονάδων (π.χ. κλάση time της Python) , η κανονικοποίηση των δεδομένων σε αποδεκτή

από το μοντέλο κλίμακα για παράδειγμα στο διάστημα [0,1], η μετατροπή κατηγορικών μεταβλητών σε ακέραιους αριθμούς ή ακολουθίες από 0,1 (π.χ. κλάσεις LabelEncoder, LabelBinarizer της preprocessing) και η διαχείριση missing data (π.χ. κλάση Imputer της preprocessing), βήματα που ακολουθήθηκαν και κατά την πειραματική διαδικασία της παρούσας εργασίας ώστε τα δεδομένα να μπορούν να εφαρμοστούν στο μοντέλο πρόβλεψης με την χρήση της Python.

3. Επιλογή και εξαγωγή χρήσιμων χαρακτηριστικών: Κατά την επόμενη διαδικασία εντοπίζονται ακόμα και δημιουργούνται ή ανακαλύπτονται χαρακτηριστικά των δεδομένων τα οποία είναι κρυμμένα μέσα στην πολύπλοκη δομή των datasets αυτών. Στη συνέχεια, εξετάζονται ποια από αυτά τα χαρακτηριστικά θεωρούνται σημαντικά και ποια όχι. Διαπιστώνεται ποια συνεισφέρουν πραγματικά στο μοντέλο ώστε να εκπαιδευτεί και ποια είναι εντελώς ασυσχέτιστα με την έξοδο ή είναι αλληλεξαρτώμενα και χρήζουν απόρριψης. Πίσω από αυτή την διαδικασία υπάρχει ένας ολόκληρος υποκλάδος της μηχανικής μάθησης που ονομάζεται Feature engineering και αποκτά όλο και μεγαλύτερη αξία και σημασία στην ανάλυση δεδομένων.

4. Επιλογή μοντέλου (model selection): Στο στάδιο αυτό γίνεται επιλογή του μοντέλου, ανάλογα με τη φύση του προβλήματος, την εμπειρία του αναλυτή και τα αποτελέσματα της αξιολόγησης. Δεν είναι απαραίτητα το μοναδικό που μπορεί ν επιλεχθεί και είναι σύνηθες να γίνεται σύγκριση μοντέλων πριν την τελική επιλογή του πλέον κατάλληλου.

5. Εκπαίδευση μοντέλου (training): Επιλέγεται ένα υποσύνολο του dataset το οποίο εφαρμόζεται (fit) πάνω στο μοντέλο ώστε να εκπαιδευτεί και να γίνει η κατάλληλη ρύθμιση των εσωτερικών του παραμέτρων. Με τον όρο εσωτερικές παράμετροι εννοούνται οι παράμετροι του μοντέλου οι οποίες μαθαίνονται κατά αυτό το στάδιο, όπως είναι για παράδειγμα η κλίση ενός μοντέλου απλής γραμμικής παλινδρόμησης. Η εκπαίδευση του μοντέλου γίνεται μέσω ελαχιστοποίησης μιας συνάρτησης σφάλματος(π.χ. mse, mae, cost functions, loss functions).

6. Αξιολόγηση μοντέλου (model evaluation): Στο βήμα αυτό γίνεται χρήση μετρικών αξιολογήσεων ώστε να διαπιστωθεί πόσο καλά λειτουργεί το μοντέλο που έγινε η εκπαίδευση με τον παρακάτω διαχωρισμό:

● Όσον αφορά στη μη επιτηρούμενη μάθηση, η αξιολόγηση μοντέλου είναι μια σχετικά διαισθητική διαδικασία εφόσον δεν έχουμε κάποιο δομημένο μέτρο εκτίμησης της απόδοσης και των σφαλμάτων του. Παρόλα αυτά, για παράδειγμα σε μεθόδους συσταδοποίησης υφίστανται μέτρα ομοιότητας, ενδοσυσταδικής και διασυσταδικής απόστασης που μπορούν να δώσουν στον μελετητή μια εικόνα των επιδόσεων του μοντέλου και των τιμών των παραμέτρων που πρέπει να τροποποιηθούν.

● Όσον αφορά στην επιτηρούμενη μάθηση, η αξιολόγηση γίνεται με μετρικές όπως για παράδειγμα το μέσο τετραγωνικό σφάλμα και το μέσο απόλυτο σφάλμα για τεχνικές παλινδρόμησης, ο πίνακας σύγχυσης (confusion matrix), τα precision - recall και η καμπύλη ROC για τα μοντέλα ταξινόμησης. Χρησιμοποιούνται ακόμη τεχνικές αξιολόγησης μοντέλων όπως αυτή της διασταυρωμένης επικύρωσης (cross validation).

7. Τροποποίηση παραμέτρων του μοντέλου (parameter tuning): Στο επόμενο αυτό βήμα γίνεται η τροποποίηση των εξωτερικών παραμέτρων του μοντέλου (hyperparameters). Με τον όρο hyperparameters γίνεται αναφορά στις παραμέτρους αυτές που παίρνουν τιμή από τον αναλυτή πριν την εκπαίδευση του μοντέλου και δεν μαθαίνονται κατά τη διάρκεια της μάθησης, ωστόσο την επηρεάζουν. Σε ένα παράδειγμα νευρωνικού δικτύου, τέτοιες παράμετροι είναι ο αριθμός νευρώνων και ο ρυθμός μάθησης, τα οποία είναι θέματα του κλάδου της βαθιάς μηχανικής μάθησης. Χαρακτηριστική είναι η υποκλάση model_selection.GridSearchCV της βιβλιοθήκης scikit-learn της Python που προσφέρει στον αναλυτή τη δυνατότητα να δοκιμάσει διαφορετικές τιμές εξωτερικών παραμέτρων επαναληπτικά πάνω σε ένα μοντέλο και επιλέξει το συνδυασμό αυτό με τα καλύτερα αποτελέσματα με μετρικές αξιολόγησης και διαδικασία επικύρωσης της επιλογής του (για παράδειγμα cross validation).

8. Πρόβλεψη (prediction): Στην περίπτωση της επιτηρούμενης μηχανικής μάθησης, σκοπός είναι να κάνουμε προβλέψεις σε νέα μη σεσημασμένα δεδομένα. Πρόκειται για το βήμα της πλέον πρακτικής εφαρμογής του μοντέλου επιλογής του μελετητή και τελικώς το λόγο της δημιουργίας του.

3.3.2 Κατηγορίες μηχανικής μάθησης

Οι κύριες κατηγορίες της μηχανικής μάθησης αναφέρονται και αναλύονται παρακάτω:

Επιτηρούμενη μηχανική μάθηση (Supervised Learning (SL)) ονομάζεται η υποκατηγορία της μηχανικής μάθησης όπου η διαδικασία μάθησης βασίζεται σε ζεύγη εισόδου και εξόδου. Το επιλεγμένο σύνολο δεδομένων (dataset) αποτελείται από μία σειρά από χαρακτηριστικά εισόδου (features) τα οποία αποτελούν τις ανεξάρτητες μεταβλητές ή αλλιώς ένα διάνυσμα ανεξάρτητων μεταβλητών X = (x1, ...xk) και μία ετικέτα (label) - έξοδος (output) που αποτελεί την εξαρτημένη μεταβλητή y. Κάθε γραμμή του dataset αποτελεί και ένα instance, πρότυπο ή αντικείμενο. Η εξαρτημένη μεταβλητή παίρνει τιμές στο συνεχή χώρο (πρόβλημα παλινδρόμησης) ή στο διακριτό χώρο (πρόβλημα ταξινόμησης). Κατά την επιτηρούμενη μάθηση, αφού έχουν ολοκληρωθεί τα αρχικά στάδια της συλλογής και προ επεξεργασίας των δεδομένων, συνήθως γίνεται ο διαχωρισμός τους σε ένα σύνολο δεδομένων εκπαίδευσης ή αλλιώς training set, σε ένα σύνολο επικύρωσης (validation set) και τέλος σε ένα test set με τα παρακάτω χαρακτηριστικά:

● Το training set, αποτελεί ένα υποσύνολο του dataset με το οποίο τροφοδοτείται το μοντέλο της μηχανικής μάθησης ή το νευρωνικό δίκτυο ώστε να εκπαιδευτούν οι εσωτερικές παράμετροι του.

● Το validation set είναι το υποσύνολο αυτό του dataset το οποίο χρησιμοποιείται προς την αξιολόγηση του μοντέλου και τον έλεγχο της επίδοσης του σε δεδομένα τα οποία δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση του μοντέλου ώστε γίνει και η ρύθμιση των εξωτερικών παραμέτρων του μοντέλου αυτού (hyperparameters).

● Το test set αποτελεί και αυτό υποσύνολο του dataset, το οποίο δε χρησιμοποιείται επίσης για την εκπαίδευση του αλγόριθμου, και θεωρητικά ούτε για τη ρύθμιση των εξωτερικών παραμέτρων του αλλά χρησιμεύει στην εφαρμογή και αξιολόγηση του μοντέλου σε νέα δεδομένα. Το test set στην πράξη προσομοιώνει νέες άγνωστες εισόδους για το μοντέλο όσο παράλληλα η επιθυμητή έξοδος είναι γνωστή ώστε να ελέγχεται η επιτυχία των προβλέψεων του μοντέλου. Θεωρητικά το test set θα έπρεπε να είναι "κρυφό" μέχρι τη διαμόρφωση του τελικού μοντέλου και να χρησιμοποιηθεί για τον έλεγχο των επιδόσεων του χωρίς επιπλέον ρύθμιση των εξωτερικών του παραμέτρων. Παρόλα αυτά, το test set και το validation set κάποιες φορές συγχέονται, δηλαδή γίνεται προσαρμογή των εξωτερικών παραμέτρων του μοντέλου έτσι ώστε να προκύψουν καλά αποτελέσματα στο test set και στην πορεία ακολουθούνται οι γνωστές διαδικασίες επικύρωσης όπως το cross-validation, για να ελεγχθεί κατά πόσο

γενικεύονται τα αποτελέσματα. Αυτό είναι λογικό και συχνό φαινόμενο να συμβαίνει γιατί δεν υπάρχει τις πιο πολλές φορές η πολυτέλεια να υπάρχουν και να ελεγχθούν μεγάλου όγκου δεδομένων. Το test set είναι κατά κανόνα μικρότερο ως προς το πλήθος των δεδομένων του από το training set. Οι αναλογίες τους κατά κανόνα κυμαίνονται από 50-50% έως 90-10% ανάλογα με το εκάστοτε πρόβλημα και τα δεδομένα που αναλύονται. Στη συνέχεια, γίνεται επιλογή ένας υποψήφιος αλγόριθμος επιτηρούμενης μάθησης ανάλογα με τη φύση του προβλήματος και η διαδικασία επιτηρούμενης μάθησης διαρθρώνεται ως εξής: έστω ένα dataset A το οποίο χωρίζουμε σε training και test sets A1 , A2. Το A1 αποτελείται από i γραμμές με συγκεκριμένες τιμές των ανεξάρτητων μεταβλητών X = (x1, ...xk) και τις αντίστοιχες τιμές yi της εξαρτημένης μεταβλητής y. Υποθέτουμε πως οι X,y συνδέονται μέσω μιας άγνωστης συνάρτησης στόχου *f* για την οποία ξέρουμε μόνο τις ακριβείς τιμές τις πάνω στα δοθέντα yi. Στόχος είναι να γίνει μια γενικευμένη εκτίμηση ώστε να ελαχιστοποιείται μια συνάρτηση σφάλματος. Η συνάρτηση σφάλματος (error function) που χρησιμοποιείται κατά την εκπαίδευση, ποικίλλει ως προς τη μορφή της ανάλογα με το είδος και τη φύση του προβλήματος που αντιμετωπίζεται. Το μοντέλο που δημιουργείται αξιολογείται στην συνέχεια ως προς τις επιδόσεις του πάνω στο validation set.

Η επιτηρούμενη μηχανική μάθηση παρουσιάζει την παρακάτω ιδιαιτερότητα: Πάντα γίνεται αναζήτηση της χρυσή τομή στην πολυπλοκότητα του μοντέλου που επιλέγουμε ανάλογα με τα δεδομένα του εκάστοτε προβλήματος. Καλές επιδόσεις ενός μοντέλου στο training set σε συνδυασμό με κακές επιδόσεις στο test set υποδηλώνουν overfitting. Αυτό σημαίνει ότι το μοντέλο παλινδρόμησης ή ταξινόμησης παρουσιάζει υψηλή μεταβλητότητα (variance) και καταλήγει να είναι υπερβολικά προσαρμοσμένο πάνω στο training set, ενσωματώνοντας μέχρι και το θόρυβο στα δεδομένα και το καθιστά ανίκανο να κάνει προβλέψεις σε νέα, πραγματικά δεδομένα. Σε τέτοια περίπτωση τόσο οι παράμετροι του μοντέλου όσο και το μοντέλο τίθενται υπό αμφισβήτηση. Ένα τέτοιο ζήτημα μπορεί να εμφανιστεί συχνά σε πολυπαραμετρικά και ευέλικτα μοντέλα τα οποία εκπαιδεύονται σε πολλή λεπτομέρεια από τα δεδομένα (π.χ. δέντρα αποφάσεων, πολυωνυμική παλινδρόμηση υψηλής τάξης). Από την άλλη, οι κακές επιδόσεις στο training set ή μεγάλο bias (οι μεγάλες και συστηματικές αποκλίσεις πρόβλεψης από την αναμενόμενη τιμή), υποδηλώνουν underfitting. Αυτό σημαίνει ότι το μοντέλο είναι υπεραπλουστευμένο και κατά κάποιο τρόπο γίνεται δύσκαμπτο. Κάτι τέτοιο μπορεί να συμβαίνει για παράδειγμα όταν γίνεται προσπάθεια πρόβλεψης σε μη γραμμικά

δεδομένα με απλό μοντέλο γραμμικής παλινδρόμησης. Σε τέτοια περίπτωση, που υπάρχει ευκολία εντοπισμού, απαιτείται αντικατάσταση του μοντέλο με ένα πιο σύνθετο και παραμετροποιήσιμο μοντέλο ως καταλληλότερο.

Κατά την επιλογή ενός μοντέλου δεν είναι απαραίτητο πως η μέθοδος που θα χρησιμοποιηθεί είναι και η μοναδική που μπορεί. Συχνά ο συνδυασμός μεθόδων βελτιώνει και τα τελικά απαιτούμενα αποτελέσματα. Οι δύο βασικές τεχνικές συνδυασμού μεθόδων είναι αυτές του bagging και boosting. Με αυτές τις τεχνικές, το κάθε μοντέλο του συνδυασμού εκπαιδεύεται ατομικά και το αποτέλεσμα της πρόβλεψης αποτελεί μια συνάρτηση των προβλέψεων των επιμέρους μοντέλων.

Το Bagging προέρχεται από τον όρο bootstrap aggregating και αποτελεί μια συνδυαστική μέθοδο της κατηγορίας του bootstrapping. Στην τεχνική αυτή ισχύουν τα παρακάτω:

- Το κάθε επιμέρους μοντέλο εκπαιδεύεται με training set ένα τυχαίο υποσύνολο του αρχικού training set και ισχύει ότι η επιλογή στοιχείων για τη δόμηση του να γίνεται ομοιόμορφα και με αντικατάσταση .
- Το τελικό αποτέλεσμα στην παλινδρόμηση είναι ο μέσος όρος των εκτιμήσεων των επιμέρους μοντέλων ενώ στην ταξινόμηση επιλέγεται η κλάση με τις περισσότερες ψήφους.

Το bagging προσφέρει τα εξής πλεονεκτήματα:

1. Περιορίζει το overfitting σε περίπτωση μοντέλων με υψηλή μεταβλητότητα

2. Περιορίζει το underfitting σε περίπτωση μοντέλων με υψηλό bias

3. Μειώνει το θόρυβο, χρησιμοποιώντας πολλές τυχαίες δειγματοληψίες

4. Βοηθάει να χτιστούν πιο ισχυρά μοντέλα ακόμη και με μικρά dataset δεδομένων

Το boosting περιορίζει το underfitting σε περίπτωση μοντέλων υψηλού bias (π.χ. σε περίπτωση ενός ρηχού δέντρου αποφάσεων) πάλι με συνδυασμό μοντέλων και με την κατασκευή ενός πιο προσαρμοστικού μοντέλου. Ως αδύναμος αλγόριθμος μάθησης θεωρείται ένας αλγόριθμος ο οποίος πετυχαίνει αποτελέσματα οριακά καλύτερα από τα τυχαία, όπως ένα δέντρο αποφάσεων ενός επιπέδου (decision stump). Το boosting

βασίζεται και αυτό στη μέθοδο του bootstrapping, όπως και το bagging, ωστόσο με τις παρακάτω διαφοροποιήσεις:

- Η εκπαίδευση στο boosting ολοκληρώνεται έπειτα από κάποιο αριθμό επαναλήψεων.
- Ο αλγόριθμος συγκρατεί ποια από τα επιμέρους datasets είχαν τα χειρότερα αποτελέσματα και προχωρά σε αντιστοίχιση τους σε μεγαλύτερα βάρη υπολογισμού για την επόμενη επανάληψη.
- Κατά την πρόβλεψη, ο αλγόριθμος, έχοντας κρατήσει αρχείο των επιδόσεων κάθε μοντέλου κατά την εκπαίδευση, δίνονται μεγαλύτερα βάρη στα μοντέλα με τα μικρότερα καταγεγραμμένα σφάλματα.

Στη βιβλιογραφία συναντώνται μεταξύ πολλών, οι παρακάτω εκδοχές του boosting:

- AdaBoost (Adaptive Boosting): Σκοπός του είναι να τροποποιεί κάθε φορά τα βάρη των δειγμάτων έτσι ώστε κάθε νέο αδύναμο μοντέλο που εκπαιδεύεται να λαμβάνει υπόψη τα λάθη των προηγούμενων. Έτσι συνδυάζονται οι αποφάσεις των επιμέρους μοντέλων ανάλογα με τις εκάστοτε επιδόσεις τους. Χρησιμοποιεί κατά κανόνα decision stumps. Είναι όμως αρκετά ευαίσθητος σε θόρυβο και outliers.
- Gradient Tree Boosting: Παρουσιάστηκε από τον Friedman μαζί με την εξέλιξη του Stochastic Gradient Boosting. Μοιάζει με την προηγούμενη τεχνική Adaboost, εφαρμόζει όμως λογική ελαχιστοποίησης τόσο σε συνάρτηση σφάλματος όσο και σε loss function, δηλαδή κάθε νέο μοντέλο εκπαιδεύεται πάνω στα σφάλματα πρόβλεψης των προηγούμενων με σκοπό την ελαχιστοποίηση τους. Το gradient boosting χρησιμοποιείται κυρίως σε προβλήματα anomaly detection.
- XGBoost: Αποτελεί την τελευταία εξέλιξη του gradient boosting. Είναι κατάλληλη για μεγάλα datasets σε καλές ταχύτητες. Επιπλέον, χρησιμοποιώντας την Python δε χρειάζεται η κανονικοποίηση των δεδομένων (feature scaling), γεγονός που εξυπηρετεί πολύ στη διαρκή διαισθητική επαφή με τα δεδομένα του προβλήματος που αναλύεται.

Είναι επίσης πολύ σημαντικό να γίνει αναφορά στα παρακάτω θέματα σχετικά με τις τεχνικές boosting γενικώς:

1. Είναι αρκετά επιρρεπείς σε overfitting, όταν υπάρχει πολύς θόρυβος ειδικά για αλγόριθμους όπως ο Adaboost,

2. Η εκπαίδευση του μοντέλου είναι ιδιαίτερα χρονοβόρα αφού εκτελείται σειριακά, πόσο μάλλον σε real-time πλατφόρμες όπου επιβάλλεται η παραλληλοποίηση,
3. Στο gradient boosting, σε σχέση με τα random forests, είναι πιο δύσκολη η ρύθμιση εξωτερικών παραμέτρων γιατί συνήθως έχουν τρεις συγκεκριμένες : τον αριθμό δέντρων, το βάθος και το ρυθμό μάθησης.

Ένα στάδιο της μηχανικής μάθησης όπως αναφέρθηκε παραπάνω, αποτελεί και η αξιολόγηση του μοντέλου. Η πιο απλή μορφή αξιολόγησης του μοντέλου είναι, όπως έχει ήδη αναφερθεί, ο διαχωρισμός των δεδομένων σε training και test set με μία αναλογία της τάξης των δύο τρίτων. Το μοντέλο εκπαιδεύεται στο training set, αξιολογείται στο test set και έπειτα γίνεται η ρύθμιση των εξωτερικών παραμέτρων του σταδιακά ώστε το μοντέλο να λειτουργεί όσο καλύτερα γίνεται και στα δύο σύνολα. Ωστόσο η μέθοδος αυτή είναι αρκετά απλοϊκή για να εξασφαλίσει ότι το μοντέλο δεν είναι overfitted πάνω στο training set ή ότι το test set που έχει επιλεχθεί δεν είναι αρκετά εύκολο έστω και τυχαία για τις προβλέψεις του μοντέλου. Αυτά τα θέματα έρχεται να λύσει η τεχνική της διασταυρωμένης επικύρωσης (cross validation). Η πιο χαρακτηριστική μορφή της είναι το k-folds cross validation που διαρθρώνεται ως εξής:

1. Ανακατεύονται τα δεδομένα, επιλέγεται μια τιμή για το k και χωρίζεται το dataset σε k τμήματα ίδιου μεγέθους.
2. Για κάθε τμήμα τα βήματα τα παρακάτω:
    - Θεωρείται ως validation set και τα υπόλοιπα k-1 ως training set
    - Προσαρμόζονται τα δεδομένα του training set στο μοντέλο και αξιολογείται το μοντέλο στο validation set.
    - Δεσμεύεται μόνο το επιθυμητό αποτέλεσμα (score) της αξιολόγησης, αγνοείται το μοντέλο και προχωράει η διαδικασία στο επόμενο τμήμα.
3. Αξιολογείται το μοντέλο βάση των επιμέρους αποτελεσμάτων αξιολόγησης που έχουν προκύψει.

Με τον παραπάνω τρόπο και βήματα μπορούν να βγουν ασφαλέστερα, πιο συνεπή και γενικευμένα συμπεράσματα για την απόδοση ενός μοντέλου, καθώς έχει δοκιμαστεί πλέον σε διαφορετικές συνθήκες μάθησης και έχει εξασφαλιστεί πως η απόδοση του δεν οφείλεται στο ότι γίνεται τροφοδότηση του από ένα συγκεκριμένο συνδυασμό training και test sets. Το cross-validation μπορεί να χρησιμοποιηθεί ακόμη και για την

επιλογή μοντέλου συγκρίνοντας τα αποτελέσματα αξιολογήσεων για διαφορετικούς αλγόριθμους μηχανικής μάθησης.

Γενικότερα η επιτηρούμενη μάθηση έχει σκοπό την πρόβλεψη των ετικετών νέων δεδομένων μέσω της μελέτης των ήδη υπαρχόντων δεδομένων με ετικέτες. Χαρακτηριστικές εργασίες της επιτηρούμενης μάθησης αποτελούν η παλινδρόμηση, η ταξινόμηση και η υποκατηγορία LDA (linear discriminant analysis) της κατηγορίας του dimensionality reduction. Κάποια παραδείγματα εφαρμογής τέτοιων εργασιών είναι:

- Προβλήματα παλινδρόμησης όπου γίνεται προσπάθεια πρόβλεψης των αναγκών προσλήψεων υπαλλήλων με κάποια συγκεκριμένα χαρακτηριστικά, μέσω της μελέτης μιας συλλογής δεδομένων με αντίστοιχες τιμές στα χαρακτηριστικά αυτά και ετικέτες που αντιστοιχούν στα δεδομένα αυτά.
- Προβλήματα ταξινόμησης για την προσπάθεια πρόβλεψης της διακριτής τιμής Ναι ή Όχι (εναλλακτικά 0 ή 1) για το αν κάποιος πελάτης ασφαλιστικής με συγκεκριμένα χαρακτηριστικά θα κάνει εισαγωγή σε νοσοκομείο, μέσω μελέτης μια συλλογής δεδομένων πελατών και μη της ασφαλιστικής με αντίστοιχες τιμές στα χαρακτηριστικά αυτά και διακριτές ετικέτες αγοράς ή όχι που τους αντιστοιχούν.

Στην κατηγορία της μη επιτηρούμενης μηχανικής μάθησης εμπίπτουν τα προβλήματα των οποίων τα δεδομένα δεν αποτελούν ζεύγη εισόδου - εξόδου. Δεν υπάρχουν δηλαδή ετικέτες οι οποίες να υποδηλώνουν κάποια έξοδο σε κάθε instance. Αντιθέτως, γίνεται προσπάθεια να αντληθούν πληροφορίες που αφορούν ομοιότητα και ανομοιότητα, κρυμμένες δομές και μοτίβα στα πρότυπα - εισόδους. Η μη επιτηρούμενη μάθηση εξυπηρετεί κυρίως στα στάδια της προ επεξεργασίας και κατά τον προσδιορισμό της δομής των δεδομένων και όχι στην πρόβλεψη όπως συμβαίνει στην κατηγορία της επιτηρούμενης μάθησης. Χαρακτηριστικές εργασίες μη επιτηρούμενης μάθησης αποτελούν η συσταδοποίηση (clustering) και η πλειονότητα των τεχνικών dimensionality reduction όπως η PCA (Principal Component Analysis).

Στην κατηγορία της ημιεπιτηρούμενης μάθησης (SSL) εμπίπτουν προβλήματα που τα δεδομένα τους είναι μερικώς σεσημασμένα με ετικέτες εξόδου. Οι διάφορες κατηγορίες σχετικών αλγόριθμων και μοντέλων αναλύονται ως εξής:

- Generative μοντέλα: αφορά μοντέλα που βασίζονται στην από κοινού συνάρτηση πιθανότητας της εξαρτημένης και ανεξάρτητης μεταβλητής. Μπορεί να αντιμετωπιστεί ως μια μορφή συσταδοποίησης με παραπάνω πληροφορίες ή μορφή ταξινόμησης με πληροφορίες οριακής πυκνότητας πιθανότητας.
- Μέθοδοι διαχωρισμού χαμηλής πυκνότητας που συμπεριλαμβάνουν μοντέλα όπως για παράδειγμα το TSVM (transductive support vector machine), η ταξινόμηση με δυαδική γκαουσσιανή διαδικασία και προσεγγίσεις της μεγιστοποίησης της εντροπίας.
- Μέθοδοι γράφων που τα δεδομένα αναπαρίστανται από κόμβους ενός γράφου και οι ακμές του γράφου είναι σεσημασμένες με πιθανοτικά βάρη. Με αυτές τις μεθόδους επιτυγχάνεται η διάδοση ετικετών από τα σεσημασμένα στα μη σεσημασμένα δεδομένα με χρήση διακριτών μαρκοβιανών πεδίων, τυχαίων γκαουσιανών πεδίων, αλλά και βαθιών συνελικτικών δικτύων.
- Μέθοδοι δύο βημάτων που πραγματοποιείται μια συσταδοποίηση στην αρχή στο σύνολο των δεδομένων και στην συνέχεια ακολουθεί ταξινόμηση στα σεσημασμένα δεδομένα.

Οι μέθοδοι αυτές έχουν στενή σχέση με εκείνες των γράφων. Η ανάγκη για SSL προκύπτει καθώς η σήμανση δεδομένων είναι συνήθως μία δύσκολη εργασία η οποία απαιτεί την έντονη συμβολή του ανθρώπινου παράγοντα πράγμα που είναι τόσο χρονοβόρο όπως και κοστοβόρο. Χαρακτηριστικές εφαρμογές του SSL είναι η ταξινόμηση ακολουθιών πρωτεϊνών στην επιστήμη της Βιολογίας και η αναγνώριση ομιλίας σε πολλές εφαρμογές ευρέως χρησιμοποιούμενες.

Η ενισχυτική μάθηση - Reinforcement Learning (RL) αποτελεί ένα είδος μάθησης (μια απεικόνιση καταστάσεων σε δράσεις) που σκοπός της είναι η μεγιστοποίηση ενός σήματος επιβράβευσης. Απαιτείται η ύπαρξη ενός πράκτορα (agent) ο οποίος διέπετε από τα παρακάτω βασικά χαρακτηριστικά:

- Έχει στόχο.
- Έχει αίσθηση του περιβάλλοντος του (όπως μέσω αισθητήρων) ώστε να μπορεί να αντιληφθεί τη συνέπεια των πράξεων του στο περιβάλλον του και εν συνεχεία την εξυπηρέτηση του ίδιου του στόχου του.
- Μπορεί να λάβει αποφάσεις και να δράσει αναλόγως με βάση τα παραπάνω χαρακτηριστικά. Ο πράκτορας μεταβαίνει από κατάσταση σε κατάσταση μέσω

της λήψης αποφάσεων. Οι αποφάσεις του αυτές επιβραβεύονται ή τιμωρούνται ανάλογα με την επίδραση που έχουν στην επίτευξη του στόχου του μέσω ενός αντίστοιχου σήματος. Η διαδικασία κινείται διαρκώς ανάμεσα στις έννοιες τις εκμετάλλευσης και της εξερεύνησης (exploitation και exploration dilemma.). Κατά συνέπεια, ο πράκτορας, στην προσπάθεια μεγιστοποίησης της ανταμοιβής έχει αφενός συμφέρον να ακολουθεί μονοπάτια αποφάσεων τα οποία ακολούθησε και στο παρελθόν και τα οποία αποδείχθηκαν αποτελεσματικά σε όρους επιβράβευσης και αφετέρου προκειμένου να ανακαλύψει τέτοια μονοπάτια οφείλει να επιλέγει δράσεις τις οποίες δεν έχει ξανά επιλέξει στο παρελθόν.

Καταλήγοντας, η ενισχυτική μάθηση δεν ορίζεται από μεθόδους επίλυσης προβλημάτων όπως ισχύει για την επιτηρούμενη μάθηση, αλλά από την παροχή μιας πλήρους περιγραφής ενός προβλήματος σε ένα πράκτορα με τα παραπάνω χαρακτηριστικά που αναλύθηκαν προηγουμένως. Το πρόβλημα συνήθως ορίζεται από μία Μαρκοβιανή στοχαστική διαδικασία αποφάσεων, η οποία ανάγεται σε πρόβλημα βελτιστοποίησης γραμμικού ή και δυναμικού προγραμματισμού. Ο πράκτορας κινείται ανάμεσα σε ένα σύνολο καταστάσεων S με ένα σύνολο δράσεων A για κάθε κατάσταση. Κάθε του επιλογή επιβραβεύεται ή τιμωρείται. Ο τελικός σκοπός είναι η εξαγωγή μιας σειράς βέλτιστων μεταβάσεων (δράσεων) από κατάσταση σε κατάσταση για τον πράκτορα.

Ένα πολύ χαρακτηριστικό παράδειγμα μπορεί να αποτελέσει η λειτουργία ενός ρομπότ-βοηθού το οποίο καθαρίζει και πρέπει να αποφασίσει αν μπορεί να μεταβεί στον επόμενο χώρο προς καθαρισμό ή πρέπει προηγουμένως να επισκεφτεί το σταθμό φόρτισης της μπαταρίας του. Η απόφαση αυτή λαμβάνεται στην αρχή βάσει της στάθμης της μπαταρίας του και βάσει του πόσο γρήγορα έχει βρεθεί σταθμός φόρτισης στο παρελθόν με βάση της τωρινής του τοποθεσίας (αίσθηση του περιβάλλοντος), στα πλαίσια του στόχου του να ολοκληρώσει τον καθαρισμό, ο οποίος απαιτεί και να μην τελειώσει η μπαταρία σε καμία περίπτωση. Γίνεται λοιπόν με αυτό τον τρόπο, εύκολα αντιληπτό πως η περίπτωση απώλειας μπαταρίας συνεπάγεται μεγάλη τιμωρία στο σύστημα επιβράβευσης του πράκτορα. Ένα ακόμη αντίστοιχο παράδειγμα στο οποίο μπορεί να εντοπιστεί το δίλημμα εκμετάλλευσης όπως έχει αναφερθεί και παραπάνω είναι το multi armed bandit problem το οποίο έχει και πολλές εφαρμογές σε προβλήματα του πραγματικού κόσμου: Σε μια απλή μορφή του μπορεί να ληφθεί ως

παράδειγμα κάποιες φαινομενικά όμοιες μηχανές τυχερών παιχνιδιών, τύπου φρουτάκια. Ο παίκτης προσπαθεί να ανακαλύψει ποια μηχανή είναι η πιο κερδοφόρα, όμως ταυτόχρονα ποντάρει χρήματα οπότε επιβάλλεται όλη αυτή η διαδικασία να γίνει με τα ελάχιστα δυνατά χαμένα χρήματα του παίκτη. Αυτό το πρόβλημα μπορεί να αντιμετωπιστεί με στοιχειώδεις αλγόριθμους ενισχυτικής μάθησης όπως ο upper confidence bound, ο thompson sampling, ο softmax, ε-greedy, των οποίων οι αποδόσεις ποικίλλουν ανάλογα με το είδος και τις παραμέτρους του εκάστοτε προβλήματος . Σε κάθε περίπτωση γίνεται προσπάθεια σταδιακής εύρεσης της μηχανής με το βέλτιστο αναμενόμενο αποτέλεσμα, χρησιμοποιώντας ταυτόχρονα όσο πιο πολύ γίνεται τις πιο κερδοφόρες μηχανές κατά την πορεία αυτή. Το multi-armed bandit problem βρίσκει εφαρμογή σε ζητήματα όπως είναι αυτό της επιλογής της καλύτερης από έναν αριθμό διαφημιστικών καμπανιών εταιρείας με την δοκιμή στην απήχηση τους στον κόσμο με βάση τα κλικ στην ιστοσελίδα και εν τέλει καταλήγοντας σε κατάργηση όλων εκτός της πιο αποτελεσματικής μεταξύ τους. Αντίστοιχη εφαρμογή υπάρχει και σε κλινικές μελέτες για την επιλογή των κατάλληλων θεραπειών σε ασθενείς. Μία ακόμη πολύ δημοφιλής περίπτωση εφαρμογής είναι τα βιντεοπαιχνίδια. Ο πράκτορας εκπαιδεύεται στο να παίζει όλο και πιο αποτελεσματικά ένα ηλεκτρονικό παιχνίδι. Συνήθης αλγόριθμος στον κλάδο αυτό είναι ο αλγόριθμος του Q-learning ενώ έναν πολύ πρόσφατο state-of-the-art αλγόριθμο αποτελεί ο Double DQN (Double Deep Q-Network) από την εταιρεία πρωτοπόρο Deepmind της Google, ο οποίος δοκιμάστηκε σε παιχνίδια της γενιάς Atari 2600 και αποτελεί εξέλιξη του DQN. Άλλοι αντίστοιχοι γνωστοί αλγόριθμοι RL είναι οι SARSA και DDPG. Να σημειωθεί πως η Google Deepmind, τροφοδοτώντας νευρωνικά δίκτυα με μετρήσεις αισθητήρων στα κεντρικά κτήρια πληροφοριών της Google (Google Data Centres) μπόρεσε να μειώσει κατά 40% την κατανάλωση ενέργειας για την ψύξη των κτηρίων αυτών . Τα νευρωνικά δίκτυα εκπαιδεύτηκαν ώστε να κάνουν πρόβλεψη του PUE, της θερμοκρασίας και της πίεσης εντός του κτηρίου υπολογιστών, ρυθμίζοντας έτσι το σύστημα ψύξης αναλόγως. Με ενεργοποιημένο τον έλεγχο από ML, η εξοικονόμηση που παρατηρείται αντιστοιχεί σε μείωση 40% στην ψυκτική ισχύ που απαιτεί το κτήριο.

Στη μηχανική μάθηση, ταξινόμηση ονομάζεται η διαδικασία στην οποία ένας αλγόριθμος (ταξινομητής-classifier) εκπαιδεύεται πάνω σε δεδομένα τα οποία χαρακτηρίζονται από συγκεκριμένες ετικέτες, οι οποίες υποδεικνύουν την κλάση τους, και εκπαιδεύεται, με αυτό τον τρόπο, να ταξινομεί νέα δεδομένα στις κλάσεις αυτές.

Στην πράξη, η ταξινόμηση αποτελεί μια διαδικασία εκτίμησης μιας συνάρτησης στόχου *f* που αντιστοιχίζει διανύσματα γνωρισμάτων (ανεξάρτητες μεταβλητές) εισόδου X = {x1 ,...,xk} σε μια διακριτή έξοδο η οποία παίρνει τιμές από ένα σύνολο y = {y1 ,...,ym} όπου m ο αριθμός των κλάσεων και k ο αριθμός των γνωρισμάτων. Όπως είναι φανερό αφορά μια εργασία επιτηρούμενης μάθησης.

Ένα τέτοιο παράδειγμα ταξινόμησης είναι η εξαγωγή απόφασης για τον αν κάποιος θα αγόραζε ένα συγκεκριμένο αυτοκίνητο το οποίο παράγει μία συγκεκριμένη εταιρία με κριτήρια το εισόδημα του, το φύλλο και την ηλικία του. Η έξοδος που περιμένουμε είναι 0 ή 1, δηλαδή αν θα το αγόραζε ή όχι και το συγκεκριμένο αποτελεί πρόβλημα δυαδικής ταξινόμησης (binary classification) αφού η διακριτή έξοδος παίρνει δύο τιμές. Ένα τέτοιου είδους dataset μπορεί να βοηθήσει την εκάστοτε εταιρεία να κατανοήσει τα χαρακτηριστικά των εν δυνάμει πελατών της και να προχωρήσει σε αποτελεσματικές και στοχευμένες διαφημίσεις και εν συνεπεία αποφάσεις ανάλογα με το αν εκείνοι έχουν τις προοπτικές να αγοράσουν κάποιο προϊόν της εταιρείας χωρίς να χρεώνεται για διαφημίσεις χωρίς αποτέλεσμα σε ομάδες που, εν γένει, δε θα δείξουν ενδιαφέρον και δεν θα προχωρήσουν σε κάποια αγορά. Υπάρχουν και προβλήματα ταξινόμησης πολλαπλών ετικετών (multi label classification) όπου κάθε παρατήρηση χρήζει αντιστοίχισης σε παραπάνω από μία ετικέτα.

Η λογιστική παλινδρόμηση αποτελεί μια επέκταση της παλινδρόμησης η οποία όμως έχει εφαρμογή σε προβλήματα δυαδικής ταξινόμησης. Η εξαρτημένη μεταβλητή λαμβάνει δυαδικές τιμές (0,1), ωστόσο κατασκευάζεται ένα μοντέλο γραμμικής παλινδρόμησης. Στη συνέχεια εφαρμόζεται ο μετασχηματισμός στην εξαρτημένη μεταβλητή έτσι ώστε εν τέλει, στον κατακόρυφο άξονα να υπάρχει πλέον αναπαράσταση της πιθανότητας της κλάσης με τιμή 1. Συνήθως τιμές από το κατώφλι πιθανότητας 0.5 και πάνω μεταφράζονται ως πρόβλεψη για την κλάση 1. Παρόλα αυτά, είναι χρήσιμο σε ορισμένα προβλήματα το γεγονός ότι είναι γνωστές οι πιθανότητες της ταξινόμησης και μπορεί να υπάρχει μια πιο ρεαλιστική οπτική της κατάστασης. Συμπληρωματικά, είναι στην ευχέρεια του μελετητή να κάνει το μοντέλο πιο αυστηρό ή πιο χαλαρό, αυξομειώνοντας το επιθυμητό κατώφλι (threshold).

Οι Instance-based τεχνικές ταξινόμησης αφορούν μια κατηγορία τεχνικών μηχανικής μάθησης που βασίζεται ατομικά σε κάθε αντικείμενο προς ταξινόμηση, πράγμα που υποδηλώνεται και από τον τίτλος της. Έτσι το κομμάτι της εκπαίδευσης απαιτεί

ελάχιστο χρόνο και υπολογιστική ισχύ και όλες οι στοιχειώδεις εργασίες γίνονται κατά τη διάρκεια της διαδικασίας ταξινόμησης, γι' αυτό και ισχύει και ο χαρακτηρισμός lazy-learning algorithms. Αντιπροσωπευτικός αλγόριθμος της κατηγορίας αυτής είναι αυτός των k-πλησιέστερων γειτόνων (k-nearest neighbors (kNN)) ο οποίος απαιτεί την επιλογή ενός μέτρου απόστασης (για παράδειγμα Ευκλείδεια απόσταση, απόσταση Chebychev, σταθμισμένες αποστάσεις).

Λαμβάνοντας την υπόθεση ότι έχουμε ένα dataset με τα δοθέντα διανύσματα γνωρισμάτων και τις ετικέτες, γίνεται επιλογή του αριθμού k των κοντινότερων γειτόνων και του επιθυμητού μέτρου απόστασης. Κάθε νέο διάνυσμα Α κατηγοριοποιείται με τον παρακάτω τρόπο:

- Γίνεται επιλογή των k κοντινότερων γειτόνων του Α με βάση του μέτρου απόστασης
- Από αυτούς γίνεται μέτρηση των πόσων ανήκουν σε κάθε κατηγορία.
- Γίνεται ανάθεση στο νέο σημείο Α της ετικέτα της κατηγορίας στην οποία ανήκουν οι περισσότεροι εκ των γειτόνων.

Τα βασικά μειονεκτήματα αυτής της κατηγορίας αλγόριθμων είναι τα παρακάτω:

- Μεγάλες απαιτήσεις στον χρόνο του υπολογισμού και της μνήμη κατά τη διαδικασία της ταξινόμησης
- Απουσία γρήγορου και συστηματικού τρόπου της εύρεσης του κατάλληλου k. Μπορεί να γίνει χρήση της cross validation τεχνικής, η οποία κατατάσσεται στις ιδιαίτερα χρονοβόρες και υπολογιστικά απαιτητικές διαδικασίες.
- Ευαισθησία στην επιλογή του μέτρου απόστασης.

Με βάση τα παραπάνω μειονεκτήματα, καλό θα είναι να αποφεύγεται η χρήση του όταν τα δεδομένα της μελέτης είναι πολλά. Το βασικό πλεονέκτημα είναι η απλότητα και η δυνατότητα να λειτουργήσει για σύνορα κλάσεων με ακαθόριστους σχηματισμούς στο χώρο σε αντίθεση για παράδειγμα με τον αλγόριθμο ταξινόμησης Naive-Bayes.

Τα δέντρα αποφάσεων αποτελούν την πιο χαρακτηριστική περίπτωση τεχνικών εκμάθησης κανόνων. Η λογική στην οποία βασίζονται είναι: αναζητείται το γνώρισμα (feature) το οποίο διαχωρίζει με τον καλύτερο τρόπο τα δεδομένα σύμφωνα με μετρικές όπως είναι το gini index ή το κέρδος πληροφορίας. Το γνώρισμα αυτό αποτελεί τη ρίζα

του δέντρου αποφάσεων και παράλληλα αποτελεί και ένα κόμβο απόφασης που οδηγεί σε διαφορετικές επιλογές ανάλογα με τις τιμές που παίρνει το γνώρισμα . Στη συνέχεια το δέντρο κατασκευάζεται με τον ανάλογο τρόπο, ορίζοντας περιοχές οι οποίες αντιστοιχούν σε κλάσεις. Τα φύλλα του δέντρου πάντα ορίζουν την περιοχή μιας συγκεκριμένης κλάσης.

Τα δέντρα αποφάσεων έχουν την κακή φήμη ως προς το overfitting, ειδικά στις περιπτώσεις που είναι πλήρως αναπτυγμένα. Γι' αυτό τον λόγο και στη βιβλιογραφία υπάρχουν πολλές και διαφορετικές μέθοδοι «κλαδέματος» (pruning) με τις οποίες από την μια περιορίζεται η απόδοση του δέντρου στο training set, από την άλλη η απόδοση του σε νέα δεδομένα βελτιώνεται αφού δεν είναι ακριβώς προσαρμοσμένη στο training set και σε ότι θόρυβο αποφέρει. Σε γενικές γραμμές, αφορά μια παρωχημένη μέθοδο που πρακτικά είχε σταματήσει να έχει εφαρμογή μέχρι πρόσφατα που επανήλθε και πάλι με αναβαθμίσεις όπως είναι το gradient boosting και το random forest που αναλύεται παρακάτω.

Η ταξινόμηση του τυχαίου δάσους αποτελεί επέκταση των απλών δέντρων αποφάσεων στα πλαίσια του συνδυασμού μεθόδων (ensemble learning) και πιο συγκεκριμένα του bagging που έχει γίνει αναφορά στην εργασία νωρίτερα. Ο αλγόριθμος του τυχαίου δάσους διαρθρώνεται ως εξής:

1. Γίνεται επιλογή k τυχαία σημεία από το training set.
2. Κατασκευάζεται δέντρο απόφασης με βάση τα k αυτά σημεία.
3. Γίνεται επιλογή του επιθυμητού αριθμού δέντρων αποφάσεων και επαναλαμβάνονται τα προηγούμενα δύο βήματα.
4. Ένα νέο σημείο-πληροφορία κατατάσσεται στην κλάση που επιτάσσει η πλειοψηφία των δέντρων αποφάσεων.

Τα πλεονεκτήματα του παραπάνω αλγόριθμου είναι:

- Προσφέρει έναν πολύ πιο ομαλό αλλά και αποτελεσματικό έλεγχο του bias-variance tradeoff σε σχέση με τα decision trees. Περιορίζεται το overfitting που εμφανίζει ένα πλήρως αναπτυγμένο δέντρο, ενώ παράλληλα παρουσιάζεται μεγαλύτερη μεταβλητότητα από συμβαίνει σε κάποιο ρηχό ή υποανάπτυκτο δέντρο.

- Πετυχαίνει μεγαλύτερη ευστοχία, εισάγοντας στην πρόβλεψη την αντικειμενικότητα παραπάνω δέντρων. Αν κάποιο δέντρο για παράδειγμα έχει "παρασυρθεί" από κάποιο outlier, το φαινόμενο εξομαλύνεται αφού η απόφαση του σταθμίζεται και με των άλλων δέντρων.
- Εξακολουθεί να είναι επιρρεπής σε overfitting, παρόλα αυτά σε αισθητά χαμηλότερο επίπεδο με μεθόδους όπως το Gradient Tree Boosting.
- Τα random forests μπορούν εύκολα να παραλληλοποιηθούν σε real-time cloud πλατφόρμες κατά το στάδιο της εκπαίδευσης, καθώς το κάθε δέντρο εκπαιδεύεται ανεξάρτητα. Μια τέτοια πρόσφατη υλοποίηση είναι αυτή σε Apache Spark.
- Ο ίδιος αλγόριθμος μπορεί να χρησιμοποιηθεί και σε προβλήματα παλινδρόμησης όπου αντί για ψηφοφορία γίνεται η χρήση της μέσης τιμής των προβλέψεων του κάθε δέντρου.

Αντίστοιχα τα μειονεκτήματα του αλγόριθμου Random Forest είναι τα εξής:

- Είναι αισθητά πιο αργή η εκτέλεση του σε περιπτώσεις μεγάλων δεδομένων, όπου εκεί κατασκευάζονται πολλά δέντρα αποφάσεων.
- Σε προβλήματα με κατηγορικές μεταβλητές πολλών επιπέδων υπάρχει τάση να είναι biased προς αυτές, γεγονός που τον κάνει αναποτελεσματικό.

Η τεχνική μάθησης συνόλου κανόνων διέπετε από την προσπάθεια περιγραφής του training set με τη χρήση λογικών συμβόλων και κανόνων ταξινόμησης. Η περιγραφή κάθε κλάσης είναι δηλαδή της μορφής: $(X1 \wedge X2 \wedge \ldots \wedge Xn) \vee (Xn+1 \wedge Xn+2 \wedge \ldots X2n) \vee \ldots \vee (X(k-1)n+1 \wedge X(k-1)n+2 \wedge \ldots \wedge Xkn)$. Ο σκοπός της είναι να κατασκευαστεί μια περιγραφή των κλάσεων με τους λιγότερους δυνατούς κανόνες διότι αν το πλήθος των κανόνων είναι πολύ μεγάλο, ο αλγόριθμος αυτό που κάνει είναι να προσπαθεί απλώς να απομνημονεύσει το training set. Η περίπτωση αυτή εμπεριέχει προφανώς τον κίνδυνο του overfitting.

Στις στατιστικές τεχνικές μάθησης χρησιμοποιούνται τα μοντέλα πιθανοτήτων. Το training set όπως έχει αναφερθεί, αποτελείται instances $(\alpha 1, \ldots, \alpha n \mid vj \in V)$ δηλαδή ζεύγη διανυσμάτων γνωρισμάτων $(\alpha 1, \ldots, \alpha n)$ και εξόδων - ετικετών αντίστοιχης κλάσης ή κλάσεων που παίρνουν τιμές από ένα δεδομένο διακριτό σύνολο κλάσεων $V = \{v1, \ldots, vk\}$. Τα instances αυτά είναι ξεχωριστά «στιγμιότυπα» ή τιμές των γνωρισμάτων $(A1, \ldots, An)$ όπου Ai τα ονόματα των γνωρισμάτων (π.χ. A1 = ηλικία,

A2 = εισόδημα, α1=30 ετών, α2=50.000 ευρώ) μαζί με τις αντίστοιχες κλάσεις τους vi ως έξοδο (όπως για παράδειγμα αν αγόρασε το αυτοκίνητο ή όχι). Οι στατιστικοί ταξινομητές επιτυγχάνουν, επεξεργαζόμενοι κάθε instance, να χτίσουν ένα πιθανοτικό μοντέλο που εξάγει μια πιθανότητα ενός νέου στοιχείου να ανήκει σε κάποια από τις δοθείσες κλάσεις. Συνήθως επιλέγεται η κλάση με τη μεγαλύτερη πιθανότητα.

Ο ταξινομητής Naive – Bayes αποτελεί την πιο απλή μορφή ενός στατιστικού ταξινομητή. Ανατρέχοντας σε όλα τα instances (α1 ,...,αn | vi), μαθαίνει την πιθανότητα P(αi | vj), κάθε γνώρισμα Ai να παίρνει τιμές δεδομένης κάποιας ετικέτας κλάσης vi . Στη συνέχεια χρησιμοποιείται ο γνωστός νόμος του Bayes για να υπολογιστεί η δεσμευμένη πιθανότητα, ένα νέο διάνυσμα γνωρισμάτων να ανήκει σε κάθε μία από γνωστές κλάσεις και επιλέγεται συνήθως αυτή με τη μεγαλύτερη πιθανότητα. Η εκμάθηση των πιθανοτήτων P(αi | vj) γίνεται ανάλογα με τον τρόπο που επιλέγει ο εκάστοτε μελετητής. Οι πιο συνηθισμένες υλοποιήσεις είναι η Gaussian Naive Bayes και η Multinomial Naive Bayes. Στη διαδικασία αυτή, γίνεται η βασική και ψευδής παραδοχή ότι τα γνωρίσματα είναι εντελώς ανεξάρτητα μεταξύ τους, και γι' αυτό προκύπτει και η ονοματοδοσία «Naive» που σημαίνει αφελής. Μια πολύ απλή περίπτωση προβλήματος που καταδεικνύει το πρόβλημα αυτό είναι όταν δύο γνωρίσματα να είναι αυτά της ηλικίας και του εισοδήματος των ατόμων. Ο αλγόριθμος Naive-Bayes θεωρεί αυτά τα δύο γνωρίσματα στατιστικά ανεξάρτητα μεταξύ τους πράγμα το οποίο προφανώς δεν ισχύει στην πραγματικότητα. Ωστόσο, τα αποτελέσματα του αλγόριθμου συχνά εκπλήσσουν. Ο αλγόριθμος Naive - Bayes:

- Διακρίνεται για την ταχύτητα του στο επίπεδο της εκπαίδευσης (training) του μοντέλου.
- Ενδείκνυται όταν τα δεδομένα εκπαίδευσης είναι μικρού μεγέθους, καθώς συγκλίνει σχετικά γρήγορα.
- Ενδείκνυται κυρίως για προβλήματα κατηγοριοποίησης κειμένου.
- Θεωρείται ο κατάλληλος αλγόριθμος όταν ισχύει σε μεγάλο βαθμό η συνθήκη της ανεξαρτησίας.

Ο αλγόριθμος Naive-Bayes είναι αρκετά αποτελεσματικός σε πολλές περιπτώσεις ωστόσο πάντα υπήρχε ένα ερωτηματικό αν μπορεί να βελτιωθεί πιο πολύ η απόδοση του, εφόσον λειτουργεί με τη μη ρεαλιστική υπόθεση της ανεξαρτησίας. Το θέμα της συνθήκης ανεξαρτησίας έρχονται να λύσουν τα Bayesian δίκτυα. Αποτελούν

ακυκλικούς κατευθυνόμενους γράφους όπου κάθε κόμβος αναπαριστά μια τυχαία μεταβλητή - γνώρισμα ή κλάση. Τα βέλη αναπαριστούν την εξάρτηση, τη σχέση αιτίου-αιτιατού. Οι κόμβοι ικανοποιούν τη Μαρκοβιανή ιδιότητα, είναι δηλαδή ανεξάρτητοι των «προγόνων» δεδομένων των «γονέων» τους. Αντίθετα με τον αλγόριθμο Naive – Bayes, στην περίπτωση αυτή επιτρέπεται η σύνδεση μεταξύ των γνωρισμάτων Ai. Συναρτήσεις (scoring functions) όπως η MDL scoring function χρησιμοποιούνται με σκοπό να βρεθεί η κατάλληλη δομή αλλά και οι παράμετροι του Bayesian δικτύου που θα ικανοποιούν με τον καλύτερο τρόπο το εκάστοτε dataset. Γνωστές υποκατηγορίες των Bayesian δικτύων που χρησιμοποιούνται ως ταξινομητές είναι:

- General Bayesian Network
- Tree augmented Naive Bayes – TAN
- BN augmented Naive Bayes - BAN

Η βασική διαφορά των Bayesian δικτύων από τους υπόλοιπους αλγόριθμους classification είναι ότι δίνουν τη δυνατότητα να αποκαλυφθεί σε ένα dataset οι από κοινού κατανομές P(Ai , Vj) και όχι μόνο οι δεσμευμένες κατανομές P(Ai | Vj). Αφορά ένα πολύ πιο δύσκολο και χρονοβόρο υπολογισμό που προσφέρει όμως τα παρακάτω πλεονεκτήματα:

- Δίνεται μια αρκετά πιο συνολική εικόνα των εξαρτήσεων για το σύνολο των δεδομένων που αναλύονται
- Φανερώνονται οι σχέσεις αιτίου και αιτιατού στα δεδομένα
- Μειώνεται η εξάρτηση από τις ελλείψεις δεδομένων (missing data), φαινόμενο το οποίο είναι για παράδειγμα πολύ σύνηθες σε ιατρικά δεδομένα που μας ενδιαφέρουν.
- Ο υπολογισμός των κοινών κατανομών γίνεται πολύ πιο εύκολος σε διακριτές μεταβλητές, όπου η χρήση τους είναι πιο συνήθης.

Η παλινδρόμηση (regression) αποτελεί ένα σύνολο από μεθόδους για την εκτίμηση της εξάρτησης μεταξύ μιας βαθμωτής ή διανυσματικής ανεξάρτητης μεταβλητής X και μιας βαθμωτής εξαρτημένης μεταβλητής y. Στόχος της είναι να προσδιοριστεί η μέση απόκριση που έχει η εξαρτημένη μεταβλητή στις επιμέρους μεταβολές της ανεξάρτητης μεταβλητής. Διαισθητικά ένα μοντέλο παλινδρόμησης αποτελεί μέσο έκφρασης των παρακάτω στοιχείων :

1. Μιας τάσης της εξαρτημένης μεταβλητής y που μεταβάλλεται με συστηματικό τρόπο σε σχέση με τις μεταβολές της ανεξάρτητης μεταβλητής.
2. Μιας διασποράς σημείων γύρω από την καμπύλη της στατιστικής σχέσης που συνδέει τις X, y.

Οι πιο συνήθεις μετρικές αξιολόγησης των μοντέλων παλινδρόμησης είναι το μέσο τετραγωνικό σφάλμα (MSE), το μέσο απόλυτο σφάλμα (MAE), το R 2 αλλά και το explained variance.

Στη βιβλιογραφία πλέον εξετάζονται πολλά διαφορετικά μοντέλα παλινδρόμησης όπως η απλή γραμμική παλινδρόμηση, η πολλαπλή γραμμική παλινδρόμηση, η πολυωνυμική παλινδρόμηση, η παλινδρόμηση με διάνυσμα στήριξης (SVR: support vector regression), η παλινδρόμηση δέντρων αποφάσεων, η παλινδρόμηση random forest και η λογιστική παλινδρόμηση η οποία έχει εφαρμογή κυρίως σε προβλήματα ταξινόμησης.

Η συσταδοποίηση είναι η εργασία καταμερισμού ενός ετερογενούς πληθυσμού σε ένα σύνολο περισσοτέρων ετερογενών συστάδων (clusters), με την ιδιότητα ότι τα στοιχεία της κάθε συστάδας να είναι πιο όμοια μεταξύ τους απ' ότι εκείνα των άλλων συστάδων. Η ομοιότητα, βέβαια, είναι μια έννοια αρκετά σχετική και εξαρτάται κάθε φορά από τη φύση του προβλήματος που καλείτε ο μελετητής να λύσει. Τα σημεία πληροφορίας αναφέρονται και ως διανύσματα ή πρότυπα. Η συσταδοποίηση αποτελεί μια εργασία μη επιτηρούμενης μάθησης καθώς οι ομάδες δεν είναι γνωστές εκ των προτέρων. Αυτό έρχεται σε αντίθεση με τον επιτηρούμενο χαρακτήρα της ταξινόμησης (classification) που έχει, εκ των προτέρων, γνώση σεσημασμένων δεδομένων (labeled data) με τα οποία εκπαιδεύεται ένα μοντέλο ώστε να κατηγοριοποιεί τα νέα δεδομένα με ετικέτες από ένα προκαθορισμένο σύνολο κλάσεων. Επιπλέον η συσταδοποίηση δεν είναι μια αυτοματοποιημένη διαδικασία, με εκ των προτέρων μοντέλα που έχουν εκπαιδευτεί και που εφαρμόζονται σε δεδομένα. Αντιθέτως πρόκειται για μια επαναληπτική διαδικασία σταδιακής «ανακάλυψης γνώσης» καθώς και βελτιστοποίησης που περιλαμβάνει δοκιμές και αποτυχίες. Πολλές φορές είναι απαραίτητη η τροποποίηση του σταδίου της προ επεξεργασίας δεδομένων (data preprocessing), η αλλαγή των παραμέτρων του μοντέλου και πιθανώς και η δοκιμή διαφορετικών αλγόριθμων συσταδοποίησης για να προκύψει το επιθυμητό αποτέλεσμα. Ο αλγόριθμος DBSCAN

(Density Based Algorithm for Discovering Clusters) στηρίζεται στην πυκνότητα των σημείων πληροφορίας με τα παρακάτω πλεονεκτήματα:

- ✓ Δε χρειάζεται ο ορισμός αριθμού συστάδων.
- ✓ Δεν επηρεάζεται από τους σχηματισμούς συστάδων στο χώρο και έτσι μπορεί να ανιχνευτούν συστάδες με αυθαίρετα σχήματα, αποστάσεις και κατανομές στο χώρο.
- ✓ Είναι αρκετά αποτελεσματικός σε σύνολα δεδομένων με έντονο θόρυβο και απομακρυσμένα σημεία (outliers). Ωστόσο απαιτείται η αρχικοποίηση δύο βασικών παραμέτρων που καθορίζουν το αποτέλεσμα συσταδοποίησης για το εκάστοτε dataset. Επιπλέον, οι τιμές που παίρνουν, για να προκύψουν τα επιθυμητά ικανοποιητικά αποτελέσματα, αποτελούν αντικείμενο διερεύνησης. Οι παράμετροι αυτές είναι: ε, είναι η μέγιστη απόσταση μεταξύ δύο σημείων ώστε να θεωρούνται αυτά «γειτονικά», minPoints, είναι ο ελάχιστος αριθμός σημείων που απαιτούνται για να συσταθεί μια "πυκνή περιοχή". Για την εκτέλεση του κατηγοριοποιούνται τα σημεία πληροφορίας σε σημεία πυρήνα, σημεία άμεσης εμβέλειας, σημεία έμμεσης εμβέλειας και σημεία περιφέρειας.

Η συσταδοποίηση, λοιπόν, διαρθρώνεται με τον παρακάτω τρόπο:

- Η κάθε συστάδα διαμορφώνεται από ένα σημείο πυρήνα και όλα τα σημεία άμεσης και έμμεσης εμβέλειας του.
- Κάθε συστάδα περιέχει τουλάχιστον ένα σημείο πυρήνα.
- Σημεία που δεν είναι πυρήνα αλλά ούτε περιφέρειας εντάσσονται σε συστάδες και αποτελούν το «σύνορο» τους .
- Τα εξωτερικά σημεία δεν ανήκουν σε καμία συστάδα.

Στη βιβλιογραφία πλέον υπάρχουν πολλές παραλλαγές του αλγόριθμου, όπως οι GDBSCAN, HDBSCAN αλλά και η ιεραρχική επέκταση του OPTICS.

Η συσταδοποίηση έχει πολλαπλές εφαρμογές, σε κλάδους όπως η βιολογία, η ιατρική, οι κοινωνικές επιστήμες, ο τομέας του Marketing αλλά και τα Συστήματα Διοίκησης. Ένα παράδειγμα εφαρμογής είναι να μπορεί μια εταιρεία να κατηγοριοποιεί τους πελάτες τις σε διαφορετικές συστάδες ανάλογα με τις καταναλωτικές τους συνήθειες ώστε να προσαρμόζονται τα προϊόντα της και οι διαφημίσεις της πάνω σε αυτές τις καταναλωτικές τους συνήθειες. Η επιλογή αλγόριθμου συσταδοποίησης είναι πάντα σε

συνάρτηση με τη φύση του προβλήματος και το αντικειμένου έρευνας αλλά και την ενασχόληση και την εμπειρίας του κάθε αναλυτή.

4 Η προτεινόμενη προσέγγιση

4.1 Βιβλιοθήκες Και Εργαλεία Που Χρησιμοποιήθηκαν

Για την επίτευξη της ανάλυσης και πρόβλεψης χρονολογικών σειρών χρησιμοποιήθηκε η γλώσσα προγραμματισμού της python και κάποιες βιβλιοθήκες που υπάρχουν για αυτήν την γλώσσα και αφορούν τα πεδία της μηχανικής μάθησης, στατιστικής, ανάλυσης δεδομένων και επιστημονικών υπολογισμών. Οι πιο σύνηθες που χρησιμοποιήθηκαν είναι οι παρακάτω:

- Pandas: είναι μια βιβλιοθήκη που παρέχει γρήγορες, εκφραστικές και ευέλικτες δομές δεδομένων. Παρέχουν αρκετά πλούσιο και ανοιχτού κώδικα ανάλυση δεδομένων, τα οποία μπορούν να είναι διαφορετικού τύπου. Οι δύο βασικές δομές δεδομένων του πακέτου είναι τα “Series” και το “Dataframe”,τα οποία μπορούν να διαχειριστούν πλήθος γνωστών εφαρμογών όπως οικονομικών, στατιστικών, κτλ. Μπορεί επίσης εύκολα να διαχειριστεί ελλιπή δεδομένα, χρονολογικές σειρές, είσοδο – έξοδο αρχείων. Η επεξεργασία και μετατροπή των δεδομένων από την μία μορφή σε άλλη, πχ. Από «dataframe» σε «Series», ώστε να βρίσκονται στην μορφή που χρειαζόμαστε κάθε φορά είναι συνηθισμένη και εύκολη.
- NumPy: Είναι μια βιβλιοθήκη που χρησιμοποιείται για θεμελιώδη επιστημονικό υπολογισμό, περιέχει μεταξύ άλλων, γραμμική άλγεβρα μετασχηματισμός Φουριέρ, κτλ. Χρησιμοποιείται στην εφαρμογή κυρίως για εργασίες με πίνακες και να φέρει τα δεδομένα στην μορφή που περιμένουν ως είσοδο οι συναρτήσεις της sklearn για εκπαίδευση και πρόβλεψη.
- Sklearn: Πλήρης ολοκληρωμένη βιβλιοθήκη για την python που αφορά την μηχανική μάθηση, περιέχει εργαλεία για ανάλυση και εξόρυξη δεδομένων, ανοιχτού κώδικα, περιέχει όλους τους γνωστούς αλγορίθμους για την κατηγοριοποίηση, παλινδρόμηση, συσταδοποίηση, προ- επεξεργασία δεδομένων και διαθέτει προ - εγκατεστημένα dataset. Στην εφαρμογή χρησιμοποιούνται για την εκπαίδευση και πρόβλεψη χρονοσειρών, καθώς και ρύθμιση παραμέτρων των αλγορίθμων.
- Matplotlib: βιβλιοθήκη η οποία διαχειρίζεται κυρίως γραφήματα, τα οποία για τις εφαρμογές και τις οικονομικές χρονολογικές σειρές αποτελούν πολύ σημαντικό κομμάτι για την ανάλυσή τους, κατανόησή τους και τέλος την πρόβλεψή τους.

4.2 Συλλογή & Επεξεργασία Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν κατά την εφαρμογή του μοντέλου είναι 1461 εγγραφές για τις εισαγωγές σε ιδιωτικά νοσοκομεία της Ελλάδας από τον Ιανουάριο του 2018 μέχρι τον Δεκέμβριο του 2021 απο τα στοιχεία Ελληνικής Στατιστικής Υπηρεσίας [21]. Για την μελέτη των παραγόντων που επιλέχθηκαν ενσωματώθηκαν στα δεδομένα η εποχή, η πληροφορία των επίσημων αργιών, ο διαχωρισμός των Σαββατοκύριακων από τις καθημερινές καθώς και η μέση θερμοκρασία που επικρατούσε ανά ημέρα στην Ελλάδα κατά τα έτη που εξετάζονται [22] και έχουν εξαχθεί σε αρχείο μορφής csv. Τα δεδομένα που αντλήθηκαν περιέχουν σειρές δεδομένων, με κάθε σειρά να αντιστοιχεί σε στήλες που αντιπροσωπεύουν συγκεκριμένα πληροφορίες σχετικά με:

- Month- Μήνας
- Day- Ημέρα της εβδομάδας
- Season- Εποχή
- Hmerominia – Ημερομηνία εισαγωγών
- Argia – Αν ημερομηνία αφορά επίσημη αργία
- Weekend –Αν η ημερομηνία αφορά Σαββατοκύριακο
- Temperature- θερμοκρασία
- PlithosEisagwgwn- Το πλήθος των εισαγωγών κατά την συγκεκριμένη ημερομηνία

Αρχικά συλλέχθηκε η πληροφορία του πλήθους εισαγωγών ανά ημέρα των ετών που μελετήθηκαν. Στην συνέχεια αυτή η ημερομηνία που αντιστοιχούσε σε κάθε ημέρα των ετών των ετών 2018-2021, αναλύθηκε και διαχωρίστηκε στις στήλες όπως περιγράφονται παρακάτω. Η στήλη Month (Μήνας) αφορά τον ημερολογιακό μήνα του έτους, ενώ η Day (Ημέρα της εβδομάδας) που καταγράφεται η πληροφορία ποια ημέρα της εβδομάδας αφορά η κάθε η ημερομηνία με την αντιστοίχιση να είναι ως εξής: 1 Κυριακή, 2 Δευτέρα, 3 Τρίτη, 4 Τετάρτη, 5 Πέμπτη, 6 Παρασκευή, 7 Σάββατο. Η αντιστοίχιση αυτή έγινε για την καλύτερη διαχείριση των δεδομένων κατά την εφαρμογή του κώδικα καθώς τα Integer αριθμητικά δεδομένα είναι πιο εύκολα διαχειριστικά από ότι τα λεκτικά δεδομένα που περιγράφουν κάτι. Η επόμενη στήλη αφορά την εποχή με βάση και πάλι την ημερομηνία εισαγωγών. Λαμβάνοντας υπόψιν και πάλι λοιπόν την ημερομηνία έχει γίνει αντιστοίχιση με την εποχή που αφορά η

ημερομηνία με τον διαχωρισμό να γίνεται ως εξής: 1 Χειμώνας (1/12 έως 28/2 ή 29/2), 2 Άνοιξη (1/3 έως 31/5), 3 Καλοκαίρι (1/6 έως 31/8), 4 Φθινόπωρο (1/9 έως 30/11). Η στήλη Hmerominia (Ημερομηνία Εισαγωγών) διατηρήθηκε όπως είχε συλλεχθεί από τα αρχικά δεδομένα που λήφθηκαν. Η στήλη Argia δημιουργήθηκε με βάση το αν η ημερομηνία ανά σειρά δεδομένων αφορούσε κάποια επίσημη αργία στην Ελλάδα με βάση τα στοιχεία του Κ.Ε.Π.Ε.Α. (Κέντρο Πληροφόρησης Εργαζομένων & Ανέργων) της Γ.Ε.Σ.Ε.Ε. (Γενική Συνομοσπονδία Εργατών Ελλάδος) [23]. Έτσι στην περίπτωση που η ημερομηνία της εκάστοτε σειράς αφορά κάποια από τις επίσημες αργίες η τιμή που παίρνει η στήλη είναι 1 διαφορετικά 0. Η επόμενη στήλη αφορά και πάλι πληροφορία που αφορά την ημερομηνία και το αν είναι Σαββατοκύριακο ή όχι, πληροφορία που λήφθηκε από το ημερολόγιο του κάθε έτους και τιμές που έλαβε και πάλι η στήλη είναι 1, 0. Η στήλη Temperature (Θερμοκρασία) αφορά την μέση θερμοκρασία στην Ελλάδα κατά την ημερομηνία της εκάστοτε σειράς, δεδομένο που λήφθηκε από το Ιστορικό Αρχείο Καιρού του meteoblue ανά ημέρα του κάθε έτους κάθε κάνοντας την αντιστοίχιση με τα δεδομένα εισαγωγών [22]. Τέλος η στήλη PlithosEisagwgwn (Πλήθος εισαγωγών) αφορά το συνολικό αριθμό καταγραμμένων εισαγωγών στα ιδιωτικά νοσοκομεία της Ελλάδος. Λέγοντας εισαγωγή συμπεριλαμβάνονται όλα τα είδη περιστατικών που μπορεί να κλήθηκαν να αντιμετωπίσουν τα νοσοκομεία κατά την συγκεκριμένη ημερομηνία, συμπεριλαμβανόμενων των ολιγόωρων εισαγωγών, των έκτακτων αλλά και προγραμματισμένων περιστατικών καθώς και των εισαγωγών ανηλίκων και ενηλίκων.

Ο στόχος της μελέτης των παραπάνω δεδομένων είναι να εντοπιστεί η πιθανή συσχέτιση των συγκεκριμένων παραγόντων ως προς την διαμόρφωση του πλήθους εισαγωγών ανά ημέρα, η εκπαίδευση του μοντέλου μέσω του train και του test set που διαμορφώνεται κατά την πορεία της εργασίας και τελικώς την πρόβλεψη τιμών που αφορούν τις εισαγωγές.

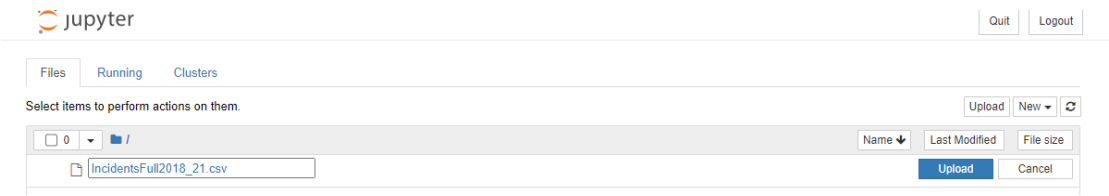
Στην επιστημονική έρευνα ο στόχος είναι η γενίκευση δηλαδή η περιγραφή μιας ή περισσοτέρων μεταβλητών του πληθυσμού καθώς και την εξήγηση των σχέσεων μεταξύ μεταβλητών του πληθυσμού. Συνεπώς χρειάζεται να έχουν συγκεντρωθεί και να αναλυθεί πληροφορίες για τις διάφορες μεταβλητές του πληθυσμού σε σχέση με το αντικείμενο που εξετάζεται. Επειδή η συγκέντρωση πληροφοριών είναι μια δύσκολη, χρονοβόρα, ακριβή και μερικές φορές αδύνατη διαδικασία, συλλέγονται πληροφορίες

από ένα δείγμα του πληθυσμού και βασιζόμενοι στα δεδομένα (στοιχεία) που επιλέξαμε από το δείγμα διεξάγουμε τις αναλύσεις. Οι αναλύσεις αυτές βασίζονται στα δεδομένα του δείγματος, που είναι εσκεμμένα και αυστηρά επιλεγμένα με επιστημονική ακρίβεια. Στην περίπτωση αυτή η έρευνα έχει περιοριστεί στα δεδομένα που αφορούν τα ιδιωτικά νοσοκομεία της Ελλάδος, το πλαίσιο των 4 προηγούμενων ετών αλλά και τους συγκεκριμένους παράγονται που επιλέχθηκαν και αφορούν την χρονική περίοδο εισαγωγής και την θερμοκρασία. Στόχος της επιλογής αυτών των παραγόντων είναι να παρατηρηθεί αν οι συγκεκριμένοι παράγοντες που είναι εύκολοι στην συλλογή τους μπορούν να βοηθήσουν στην πρόβλεψη των εισαγωγών αλλά και στην λήψη αποφάσεων σχετικών με διαχειριστικά θέματα ενός νοσοκομείου όπως η κατανομή του προσωπικού ή άδειες του προσωπικού, τον προγραμματισμό προσλήψεων ή η μείωση προσωπικού αντίστοιχα. Αυτές είναι μόνο μερικές από τις αποφάσεις που θα μπορούσαν να διευκολυνθούν από προβλέψεις που αφορούν τις εισαγωγές ασθενών καθώς όλη η διαχείριση και διοίκηση ενός φορέα παροχής υπηρεσιών υγείας επηρεάζεται από μια τέτοια πρόβλεψη.

Αρχικώς έχει γίνει αναφορά στα Μεγάλα Δεδομένα και στα μοντέλα που βασίζονται στην ανάλυση αυτών, παρόλα αυτά το δείγμα που μελετάται στην προκειμένη περίπτωση δεν είναι παράδειγμα Big Data με βάση το μέγεθος του καθώς η λήψη τέτοιων μεγεθών προϋποθέτει την πρόσβαση μεγάλο πλήθος δεδομένων και λήψη επιπλέον παραγόντων και εφαρμογή τους σε πολύπλοκα μοντέλα, γεγονός δύσκολο στο επίπεδο της διπλωματικής εργασίας. Τα δεδομένα που αφορούν τον τομέα της Υγείας είναι πάντα δύσκολα προσβάσιμα λόγω του ευαίσθητου χαρακτήρα τους και των κανονισμών GDPR που εφαρμόζονται στην Ελλάδα. Παρόλα αυτά οφείλουμε να δηλώσουμε τα οφέλη που λογικά θα προκύπταν από την αντίστοιχη ανάλυση και μελέτη με Big Data. Θα μπορούσαν να αναλυθούν επιπλέον παράγοντες που μπορεί να επηρέαζαν τις αποφάσεις και τα συμπεράσματα μας, καθώς επίσης θα υπήρχε μεγαλύτερη ακρίβεια των προβλέψεων και των μοντέλων που θα μπορούσαν να αναπτυχθούν. Οι επιπλέον παράγοντες που θα μπορούσαν να προστεθούν είναι άλλες καιρικές συνθήκες ανά ημέρα του έτους, για παράδειγμα υγρασία, χιόνι, βροχή, οι ώρες κατά τις οποίες γίνονται οι εισαγωγές, παράγοντας που μπορεί να υποδείξει τις ώρες αιχμής των μονάδων παροχής υπηρεσιών υγείας και να προσφέρει και την αντίστοιχη πρόβλεψη. Πέρα από τα δεδομένα που αφορούν ημερολογιακή πληροφορία και καιρικές συνθήκες, μεγάλο ενδιαφέρον θα υπήρχε και στην μελέτη παραγόντων όπως

ο τύπος περισταστικού που αφορά η κάθε εισαγωγή (παθολογικό, χειρουργικό, ογκολογικό κ.τ.λ) αλλά και ο χαρακτηρισμός της εισαγωγής σε επείγουσα ή προγραμματισμένη και σε ασθένεια ή ατύχημα. Αυτές η κατηγοριοποιήσεις των εισαγωγών μπορούν να αναδείξουν τις τάσεις ανά περίοδο, την αύξηση συγκεκριμένων ασθενειών ή τυχόν προβλήματα υγείας που επηρεάζουν όλο και περισσότερο μέρος του πληθυσμού.

Η δημιουργία αρχείων DataFrames από αρχεία CSV (με κόμμα διαχωρισμένα) γίνεται εξαιρετικά απλή με τη λειτουργία read_csv () στα Pandas, αφού είναι γνωστή η διαδρομή προς το αρχείο. Ένα αρχείο CSV είναι ένα αρχείο κειμένου που περιέχει δεδομένα σε μορφή πίνακα, όπου οι στήλες διαχωρίζονται με το χαρακτήρα ',' και οι γραμμές βρίσκονται σε ξεχωριστές γραμμές.
Εάν τα δεδομένα είναι σε κάποια άλλη μορφή, όπως μια βάση δεδομένων SQL ή ένα αρχείο Excel (XLS / XLSX), μπορούν να χρησιμοποιηθούν άλλες συναρτήσεις που διαβάζονται από αυτές τις πηγές σε DataFrames, δηλαδή read_xlsx, read_sql. Ωστόσο, για απλότητα, η εξαγωγή δεδομένων απευθείας στο CSV και η χρήση τους, είναι προτιμότερη.



6. Ανέβασμα αρχείου δεδομένων

4.3 Εφαρμογή μοντέλου- κώδικας

Στην ενότητα αυτή περιγράφεται ο κώδικας που χρησιμοποιήθηκε για την υλοποίηση της εφαρμογής καθώς και τα αποτελέσματα του κώδικα όπως εμφανίστηκαν κατά την χρήση του κώδικα με τα δεδομένα που εφαρμόστηκαν.

Για να διαβαστεί το αρχείο διαβάζεται το όνομα με το οποίο το αρχείο είναι αποθηκευμένο στον υπολογιστή και χρησιμοποιείται το encoding που είναι ο τρόπος για να γίνει η ερμηνεία των bytes (data= pd.read_csv). Χρησιμοποιώντας το Jupiter Notebook, η έξοδος είναι το όνομα του DataFrame. Η εκτύπωση είναι ένας βολικός

τρόπος για την προεπισκόπηση των δεδομένων που έχουν φορτωθεί, επιβεβαιώνεται ότι τα ονόματα των στηλών έχουν εισαχθεί σωστά, ότι οι μορφές δεδομένων είναι όπως αναμένεται και αν υπάρχουν τιμές που λείπουν οπουδήποτε. Παρατηρείται ότι το Pandas εμφανίζει μόνο 20 στήλες από προεπιλογή για δεδομένα πλαισίων δεδομένων και μόνο 60 περίπου σειρές, περικόπτοντας το μεσαίο τμήμα.

Για τη διαγραφή σειρών και στηλών από το DataFrames, το Pandas χρησιμοποιεί τη συνάρτηση "drop". Για τη διαγραφή μιας στήλης ή πολλών στηλών, χρησιμοποιείται το όνομα της στήλης και ορίζεται ο "axis" ως 1. Εναλλακτικά, όπως στο παρακάτω παράδειγμα, η παράμετρος "columns" έχει προστεθεί στο Pandas, το οποίο κόβει την ανάγκη για "axis". Η συνάρτηση drop επιστρέφει ένα νέο DataFrame, με τις στήλες να καταργούνται.

In [19]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import math
import seaborn as sns
from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
```

In [7]:
```
data = 'IncidentsFull2018_21.csv'

incidents_data = pd.read_csv(data,sep='\t')

incidents_data.head()
```

Out[7]:

	Month	Day	Season	Hmerominia	Argia	Weekend	Plithos	Temperature
0	1	2	1	1/1/2018	0	0	70	143.054
1	1	3	1	2/1/2018	0	0	141	125.816
2	1	4	1	3/1/2018	0	0	228	91.848
3	1	5	1	4/1/2018	0	0	210	73.128
4	1	6	1	5/1/2018	0	0	189	73.152

In [15]:
```
#Prep Analysis and Feature Selection
incidents_data = incidents_data.drop (['Hmerominia'],axis=1)
```

7. *Προετοιμασία δεδομένων και συλλογής χαρακτηριστικών 1*

Η εντολή shape δίνει πληροφορίες σχετικά με το μέγεθος του συνόλου δεδομένων - το shape επιστρέφει μια πλειάδα με τον αριθμό των γραμμών και τον αριθμό των στηλών για τα δεδομένα στο DataFrame. Μια άλλη περιγραφική ιδιότητα είναι το ndim που

δίνει τον αριθμό των διαστάσεων στα δεδομένα, συνήθως 2. Τα δεδομένα περιέχουν 1461 σειρές, το καθένα με 5 στήλες όπως φαίνεται από την έξοδο του .shape. Υπάρχουν δύο διαστάσεις, DataFrame 2D με ύψος και πλάτος.

```
In [60]: y= data.Plithos
         x= data.drop ('Plithos',axis=1)

In [61]: x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2)
         x_train.head()
```

Out[61]:

	Hmeromina	Argia	Weekend	Temperature
1308	1627776000	0	1	24.3997
1452	1640217600	0	0	12.5351
1422	1637625600	0	0	11.9724
819	1585526400	0	0	12.7632
1175	1616284800	0	1	13.7351

```
In [62]: x_train.shape
Out[62]: (1168, 4)
```

8. Προετοιμασία δεδομένων και συλλογής χαρακτηριστικών 2

Η μέθοδος DataFrame.head() στο Pandas, από προεπιλογή, δείχνει τις 5 πρώτες σειρές δεδομένων στο DataFrame. Το αντίθετο είναι το DataFrame.tail(), το οποίο σας δίνει τις τελευταίες 5 σειρές. Δίνοντας έναν αριθμό τα Pandas θα εκτυπώσουν τον καθορισμένο αριθμό γραμμών όπως φαίνεται στο παρακάτω παράδειγμα. Το head() και το tail() αποτελούν βασικά τμήματα των λειτουργιών Python Pandas για διερεύνηση των συνόλων δεδομένων.

```
In [59]: data.info ()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1461 entries, 0 to 1460
Data columns (total 5 columns):
Hmeromina      1461 non-null int32
Argia          1461 non-null int64
Weekend        1461 non-null int64
Plithos        1461 non-null int64
Temperature    1461 non-null float64
dtypes: float64(1), int32(1), int64(3)
memory usage: 51.4 KB
```

9. Προετοιμασία δεδομένων και συλλογής χαρακτηριστικών 3

Πολλά DataFrames έχουν διαφορετικούς τύπους δεδομένων, δηλαδή μερικές στήλες είναι αριθμοί, μερικές είναι συμβολοσειρές και μερικές είναι ημερομηνίες κλπ. Εσωτερικά, τα αρχεία CSV δεν περιέχουν πληροφορίες για τους τύπους δεδομένων που περιέχονται σε κάθε στήλη, όλα τα δεδομένα είναι μόνο χαρακτήρες. Το Pandas συνάγει τους τύπους δεδομένων κατά τη φόρτωση των δεδομένων, π.χ. εάν μια στήλη περιέχει μόνο αριθμούς, οι pandas θα ορίσουν τον τύπο δεδομένων αυτής της στήλης σε αριθμητικό. Για τον τύπο κάθε στήλης χρησιμοποιείται το df.dtypes.

```
In [16]: incidents_data.head()
```

Out[16]:

	Month	Day	Season	Argia	Weekend	Plithos	Temperature
0	1	2	1	0	0	70	143.054
1	1	3	1	0	0	141	125.816
2	1	4	1	0	0	228	91.848
3	1	5	1	0	0	210	73.128
4	1	6	1	0	0	189	73.152

10. Προετοιμασία δεδομένων και συλλογής χαρακτηριστικών 4

Οι τιμές που μπορεί να λείπουν είναι σύνηθες όταν εργαζόμαστε με σύνολα δεδομένων πραγματικά και όχι με τα καθαρισμένα που είναι διαθέσιμα στο διαδίκτυο για παράδειγμα. Τα δεδομένα που λείπουν μπορεί να προκύψουν από έναν ανθρώπινο παράγοντα (για παράδειγμα, ένα άτομο που σκόπιμα δεν έχει απαντήσει σε μια ερώτηση έρευνας), ένα πρόβλημα στους ηλεκτρονικούς αισθητήρες ή άλλους παράγοντες. Όταν συμβεί αυτό, μπορούν να χαθούν σημαντικές πληροφορίες.

Δεν υπάρχει τέλειος τρόπος για να γίνει χειρισμός αυτών των τιμών που λείπουν και που θα δώσουν ένα ακριβές αποτέλεσμα ως προς το ποια είναι η τιμή που λείπει. Υπάρχουν όμως αρκετές τεχνικές που μπορούν να αξιοποιηθούν και που θα δώσουν αξιοπρεπή απόδοση.

Στη συγκεκριμένη εργασία θα ακολουθηθεί η μέθοδο διαγραφής αφαιρώντας μια εγγραφή ή μια παρατήρηση στο σύνολο δεδομένων εάν περιέχει κάποιες τιμές που λείπουν. Για να εντοπιστούν αν υπάρχουν τιμές που λείπουν εκτελείται η παρακάτω εντολή που τις εντοπίζει:

```
In [17]: #Basic Checks for missing values
         incidents_data.isnull().sum()

Out[17]: Month          0
         Day            0
         Season         0
         Argia          0
         Weekend        0
         Plithos        0
         Temperature    0
         dtype: int64
```

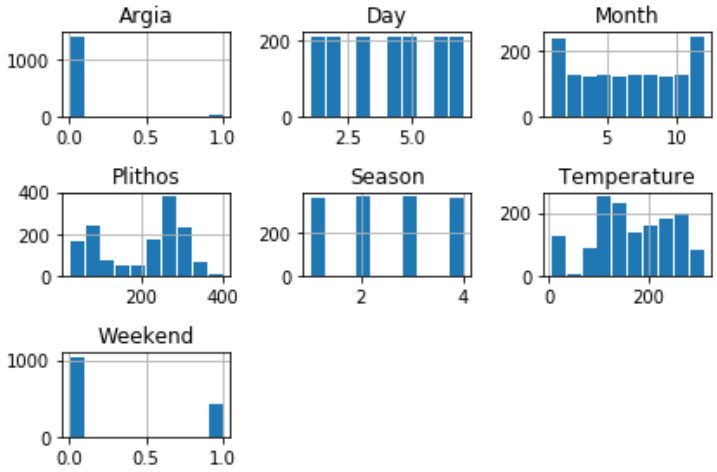
11. Βασικοί έλεγχοι κενών τιμών

Στο σύνολο των δεδομένων που γίνεται η παρούσα μελέτη δεν εντοπίζονται κενές τιμές ώστε να αφαιρεθούν.

Ως επόμενο βήμα επιλέχθηκε μια αρχική απεικόνιση με ιστογράμματα ώστε να υπάρξει καλύτερη αντίληψη των δεδομένων. Το ιστόγραμμα είναι μια γραφική αναπαράσταση που χρησιμοποιείται συνήθως για την οπτικοποίηση της κατανομής των αριθμητικών δεδομένων. Όταν εξερευνάται ένα σύνολο δεδομένων, συχνά χρειάζεται να κατανοηθεί γρήγορα η κατανομή ορισμένων αριθμητικών μεταβλητών μέσα σε αυτό. Μπορεί λοιπόν να γίνει αυτό χρησιμοποιώντας ένα ιστόγραμμα. Ένα ιστόγραμμα διαιρεί τις τιμές μέσα σε μια αριθμητική μεταβλητή σε "bins" και μετράει τον αριθμό των παρατηρήσεων που εμπίπτουν σε κάθε bin. Οπτικοποιώντας αυτές τις δεσμευμένες μετρήσεις σε στήλη, μπορούμε να αποκτήσουμε μια πολύ άμεση και διαισθητική αίσθηση της κατανομής των τιμών μέσα σε μια μεταβλητή.

Με την παρακάτω εντολή δημιουργήθηκαν τα ιστογράμματα χρησιμοποιώντας Python. Συγκεκριμένα, θα χρησιμοποιήθηκε η μέθοδος pandas hist(), η οποία είναι απλώς ένα περιτύλιγμα για το matplotlib pyplot API.

```
In [18]: #Visualise the data using Pandas Histogram
         incidents_data.hist(rwidth=0.9)
         plt.tight_layout()
```



12. Οπτικοποίηση δεδομένων με ιστογράμματα Pandas

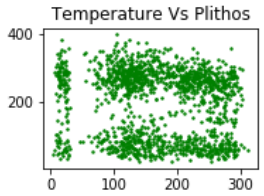
Η παραπάνω απεικόνιση δείχνει την μεταβλητότητα των παραγόντων-στηλών των δεδομένων επιλέγοντας το μήκος των απεικονιζόντων δεδομένων να είναι το 90% του συνόλου των δεδομένων που έχουν εισαχθεί στο data set που εξετάζεται. Στον άξονα x εμφανίζεται το ενδεικτικό δείγμα που έχει ληφθεί για τα ιστογράμματα από κάθε στήλη που εξετάζεται, ενώ στον δείκτη y παρουσιάζεται η διακύμανση που εμφανίζουν με βάση τις τιμές που λαμβάνει η εκάστοτε μεταβλητή. Έτσι στο ιστόγραμμα της αργίας ο άξονας y σημειώνει τιμές μόνο στο 0 και στο 1 όπως οι τιμές που λαμβάνει αν είναι αργία ή όχι. Αντίστοιχα στο ιστόγραμμα του μήνα, βλέπουμε στον άξονα y 12 στήλες του ιστογράμματος όσες και οι τιμές της μεταβλητής δηλαδή οι 12 μήνες του έτους.

Από τα παραπάνω ιστογράμματα παρατηρείται ότι η θερμοκρασία και το πλήθος ακολουθούν την κανονική κατανομή λόγω ότι είναι ευδιάκριτη μια εμφανής συμμετρία της κατανομής γύρω από συγκεκριμένες τιμές που παρουσιάζουν τη μεγαλύτερη συχνότητα, στην θερμοκρασία κατά το κέντρο του γραφήματος και στο πλήθος σε δύο ευδιάκριτα σημεία του γραφήματος, όπως επίσης παρατηρείται ότι η μόνη συνεχής μεταβλητή των δεδομένων που χρησιμοποιούνται είναι η θερμοκρασία. Συνεχής είναι η μεταβλητή που μπορεί να πάρει οποιαδήποτε τιμή σε κάποιο διάστημα.

Στην συνέχεια αφού έχει εντοπιστεί η συνεχής μεταβλητή γίνεται απεικόνιση της σε σχέση με το δεδομένο που έχει επιλεχθεί να γίνει η πρόβλεψη, στην συγκεκριμένη περίπτωση το πλήθος των εισαγωγών. Στο παρακάτω γράφημα απεικονίζεται αυτή η σχέση με την Θερμοκρασία (Temperature) στον άξονα x και το πλήθος των εισαγωγών (Plithos) στον άξονα y. Επόμενο είναι να αλλάξουμε το μέγεθος των σημείων μας στο διάγραμμα διασποράς. Πρώτα ορίζονται όλα στο ίδιο μέγεθος, αλλά μικρότερα. Αυτό γίνεται περνώντας ένα scaler (μονή τιμή) στην παράμετρο "s=" Στο γράφημα διασποράς που παρουσιάζεται έχει οριστεί ως scaler, s=2 για να λάβουμε ένα δείγμα των τιμών καθώς και ως χρώμα του γραφήματος color, c=g για το πράσινο χρώμα του γραφήματος όπως φαίνεται παρακάτω (green). Το γράφημα διασποράς παρουσιάζει πως κατανέμεται η μέση τιμή της θερμοκρασίας στην χώρα με τα αντίστοιχα πλήθη εισαγωγών των δεδομένων.

```
In [23]: #Data Visualisation
         #Visualise the continuous feature Vs Plithos
         plt.subplot(2,2,1)
         plt.title('Temperature Vs Plithos')
         plt.scatter(incidents_data['Temperature'], incidents_data['Plithos'], s=2, c='g')
```

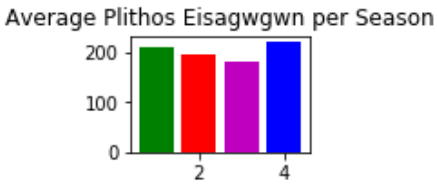
```
Out[23]: <matplotlib.collections.PathCollection at 0x14932595080>
```



13. Οπτικοποιώντας συνεχή μεταβλητή

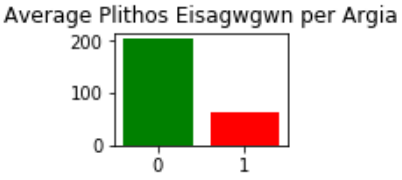
Οι κατηγορικές μεταβλητές των δεδομένων είναι η αργία, η ημέρα της εβδομάδας, ο μήνας, η εποχή του χρόνου και το Σαββατοκύριακο. Παρακάτω γίνεται αντίστοιχα η απεικόνιση των μεταβλητών αυτών σε σχέση με το πλήθος των εισαγωγών και πάλι. Σε κάθε γράφημα από τα παρακάτω ο άξονας x απεικονίζει την μέση τιμή του πλήθους των εισαγωγών και ο άξονας y την εκάστοτε μεταβλητή που λαμβάνει χρώμα με βάση τα χρώματα που έχουν οριστεί στον κώδικά για κάθε τιμή που λαμβάνουν οι στήλες του γραφήματος.

```
In [37]: #Plot the categorical features Vs Plithos
         #Create a 3x3 subplot
         plt.subplot(3,3,1)
         plt.title('Average Plithos Eisagwgwn per Season')
         #1.Create a list of unique season's values
         cat_list= incidents_data['Season'].unique()
         #2.Create average plithos per season by using Group by
         cat_average= incidents_data.groupby('Season').mean()['Plithos']
         colours= ('g', 'r','m','b')
         plt.bar(cat_list,cat_average,color=colours)

Out[37]: <BarContainer object of 4 artists>
```



14. Σχεδιασμός κατηγορικών μεταβλητών 1

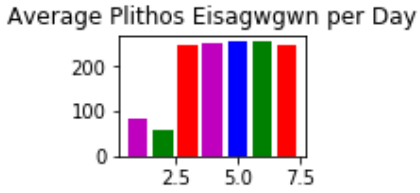
```
In [34]: #Create a 3x3 subplot
         plt.subplot(3,3,2)
         plt.title('Average Plithos Eisagwgwn per Argia')
         #1.Create a list of unique Argia's values
         cat_list= incidents_data['Argia'].unique()
         #2.Create average plithos per argia by using Group by
         cat_average= incidents_data.groupby('Argia').mean()['Plithos']
         colours= ('g', 'r')
         plt.bar(cat_list,cat_average,color=colours)

Out[34]: <BarContainer object of 2 artists>
```



15. Σχεδιασμός κατηγορικών μεταβλητών 2

```
In [35]:
#Create a 3x3 subplot
plt.subplot(3,3,3)
plt.title('Average Plithos Eisagwgwn per Day')
#1.Create a list of unique day's values
cat_list= incidents_data['Day'].unique()
#2.Create average plithos per day by using Group by
cat_average= incidents_data.groupby('Day').mean()['Plithos']
colours= ('g', 'r','m','b')
plt.bar(cat_list,cat_average,color=colours)
```

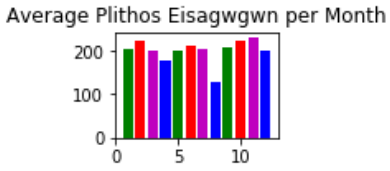
```
Out[35]: <BarContainer object of 7 artists>
```



16. Σχεδιασμός κατηγορικών μεταβλητών 3

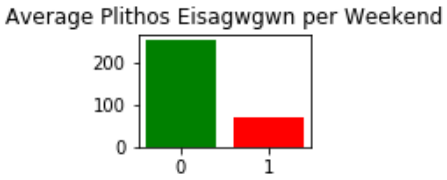
```
In [36]:
#Create a 3x3 subplot
plt.subplot(3,3,4)
plt.title('Average Plithos Eisagwgwn per Month')
#1.Create a list of unique Month's values
cat_list= incidents_data['Month'].unique()
#2.Create average plithos per Month by using Group by
cat_average= incidents_data.groupby('Month').mean()['Plithos']
colours= ('g', 'r','m','b')
plt.bar(cat_list,cat_average,color=colours)
```

```
Out[36]: <BarContainer object of 12 artists>
```



17. Σχεδιασμός κατηγορικών μεταβλητών 4

```
In [38]: #Create a 3x3 subplot
         plt.subplot(3,3,5)
         plt.title('Average Plithos Eisagwgwn per Weekend')
         #1.Create a list of unique Weekend's values
         cat_list= incidents_data['Weekend'].unique()
         #2.Create average plithos per Weekend by using Group by
         cat_average= incidents_data.groupby('Weekend').mean()['Plithos']
         colours= ('g', 'r','m','b')
         plt.bar(cat_list,cat_average,color=colours)

Out[38]: <BarContainer object of 2 artists>
```



18. Σχεδιασμός κατηγορικών μεταβλητών 5

Το συμπέρασμα που προκύπτει από τα διαγράμματα που παρήχθησαν και που μπορεί να είναι ένα δεδομένο σημαντικό για τις αποφάσεις που λαμβάνονται κατά την διαδικασία αυτή και τα επόμενα βήματα είναι ότι θα μπορούσε να εξαιρεθεί από την διαδικασία ο παράγοντας Εποχή καθώς παρατηρείται ότι δεν παρουσιάζει το πλήθος των εισαγωγών μεγάλη διαφοροποίηση από εποχή σε εποχή του χρόνου. Άρα θα μπορούσε να θεωρηθεί ένας παράγοντας μη σημαντικός για την μελέτη του. Παρόλα αυτά πάρθηκε η απόφαση να συνεχίσει η διαδικασία με το σύνολο των μεταβλητών. Το συγκεκριμένο γράφημα της Εποχής δείχνει ότι υπάρχει μια αύξηση των εισαγωγών το Φθινόπωρο σε σχέση με τις υπόλοιπες εποχές παρόλα αυτά δεν είναι σημαντική, ακολουθεί ο Χειμώνας και στην συνέχεια υπάρχει πτωτική πορεία των εισαγωγών πρώτα την Άνοιξη και έπειτα το Καλοκαίρι με την μεγαλύτερη πτώση των εισαγωγών ασθενών. Συζητώντας την παραπάνω διαπίστωση που προκύπτει από το γράφημα με επαγγελματίες του χώρου υγείας θεωρήθηκε λογικό αφού οι περισσότερες προγραμματισμένες εισαγωγές αποφεύγονται το καλοκαίρι που είναι περίοδος διακοπών και ξεκούρασης για το μεγαλύτερο ποσοστό πληθυσμού στην Ελλάδα. Αντιθέτως το φθινόπωρο είναι η περίοδος που προτιμάτε μετά τις καλοκαιρινές διακοπές ως περίοδος προγραμματισμού οποιασδήποτε απαιτούμενης επέμβασης. Επιπλέον το φθινόπωρο και ο χειμώνας υπάρχει έξαρση των ιώσεων κάτι που αυξάνει και τις εισαγωγές στα νοσοκομεία και τα έκτακτα περιστατικά που προκύπτουν.

Στο επόμενο γράφημα που αφορά τον παράγοντα Αργία σε σχέση με το πλήθος εισαγωγών, παρατηρείται μεγάλη διαφοροποίηση τις ημέρες που δεν είναι κάποια επίσημη αργία σε σχέση με τις ημέρες που ήταν κάποια αργία. Οι εισαγωγές μειώνονται αισθητά κατά τις ημέρες αργιών, γεγονός δικαιολογημένο καθώς οι ασθενείς αλλά και οι θεράποντες ιατροί αποφεύγουν τον προγραμματισμό εισαγωγής εκείνες τις ημέρες, εκτός αν αφορά κάποιο έκτακτο περιστατικό που δεν μπορεί να αποφευχθεί η εισαγωγή. Άρα ο συγκεκριμένος παράγοντας επηρεάζει κατά πολύ την διαμόρφωση του πλήθους εισαγωγών κατά την διάρκεια του έτους. Επίσης στην Ελλάδα οι επίσημες αργίες είναι κατά βάση σταθερές και επαναλαμβανόμενες κάτι που βοηθάει στην πρόβλεψη και λήψη αντίστοιχων αποφάσεων με βάση αυτό τον παράγοντα.

Στο γράφημα που αφορά την ημέρα της εβδομάδας, υπάρχει μια αισθητή διαφοροποίηση κατά το Σαββατοκύριακο αλλά τις καθημερινές οι εισαγωγές παραμένουν σε παρόμοια επίπεδα με την Τετάρτη και την Πέμπτη στα υψηλότερα επίπεδα και την Κυριακή στο χαμηλότερο. Το ίδιο συμπέρασμα προκύπτει και με το γράφημα για το Σαββατοκύριακο που φαίνεται και πιο συγκεκριμένα η μεγάλη διαφορά εισαγωγών από καθημερινή σε Σαββατοκύριακο. Και τα συγκεκριμένα αποτελέσματα δικαιολογούνται από τους επαγγελματίες του χώρου καθώς και πάλι οι προγραμματισμένες εισαγωγές κατά το Σαββατοκύριακο αποφεύγονται κατά βάση και υπάρχει προτίμηση στο να προγραμματίζεται μια εισαγωγή κατά τις υπόλοιπές ημέρες τις εβδομάδας, γεγονός που βασίζεται κυρίως στην προτίμηση αυτή των θεραπόντων ιατρών να μην επιβαρύνουν το πρόγραμμα του Σαββατοκύριακου με προγραμματισμό εισαγωγών των ασθενών τους που δεν αφορά επείγον περιστατικό.

Το γράφημα της μεταβλητής Μήνας παρουσιάζει παρόμοια εικόνα με το γράφημα της Εποχής καθώς επιβεβαιώνει τα συμπεράσματα που είχαν προκύψει εκεί αλλά πιο αναλυτικά με τον Αύγουστο να είναι ο μήνας με τις λιγότερες εισαγωγές και ο Νοέμβριο με τις περισσότερες. Η συγκεκριμένη απεικόνιση είναι πολύ χρήσιμη για λήψη σχετικών αποφάσεων προγραμματισμού καθώς μπορούν να εντοπιστούν πιο ξεκάθαρα οι ανάγκες μηνιαίως σε σχέση μόνο με την εποχή του χρόνου ή την εβδομαδιαία.

Συνεχίζοντας την μελέτη των δεδομένων, υπάρχουν ορισμένα πράγματα τα οποία, εάν δεν γίνουν στη φάση αυτή, μπορούν να επηρεάσουν την περαιτέρω μοντελοποίηση

στατιστικής/Μηχανικής Μάθησης. Ένα από αυτά είναι η εύρεση των “outliers”, δηλαδή των ακραίων τιμών. Στην στατιστική, το ακραίο σημείο είναι ένα σημείο παρατήρησης που απέχει από άλλες παρατηρήσεις. Ο παραπάνω ορισμός υποδηλώνει ότι το outlier είναι κάτι που είναι ξεχωριστό/διαφορετικό από το πλήθος. Ωστόσο, δεν είναι καθόλου φανερά τα ακραία στοιχεία στη φάση της συλλογής. Οι ακραίες τιμές μπορεί να είναι αποτέλεσμα λάθους κατά τη συλλογή δεδομένων ή μπορεί να είναι απλώς μια ένδειξη διακύμανσης των δεδομένων. Γνωρίζοντας ότι οι ακραίες τιμές μπορεί να είναι είτε λάθος είτε απλώς διακύμανση, θα πρέπει να αποφασιστεί εάν είναι σημαντικά ή όχι. Αν είναι αποτέλεσμα λάθους, τότε μπορούν να αγνοηθούν, αλλά αν είναι απλώς μια απόκλιση στα δεδομένα, θα πρέπει να μελετηθούν λίγο περισσότερο. Πριν γίνει προσπάθεια απόφασης αν πρέπει να αγνοηθούν τα ακραία στοιχεία ή όχι, πρέπει να αναγνωριστούν όπως γίνεται παρακάτω:

```
In [39]: #Check for outliers

         incidents_data['Plithos'].describe()

Out[39]: count    1461.000000
         mean      200.793977
         std        99.373259
         min        23.000000
         25%        88.000000
         50%       242.000000
         75%       281.000000
         max       399.000000
         Name: Plithos, dtype: float64
```

```
In [44]: incidents_data['Plithos'].quantile([0.05,0.1,0.15,0.9,0.95,0.99])

Out[44]: 0.05     47.0
         0.10     58.0
         0.15     66.0
         0.90    307.0
         0.95    325.0
         0.99    356.4
         Name: Plithos, dtype: float64
```

19. Έλεγχος ακραίων τιμών

Στην συνέχεια χρησιμοποιήθηκε η μέθοδος “quantile” που αναφέρεται σε έναν αριθμό όπου ορισμένα ποσοστά ορίζονται και για αυτά τα ποσοστά παράγονται οι τιμές που λαμβάνει το πλήθος των εισαγωγών, δεδομένο χρήσιμο για την εκπαίδευση του μοντέλου.

Στο επόμενο βήμα ελέγχεται η Πολλαπλή Γραμμική Παλινδρόμηση (Multiple Linear Regression) Θερμοκρασίας-Πλήθους εισαγωγών. Επιλέγεται η πολλαπλή αντί για την απλή λόγω των πολλών μεταβλητών του data set. Οι συντελεστές συσχέτισης ποσοτικοποιούν τη συσχέτιση μεταξύ μεταβλητών ή χαρακτηριστικών ενός συνόλου δεδομένων. Αυτά τα στατιστικά στοιχεία έχουν μεγάλη σημασία και η Python διαθέτει εξαιρετικά εργαλεία που μπορούν να χρησιμοποιηθούν για να τα υπολογιστούν. Οι μέθοδοι συσχέτισης SciPy, NumPy και Pandas είναι γρήγορες, περιεκτικές και καλά τεκμηριωμένες. Όταν αναλύεται τη συσχέτιση, θα πρέπει πάντα να υπάρχει κατά νου η σκέψη ότι η συσχέτιση δεν υποδηλώνει αιτιότητα. Ποσοτικοποιεί την ισχύ της σχέσης μεταξύ των χαρακτηριστικών ενός συνόλου δεδομένων. Μερικές φορές, η συσχέτιση προκαλείται από έναν παράγοντα κοινό σε πολλά χαρακτηριστικά ενδιαφέροντος. Η συσχέτιση είναι στενά συνδεδεμένη με άλλα στατιστικά μεγέθη όπως ο μέσος όρος, η τυπική απόκλιση, η διακύμανση και η συν διακύμανση. Ο προκύπτων πίνακας συσχέτισης είναι μια νέα παρουσία του DataFrame και διατηρεί τους συντελεστές συσχέτισης για τις στήλες Θερμοκρασία και Πλήθος εισαγωγών. Τέτοια αποτελέσματα με ετικέτα είναι συνήθως πολύ βολικά για εργασία, επειδή μπορείτε να υπάρχει πρόσβαση σε αυτά είτε με τις ετικέτες τους είτε με τους δείκτες ακέραιων θέσεων. Μια μεγαλύτερη απόλυτη τιμή του r δείχνει ισχυρότερη συσχέτιση, πιο κοντά σε μια γραμμική συνάρτηση. Μια μικρότερη απόλυτη τιμή του r δείχνει ασθενέστερη συσχέτιση.

```
In [47]: #Check Multiple Linear Regression Assumptions
         #Linearity using correlation coefficient matrix using corr

         correlation = incidents_data[['Temperature','Plithos']].corr()

         print(correlation)
```

```
             Temperature   Plithos
Temperature     1.000000 -0.102104
Plithos        -0.102104  1.000000
```

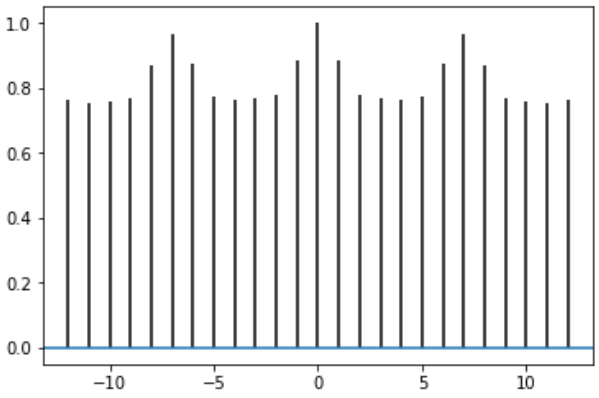
20. Πολλαπλή Γραμμική Παλινδρόμηση

Η Matplotlib είναι μια βιβλιοθήκη της python που παρέχει δυνατότητες σχεδίασης γραφικών παραστάσεων υψηλής ποιότητας. Η pyplot είναι ένα τμήμα της Matplotlib που αποτελεί μια συλλογή εντολών με λειτουργίες παρόμοιες με αυτές του Matlab. Με

την εντολή import matplotlib.pyplot as plt γίνεται η εγκατάσταση της βιβλιοθήκης για να μπορέσει να γίνει χρήση της. Υπάρχει μια τελευταία διαμόρφωση για να ολοκληρωθεί πριν εμφανιστούν plots. Πρέπει να επισημανθεί στο jupyter να τα εμφανίσει ως εικόνες στο ίδιο το notebook. Από την βιβλιοθήκη λοιπόν αυτή γίνεται ο έλεγχος και αντίστοιχή απεικόνιση για το Autocorrelation των δεδομένων.

```
In [49]: #Check the autocorrelation in Plithos using acorr

         df1= pd.to_numeric(incidents_data['Plithos'],downcast='float')
         plt.acorr(df1,maxlags=12)
```

```
Out[49]: (array([-12, -11, -10,  -9,  -8,  -7,  -6,  -5,  -4,  -3,  -2,  -1,   0,
                   1,   2,   3,   4,   5,   6,   7,   8,   9,  10,  11,  12]),
          array([0.76474565, 0.7555082 , 0.75622135, 0.7665484 , 0.8690126 ,
                 0.9662519 , 0.8734938 , 0.7753594 , 0.7653677 , 0.7661951 ,
                 0.7790372 , 0.8827577 , 1.        , 0.8827577 , 0.7790372 ,
                 0.7661951 , 0.7653677 , 0.7753594 , 0.8734938 , 0.9662519 ,
                 0.8690126 , 0.7665484 , 0.75622135, 0.7555082 , 0.76474565],
                dtype=float32),
          <matplotlib.collections.LineCollection at 0x149326054a8>,
          <matplotlib.lines.Line2D at 0x149325fd390>)
```



21. Έλεγχος αυτοσυσχέτισης

Η αυτοσυσχέτιση (ACF) είναι μια υπολογισμένη τιμή που χρησιμοποιείται για να αναπαραστήσει πόσο παρόμοια είναι μια τιμή σε μια χρονοσειρά με μια προηγούμενη τιμή. Η βιβλιοθήκη κάνει τον υπολογισμό της αυτοσυσχέτισης στην Python πολύ απλοποιημένο. Με μερικές γραμμές κώδικα, μπορεί κανείς να αντλήσει χρήσιμες πληροφορίες σχετικά με τις παρατηρούμενες τιμές σε δεδομένα χρονοσειρών. Η ACF μπορεί να χρησιμοποιηθεί για τον προσδιορισμό των τάσεων στα δεδομένα και την επιρροή των τιμών που είχαν παρατηρηθεί προηγουμένως σε μια τρέχουσα παρατήρηση. Η αυτοσυσχέτιση, που ονομάζεται επίσης σειριακή συσχέτιση, χρησιμοποιείται από εμπόρους μετοχών, μετεωρολόγους, χημικούς και πολλά άλλα για την πρόβλεψη μελλοντικών τιμών δεδομένων των ιστορικών δεδομένων χρονοσειρών.

Αυτός ο τύπος παλινδρομικής ανάλυσης χρησιμοποιείται για να βοηθήσει στην πρόβλεψη μελλοντικών τιμών εντός ενός διαστήματος εμπιστοσύνης (συνήθως 95%) και συσχετίζει μια τρέχουσα τιμή με προηγούμενες. Η αυτοσυσχέτιση εκτιμά την επίδραση όλων των προηγούμενων παρατηρούμενων τιμών στην τρέχουσα παρατηρούμενη τιμή. Αυτό διαφέρει από τη μερική αυτοσυσχέτιση στην οποία μετριέται μόνο μία προηγούμενη παρατηρούμενη τιμή για επιρροή στην τρέχουσα παρατηρούμενη τιμή.

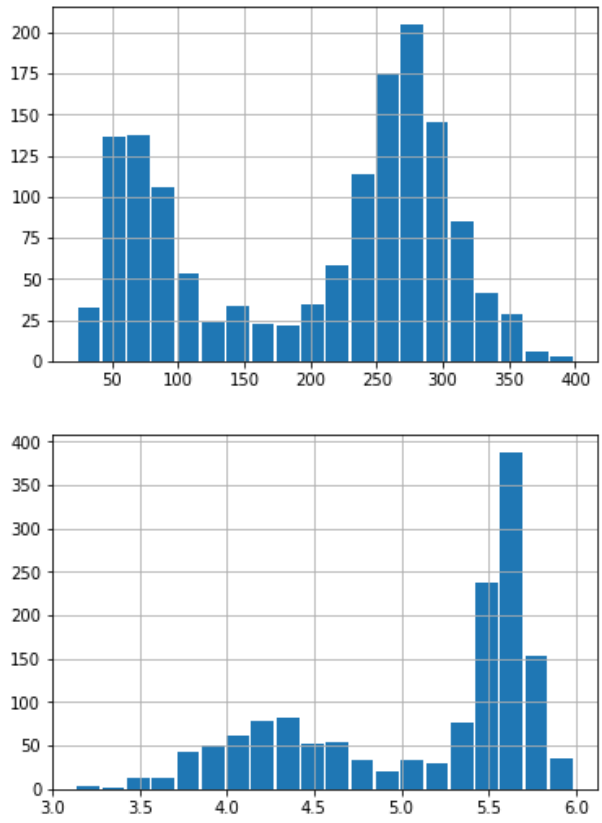
Το διάγραμμα που προκύπτει φαίνεται αρκετά ενδιαφέρον, αλλά τελικά δεν προσφέρει μεγάλη χρησιμότητα χωρίς να μπορεί να αναλυθεί. Οι κάθετες γραμμές με δείκτες στην κορυφή τους είναι οι «υστερήσεις» που αντιπροσωπεύουν έναν συγκεκριμένο αριθμό προηγούμενων τιμών. Αυτά αντιπροσωπεύουν την τιμή συσχέτισης (που εμφανίζεται στον άξονα y) και αυξομειώνονται με σταθερό ρυθμό καθώς αυξάνεται η εγγύτητά τους από την τρέχουσα τιμή. Αυτό μας ενημερώνει ότι οι προηγούμενες τιμές επηρεάζουν την τρέχουσα τιμή με σταθερό ρυθμό, αλλά η σημασία αυτής της επιρροής αυξομειώνεται σταθερά με το χρόνο. Η ισχύς αυτής της σχέσης μετριέται σε μια κλίμακα από 0 έως 10 όταν το 10 είναι 100% θετική συσχέτιση. Αυτό το μέτρο εμφανίζεται στον άξονα y.

Ένα σημαντικό σημείο απόφασης κατά την εργασία με ένα δείγμα δεδομένων είναι εάν θα χρησιμοποιηθούν παραμετρικές ή μη παραμετρικές στατιστικές μέθοδοι. Οι παραμετρικές στατιστικές μέθοδοι υποθέτουν ότι τα δεδομένα έχουν μια γνωστή και συγκεκριμένη κατανομή, συχνά μια κατανομή Gauss. Εάν ένα δείγμα δεδομένων δεν είναι Gaussian, τότε παραβιάζονται οι παραδοχές των παραμετρικών στατιστικών δοκιμών και πρέπει να χρησιμοποιηθούν μη παραμετρικές στατιστικές μέθοδοι. Υπάρχει μια σειρά τεχνικών που μπορείτε να χρησιμοποιήσετε για να ελέγξετε εάν το δείγμα δεδομένων σας αποκλίνει από μια κατανομή Gauss, που ονομάζονται τεστ κανονικότητας. Σε αυτό το σημείο ελέγχεται η κανονικότητα, δηλαδή εάν το δείγμα δεδομένων αποκλίνει από την κανονική κατανομή. Ένα απλό και ευρέως χρησιμοποιούμενο διάγραμμα για τον γρήγορο έλεγχο της κατανομής ενός δείγματος δεδομένων είναι το ιστόγραμμα. Στο ιστόγραμμα, τα δεδομένα χωρίζονται σε έναν προκαθορισμένο αριθμό ομάδων που ονομάζονται bins. Στη συνέχεια, τα δεδομένα ταξινομούνται σε κάθε bin και διατηρείται η καταμέτρηση του αριθμού των παρατηρήσεων σε κάθε bin. Η γραφική παράσταση δείχνει τα bins κατά μήκος του

άξονα x διατηρώντας τη σειρά τους και την καταμέτρηση σε κάθε bin στον άξονα y. Ένα δείγμα δεδομένων έχει μια κατανομή Gauss της γραφικής παράστασης ιστογράμματος, που δείχνει το γνωστό σχήμα καμπάνας. Από προεπιλογή, ο αριθμός των δοχείων υπολογίζεται αυτόματα από το δείγμα δεδομένων. Το αποτέλεσμα που δείχνει την γραφική παράσταση ιστογράμματος στα δεδομένα του πλήθους εισαγωγών παρατίθεται παρακάτω. Μπορούμε να δούμε ένα σχήμα Gaussian στα δεδομένα, που αν και δεν είναι έντονα το γνωστό σχήμα καμπάνας, είναι κατά προσέγγιση. Άρα δεν χρειάζεται να διερευνηθεί γιατί τα δεδομένα δεν είναι κανονικά και να χρησιμοποιηθεί το δεύτερο βήμα που απεικονίζεται για τις τεχνικές προετοιμασίας δεδομένων για να γίνουν τα δεδομένα πιο κανονικά

```
In [51]: #Normality of Plithos

         df1= incidents_data['Plithos']
         df2= np.log(df1)

         plt.figure()
         df1.hist(rwidth=0.9,bins=20)

         plt.figure()
         df2.hist(rwidth=0.9,bins=20)
```

```
Out[51]: <matplotlib.axes._subplots.AxesSubplot at 0x14932434358>
```



22. Έλεγχος κανονικότητας

Προχωρώντας στο επόμενο βήμα, η διαδικασία διαχωρισμού train & test data set χρησιμοποιείται για την εκτίμηση της απόδοσης των αλγορίθμων μηχανικής μάθησης όταν χρησιμοποιούνται για να κάνουν προβλέψεις σε δεδομένα. Είναι μια γρήγορη και εύκολη διαδικασία στην εκτέλεση, τα αποτελέσματα της οποίας επιτρέπουν την σύγκριση της απόδοσης των αλγορίθμων μηχανικής μάθησης για το πρόβλημά της προγνωστικής μοντελοποίησης. Αν και είναι απλή στη χρήση και την ερμηνεία, υπάρχουν φορές που η διαδικασία δεν πρέπει να χρησιμοποιείται, όπως όταν ένα σύνολο δεδομένων είναι πολύ μικρό και περιπτώσεις όπου απαιτείται πρόσθετη διαμόρφωση, όπως όταν χρησιμοποιείται ταξινόμηση και το σύνολο δεδομένων δεν

είναι ισορροπημένο. Η διαδικασία περιλαμβάνει τη λήψη ενός συνόλου δεδομένων και τη διαίρεση του σε δύο υποσύνολα. Το πρώτο υποσύνολο χρησιμοποιείται για να ταιριάζει στο μοντέλο και αναφέρεται ως το σύνολο δεδομένων εκπαίδευσης. Το δεύτερο υποσύνολο δεν χρησιμοποιείται για την εκπαίδευση του μοντέλου. Αντίθετα, το στοιχείο εισόδου του συνόλου δεδομένων παρέχεται στο μοντέλο, στη συνέχεια γίνονται προβλέψεις και συγκρίνονται με τις αναμενόμενες τιμές. Αυτό το δεύτερο σύνολο δεδομένων αναφέρεται ως το σύνολο δεδομένων δοκιμής (test data set).

```
In [101]: #Create train & test split
          #Splitting Y & X dataset into training & testing set

          #Plithos is time dependent or time series

          Y= incidents_data [['Plithos']]
          X= incidents_data.drop(['Plithos'],axis=1)

          print (Y)
          print (X)
```

```
      Plithos
0          70
1         141
2         228
3         210
4         189
5          49
6          61
7         220
8         245
9         229
10        253
11        250
12         82
13         59
14        247
15        261
16        275
17        250
```

23. Δημιουργία train & test set 1

[1461 rows x 1 columns]

	Month	Day	Season	Argia	Weekend	Temperature
0	1	2	1	0	0	143.054
1	1	3	1	0	0	125.816
2	1	4	1	0	0	91.848
3	1	5	1	0	0	73.128
4	1	6	1	0	0	73.152
5	1	7	1	0	1	93.149
6	1	1	1	0	1	132.505
7	1	2	1	0	0	115.599
8	1	3	1	0	0	75.031
9	1	4	1	0	0	100.793
10	1	5	1	0	0	111.343
11	1	6	1	0	0	106.836
12	1	7	1	0	1	113.807
13	1	1	1	0	1	106.818
14	1	2	1	0	0	109.814
15	1	3	1	0	0	117.271

24. Δημιουργία train & test set 2

Η διαδικασία έχει μία κύρια παράμετρο διαμόρφωσης, η οποία είναι το μέγεθος των δύο αυτών data sets. Αυτό εκφράζεται πιο συχνά ως ποσοστό μεταξύ 0 και 1. Δεν υπάρχει βέλτιστο ποσοστό διαίρεσης. Πρέπει να επιλεχθεί ένα διαιρεμένο ποσοστό που να ανταποκρίνεται στους στόχους του έργου με κριτήριο που περιλαμβάνει την αντιπροσωπευτικότητα των συνόλων δεδομένων. Στην συγκεκριμένη περίπτωση επιλέχθηκε: Train: 70%, Test: 30%.

```
In [102]: #Create the size for the 70% of the data
          tr_size= 0.7 * len(X)

In [103]: tr_size = int(tr_size)

In [104]: X_train = X.values[0: tr_size]
          X_test = X.values[tr_size: len(X)]

          Y_train = Y.values[0:tr_size]
          Y_test = Y.values[tr_size : len(Y)]
```

25. Δημιουργία train & test set 3

Μια άλλη σημαντική παράμετρος είναι ότι οι σειρές που αντιστοιχίζονται στα train & test data sets τυχαία. Αυτό γίνεται για να διασφαλιστεί ότι τα σύνολα δεδομένων είναι ένα αντιπροσωπευτικό δείγμα (π.χ. τυχαίο δείγμα) του αρχικού συνόλου δεδομένων, το οποίο με τη σειρά του θα πρέπει να είναι ένα αντιπροσωπευτικό δείγμα παρατηρήσεων από τον τομέα του προβλήματος.

Πλέον μπορεί να γίνει αξιολόγηση μοντέλου χρησιμοποιώντας τον διαχωρισμό train & test. Πρώτον, το uploaded σύνολο δεδομένων πρέπει να χωριστεί σε στοιχεία εισόδου και εξόδου. Στη συνέχεια, μπορούμε να ορίσουμε και να προσαρμόσουμε το μοντέλο στο σύνολο δεδομένων εκπαίδευσης. Έπειτα χρησιμοποιείται το μοντέλο προσαρμογής για να γίνουν προβλέψεις και να αξιολογηθούν οι προβλέψεις χρησιμοποιώντας τη μέτρηση απόδοσης ακρίβειας ταξινόμησης. Έχοντας έτοιμα τα σετ, η Scikit-Learn έχει μια πληθώρα τύπων μοντέλων που μπορούν εύκολα να εισαχθούν και να εκπαιδευτούν, με τη LinearRegression να είναι ένας από αυτούς, όπως φαίνεται και στην χρήση της παρακάτω. Ένας πολύ εύχρηστος τρόπος για να προβλέψουμε νέες τιμές χρησιμοποιώντας το μοντέλο είναι να καλέσουμε τη συνάρτηση predict(). Έτσι μπορούμε να προβλέψουμε χρησιμοποιώντας τα δεδομένα των δοκιμών και να συγκριθούν τα προβλεπόμενα με τα πραγματικά αποτελέσματα. Για να γίνουν προβλέψεις στα δεδομένα δοκιμής, περνούν οι τιμές X_test στη μέθοδο predict(). Μπορούν να αντιστοιχηθούν τα αποτελέσματα στη μεταβλητή y_predict. Η μεταβλητή y_predict περιέχει τώρα όλες τις προβλεπόμενες τιμές για τις τιμές εισόδου στο X_test. Μπορούμε τώρα να συγκρίνουμε τις πραγματικές τιμές εξόδου για το X_test με τις προβλεπόμενες τιμές, τοποθετώντας τες δίπλα-δίπλα σε μια δομή πλαισίου δεδομένων. Αν και το μοντέλο δεν φαίνεται να είναι πολύ ακριβές, τα προβλεπόμενα ποσοστά είναι κοντά στα πραγματικά.

```
In [107]: #Fit & score the model

          #Linear Regression

          from sklearn.linear_model import LinearRegression
          std_reg = LinearRegression()
          std_reg.fit(X_train, Y_train)

          r2_train= std_reg.score(X_train, Y_train)
          r2_test= std_reg.score(X_test, Y_test)



          print (Y_predict)
```

```
[[ 248.74825086]
 [ 248.49638063]
 [ 249.53436013]
 [ 273.67203446]
 [ 253.73891443]
 [  76.40708066]
 [  67.53086067]
 [ 250.01906175]
 [ 249.54235176]
 [  74.3560911 ]
 [ 253.12859673]
 [ 255.0808296 ]
 [  78.61144281]
 [  66.94709728]
 [ 250.24127636]
 [ 252.13859446]
 [ 252.57005059]
 [ 273.50298127]
 [ 275.21794085]
```

26. Έλεγχος μοντέλου και πρόβλεψη 1

Αφού εξεταστούν τα δεδομένα, δούμε μια γραμμική σχέση, εκπαιδευτεί και δοκιμαστεί το μοντέλο, μπορεί να γίνει αντιληπτό πόσο καλά προβλέπει χρησιμοποιώντας ορισμένες μετρήσεις. Για τα μοντέλα παλινδρόμησης, χρησιμοποιούνται συγκεκριμένες μετρήσεις αξιολόγησης, δύο από αυτές χρησιμοποιούνται και εδώ. Το μοντέλο αξιολογείται στο σύνολο δοκιμών χρησιμοποιώντας το σφάλμα ριζικού μέσου τετραγώνου (RMSE) αρχικά. Δείχνει πόσο μπορεί να διαφέρουν τα δεδομένα, επομένως, έχοντας RMSE 53,02 το μοντέλο μας μπορεί να κάνει σφάλμα είτε επειδή πρόσθεσε 53,02 στην πραγματική τιμή είτε χρειάστηκε 53,02 για να φτάσει στην πραγματική τιμή. Όταν γίνεται επιλογή μεταξύ των μοντέλων, αυτά με τα μικρότερα σφάλματα, συνήθως αποδίδουν καλύτερα. Κατά την παρακολούθηση μοντέλων, εάν οι μετρήσεις χειροτέρευαν, τότε μια προηγούμενη έκδοση του μοντέλου ήταν καλύτερη ή υπήρξε κάποια σημαντική αλλαγή στα δεδομένα ώστε το μοντέλο να έχει χειρότερη απόδοση από αυτή που είχε.

```
In [112]:

#Create Y predictions

from sklearn.metrics import mean_squared_error
rmse = math.sqrt(mean_squared_error(Y_test,Y_predict))

print (rmse)
```

```
53.02332250969289
```

27. Έλεγχος μοντέλου και πρόβλεψη 2

Το ριζικό μέσο τετράγωνο σφάλμα (RMSE) και το ριζικό μέσο τετράγωνο λογαριθμικό σφάλμα (RMSLE) είναι και οι δύο τεχνικές για να βρεθεί η διαφορά μεταξύ των τιμών που προβλέπονται από το μοντέλο μηχανικής εκμάθησης και των πραγματικών τιμών. Το RMSE έχει δραστική επίδραση των ακραίων τιμών στις τιμές του. Αλλά στην περίπτωση του RMSLE μπορεί να μειωθεί η επίδραση των ακραίων τιμών κατά μεγάλο ποσοστό. Η τιμή RMSLE θα εξετάσει μόνο το σχετικό σφάλμα μεταξύ της προβλεπόμενης και της πραγματικής τιμής παραβλέποντας την κλίμακα των δεδομένων. Αλλά η τιμή RMSE θα αυξηθεί σε μέγεθος εάν η κλίμακα του σφάλματος αυξηθεί. Επίσης σε περίπτωση υποεκτίμησης τα αποτελέσματα από το RMSLE επηρεάζονται σε μεγάλο βαθμό. Έτσι μπορεί κανείς εύκολα να καταλάβει ότι είναι καλύτερο το RMSE σε ορισμένα σενάρια, αλλά το RMSE λειτουργεί καλύτερα για γενικευμένες περιπτώσεις.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

$$\text{RMSLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log y_i - \log \hat{y}_i)^2},$$

28. RMSE & RMSLE σφάλματα

In [114]:
```
#Final step
# Calculate RMSLE & Compare results

Y_test_e = []
Y_predict_e = []
```

In [119]:
```
#
for i in range (0, len(Y_test)):
 Y_test_e.append(math.exp(Y_test[i]))
 Y_predict_e.append(math.exp(Y_predict[i]))
```

In [120]:
```
#Do the sum of the logs and the squares
log_sq_sum = 0.0

for i in range (0, len (Y_test_e)):
    log_a = math.log(Y_test_e[i] +1)
    log_p = math.log(Y_predict_e[i] +1)

    log_diff = (log_p - log_a)**2
    log_sq_sum = log_sq_sum + log_diff

rmsle = math.sqrt(log_sq_sum/ len(Y_test) )

print ("")
print (rmsle)
```

```
70.778657525986
```

29. Έλεγχος μοντέλου και πρόβλεψη 3

Ολοκληρώνοντας την πειραματική διαδικασία και την εφαρμογή του μοντέλο, πρέπει να αναφερθεί ότι μπορεί πάντα να ελέγχεται εάν συμπεριλαμβάνεται ή αν πρέπει να συμπεριλαμβάνεται μια συγκεκριμένη μεταβλητή ή μπορεί να αφαιρεθεί μια ακραία τιμή ή να απορριφθεί από τα χαρακτηριστικά της μια επιλεγμένη μεταβλητή και γενικότερα να ελέγχεται τι μπορεί να κάνει τη διαφορά ή όχι στο μοντέλο.

5 Συμπεράσματα

Στόχος της παρούσας διπλωματικής εργασίας ήταν η χρήση μοντέλων ανάλυση δεδομένων για την πρόβλεψη μελλοντικών δεδομένων απαραίτητων για την λήψη αποφάσεων στο χώρο της υγείας, με εφαρμογή σε οποιοδήποτε επίπεδο και υπηρεσίας φροντίδας υγείας. Παραδοσιακά, τέτοιες τεχνικές πρόβλεψης δε χρησιμοποιούνται ακόμη ευρέως από τον συγκεκριμένο τομέα παρόλο που θα μπορούσε να βελτιώσει κατά πολύ το παρεχόμενο επίπεδο υπηρεσιών αλλά και την ποιότητα και βιωσιμότητα στο επίπεδο των ίδιων των οργανισμών εσωτερικά . Επίσης έγινε μία προσπάθεια να αναδειχθούν οι βασικές κατηγορίες μηχανικής μάθησης και οι αλγόριθμοι που έχουν αναπτυχθεί καθώς και η σημαντικότητα των ποιοτικών προβλέψεων που παρέχουν. Η αποτελεσματική λήψη αποφάσεων με βάση τις προβλέψεις που παράγονται από τις μεθόδους που έχουν αναλυθεί, διαδραματίζουν σημαντικό ρόλο στη λειτουργία μιας επιχείρησης ακόμη και υγειονομικού ενδιαφέροντος και συνεπώς η σωστή παρακολούθησή τους είναι αναγκαία για την επιβίωση της, ιδιαίτερα στο σημερινό ανταγωνιστικό οικονομικό κλίμα, όπου υπάρχει όχι μόνο ανταγωνισμός μεταξύ τους στο επίπεδο εξυπηρέτησης και ποιότητας των προσφερόμενων υπηρεσιών, αλλά και στην προσπάθεια επιβίωσης σε ένα διαρκώς μεταβαλλόμενο και αβέβαιο περιβάλλον με τις εξελίξεις και τα νέα δεδομένα στον χώρο της υγείας να προκύπτουν συνεχώς. Στην εργασία αυτή έγινε μια προσπάθεια να αναδειχθούν αυτά τα θετικά αποτελέσματα που προκύπτουν από την ενσωμάτωση προηγμένων μεθόδων πρόβλεψης μέσω και του παραδείγματος που παρουσιάστηκε όπου δεδομένα που αφορούν τις εισαγωγές ασθενών εφαρμόστηκαν σε μοντέλο πρόβλεψης ώστε να αποτυπωθούν και στην πράξη τα παραπάνω ζητούμενα. Συγκεκριμένα για το παράδειγμα της πειραματικής διαδικασίας του προηγούμενου κεφαλαίου , η ανάλυση χρονολογικών σειρών φαίνεται πως μπορεί να ακολουθήσει την τάση της πορείας μίας τιμής και οι προβλέψεις της να είναι ικανοποιητικές λαμβάνοντας υπόψιν περαιτέρω παράγοντες που φαινομενικά θεωρούνται μη σχετικοί με την τιμή που θέλουμε να προβλέψουμε όπως θερμοκρασία ημέρας, εποχή. Επίσης η εξάρτηση της με τον παράγοντα «Αργία» και «Σαββατοκύριακο» φαίνεται να διαδραματίζουν σημαντικό ρόλο στην απόδοση του μοντέλου. Η ανάλυση χρονολογικών σειρών φαίνεται πως μπορεί να ακολουθήσει την τάση της πορείας μίας τιμής και οι προβλέψεις της να είναι ικανοποιητικές λαμβάνοντας υπόψιν περαιτέρω παράγοντες που φαινομενικά θεωρούνται μη σχετικοί με την τιμή που θέλουμε να προβλέψουμε όπως θερμοκρασία ημέρας, εποχή. Επίσης η

εξάρτηση της με τον παράγοντα «Αργία» και «Σαββατοκύριακο» φαίνεται να διαδραματίζουν σημαντικό ρόλο στην απόδοση του μοντέλου. Η εφαρμογή τεχνικών μηχανικής μάθησης σε δεδομένα χρονολογικών σειρών φαίνεται πως και με μη εξιδεικευμένα μοντέλα μπορεί να αποδώσει αρκετά καλά ακόμα και σε δεδομένα αρκετά πολύπλοκα όπως οι εισαγωγές ενός νοσοκομείου. Ο υπολογισμός και η οπτικοποίηση, τουλάχιστον των θεμελιωδών στατιστικών μεγεθών θα μπορούσε να υποστηρίξει την ανάπτυξη στρατηγικής για την διοίκηση ενός νοσοκομείου για παράδειγμα καθώς την βοηθάει να δει γραφικά την πορεία των εισαγωγών και των παραγόντων που την επηρεάζουν. Με την στρατηγική που θα αναπτύξει θα συμβουλευτεί το μοντέλο ώστε να «χτίσει» τις επιλογές της ή να πάρει τις κατάλληλες αποφάσεις όσον αφορά την διαχείριση και διοίκηση ενός νοσοκομείου, ενός χώρου πολύ ιδιαίτερο. Επίσης θα μπορούσε, αναλόγως του τι χρειαζόταν, να εισάγει και διαφορετικές μεταβλητές ως είσοδο στο μοντέλο και είτε συνυπολογίζοντας με τις γνωστές ιστορικές τιμές ή όχι να έχει διαφορετικά αποτελέσματα στις προβλέψεις, είτε να προβλέψει άλλα μεγέθη. Τέλος, εφαρμόζοντας τις παραπάνω τεχνικές και μελετώντας τα αποτελέσματα, τα οποία θα πάρει κάποιος να τα θεωρήσει ως συμβουλευτικά και σε συνεργασία με επαρκείς γνώσεις που ενδεχομένως να έχει να αντιληφθεί ακόμα καλύτερα πως λειτουργεί η «αγορά», πως μεταβάλλεται ο ρυθμός εισαγωγών σε έναν χώρο ευαίσθητο όπως αυτός της υγείας και πως μπορεί να έχει ενδεχομένως να αποφέρει καλύτερα αποτελέσματα απόδοσης στο προσωπικό που απασχολείται στον χώρο αυτό από ότι θα είχε αν δεν χρησιμοποιούσε τις ανωτέρω εφαρμογές. Παρόλα αυτά, η χρήση οποιουδήποτε μοντέλου πρόβλεψης πρέπει να εφαρμόζεται σε συνδυασμό με την υποκειμενική κρίση του χρήστη ή της διοίκησης προκειμένου να προσαρμοστούν στα πραγματικά δεδομένα της οικονομίας και της πορείας του τομέα και του εκάστοτε οργανισμού ή επιχείρησης.

Βιβλιογραφία

[1] A. D. Mauro, «What is big data? A consensual definition and a review of key research topics,» 2015.

[2] J.-R. Lee, «Big Data: Survey, Technologies, Opportunities, and Challenges,» *The Scientific World Journal,* 2014.

[3] M. G. Reem Alshamy, «A Review of Big Data in Network Intrusion Detection System:Challenges, Approaches, Datasets, and Tools,» *International Journal of Computer Sciences and Engineering,* 2020.

[4] M. A. H. S. G. B. A. A. A. S. S. Nawsher Khan, «The 10 Vs, Issues and Challenges of Big Data,» σε *Proceedings of the 2018 International Conference on Big Data and Education,* 2018.

[5] B. H. F. &. C. N. Emad A Mohammed, «Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends,» *BioData Mining,* 2014.

[6] X. Li, «Big Data Analytics in Healthcare,» *BioMed Research International,* 2015.

[7] S. K. S. M. S. &. S. K. Sabyasachi Dash, «Big data in healthcare: management, analysis and future prospects,» *Journal of Big Data,* 2019.

[8] S. K. S. M. S. &. S. K. Sabyasachi Dash, «Big data in healthcare: management, analysis and future prospects,» *Journal of Big Data,* 2019.

[9] W. R. &. V. Raghupathi, «Big data analytics in healthcare: promise and potential,» *Health Information Science and Systems,* 2014.

[10] R. A. a. R. Agrawal., «An introductory study on time series,» 2013.

[11] G. M. J. G. C. R. a. G. M. George EP Box, «Time series analysis: forecasting and control,» John Wiley & Sons, 2015.

[12] «https://www.wikiwand.com/el/Χρονολογικές_Σειρές,» [Ηλεκτρονικό].

[13] D. C. Montgomery, «Introduction to Time Series Analysis and Forecasting,» 2008.

[14] «A Course in Time Series Analysis,» New York, John Wiley, 2001.

[15] «ΥΔΡΟΛΟΓΙΚΗ ΠΡΟΣΟΜΟΙΩΣΗ ΚΑΙ ΠΡΟΒΛΕΨΗ,» Τμήμα Πολιτικών Μηχανικών Πανεπιστήμιο Θεσσαλίας.

[16] C. a. s. t. series, «Cyclic and seasonal time series,» [Ηλεκτρονικό]. Available: https://robjhyndman.com/hyndsight/cyclicts/.

[17] σε *Introduction to Time Series Analysis and Forecasting*, Wiley, 2008.

[18] R. T. Olszewski, « Generalized feature extraction for structural pattern,» Technical report, DTIC Document, 2001.

[19] «R vs python for data science: The winner is ….,» 2015. [Ηλεκτρονικό]. Available: https://www.kdnuggets.com/2015/05/r-vs-python-data-science.html.

[20] «https://www.datacamp.com/tutorial/r-or-python-for-data-analysis,» 1 2020. [Ηλεκτρονικό].

[21] [Ηλεκτρονικό]. Available: https://www.statistics.gr/el/statistics.

[22] [Ηλεκτρονικό]. Available: https://www.meteoblue.com/.

[23] «https://www.kepea.gr/,» [Ηλεκτρονικό].

[24] G. M. J. G. C. R. a. G. M. George EP Box, «Time series analysis: forecasting and control,» John Wiley & Sons, 2015.