ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΝΟΣΗΛΕΥΤΙΚΗΣ

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΑΡΧΕΙΑ ΜΕΓΑΛΟΥ ΟΓΚΟΥ ΔΕΔΟΜΕΝΩΝ ΥΓΕΙΑΣ-BIG DATA-ΜΕ ΧΡΗΣΗ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΑΛΓΟΡΙΘΜΩΝ ΑΝΑΛΥΣΗΣ-HEALTH ANALYTICS

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ
ΙΩΑΝΝΗ Δ. ΜΗΝΟΥ

ΑΘΗΝΑ 2021

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΑΡΧΕΙΑ ΜΕΓΑΛΟΥ ΟΓΚΟΥ ΔΕΔΟΜΕΝΩΝ ΥΓΕΙΑΣ-BIG DATA-ΜΕ ΧΡΗΣΗ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΑΛΓΟΡΙΘΜΩΝ ΑΝΑΛΥΣΗΣ-HEALTH ANALYTICS

Επταμελής Εξεταστική Επιτροπή:

Καθηγητής Ιωάννης Μαντάς (επιβλέπων)

Καθηγήτρια Φλώρα Μαλαματένιου

Αναπληρώτρια Καθηγήτρια Δάφνη Καϊτελίδου

Καθηγητής Γεώργιος Φιλντίσης

Καθηγήτρια Αθηνά Καλοκαιρινού

Καθηγητής Παναγιώτης Μπαμίδης

Καθηγήτρια Μαριάννα Διομήδους

«Η έγκριση διδακτορικής διατριβής από το τμήμα Νοσηλευτικής του Πανεπιστημίου Αθηνών δεν σημαίνει και αποδοχή των γνωμών του συγγραφέα» (Σχετικές οι διατάξεις του άρθρου 50 του νόμου 1268/82 σε συνδυασμό με τις διατάξεις του οργανισμού του Πανεπιστημίου Αθηνών, άρθρου 202 παρ. 2 του νόμου 5343)

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΝΟΣΗΛΕΥΤΙΚΗΣ

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΑΡΧΕΙΑ ΜΕΓΑΛΟΥ ΟΓΚΟΥ ΔΕΔΟΜΕΝΩΝ ΥΓΕΙΑΣ-BIG DATA-ΜΕ ΧΡΗΣΗ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΑΛΓΟΡΙΘΜΩΝ ΑΝΑΛΥΣΗΣ-HEALTH ANALYTICS

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ
ΙΩΑΝΝΗ Δ. ΜΗΝΟΥ

ΑΘΗΝΑ 2021

Αφιερώνεται στη Φωτεινή και στην οικογένειά μου…

Ευχαριστίες

Θα ήθελα να ευχαριστήσω όλους όσους συμμετείχαν στην έρευνα και που ήταν αρωγοί στην προσπάθεια ολοκλήρωσης της παρούσας διατριβής.

Αρχικά θα ήθελα να ευχαριστήσω τον Επιβλέποντα Καθηγητή μου κ. Ιωάννη Μαντά για την πολύτιμη συμβολή του στην ολοκλήρωση του έργου. Η στοχευμένη καθοδήγησή του και οι παραινέσεις του βοήθησαν σε όλα τα στάδια και επίπεδα της έρευνας.

Ευχαριστώ, επίσης, τα δύο μέλη της τριμελούς συμβουλευτικής επιτροπής, την Καθηγήτρια κ. Φλώρα Μαλαματένιου και την Αναπληρώτρια Καθηγήτρια Δάφνη Καϊτελίδου για τη στήριξη τους κατά τη διάρκεια της διατριβής.

Ένα μεγάλο ευχαριστώ στον Δρ. Ιωσήφ Λιάσκο για τις συμβουλές του και τις προτάσεις του που συνέβαλαν στην περάτωση καθώς και τη βελτίωση της δομής της διατριβής.

Ιδιαίτερα θέλω να ευχαριστήσω τον φίλο μου και συνεργάτη Δρ. Παρίση Γάλλο για τις συμβουλές του και την πολύτιμη βοήθειά του στην εκπόνηση του ερωτηματολογίου και της στατιστικής ανάλυσης.

Τέλος ένα μεγάλο ευχαριστώ το οφείλω στη Φωτεινή και στην οικογένειά μου για την αμέριστη συμπαράστασή τους και υπομονή τους κατά το χρονικό διάστημα εκπόνησης της διατριβής.

Περιεχόμενα

Εισαγωγή

Ο τομέας της υγειονομικής περίθαλψης ανέκαθεν δημιουργούσε μεγάλες ποσότητες δεδομένων, γεγονός που έκανε επιτακτική την τήρηση αρχείων, για τη φροντίδα των ασθενών [1]. Ενώ τα περισσότερα δεδομένα είναι αποθηκευμένα σε έντυπη μορφή, η τρέχουσα τάση είναι προς την ταχεία ψηφιοποίηση αυτών των μεγάλων ποσοτήτων δεδομένων[2]. Επιπλέον οι απαιτήσεις για βελτίωση της ποιότητας των παρεχόμενων υπηρεσιών υγείας σε συνδυασμό με την ταυτόχρονη μείωση του κόστους, οδηγεί στη χρήση τεράστιων ποσοτήτων δεδομένων (γνωστών ως «big data») τα οποία θα υποστηρίζουν ένα ευρύ φάσμα ιατρικών και νοσηλευτικών λειτουργιών, συμπεριλαμβανομένων, μεταξύ άλλων, την υποστήριξη στη διαδικασία λήψης κλινικών αποφάσεων, την παρακολούθηση της νόσου, και τη διαχείριση της υγείας του πληθυσμού [3, 4].

Τα μεγάλα δεδομένα προέρχονται μέσα από την εξέλιξη των βάσεων δεδομένων και της επιστήμης των υπολογιστών. Με την πάροδο του χρόνου, αυτές οι εξελίξεις περιελάμβαναν την ανάπτυξη σχεσιακών βάσεων δεδομένων, την παραγωγή των δεδομένων μέσω διαδικτύου, την ανάπτυξη μεγάλων αποθηκών δεδομένων (data warehouses), την εισαγωγή μεγάλων, μη δομημένων η ημιδομημένων συνόλων δεδομένων, την ανάπτυξη υπολογιστικών μεθόδων για παράλληλη επεξεργασία δεδομένων και, τέλος, την ανάπτυξη της τεχνολογίας του υπολογιστικού νέφους (cloud computing). Αυτές οι τεχνολογικές εξελίξεις, έχουν καταστήσει δυνατή την αποτελεσματική αντιμετώπιση μεγάλων δεδομένων τόσο σε ερευνητικό όσο και σε επιχειρησιακό περιβάλλον, συμπεριλαμβανομένης της υγειονομικής περίθαλψης.

Α.ΓΕΝΙΚΟ ΜΕΡΟΣ

Κεφάλαιο 1 – Μεγάλα Δεδομένα (Big Data)

1.1 Ορισμός Big Data

Τα μεγάλα δεδομένα (big data) στον τομέα της υγείας αναφέρονται σε ηλεκτρονικά δεδομένα υγείας τόσο μεγάλα και πολύπλοκα που είναι δύσκολη η διαχείρισή τους με κλασικές εφαρμογές λογισμικού, ούτε μπορούν εύκολα να αντιμετωπιστούν με τις παραδοσιακές μεθόδους διαχείρισης δεδομένων [7, 8].

1.2 Ιστορική αναδρομή

Τα τελευταία χρόνια υπάρχει ένα αυξανόμενο ενδιαφέρον για την διαχείριση μεγάλων δεδομένων κινούμενο από πραγματικές ανάγκες επεξεργασίας αυτών. Η πρώτη εμφάνιση του όρου έγινε το 1997 από τους επιστήμονες της NASA [5]. Ανέφεραν ότι αδυνατούσαν να αναπαραστήσουν γραφικά (visualization) τα σύνολα δεδομένων (data sets) που κατείχαν καθώς ήταν τόσο μεγάλα που ήταν ακατόρθωτο να τα αποθηκεύσουν στη κύρια μνήμη, στον τοπικό δίσκο και σε εξωτερικό σκληρό δίσκο.. Οι τελευταίες τεχνολογικές εξελίξεις κυρίως στον τομέα των επικοινωνιών και των ολοκληρωμένων κυκλωμάτων έχουν δώσει την δυνατότητα να δημιουργηθούν μηχανισμοί παρακολούθησης των λειτουργιών ενός οργανισμού σε πολύ λεπτομερές επίπεδο. Η λεπτομερής αυτή ψηφιοποίηση των διαδικασιών παραγωγής έχουν καταστήσει μεγάλους οργανισμούς αλλά και εταιρείες μικρού μεγέθους ικανούς να παράγουν τεράστιους όγκους δεδομένων με πολύ ταχείς ρυθμούς. Τα δεδομένα αυτά κρύβουν πολύτιμη γνώση καθώς η ανάλυση τους μπορεί να οδηγήσει σε σημαντικές βελτιστοποιήσεις της παραγωγής αλλά και σε προβλήματα, αφού οι υπάρχουσες τεχνολογικές λύσεις για την διαχείριση δεδομένων δεν ανταποκρίνονται πλήρως στον όγκο αλλά και στην φύση τους.

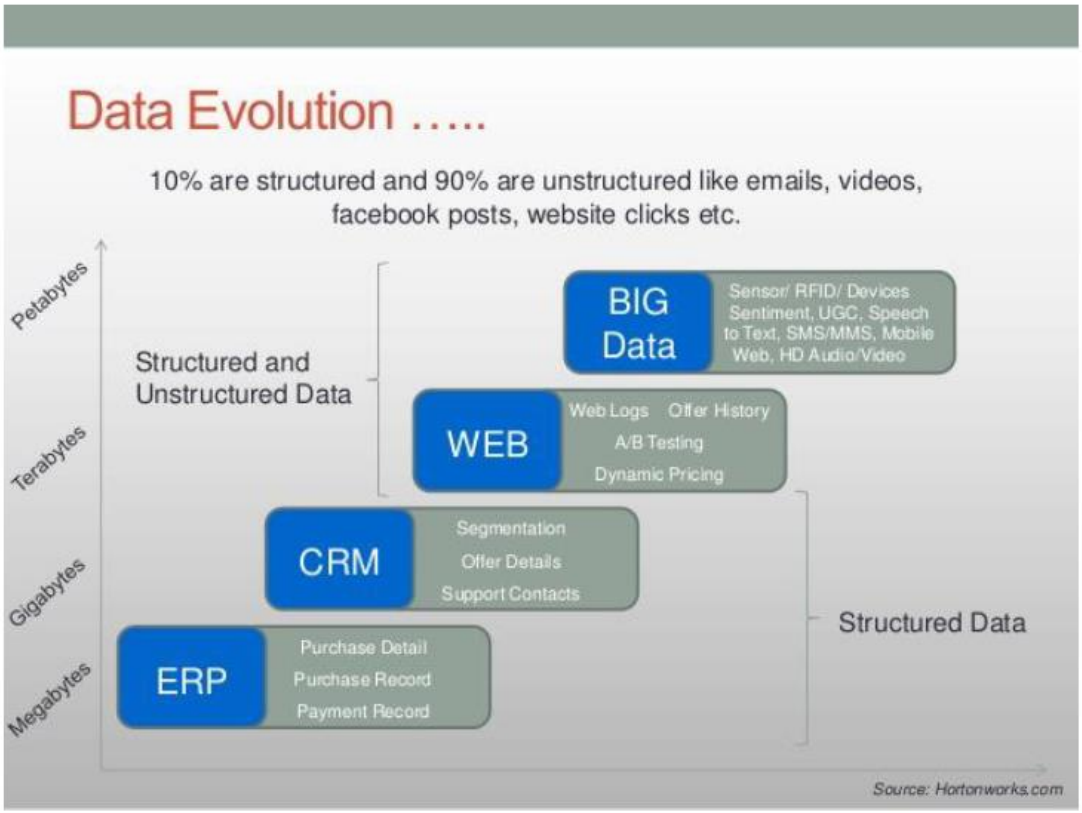
Πολλές τεχνολογικές καινοτομίες έχουν οδηγήσει σε δραματική αύξηση των δεδομένων και στη συλλογή δεδομένων. Αυτός είναι ο λόγος που τα μεγάλης κλίμακας δεδομένα έχουν γίνει πρόσφατη περιοχή των στρατηγικών επενδύσεων για τους οργανισμούς πληροφορικής (IT). Δεν είναι όμως μόνο οι οργανισμοί που παράγουν τεράστιους όγκους δεδομένων. Ακόμη και σε μικρότερη κλίμακα οργάνωσης, στο επίπεδο του ατόμου, η παραγωγή

δεδομένων είναι πρωτόγνωρη. Οι περισσότεροι άνθρωποι διαθέτουν έναν ψηφιακό εαυτό, ως προβολή των δραστηριοτήτων τους στα κοινωνικά δίκτυα. Η Google εκτιμά ότι κάθε δύο μέρες το ψηφιακό υλικό που δημιουργείται από τους χρήστες είναι ισομεγέθες με το έντυπο υλικό που παρήγαγε η ανθρωπότητα από την αρχή της γραφής μέχρι το 2003. Έκρηξη στον όγκο των παραγόμενων δεδομένων παρατηρείται ακόμη στην επιστημονική έρευνα. Τομείς, όπως η ιατρική, η αστρονομία, η μετεωρολογία αλλά και η βιολογία χάρη στις νέες τεχνολογίες, τα νέα τηλεσκόπια, τους νέους και φτηνούς αισθητήρες και τα νέα μηχανήματα για την αποκωδικοποίηση DNA μπορούν και παράγουν όγκους δεδομένων που δεν είναι δυνατόν να αντιμετωπιστούν με τις υπάρχουσες υποδομές. Μάλιστα οι ρυθμοί αύξησης παρατηρούμε ότι είναι εκθετικής κατανομής. Έτσι προβλέπεται για τα επόμενα χρόνια μια ακόμη μεγαλύτερη “έκρηξη πληροφορίας”.

Το Πρόγραμμα σχετικά με το Ανθρώπινο Γονιδίωμα ήταν ένα από τα πρώτα έργα που εισήγαγαν την έννοια των μεγάλων δεδομένων στους επιστήμονες και αναλυτές της υγειονομικής περίθαλψης. Από τότε, οι τεχνολογικές εξελίξεις, έργα και πρωτοβουλίες έχουν επίσης ωθήσει στη χρήση μεγάλου όγκου δεδομένων της υγειονομικής περίθαλψης. Για παράδειγμα, η μαζική ανάπτυξη ηλεκτρονικών ιατρικών φακέλων (Electronic Health Record) έχει δημιουργήσει μια ευκαιρία για τη συγκέντρωση κλινικών δεδομένων από διάφορους παρόχους υπηρεσιών υγείας, δίνοντας την ευκαιρία ενσωμάτωσης νέων πηγών δεδομένων, όπως δεδομένα γονιδιωματικής και κινητής υγείας (mHealth), με κλινικά σύνολα δεδομένων για την ανάλυσης και τη βελτίωση της υγείας του πληθυσμού. Πράγματι, η ποσότητα των δεδομένων υγειονομικής περίθαλψης που πρόκειται να συγκεντρωθεί μέχρι το 2020 αναμένεται να ξεπεράσει περισσότερα από 25.000 petabytes, πράγμα που ισοδυναμεί με 25 δισεκατομμύρια gigabytes πληροφοριών.

Πρόσφατες αναφορές δείχνουν ότι τα δεδομένα από το σύστημα υγειονομικής περίθαλψης των ΗΠΑ, και μόνο αυτού, έφθασαν, το 2011, τα 150 exabytes(βλ. Εικόνα 1). Με αυτόν τον ρυθμό ανάπτυξης, τα Big Data για την υγεία στις ΗΠΑ θα φτάσουν σύντομα την τάξη των zettabyte (1.024 exabytes), και όχι πολύ καιρό μετά, τη yottabyte (1.024 zetabytes) κλίμακα[5].

Επιπλέον το μοντέλο αποζημίωσης της υγειονομικής περίθαλψης αλλάζει ως προς τη χρήση και την αμοιβή των παρεχόμενων υπηρεσιών υγείας [6]. Παρά το γεγονός ότι το κέρδος δεν είναι και δεν πρέπει να είναι ο πρωταρχικός στόχος, είναι ζωτικής σημασίας για τους φορείς και οργανισμούς υγειονομικής περίθαλψης στο να αποκτήσουν τα διαθέσιμα εργαλεία, τις υποδομές και τις τεχνικές για την αποτελεσματική διαχείριση των Big Data, αλλιώς κινδυνεύουν να χάσουν εκατομμύρια δολάρια από έσοδα και κέρδη [6].



Εικόνα 1:Εξέλιξη των Big Data

Πηγή: www.hortonworks.com

Παρακάτω αναφέρονται ενδεικτικά κάποιες εκτιμήσεις και προβλέψεις [6] σχετικές με την εξέλιξη των Big Data:

- Το 2011, η ανθρωπότητα δημιούργησε πάνω από 1,2 τρισεκατομμύρια GB δεδομένων.
- Ο όγκος των δεδομένων αναμένεται να αυξηθεί 50 φορές μέχρι το 2020.

- Η Google λαμβάνει πάνω από 2.000.000 ερωτήματα αναζήτησης κάθε λεπτό.
- 72 ώρες βίντεο προστίθενται στο YouTube κάθε λεπτό.
- Υπάρχουν 217 νέοι χρήστες του Ίντερνετ κάθε λεπτό.
- Οι χρήστες του Twitter στέλνουν πάνω από 100.000 tweets κάθε λεπτό (που είναι πάνω από 140 εκατομμύρια ανά ημέρα).
- Εταιρείες, και οργανισμοί λαμβάνουν 34.000 “likes” σε κοινωνικά δίκτυα κάθε λεπτό.
- Διεθνή δεδομένα Corporation (IDC) προβλέπουν ότι η αγορά για την τεχνολογία των μεγάλης κλίμακας δεδομένων και υπηρεσιών θα φτάσει τα16,9 εκατομμύρια δολάριο.

1.3 Πηγές δεδομένων στην υγεία

Τα Big Data στον τομέα της υγείας είναι πολύπλοκα, όχι μόνο λόγω του όγκου τους, αλλά και λόγω της ποικιλίας των τύπων δεδομένων και απαιτούν προηγμένες τεχνικές και τεχνολογίες που επιτρέπουν την αποθήκευση, τη διανομή, τη διαχείριση και την ανάλυση των πληροφοριών [8]. Περιλαμβάνουν:

- Κλινικά δεδομένα από τα συστήματα CPOE (Computerized Provider Order Entry) και τα συστήματα υποστήριξης κλινικών αποφάσεων (γραπτές σημειώσεις και ιατρικές συνταγές, ιατρικές απεικονίσεις, εξετάσεις εργαστηρίων, δεδομένα φαρμακείου, πληροφορίες ασφαλιστικού φορέα και άλλα διοικητικά δεδομένα).
- Πληροφορίες που προέρχονται από Ηλεκτρονικούς Φακέλους Ασθενών
- Λιγότερο σημαντικές πληροφορίες όπως ειδήσεις και άρθρα σε επιστημονικά περιοδικά.
- Συστήματα κλινικών πληροφοριών.
- Δεδομένα ασφαλιστικών οργανισμών.

- Δημογραφικές και επιδημιολογικές εγγραφές.
- Δεδομένα ιατρικών συσκευών και φαρμάκων.
- Χρηματοδοτικές και οικονομικές συναλλαγές.
- Δεδομένα βιοϊατρικών μετρήσεων.
- Βάσεις γενετικών δεδομένων.
- Αναζητήσεις διαδικτύου.
- Μέσα κοινωνικής δικτύωσης.
- Μη δομημένα δεδομένα σημειώσεων και αλληλογραφίας.

1.4 Χαρακτηριστικά των Big Data στην υγεία

Τα Big Data στον χώρο της υγείας έχουν τα εξής χαρακτηριστικά[9]:

1. Όγκος(Volume)

Η ποσότητα της πληροφορίας που δημιουργείται είναι πολύ σημαντική στο πεδίο των Big Data. Είναι το μέγεθος των δεδομένων που καθορίζει την αξία των υπό εξέταση στοιχείων και κατά πόσο αυτά μπορούν να χαρακτηριστούν ως Big Data. Επίσης, η αύξηση της ποσότητας των δεδομένων είναι αποτέλεσμα τόσο ψηφιοποίησης ήδη διαθέσιμων δεδομένων, όσο και δημιουργίας νέων μορφών δεδομένων. Ο όγκος των διαθέσιμων δεδομένων αποτελείται από προσωπικά ιατρικά αρχεία, εικόνες ραδιολογίας και ακτινολογίας, κλινικές δοκιμές, έρευνες, δημογραφικά στοιχεία, ανθρώπινα γονιδιώματα, γενετικές ακολουθίες κ.ά. Οι νέες μορφές Big Data, όπως εικόνες τριών διαστάσεων (3D), δεδομένα γονιδιωματικών και βιομετρικών αισθητήρων, συμβάλλουν στην εκθετική αύξηση των δεδομένων στον κλάδο της υγείας. Ο όγκος των διαθέσιμων ιατρικών δεδομένων ήταν 500 petabytes (1015 bytes) το 2012 και αναμένεται να ξεπεράσει τα 25000 petabytes έως το 2020. Οι εξελίξεις στη διαχείριση των δεδομένων, κυρίως στην οπτική απεικόνιση, και η χρήση του υπολογιστικού νέφους (cloud computing), έχουν οδηγήσει στην διαμόρφωση εφαρμογών που στοχεύουν στην αποτελεσματικότερη λήψη, αποθήκευση και διαχείριση των Big Data.

2. Ποικιλία (Variety)

Η ποικιλία είναι ένα σημαντικό χαρακτηριστικό των Big Data καθώς η κατηγορία στην οποία ανήκουν πρέπει να είναι γνωστή στους αναλυτές δεδομένων ώστε να χρησιμοποιούν τα δεδομένα αποτελεσματικότερα. Τα δεδομένα του τομέα υγείας έχουν το χαρακτηριστικό ότι αποτελούν αδόμητα δεδομένα δεδομένου ότι περιλαμβάνουν ηλεκτρονικές και χειρόγραφες σημειώσεις, ιατρικές εικόνες κ.ά. Ωστόσο παρατηρείται μία αύξηση της χρήσης δομημένων δεδομένων υγείας όπως ποσοτικά δεδομένα μετρήσεων ιατρικών οργάνων, δεδομένα ηλεκτρονικής συνταγογράφησης, σε μία ενιαία δομή για περαιτέρω αξιοποίηση και επεξεργασία. Επίσης, ιδιαίτερα χρήσιμα και είναι δεδομένα νέων πηγών, όπως συσκευών ευεξίας που καταγράφουν τους παλμούς ή τη διάρκεια ύπνου των ασθενών, κοινωνικών δικτύων, γονιδιωματικής έρευνας κ.ά. Τα δεδομένα αυτά, αν και ιδιαίτερα χρήσιμα, δεν είναι προς το παρόν αξιοποιήσιμα στο έπακρο, καθώς δεν είναι εύκολη η μετατροπή από μία αδόμητη σε μία δομημένη μορφή.

3. Ταχύτητα (Velocity)

Ο όρος ταχύτητα αναφέρεται στο πόσο γρήγορα παράγονται και υφίστανται επεξεργασία τα δεδομένα. Τα περισσότερα δεδομένα στον τομέα της υγείας προέρχονται παραδοσιακά από στατικές πηγές, όπως ακτινογραφίες, έγγραφα νοσοκομείων, δελτία ασθενών, βιβλιάρια υγείας κ.ά. Σε κάποιες εφαρμογές, όμως, κρίνεται απαραίτητη η επεξεργασία και αξιοποίηση των δεδομένων σε πραγματικό χρόνο, όπως για παράδειγμα η επίβλεψη της αρτηριακής πίεσης και της καρδιακής λειτουργίας σε κάποια εγχείρηση. Υπάρχουν περιπτώσεις επίσης που απαιτείται επεξεργασία των δεδομένων σε πιο αργούς σχετικά ρυθμούς, όπως ο προσδιορισμός των επιπέδων της γλυκόζης διαβητικών ατόμων σε καθημερινή βάση.

4. Εγκυρότητα (Veracity)

Η ποιότητα των δεδομένων μπορεί να ποικίλει σε μεγάλο βαθμό. Η ακρίβεια της ανάλυσης εξαρτάται από την ακρίβεια της πηγής των δεδομένων. Ιδιαίτερα στον κλάδο της υγείας, η εγκυρότητα των δεδομένων

είναι υψίστης σημασίας για δύο κυρίως λόγους: Σε αυτή βασίζονται αποφάσεις που καθορίζουν την ανθρώπινη ύπαρξη. Επίσης τα μη-δομημένα δεδομένα, αν και είναι μεγάλης σημασίας, συχνά εμπεριέχουν σφάλματα και λάθη ή είναι αδύνατο να αξιοποιηθούν για διάφορους λόγους.

Η αξιοπιστία των δεδομένων στην υγεία αφορά στην εγκυρότητα των στοιχείων του ασθενούς, τα ορθώς συμπληρωμένα δεδομένα στα πεδία που αφορούν το νοσοκομείο ή την κλινική, την ασφάλιση του ασθενούς, σύνδεση με τραπεζικούς λογαριασμούς, καταγραφή χρηματικών ποσών πληρωμής, εκκρεμών οφειλών κ.ά. Αν και στον κλάδο της υγείας υπάρχουν δεδομένα τα οποία δεν παρατηρούνται σε άλλους τομείς, όπως πληροφορίες σχετικές με, τη θεραπεία, τη χορήγηση φαρμάκων, την περίθαλψη, η εγκυρότητα των στοιχείων αυτών είναι, σε κάθε περίπτωση, εξίσου σημαντική με αυτή των προαναφερθέντων.

5. Πολυπλοκότητα (Complexity)

Η διαχείριση των δεδομένων μπορεί να γίνει μία πολύ σύνθετη διαδικασία, ειδικότερα στην περίπτωση όπου ο μεγάλος όγκος των δεδομένων προέρχεται από διαφορετικές πηγές. Τα στοιχεία αυτά συνδέονται και συσχετίζονται μεταξύ τους προκειμένου οι αναλυτές να είναι σε θέση να κατανοήσουν τις πληροφορίες που μεταφέρουν τα δεδομένα. Η κατάσταση αυτή ονομάζεται πολυπλοκότητα των Big Data.

6. Μεταβλητότητα (Volatility)

Η μεταβλητότητα αναφέρεται στο χρονικό διάστημα στο οποίο τα δεδομένα ισχύουν και πρέπει να αποθηκεύονται. Σε περίπτωση που τα δεδομένα δεν είναι πλέον έγκυρα, δεν χρησιμοποιούνται για περεταίρω ανάλυση.

1.5 Πλεονεκτήματα-Μειονεκτήματα Big Data

Η ψηφιοποίηση, ο συνδυασμός και η αποτελεσματική χρήση των Big Data προσφέρει σημαντικά πλεονεκτήματα στους οργανισμούς υγειονομικής περίθαλψης στα ακόλουθα πεδία[10]:

- Δημόσια υγεία: με την ανάλυση των ασθενειών και την καταγραφή των επιδημιών, τα ζητήματα της δημόσιας υγείας μπορούν να βελτιωθούν. Τα Big Data βοηθούν στον προσδιορισμό των αναγκών, στην παροχή απαιτούμενων υπηρεσιών υγείας, στην πρόβλεψη και πρόληψη των μελλοντικών κρίσεων προς όφελος του πληθυσμού.

- Ηλεκτρονικός Φάκελος Ασθενών: ο Ηλεκτρονικός Φάκελος Ασθενών περιέχει τυποποιημένα(δομημένα και μη δομημένα) ιατρικά δεδομένα που μπορούν να αξιολογηθούν μέσω της ανάλυσης των δεδομένων για την πρόβλεψη ασθενών που βρίσκονται σε κίνδυνο και την παροχή αποτελεσματικής φροντίδας.

- Ανάλυση των χαρακτηριστικών του ασθενούς: εφαρμογή εξελιγμένων εργαλείων ανάλυσης του προφίλ του ασθενούς (όπως ο κατακερματισμός και η προγνωστική μοντελοποίηση) για τον εντοπισμό των ατόμων που θα ωφεληθούν από την ενεργητική φροντίδα ή τις αλλαγές στον τρόπο ζωής τους.

- Απομακρυσμένη παρακολούθηση: μέσω ηλεκτρονικών υπολογιστών, έξυπνων κινητών συσκευών (smart phones) ή άλλων φορητών συσκευών οι οποίες στέλνουν, λαμβάνουν και αναλύουν, σε πραγματικό χρόνο, μεγάλες ποσότητες δεδομένων από και προς το νοσοκομείο ή το σπίτι του ασθενούς. Τα δεδομένα αυτά χρησιμοποιούνται για την εξαγωγή συμπερασμάτων σχετικά με την κατάσταση της υγείας του ασθενούς καθώς και για τη χορήγηση της κατάλληλης θεραπείας.

- Γονιδιωματική ανάλυση: η προσέγγιση της ανάλυσης δεδομένων μπορεί να συμπεριληφθεί αποτελεσματικά στη γονιδιωματική ανάλυση για να καταστεί αυτή η προσέγγιση μέρος της τακτικής διαδικασίας λήψης αποφάσεων ιατρικής περίθαλψης.

- Πρόσβαση στην πληροφορία: με τις τεχνολογίες Internet και cloud αποθήκευσης να βελτιώνονται και να επεκτείνονται συνεχώς, τα Big Data μπορούν να είναι προσβάσιμα από σχεδόν οπουδήποτε. Αυτό είναι ένα τεράστιο όφελος για τους γιατρούς που μπορούν να έχουν

πρόσβαση σε ηλεκτρονικά ιατρικά αρχεία οποιαδήποτε στιγμή, οπουδήποτε για να βελτιώσουν την περίθαλψη των ασθενών. Είναι επίσης επωφελής για τη διαχείριση εγκαταστάσεων, χειριστές και μηχανικούς, οι οποίοι μπορούν να εντοπίσουν και να αντιμετωπίσουν τα προβλήματα εξ αποστάσεως.

- Τεκμηριωμένη ιατρική: μέσω του συνδυασμού και της ανάλυσης δομημένων και αδόμητων δεδομένων, κλινικών και γονιδιακών στοιχείων τα big data μπορούν να βοηθήσουν τους επαγγελματίες υγείας να προβλέπουν ποιοι ασθενείς βρίσκονται σε κίνδυνο να νοσήσουν και να παρέχουν πιο αποτελεσματική φροντίδα.

Γενικότερα τα κυριότερα πλεονεκτήματα των Big Data είναι τα εξής[11]:

- Ιατρική βασισμένη σε ενδείξεις (evidence- based medicine).
- Αποδοτικότερη διαχείριση και κατανομή του ανθρωπίνου δυναμικού.
- Σύγκριση της αποτελεσματικότητας και της αποδοτικότητας των επαγγελματιών και φορέων υγείας.
- Έλεγχος των πόρων που καταναλώνονται.
- Άμεση πρόσβαση στη γνώση από ασθενείς και επαγγελματίες υγείας.
- Δυνατότητα παροχής υπηρεσιών φροντίδας προσαρμοσμένων στις ανάγκες του ασθενή.
- Αποτελεσματικότερος και ταχύτερος σχεδιασμός νέων φαρμάκων και κλινικών δοκιμών.
- Μείωση κόστους μέσω εντοπισμού των μεθόδων που απαιτούν τις υψηλότερες δαπάνες.
- Άμεση σύγκριση αποτελεσμάτων διαφορετικών μεθόδων θεραπείας και πρόληψης.
- Συνδυασμός δεδομένων ετερογενών πηγών για ενιαία αποτελέσματα.
- Αποτελεσματικότερη πρόληψη στους πληθυσμούς υψηλού κινδύνου.

Τα κύρια μειονεκτήματα των Big Data είναι τα εξής:

- Απόρρητο-Προσωπικά δεδομένα: ένα από τα αρνητικά που σχετίζονται με τα μεγάλα δεδομένα είναι η έλλειψη ιδιωτικότητας,

ειδικά όταν πρόκειται για εμπιστευτικά ιατρικά αρχεία. Για να είναι αποτελεσματική τα μεγάλα δεδομένα πρέπει να έχουν πρόσβαση σε όλα, συμπεριλαμβανομένων των ιδιωτικών αρχείων και των δεδομένων των μέσων κοινωνικής δικτύωσης. Σύμφωνα με πολλούς αναλυτές δεδομένων αν και τα μεγάλα δεδομένα επιτρέπουν στους γιατρούς να παρακολουθούν την υγεία ενός ασθενούς σχεδόν από οπουδήποτε χωρίς να διασφαλίζουν την ιδιωτικότητα των προσωπικών του δεδομένων.

- Αντικατάσταση των γιατρών: ενώ πολλοί βλέπουν την ικανότητα πρόβλεψης μελλοντικών ιατρικών ζητημάτων ως θετική, τα μεγάλα δεδομένα θέτουν επίσης τον κίνδυνο αντικατάστασης των γιατρών Ορισμένοι εμπειρογνώμονες φοβούνται ότι η ανάπτυξη μεγάλων δεδομένων θα μπορούσε ενδεχομένως να υπονομεύσει τους γιατρούς και να αφήσει τους ασθενείς να στραφούν προς την τεχνολογία για απαντήσεις αντί να ζητήσουν τη γνώμη ενός εξουσιοδοτημένου ιατρού.

Κεφάλαιο 2 - Ζητήματα υλοποίησης και αποθήκευσης Big Data

2.1 Υλοποίηση των Big Data

Η υλοποίηση των Big Data περιλαμβάνει τα παρακάτω βήματα [12] :

- Δημιουργία Βάσης Δεδομένων
- Απόκτηση Δεδομένων
- Συστήματα Αποθήκευσης Δεδομένων
- Ανάλυση Big Data
- Εξόρυξη Δεδομένων

2.1.1 Δημιουργία Βάσης Δεδομένων

Η παραγωγή των δεδομένων καθώς και η δημιουργία Βάσης Δεδομένων είναι το πρώτο βήμα υλοποίησης των Big Data[12]. Λαμβάνοντας υπόψη τα δεδομένα του διαδικτύου ως παράδειγμα, προκύπτει ότι δημιουργείται μια τεράστια ποσότητα δεδομένων από αναζητήσεις, καταχωρήσεις, δημοσιεύσεις, αρχεία και μηνύματα σε ιστοσελίδες και ιστολόγια (blogs). Τα εν λόγω δεδομένα μπορεί να είναι μεμονωμένα και άνευ αξίας αλλά, μέσω της αξιοποίησης με τα Big Data μπορεί να προσφέρουν χρήσιμες πληροφορίες σχετικά με συνήθειες και χαρακτηριστικά των χρηστών. Επιπλέον, δημιουργούνται μέσω κατανεμημένων πηγών δεδομένων, σύνολα μεγαλύτερης κλίμακας, ποικιλίας και πολυπλοκότητας. Τέτοιες πηγές δεδομένων περιλαμβάνουν βίντεο, κλικ χρηστών, μηνύματα κλπ.

Η δημιουργία των δεδομένων (data generation) αποτελεί τη βάση σύνθεσης του οικοδομήματος των Big Data απαιτώντας αυξημένη προσοχή στη συγκέντρωσή τους [13]. Ειδικότερα, για τη δημιουργία των δεδομένων (data generation) αξιοποιούνται τα εξής:

1. Δεδομένα επιχειρήσεων: τα εσωτερικά δεδομένα των επιχειρήσεων είναι η κύρια πηγή δημιουργίας Big Data. Τα δεδομένα των επιχειρήσεων είναι αποτέλεσμα των εμπορικών της συναλλαγών, της παραγωγής, στοιχείων αποθεμάτων, δεδομένα πωλήσεων, οικονομικά στοιχεία και δεδομένα με γνώμονα τις δραστηριότητες.

2. Διαδίκτυο των Πραγμάτων (Internet of Things-IoT): το IoT είναι μια σημαντική πηγή των Big Data. Αξιοποιώντας τις διεργασίες απόκτησης και μετάδοσης των δεδομένων στο IoT, η αρχιτεκτονική του δικτύου μπορεί να διαστρωματωθεί ως εξής: ανίχνευση, συγκέντρωση και αξιοποίηση των δεδομένων. Τα δεδομένα που δημιουργούνται από το IoT έχουν μεγάλη κλίμακα, ετερογένεια, ισχυρή συσχέτιση χώρου και χρόνου.

3. Βιοϊατρικά δεδομένα: στοιχεία βιοϊατρικών καταγραφών, μετρήσεις, ερευνητικά στοιχεία στον τομέα της βιοϊατρικής έχουν εισέλθει στην εποχή των Big Data.
4. Διάφορες πηγές: η δημιουργία των Big Data βασίζεται σε στοιχεία επιστημονικών εφαρμογών, στοιχεία αστρονομίας, φυσικών καταγραφών, εμπορικών συναλλαγών, διαδίκτυο κλπ. Τα στοιχεία από διαφορετικές πηγές παρουσιάζουν αυξημένη ετερογένεια ενώ είναι πολλές φορές αρκετά πολύπλοκα.

2.1.2 Απόκτηση Δεδομένων

Η απόκτηση των δεδομένων (data acquisition) είναι το επόμενο βήμα στη δημιουργία των Big Data [14]. Περιλαμβάνει τη συλλογή, τη μετάδοση και την προεπεξεργασία των δεδομένων. Κατά τη διάρκεια της συλλογής των δεδομένων, θα πρέπει ταυτόχρονα να εφαρμόζεται και ο κατάλληλος μηχανισμός μετάδοσης για να μεταφερθούν τα δεδομένα αποτελεσματικά σε ένα κατάλληλο σύστημα διαχείρισης και αποθήκευσης που υποστηρίζει τις διαφορετικές εφαρμογές. Η συγκέντρωση δεδομένων μπορεί να περιλαμβάνει μερικές φορές πολύ περιττά ή άχρηστα δεδομένα, γεγονός που αυξάνει άσκοπα το χώρο αποθήκευσης και επηρεάζει τη μετέπειτα ανάλυση των δεδομένων. Ως εκ τούτου, καθίστανται αναγκαίες ενέργειες επεξεργασίας των δεδομένων για να εξασφαλιστεί η αποτελεσματική αποθήκευση και αξιοποίησή τους.

2.1.3 Συστήματα Αποθήκευσης Δεδομένων

Η ανάπτυξη των δεδομένων έχει οδηγήσει σε πιο αυστηρές απαιτήσεις σε θέματα αποθήκευσης και διαχείρισης των δεδομένων[15]. Η αποθήκευση των Big Data επικεντρώνεται και στη διαχείριση δεδομένων μεγάλης

κλίμακας, επιτυγχάνοντας παράλληλα την αξιοπιστία, τη διαθεσιμότητα της πρόσβασης καθώς και την ακεραιότητα των δεδομένων. Η υποδομή αποθήκευσης πρέπει να παρέχει πληροφορίες αποθήκευσης και από την άλλη να εξασφαλίζει την άμεση και σωστή πρόσβαση. Για την αποθήκευση των Big Data χρησιμοποιείται βοηθητικός εξοπλισμός αποθήκευσης δεδομένων. Η αύξηση των δεδομένων έχει ανάλογα αυξήσει και τις ανάγκες για συσκευές αποθήκευσης δεδομένων επιδιώκοντας τη δημιουργία όλο και μεγαλύτερων μονάδων αποθήκευσης. Για την αποθήκευση δεδομένων χρησιμοποιούνται συστήματα όπως:

1. **Κατανεμημένα συστήματα αποθήκευσης:** στα κατανεμημένα συστήματα αποθήκευσης δεδομένων θα πρέπει να λαμβάνεται υπόψη η σύνδεση των διακομιστών (server) για αποφυγή αποτυχιών επικοινωνίας και ασυνέπειας μεταξύ διαφορετικών αντιγράφων των ίδιων δεδομένων, να εξασφαλίζεται η διαθεσιμότητα και να διαθέτει κατάλληλο επίπεδο ανοχής για τα προβλήματα που προκαλούνται από βλάβες του δικτύου. Τα συγκεκριμένα συστήματα θεωρούνται γενικά ιδιαίτερα ασφαλή και χρήσιμα στην αποθήκευση δεδομένων των Big Data ενώ επιτυγχάνουν υψηλά επίπεδα συνοχής.
2. **Σύστημα αποθήκευσης για δεδομένα μεγάλου όγκου:** Οι τωρινές τεχνολογίες αποθήκευσης μπορεί να είναι άμεσης αποθήκευσης (Direct Attached Storage, DAS) είτε αποθήκευσης στο δίκτυο (Network Attached Storage, NAS ή Storage Area Network, SAN). Το σύστημα αποθήκευσης στα Big Data επιλέγεται βάση του σκοπού και των εκάστοτε αναγκών. Ειδικότερα, τα συστήματα DAS είναι κατάλληλα για διασύνδεση με servers μικρής κλίμακας, καθώς λόγω της χαμηλής επεκτασιμότητας του συστήματος υπάρχει ενδεχόμενο η αναβάθμιση του χώρου αποθήκευσης να είναι δύσκολη. Το σύστημα NAS από την άλλη χαρακτηρίζεται από υψηλή επεκτασιμότητα καθώς στην πραγματικότητα λειτουργεί ως βοηθητικός εξοπλισμός αποθήκευσης ενός δικτύου. Είναι άμεσα συνδεδεμένος με το δίκτυο μέσω ενός hub ή μέσω εναλλαγής πρωτοκόλλων TCP/IP.

2.1.4 Ανάλυση Big Data

Η ανάλυση των δεδομένων είναι το τελευταίο και το πιο σημαντικό στάδιο στα Big Data καθώς καλύπτει μια ευρεία περιοχή η οποία είναι εξαιρετικά πολύπλοκη. Για την ανάλυση των Big Data αξιοποιούνται μέθοδοι, αρχιτεκτονικές και εργαλεία τα οποία παράγουν κάθε φορά διαφορετικές πληροφορίες. Στην παραδοσιακή ανάλυση των δεδομένων απαιτείται να εφαρμόζονται οι κατάλληλες στατιστικές μέθοδοι ώστε να μεγιστοποιείται η αξία των πληροφοριών. Η ανάλυση των δεδομένων διαδραματίζει σημαντικό ρόλο στη λήψη αποφάσεων, στην κατανόηση απαιτήσεων και στον καθορισμό προβλέψεων. Για την ανάλυση των Big Data μπορούν να χρησιμοποιηθούν παραδοσιακές μέθοδοι με τις στατιστικές να ξεχωρίζουν (π.χ. ανάλυση συστάδων (cluster analysis), παραγοντική ανάλυση (factor analysis), ανάλυση συσχέτισης (correlation analysis), ανάλυση παλινδρόμησης (regression analysis), αλγόριθμοι εξόρυξης γνώσης (data mining algorithms)) [16].

Η ανάλυση των Big Data μπορεί να ταξινομηθεί ανάλογα με τους επιχειρησιακούς σκοπούς και δραστηριότητα της εκάστοτε εταιρείας ή οργανισμού αλλά και με βάση την πολυπλοκότητα των αλγορίθμων που χρησιμοποιούνται.

2.2 Μηχανισμοί αποθήκευσης των Big Data

Η τεχνολογία των Big Data προσφέρει μηχανισμούς αποθήκευσης οι οποίοι είναι αξιόπιστοι και αποδοτικοί για την πρόσβαση στα δεδομένα [17]. Οι όποιες βελτιώσεις στην πρόσβαση στα δεδομένα είναι ιδιαίτερα σημαντικές καθώς προσφέρουν καλύτερη ποιότητα στην ανάλυση των δεδομένων. Οι υφιστάμενοι μηχανισμοί αποθήκευσης των Big Data μπορούν να ταξινομηθούν σε τρία διαφορετικά επίπεδα από κάτω προς τα πάνω: α) τα συστήματα αρχείων, β) βάσεις δεδομένων και γ) μοντέλα προγραμματισμού. Τα συστήματα αρχείων είναι η βάση για τα ανώτερα επίπεδα εξασφαλίζοντας τη λειτουργία του μηχανισμού αποθήκευσης. Οι μηχανισμοί αποθήκευσης σε πολλές περιπτώσεις χαρακτηρίζονται από υψηλές επιδόσεις υποστηρίζοντας εφαρμογές μεγάλης κλίμακας. Ωστόσο, μπορεί να προκύψουν αδυναμίες και περιορισμοί όπως η δυσκολία παρουσίασης δεδομένων μικρής κλίμακας.

Το Google File System (GFS) είναι ένα χαρακτηριστικό παράδειγμα μηχανισμού αποθήκευσης αρχείων το οποίο μπορεί να υποστηρίζει εφαρμογές υψηλών απαιτήσεων. Αναπτύχτηκε για τη διαχείριση και αποθήκευση μεγάλου όγκου δεδομένων χρησιμοποιώντας τεχνικές και προδιαγραφές κατανεμημένων συστημάτων. Το Hadoop Distributed File System (HDFS) δημιουργήθηκε με βάση το Google GFS. Το HDFS είναι ένα ανοιχτού κώδικα, κατανεμημένο σύστημα αποθήκευσης σχεδιασμένο για να δουλεύει με την πλατφόρμα Apache Hadoop. Τέλος η Microsoft ανέπτυξε το Cosmos για την υποστήριξη της αναζήτησης και της διαφήμισης των επιχειρήσεων ενώ το Facebook χρησιμοποιεί το Haystack για να αποθηκεύει το μεγάλο όγκο των φωτογραφιών μικρού μεγέθους.

2.3 Τεχνολογίες αποθήκευσης Big Data

Μέχρι τώρα η αποθήκευση, η διαχείριση και η ανάκτηση των δεδομένων γινόταν χρησιμοποιώντας συστήματα διαχείρισης βάσεων δεδομένων, τα οποία αφορούσαν τη διαχείριση σχεσιακών βάσεων δεδομένων.[18] Λόγω του ότι οι σχεσιακές βάσεις δεδομένων δεν μπορούν να καλύψουν τις ανάγκες αποθήκευσης των Big Data, έχουν αναπτυχθεί οι NoSQL βάσεις δεδομένων οι οποίες διαθέτουν χαρακτηριστικά και ενσωματώνουν τεχνολογίες που συμβάλλουν στον αποδοτικό χειρισμό και αποθήκευση μεγάλου όγκου δεδομένων. Οι NoSQL βάσεις δεδομένων περιλαμβάνουν τις παρακάτω κατηγορίες:

- **Βάσεις δεδομένων κλειδιού - τιμής (key - value databases)**: οι βάσεις δεδομένων κλειδιού - τιμής αποτελούν ένα απλό μοντέλο δεδομένων στην οποία το κάθε αντικείμενο έχει ένα μοναδικό κλειδί καθώς και ένα σύνολο από ζεύγη χαρακτηριστικού-τιμής [19]. Οι συγκεκριμένες βάσεις δεδομένων διαθέτουν απλή δομή και χαρακτηρίζονται από υψηλή επεκτασιμότητα και μικρότερο χρόνο απόκρισης ερωτήματος σε σύγκριση με τις σχεσιακές βάσεις δεδομένων.
- **Βάση προσανατολισμένη σε στήλες**: Η βάση δεδομένων σε στήλες αποθηκεύει και επεξεργάζεται τα δεδομένα σύμφωνα με τις στήλες εκτός από σειρές. Οι συγκεκριμένες βάσεις δεδομένων είναι

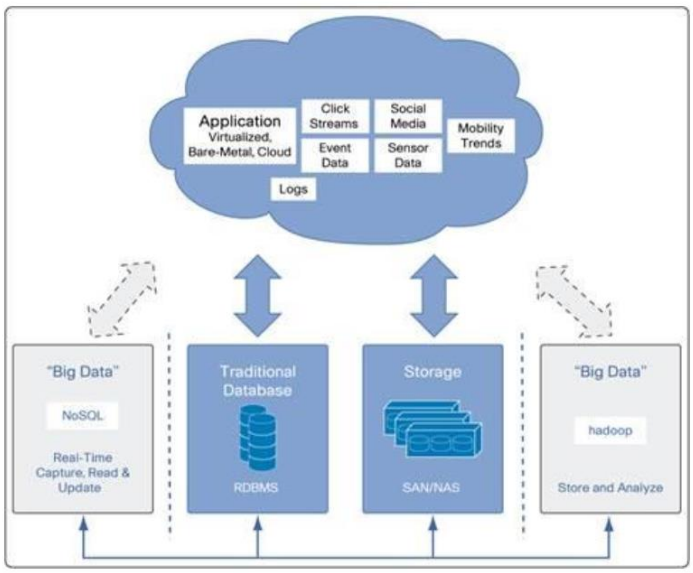
εμπνευσμένες κυρίως από το Google Big Table. Τα δεδομένα είναι κατανεμημένα και δομημένα σε ένα σύστημα αποθήκευσης δεδομένων, το οποίο έχει σχεδιαστεί για την επεξεργασία δεδομένων μεγάλης κλίμακας.

- **Βάση δεδομένων εγγράφων**: Η συγκεκριμένη τεχνολογία μπορεί να υποστηρίξει πιο σύνθετες μορφές δεδομένων [20]. Οι βασικότεροι εκπρόσωποι των συστημάτων αποθήκευσης του εγγράφου είναι το MongoDB, το SimpleDB και το CouchDB.
- **Βάσεις Δεδομένων Γραφημάτων:** Για την αποθήκευση χρησιμοποιούνται γραφήματα ως δομές τα οποία έχουν άκρες, ιδιότητες, κόμβους για τη μοντελοποίηση και την αποθήκευση των δεδομένων.

Κεφάλαιο 3 – Ανάλυση και επεξεργασία Big data

3.1 Συστήματα ανάλυσης και επεξεργασίας μεγάλου όγκου δεδομένων

Η αύξηση του όγκου των δεδομένων και η εξάπλωση της τεχνολογίας των Big Data αναγκάζει τις επιχειρήσεις και τους οργανισμούς να κάνουν αλλαγές στις τεχνολογικές τους υποδομές καθώς τα παραδοσιακά μοντέλα και εργαλεία διαχείρισης και αποθήκευσης δεδομένων δεν μπορούν να ανταπεξέλθουν (βλ. Εικόνα 2). Οι νέες τεχνολογίες δημιουργούν νέες προκλήσεις καθώς οι επιχειρήσεις και οι οργανισμοί θα πρέπει να αποφασίσουν ποιες είναι οι αλλαγές που πρέπει να γίνουν, εφόσον τα νέα μοντέλα θα πρέπει να πληρούν κάποια κριτήρια τα οποία θα πρέπει να προσαρμοστούν στις τρέχουσες επιχειρησιακές ανάγκες, στις στρατηγικές αποφάσεις των οργανισμών καθώς και στις υπάρχουσες υλικές και δικτυακές υποδομές [21]. Όσον αφορά τα συστήματα ανάλυσης και χειρισμού Big Data υπάρχει η δυνατότητα επιλογής μεταξύ ελεύθερου λογισμικού, αλλά και εμπορικών λύσεων, οι οποίες απαιτούν την αγορά κάποιο προϊόντος. Η πλατφόρμα που θα επιλεχθεί, σε κάθε περίπτωση, θα πρέπει να χειρίζεται την εισαγωγή, την επεξεργασία, την αποθήκευση και την αναζήτηση των δεδομένων, καθώς επίσης να παρέχει δυνατότητες ανάλυσής τους.



Εικόνα 2: Επιχειρησιακό Μοντέλο Big Data
Πηγή: www.bigdataframework.com

3.1.1 Map Reduce

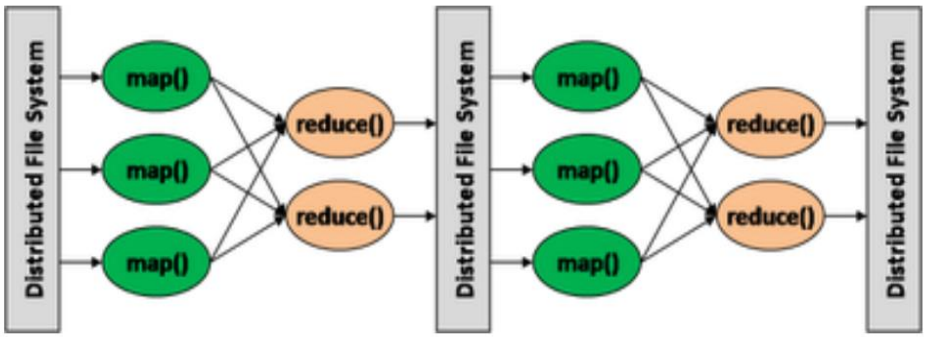
Το Map Reduce είναι ένα μοντέλο προγραμματισμού για την επεξεργασία και παραγωγή μεγάλων συνόλων δεδομένων πάνω σε συστάδες υπολογιστών (clusters) που αρχικά προτάθηκε από τη Google το 2004 [22]. Το μοντέλο είναι αρκετά απλό στη χρήση και ευρέως διαδεδομένο. Η Google έχει υλοποιήσει πάνω από 10000 προγράμματα και κατά μέσο όρο κάθε μέρα τρέχουν στα clusters της 1000 ξεχωριστές εργασίες Map Reduce που συνολικά επεξεργάζονται πάνω από 20 petabytes δεδομένων.

Είναι σχεδιασμένο για να πετύχει σημαντικές επιδόσεις σε μεγάλες συστάδες εμπορικών ηλεκτρονικών υπολογιστών**.** Βασίζεται στη μέθοδο του διαίρει και βασίλευε και λειτουργεί αναδρομικά διασπώντας ένα σύνθετο πρόβλημα σε πολλά υποπροβλήματα μέχρι που αυτά τα υποπροβλήματα να μπορούν να επιλυθούν άμεσα (βλ. Εικόνα 3)**.** Στη συνέχεια τα υποπροβλήματα εκχωρούνται σε ένα κόμβο**-**εργάτη και επιλύονται ξεχωριστά και παράλληλα**.** Τελικά οι λύσεις τους συνδυάζονται και δίνουν μια ολοκληρωμένη λύση στο αρχικό πρόβλημα**.**

Μια τυπική εργασία του Map Reduce αποτελείται από τρεις φάσεις ή συναρτήσεις που κατά σειρά είναι το map, το shuffle και το reduce παίρνοντας μια λίστα από ζεύγη κλειδιού-τιμής ως είσοδο[22]. Στη φάση map κάθε ζεύγος εισόδου εκτελείται μέσα από την λειτουργία map, και μηδέν ή περισσότερα νέα ζεύγη κλειδιού-τιμής παράγονται ως αποτέλεσμα εξόδου. Στη φάση shuffle το πλαίσιο ταξινομεί τις εξόδους της πρώτης φάσης, ομαδοποιώντας τα ζεύγη με βάση τα κλειδιά πριν στείλει το καθένα από αυτά στην επόμενη φάση. Στη φάση reduce κάθε ομάδα τιμών επεξεργάζεται και το αποτελέσμα είναι μια λίστα από νέες τιμές που συλλέγονται. Το Map Reduce παίρνει σαν είσοδο ένα σύνολο από ζευγάρια κλειδί εισόδου – τιμή και να παράγει στην έξοδο ένα σύνολο από ζευγάρια κλειδί εξόδου– αποτέλεσμα.. Οι χρήστες χρειάζεται μόνο να εφαρμόσουν τις διεπαφές των λειτουργιών map και reduce και μπορούν να αφήσουν στο σύστημα που υιοθετεί το Map Reduce να διαχειριστεί όλες τις επικοινωνίες δεδομένων και την παράλληλη επεξεργασία.

Ένα από τα πλεονεκτήματα του Map Reduce είναι η απλότητά του αφού επιτρέπει την παράλληλη εκτέλεση και επανεκτέλεση μεγάλων υπολογισμών προσφέροντας ταυτόχρονα ανοχή στα σφάλματα [22]. Αυτό

είχε ως αποτέλεσμα να παρακινήσει πολλούς μη εξειδικευμένους χρήστες να ασχοληθούν με τα Big Data. Επίσης, η σχετικά απλή αρχιτεκτονική του παρακίνησε πολλούς προγραμματιστές να αναπτύξουν προηγμένες ικανότητες όπως η υποστήριξη κατανεμημένων συστημάτων, η κατάτμηση των δεδομένων και η επεξεργασία των ροών.



Εικόνα 3: Αναπαράσταση λειτουργίας Map Reduce

Πηγή: www.towardsdatascience.com

3.1.2 Apache Hadoop

Το Hadoop είναι ένα λογισμικό ανοιχτού κώδικα που υποστηρίζει κατανεμημένη επεξεργασία μεγάλου όγκου δεδομένων (petabytes) και παρέχει μια υλοποίηση του Map Reduce. Το Hadoop βασίστηκε στο Google Map Reduce framework και το Google File System (GFS). Είναι ένα έργο του Apache Software Foundation που αναπτύσσεται και χρησιμοποιείται από ανθρώπους από όλο τον κόσμο και κυρίως την Yahoo [23].

Είναι μια από τις δημοφιλέστερες πλατφόρμες λογισμικού που υποστηρίζει κατανεμημένες εφαρμογές για Big Data. Παρέχει δυνατότητες για το χειρισμό μεγάλου όγκου δεδομένων είτε για μετατροπή τους σε μια πιο εύχρηστη δομή και μορφή, είτε για ανάλυση και εξαγωγή πολύτιμων πληροφοριών από αυτά. Σχεδιάστηκε με αρχικό στόχο την επεξεργασία δεδομένων σε δέσμες (batch data processing) με δυνατότητα κλιμάκωσης και εφαρμογής σε περιβάλλοντα με χιλιάδες μηχανήματα. Υποστηρίζοντας τόσο μεγάλα υπολογιστικά περιβάλλοντα έχει να αντιμετωπίσει και πολλά πιθανά προβλήματα. Ένα από αυτά είναι και μια πιθανή αποτυχία των

υπολογιστικών κόμβων, έχοντας τόσο μεγάλο αριθμό μηχανών σε ένα περιβάλλον.

Η στρατηγική του διαίρει και βασίλευε είναι αρκετά αποτελεσματική για διάφορα είδη φορτίων εργασίας που ασχολούνται με τεράστιες ποσότητες δεδομένων. Ένα μεγάλο ενιαίο φορτίο εργασίας μπορεί όμως να διαιρεθεί ή χαρτογραφηθεί σε μικρότερα υποφορτία και τα αποτελέσματα αυτών να συγχωνευτούν, συμπυκνωθούν και να μειωθούν ώστε να παρθεί το τελικό αποτέλεσμα. Αυτό θέλει να εκμεταλλευθεί και το Hadoop και να εκχωρήσει μικρότερα φορτία εργασίας σε ένα μεγάλο σύμπλεγμα από κόμβους δομημένους από hardware γενικού σκοπού, αντί για κάτι ακριβό, και ανεκτικό στα σφάλματα. Επιπλέον ο χειρισμός τεράστιων ποσοτήτων δεδομένων απαιτεί και την αποθήκευση μεγάλου όγκου δεδομένων.

Αποτελείται από τα εξής μέρη [24]:

- Το Hadoop Common Utilities που περιέχει βασικές βιβλιοθήκες και λειτουργίες που απαιτούνται από τα υπόλοιπα στοιχεία. Το Hadoop Distributed File System (HDFS) που διαχειρίζεται την αποθήκευση κατανεμημένων δεδομένων.

- Το Hadoop YARN Framework, το οποίο αποτελεί μία πλατφόρμα διαχείρισης πόρων. Είναι υπεύθυνο για τη διαχείριση των υπολογιστικών πόρων σε συστάδες και για τον προγραμματισμό των εφαρμογών των χρηστών.

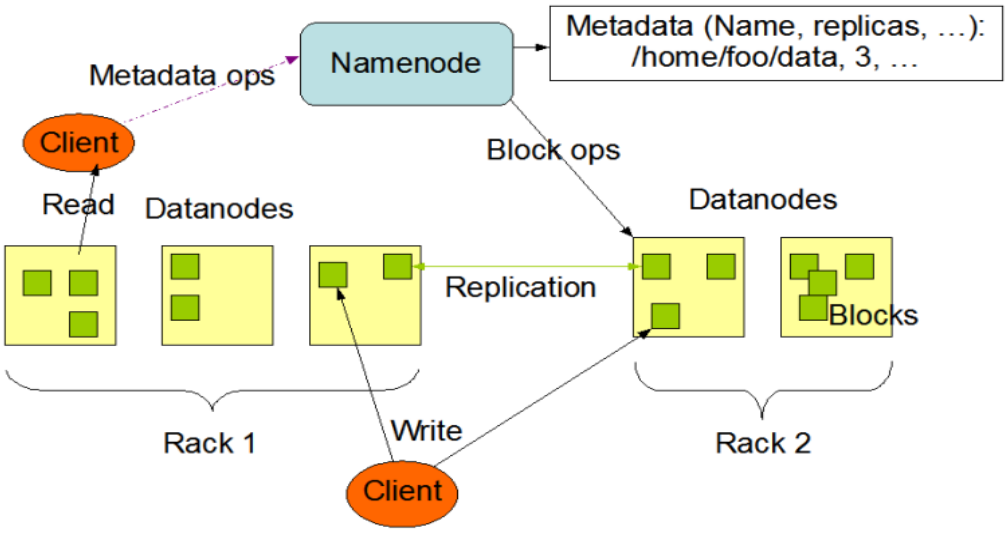
- Το Hadoop Map Reduce που αποτελεί υλοποίηση του μοντέλου Map Reduce για κατανεμημένη επεξεργασία μεγάλης κλίμακας δεδομένων.

Ένα υπολογιστικό σύστημα που εκτελεί την εφαρμογή Hadoop αποτελείται από υπολογιστικές συστάδες (clusters) οι οποίες απαρτίζονται από εμπορικό υλικό (commodity hardware). Η δομή του Hadoop βασίζεται στην υπόθεση ότι οι αστοχίες υλικού (hardware failures), δηλαδή οι δυσλειτουργίες στα ηλεκτρονικά στοιχεία των υπολογιστικών συστημάτων - είναι συχνές κατά τη διαχείριση μεγάλου όγκου δεδομένων και οφείλει η ίδια η εφαρμογή να τις διαχειρίζεται αποδοτικά.

Το Hadoop έχει ένα κατανεμημένο σύστημα αρχείων, γνωστό ως Hadoop Distributed File System (HDFS) που βοηθάει στην αποθήκευση μεγάλων ποσοτήτων δεδομένων [25]. Το Hadoop χωρίζει τα δεδομένα σε μεγάλα τμήματα (blocks) και τα κατανέμει μεταξύ διαφόρων υπολογιστικών κόμβων που συνιστούν το υπολογιστικό σύστημα. Στη συνέχεια, μεταφέρει τον κώδικα που πρόκειται να εκτελεστεί στους κόμβους ώστε να πραγματοποιηθεί παράλληλη, δηλαδή ταυτόχρονη επεξεργασία των δεδομένων στους κόμβους αυτούς. Ουσιαστικά, διενεργείται αξιοποίηση της ιδιότητας της τοπικότητας των δεδομένων (data locality) και οι κόμβοι διαχειρίζονται τα επιμέρους δεδομένα στα οποία έχουν πρόσβαση.

Αναλυτικότερα, τα δεδομένα εισόδου ρέουν ή φορτώνονται στο HDFS και επεξεργάζονται από το πλαίσιο του Map Reduce και ό,τι αποτελέσματα προκύψουν αποθηκεύονται πάλι πίσω στο HDFS. Τα αρχικά δεδομένα εισόδου δεν μεταβάλλονται κατά τη διάρκεια της ζωής τους στο HDFS. Για να αυξηθεί η αξιοπιστία και η διαθεσιμότητα των δεδομένων στο HDFS, τα δεδομένα που έχουν εκχωρηθεί σε ένα κόμβο αναπαράγονται μεταξύ των άλλων κόμβων κάτι που βοηθά στο να διασφαλιστεί ότι θα μπορούν να επιβιώσουν από πιθανή αποτυχία ή μη διαθεσιμότητα κάποιου κόμβου.

Το σύστημα αρχείων HDFS χρησιμοποιεί ένα κεντρικό κόμβο, τον name node, ο οποίος είναι και το μοναδικό σημείο αποτυχίας (single point of failure), που κρατά τις πληροφορίες για το που βρίσκεται κάθε δεδομένο στο HDFS [25]. Αν αυτός δεν είναι διαθέσιμος τότε δεν υπάρχει πρόσβαση στο σύστημα αρχείων (βλ. Εικόνα 4). Επιπλέον χρησιμοποιεί ακόμα ένα ακόμα κόμβο, τον Secondary Namenode, ο οποίος κρατά snapshots των φακέλων του name node και μαζί με τα αρχεία ιστορικού (logs) του name node επαναφέρει το σύστημα αρχείων μετά από αποτυχία. Οι υπόλοιποι κόμβοι ονομάζονται datanodes και απλά αποθηκεύουν δεδομένα.



Εικόνα 4: Αρχιτεκτονική του HDFS

Πηγή: www.hadoop.apache.org

Τέλος, αν και η βασική δομή του Hadoop συνίσταται από τα στοιχεία που ήδη αναφέρθηκαν, συχνά χρησιμοποιούνται επεκτάσεις από την Apache που εμπλουτίζουν τις δυνατότητες του Hadoop, αναλόγως την περίσταση, οι σημαντικότερες από τις οποίες είναι: Apache HBase, Apache Pig, Apache Hive, Apache Phoenix, Apache Spark, Apache ZooKeeper, Apache Flume, Apache Sqoop, Apache Storm.

3.1.3 Yarn (Yet Another Resource Negotiator)

Αναπτύχτηκε έχοντας ως στόχο να λύσει κάποια μειονεκτήματα που είχε το Hadoop [26]. Πιο συγκεκριμένα πρόκειται για το περιορισμό της επεκτασιμότητας του MapReduce σε cluster με πάρα πολύ μεγάλο αριθμό κόμβων και τη χαμηλή απόδοση των διεργασιών επειδή το σύστημα εκτελεί παράλληλα εργασίες προγραμματισμού και παρακολούθησης των εργασιών αλλά και των υπολογιστικών πόρων

Το Yarn, διαχωρίζει τις λειτουργίες διαχείρισης των πόρων από το μοντέλο προγραμματισμού. Τα καθήκοντα του JobTracker μοιράζονται μεταξύ του διαχειριστή πόρων ο οποίος διαχειρίζεται τους πόρους του cluster και του Application Master [26]. Επίσης, χρησιμοποιεί το NodeManager αντί για το TaskTracker του Hadoop για να αντικαταστήσει τον μηχανισμό σταθερού χρόνου. Ο NodeManager δημιουργεί τα containers της εφαρμογής, επιβλέπει τη χρήση πόρων από κάθε εφαρμογή και

αναφέρει στο διαχειριστή πόρων. Ο Application Master προσπαθεί να βλέπει τη θέση των containers και την εξέλιξη που έχουν.

3.1.4 Apache Spark

Το Hadoop ενώ ήταν κάτι το πρωτοπόρο στο κομμάτι των συστημάτων επεξεργασίας δεδομένων μεγάλης κλίμακας γενικού σκοπού, έχει αρκετούς περιορισμούς απόδοσης καθώς υλοποιεί τα ενδιάμεσα αποτελέσματα πριν από κάθε βήμα. Για αυτό το λόγο αναπτύχθηκαν μερικά συστήματα με στόχο να δώσουν λύση σε αυτό το πρόβλημα απόδοσης.

Το Apache Spark είναι ένα από αυτά και αποτελεί ελεύθερο λογισμικό ανοιχτού κώδικα το οποίο δημιουργήθηκε στο Πανεπιστήμιο Berkeley, της California και στη συνέχεια παραχωρήθηκε αφιλοκερδώς στην Apache Software Foundation. Αναπτύχθηκε ως μια μεγάλη μηχανή επεξεργασίας δεδομένων γενικής χρήσης που μπορεί να χρησιμοποιηθεί για πολλές διαφορετικές περιπτώσεις [27]. Σχεδιάστηκε αρχικά για να έχει καλή απόδοση σε διαδραστικά ερωτήματα, επαναληπτικούς αλγόριθμους και επεξεργασία δεδομένων συνεχούς ροής, κάτι που δεν υποστηριζόταν καλά από το MapReduce.

Το Spark έχει μεγαλύτερες δυνατότητες σε σχέση με το Hadoop καθώς φορτώνει τα δεδομένα σε μια κατανεμημένη μνήμη. Επιπλέον προσφέρει στον προγραμματιστή μία διεπαφή (Interface) επικεντρωμένη σε μία δομή δεδομένων, γνωστή ως Ελαστικό Κατανεμημένο Σύνολο Δεδομένων (Resilient Distributed Dataset ή RDD) και πρόκειται για μια συλλογή κατανεμημένων αντικειμένων σε ένα σύνολο υπολογιστικών κόμβων η οποία διασφαλίζει αποτελεσματική διαχείριση αστοχιών υλικού, όπως ακριβώς το Hadoop [28].

Το Spark εφαρμόζει μια συνάρτηση μετασχηματισμού φίλτρου σε κάθε στοιχείο στο αρχικό RDD και επιστρέφει ως αποτέλεσμα ένα νέο RDD που περιλαμβάνει τα νέα στοιχεία που προκύπτουν από τον μετασχηματισμό. Επιπλέον ο μετασχηματισμός ένωσης του Spark επιστρέφει όλα τα στοιχεία από δύο RDD σε ένα καινούριο. Οι δομές δεδομένων RDD είναι διαχειρίσιμες μέσω των εντολών Map, Filter και Reduce οι οποίες λαμβάνουν λειτουργίες, γραμμένες σε κάποια γλώσσα προγραμματισμού υψηλού επιπέδου, και τα αποστέλλουν στους κόμβους της συστάδας. Αυτό

απλοποιεί την πολυπλοκότητα του προγραμματισμού , επειδή ο τρόπος με τον οποίο οι εφαρμογές διαχειρίζονται τα RDD είναι παρόμοιος με αυτό που διαχειρίζονται τις τοπικές συλλογές δεδομένων [29] .

Με τα RDD το Spark θυμάται την ακολουθία των ενεργειών που έγιναν για να οδηγήσουν σε ένα συγκεκριμένο σύνολο δεδομένων. Έτσι επιτυγχάνει μεγάλη ανοχή σφαλμάτων καθώς ένα σύνολο δεδομένων μπορεί να ανακατασκευαστεί σε περίπτωση που χαθεί. Αν λοιπόν ένα τμήμα από κάποιο RDD χαθεί, το συγκεκριμένο RDD έχει αρκετές πληροφορίες για το πως προήλθε από άλλα RDD και μπορεί να το ξαναφτιάξει.

Κατά τη διάρκεια της εκτέλεσης, το Spark υιοθετεί έναν μηχανισμό ο οποίος αξιολογεί τις δραστηριότητες του προγράμματος και σύμφωνα με τον οποίο τα RDD δεν υλοποιούνται πάντα αμέσως. Αντίθετα, τα δεδομένα σε αυτό δεν υπόκεινται επεξεργασία και δεν υλοποιούνται στη μνήμη μέχρι να εκτελεστεί σε αυτά μια συνάρτηση δράσης. Τότε η μηχανή αρχίζει την υλοποίηση του νέου RDD.

Αυτό που κάνει το Spark να ξεχωρίζει είναι το caching. Το caching είναι η δυνατότητα που παρέχει στις εφαρμογές να αποθηκεύουν ενδιάμεσα αποτελέσματα στη κύρια μνήμη των κόμβων ενός cluster. Αντίθετα σε άλλα μοντέλα επιτρέπεται μόνο η διατήρηση δεδομένων στη κύρια μνήμη και μόνο όταν αυτά αποτελούν αντικείμενο επεξεργασίας από κάποια διεργασία, διαφορετικά πρέπει τα δεδομένα να έρθουν από τον δίσκο κάτι που καθυστερεί όλη τη διαδικασία επεξεργασίας. Βέβαια αυτό δεν σημαίνει ότι δεν χρησιμοποιείται καθόλου ο δίσκος του κάθε κόμβου για αποθήκευση καθώς όταν τα δεδομένα ξεπερνούν το μέγεθος της μνήμης, το Spark επιλέγει να γράψει κάποια από αυτά στον δίσκο.

Για να μπορούν οι προγραμματιστές να χρησιμοποιήσουν το Spark πρέπει να γράψουν ένα πρόγραμμα οδήγησης που υλοποιεί τον έλεγχο ροής της εφαρμογής σε υψηλό επίπεδο και τρέχει διάφορες διαδικασίες παράλληλα. Για παράλληλο προγραμματισμό χρησιμοποιεί ελαστικά σύνολα δεδομένων που διανέμονται εύκολα και παράλληλες διεργασίες αυτών των συνόλων.

Το Spark αποτελείται από τα παρακάτω υποσυστήματα και επεκτάσεις [30]:

- Spark SQL: επιτρέπει ερωτήματα (queries) σε δεδομένα με χρήση SQL, σε συνδυασμό με τις γλώσσες προγραμματισμού Java, Scala, Python και R.

- Spark Streaming: επιτρέπει την επεξεργασία δεδομένων σε ροή, δηλαδή δεδομένων που εισέρχονται στο σύστημα ενώ βρίσκονται ήδη σε εξέλιξη υπολογισμοί στα προηγούμενα δεδομένα. Αυτό το χαρακτηριστικό είναι πολύ σημαντικό, καθώς στο Hadoop δεν μπορούν να προστίθενται νέα δεδομένα κατά τη διάρκεια της επεξεργασίας, αλλά πρέπει να είναι διαθέσιμο όλο το σύνολό τους όταν εκκινεί μία Map-Reduce διαδικασία. Υποστηρίζονται οι γλώσσες προγραμματισμού Java, Scala και Python.

- MLlib: είναι μία βιβλιοθήκη μηχανικής μάθησης (Machine Learning Library) η οποία δίνει τη δυνατότητα εκτέλεσης αλγορίθμων αυτού του είδους έως και 100 φορές ταχύτερα από το Hadoop.

- GraphX: παρέχει το API (Application Programming Interface) για τα δεδομένα σε μορφή γραφημάτων, επιτρέποντας μάλιστα υπολογισμούς με χρήση επαναληπτικών αλγορίθμων με αποδοτικό τρόπο.

3.1.5 Talend

Το Talend αποτελεί ελεύθερο λογισμικό ανοιχτού κώδικα. Χρησιμοποιείται για διαχείριση δεδομένων μεγάλου όγκου ενώ η υλοποίησης και οι λειτουργίες του βασίζονται στο Hadoop [31]. Το βασικότερο στοιχείο είναι το Master Data Management (MDM), το οποίο έχει τη δυνατότητα να επεξεργάζεται δεδομένα σε πραγματικό χρόνο, να αξιοποιεί άλλες εφαρμογές, να ενσωματώνει τα δεδομένα τους και να εκτελεί διάφορες διαδικασίες, όπως εκτιμήσεις της ποιότητας των Big Data.

3.1.6 IBM SPSS Modeler

Το IBM SPSS Modeler είναι μια εφαρμογή λογισμικού εξόρυξης δεδομένων(data mining) και ανάλυσης κειμένου(text analytics) από την

IBM. Χρησιμοποιείται για την κατασκευή μοντέλων πρόγνωσης και την εκτέλεση άλλων εργασιών ανάλυσης [32]. Το περιβάλλον του επιτρέπει στους χρήστες να χρησιμοποιούν αλγόριθμους στατιστικής και εξόρυξης δεδομένων χωρίς προγραμματισμό. Ένας από τους κύριους στόχους της πλατφόρμας ήταν να απαλλαγούν οι χρήστες από την περιττή πολυπλοκότητα των μετασχηματισμών των δεδομένων καθιστώντας παράλληλα τα πολύπλοκα μοντέλα πρόβλεψης πολύ εύχρηστα. Η πρώτη έκδοση ενσωμάτωνε δέντρα αποφάσεων (ID3) και νευρωνικά δίκτυα (backprop), τα οποία θα μπορούσαν να εκπαιδεύονται και χωρίς να έχουν υποκείμενη γνώση του τρόπου με τον οποίο οι τεχνικές αυτές δουλεύουν. Το SPSS Modeler χρησιμοποιείται στους παρακάτω τομείς [32]:

- Συστήματα CRM (Customer Relationship Manager)
- Βελτιστοποίηση ασφαλιστικών απαιτήσεων
- Διαχείριση κινδύνων
- Βελτίωση της ποιότητας της υγειονομικής περίθαλψης.
- Πρόβλεψη ζήτησης ή πωλήσεων
- Εκπαίδευση
- Τηλεπικοινωνίες
- Μέσα κοινωνικής δικτύωσης

3.2 NoSQL Βάσεις Δεδομένων

Οι σχεσιακές βάσεις δεδομένων αποτελούν τον πιο δημοφιλή τρόπο διαχείρισης δεδομένων για οργανισμούς και επαγγελματίες .Με την έλευση των Big Data, τα οποία χαρακτηρίζονται τόσο από το μεγάλο μέγεθος, όσο και από την ποικιλομορφία στη δομή τους, καθίσταται απαραίτητη η δυνατότητα επεξεργασίας δεδομένων σε μεγάλη κλίμακα με σκοπό την εξαγωγή ενιαίων συμπερασμάτων [33]. Στο πρόβλημα της διαχείρισης των δεδομένων αυτών, τα συστήματα που βασίζονται στις σχεσιακές βάσεις δεδομένων και δεν είναι δυνατόν να προσφέρουν λύση καθώς δεν μπορούν να χειρίζονται τέτοιου τύπου δεδομένα με τη μορφή των παραδοσιακών ερωτημάτων γραμμένα σε γλώσσα SQL.

Οι βάσεις δεδομένων που είναι κατάλληλες για αποθήκευση και χειρισμό των Big Data είναι οι NoSQL βάσεις δεδομένων καθώς παρέχουν τεχνικές δυναμικής διαχείρισης δεδομένων αυξημένης κλιμάκωσης και

ευελιξίας. Τα χαρακτηριστικά τους, τις καθιστούν ιδανικές για διαχείριση μεγάλου όγκου δεδομένων τα οποία ανανεώνονται συχνά και πολλές φορές μεταβάλλονται οι τύποι των πεδίων των δεδομένων, πέραν των ίδιων των δεδομένων. Η διαδικασία αυτή είναι εξαιρετικά χρονοβόρα και σε πολλές περιπτώσεις ανέφικτη σε σχεσιακά συστήματα διαχείρισης βάσεων δεδομένων.

Οι 10Gen, Cloudera και Amazon ήταν οι πρώτες εταιρίες που διαμόρφωσαν πλατφόρμες χειρισμού Big Data με δυνατότητα υποστήριξης του Apache Hadoop και τεχνολογιών για μη σχεσιακές βάσεις δεδομένων ενώ οι κυρίαρχες μη σχεσιακές βάσεις δεδομένων είναι οι DynamoDB. Neo4j, CouchBase, MongoDB, HBase και Cassandra. Για τον χώρο της υγείας ειδικότερα έχει επικρατήσει η χρήση της MongoDB [33].

3.2.1 Cassandra

Η Cassandra είναι μία βάση η οποία συνδυάζει την κατανεμημένη αρχιτεκτονική του Dynamo της Amazon. Είναι ανοιχτού λογισμικού γραμμένη σε γλώσσα Java και αναπτύχθηκε αρχικά από το Facebook για να υλοποιήσει το χαρακτηριστικό της αναζήτησης στο inbox των χρηστών [34]. Έχει σχεδιαστεί για να διαχειρίζεται τεράστιες ποσότητες δεδομένων σε πολλαπλά datacenters καθώς και στο cloud με ασύγχρονη αντιγραφή επιτρέποντας υψηλή απόδοση και χαμηλή καθυστέρηση για τους χρήστες. Χρησιμοποιείται σε πάρα πολλές εταιρίες ανάμεσα στις οποίες οι εξής: Facebook, Twitter, Cisco, Hulu, Reddit, CERN, eBay και Instagram.

3.2.2 Mongo DB

Η MongoDB είναι μη σχεσιακή βάση δεδομένων που βασίζεται σε έγγραφα (cross-platform document- oriented database system). Η MongoDB δεν έχει την παραδοσιακή δομή μίας σχεσιακής βάσης δεδομένων με πίνακες, αλλά χρησιμοποιεί JSON έγγραφα με δυναμικά schemas, καθιστώντας την ενσωμάτωση των δεδομένων σε ορισμένους τύπους εφαρμογών ευκολότερη και ταχύτερη [35]. Είναι γραμμένη σε C++ και είναι σχεδιασμένη για να προσφέρει υψηλή απόδοση στις εφαρμογές, επεκτασιμότητα, υψηλή διαθεσιμότητα και δυνατότητα υποβολής σύνθετων ερωτημάτων.

Αρχικά, αναπτύχθηκε από την εταιρεία 10gen. Στη συνέχεια, η εταιρεία στράφηκε σε ένα μοντέλο ανάπτυξης ανοιχτού κώδικα με την 10gen να προσφέρει εμπορική υποστήριξη και άλλες υπηρεσίες. Από τότε η MongoDB έχει υιοθετηθεί ως λογισμικό backend από μια σειρά σημαντικών ιστοσελίδων και υπηρεσιών, όπως οι eBay, Foursquare, SourceForge και The New York Times, μεταξύ άλλων. Η MongoDB είναι ίσως το πιο δημοφιλές NoSQL σύστημα βάσεων δεδομένων.

Μερικά ενδεικτικά παραδείγματα λύσεων που παρέχει η MongoDB στον τομέα της υγείας είναι:

- Έγκαιρη διάγνωση σπάνιων ασθενειών. Γίνεται εφικτό να εντοπιστούν σπάνιες ασθένειες που μπορεί να έχουν ένα κοινό σύνολο συμπτωμάτων, αλλά το κάθε ένα εξ αυτών ή κάποιο υποσύνολό τους δεν αποτελούν επίφοβη ένδειξη. Η παρατήρηση αυτή αποκτά μεγαλύτερη σημασία καθώς οι ιατροί πραγματοποιούν διαγνώσεις στηριζόμενοι κυρίως στην εμπειρία και στο ιστορικό των ασθενών που έχουν εξετάσει στο παρελθόν, καθιστώντας τη διαδικασία διάγνωσης σπάνιων ασθενειών σε πρώιμο στάδιο εξαιρετικά δύσκολη, λόγω της φύσεως της ανθρώπινης συλλογιστικής. Εφαρμογές που διαθέτουν μεγάλο σύνολο στατιστικών δεδομένων είναι πολύ εύκολο να εξάγουν δείκτες ταύτισης με ασθένειες και να προσφέρουν ένα εξαιρετικά χρήσιμο εργαλείο για το ιατρικό προσωπικό.
- Δημιουργία πλήρους προφίλ ασθενούς αποτελούμενο από όλες τις εξετάσεις που του έχουν διενεργηθεί και εξαγωγή χρήσιμων σχέσεων μεταξύ αυτών μέσω τεχνικών εξόρυξης δεδομένων.
- Άμεση γνωμάτευση σε πραγματικό χρόνο. Σε περιπτώσεις εργαστηριακών δεδομένων από εξετάσεις που διενεργούνται σε ασθενείς, είναι δυνατή η άμεση εξαγωγή ποσοτικών συμπερασμάτων από το ιατρικό προσωπικό, καθώς παρέχεται η δυνατότητα οπτικής απεικόνισης μετρήσεων πάσης φύσεως και πηγής δεδομένων σε ενιαία διαγράμματα μέσω γραφημάτων, δίχως να απαιτείται ανεξάρτητη μελέτη των επιμέρους εξετάσεων από τον θεράποντα ιατρό.

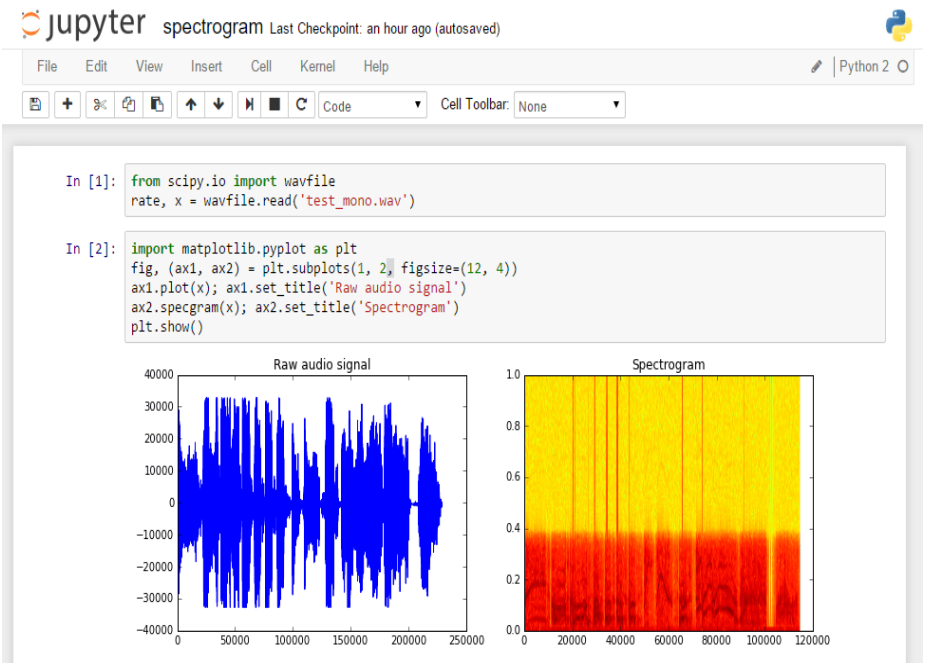
3.3 Big Data Visualization

Η συλλογή του τεράστιου όγκου των δεδομένων στην υγεία σε συνδυασμό με την ποικιλομορφία τους και τις ιδιαιτερότητες τους αναδεικνύουν ένα ακόμη σημαντικό πρόβλημα στην επεξεργασία τους. Το πρόβλημα αυτό αφορά στην κατάλληλη απεικόνιση των αποτελεσμάτων προκειμένου να είναι άμεσα κατανοητά και να βοηθούν τους ενδιαφερόμενους στην εξαγωγή συμπερασμάτων καθώς και στη λήψη αποφάσεων [36].

Ο τομέας της Οπτικής Αναλυτικής ή Visual Analytics ασχολείται με την ορθή παρουσίαση των αποτελεσμάτων από την επεξεργασία και την ανάλυση των Big Data. Ουσιαστικά αφορά στην αναλυτική συλλογιστική η οποία διευκολύνεται από προηγμένες γραφικές διεπαφές με αυξημένη αλληλεπίδραση. Ενώ η δυνατότητα συλλογής και αποθήκευσης δεδομένων αυξάνεται ραγδαία η εκμετάλλευση και η ανάλυση των Big Data έχουν καταστεί εξαιρετικά δύσκολες. Αυτό το κενό οδηγεί σε νέες προκλήσεις στην αναλυτική διαδικασία, καθώς οι αναλυτές εξαρτώνται άμεσα από την γνώση που υπάρχει στα δεδομένα. Ο τομέας της οπτικής αναλυτικής απασχολείται συστηματικά με την επεξεργασία αυτών των μαζικών, ετερογενών όγκων πληροφορίας ενσωματώνοντας στην αναλυτική διαδικασία την ανθρώπινη κρίση με την βοήθεια οπτικών αναπαραστάσεων και τεχνικών αλληλεπίδρασης [36]. Τα εργαλεία και οι τεχνικές της οπτικοποίησης (visualization), υστερούν σήμερα συγκριτικά με τους μεγάλους όγκους δεδομένων και τις πολύπλοκες δομές τους. Αυτό συμβαίνει διότι το βασικό εργαλείο ανάλυσης στην οπτική εξερεύνηση και ανάλυση είναι το ανθρώπινο μυαλό. Το ανθρώπινο μυαλό όμως έχει φυσικούς περιορισμούς ως προς το μέγεθος της πληροφορίας που μπορεί να αντιληφθεί αποτελεσματικά. Επομένως, είναι συχνά αδύνατον να οπτικοποιήσουμε όλα τα δεδομένα που χρειάζονται να αναλυθούν σε τέτοιο βαθμό, που ο αναλυτής να μπορέσει να τα αντιληφθεί, χωρίς να επέλθουν ουσιαστικές απώλειες. Δημοφιλή εργαλεία απεικόνισης μεγάλων δεδομένων είναι τα παρακάτω: JupyterR, Google Charts, Tableau, D3.js, CartoDB, Chartio κ.ά.

3.3.1 JupyterR

Το JupyterR είναι λογισμικό ανοικτού κώδικα το οποίο χρησιμοποιείται για ανάλυση μεγάλου όγκου δεδομένων (βλ. Εικόνα 5). Επίσης περιλαμβάνει και επιπλέον δυνατότητες όπως είναι ο καθαρισμός και ο μετασχηματισμός δεδομένων, η αριθμητική προσομοίωση, η στατιστική μοντελοποίηση, η μηχανική μάθηση και πολλά άλλα [37]. Υποστηρίζει πάνω από 40 γλώσσες προγραμματισμού, συμπεριλαμβανομένων των δημοφιλών γλωσσών στην Επιστήμη των δεδομένων, όπως η Python, η R, η Julia και η Scala. Ακόμα προσφέρει την δυνατότητα συγχρονισμού και κοινής χρήσης των σημειώσεων μέσω email, Dropbox, GitHub ή η του Jupyter Notebook Viewer, έχει την δυνατότητα εισαγωγής εικόνων, βίντεο, κείμενο σε Latex και JavaScript και προσφέρει διαδραστικά widgets που μπορούν να χρησιμοποιηθούν για τον χειρισμό και την οπτικοποίηση των δεδομένων σε πραγματικό χρόνο.



Εικόνα 5: Περιβάλλον εργασίας JupyterR

Πηγή:www.jupyter.org

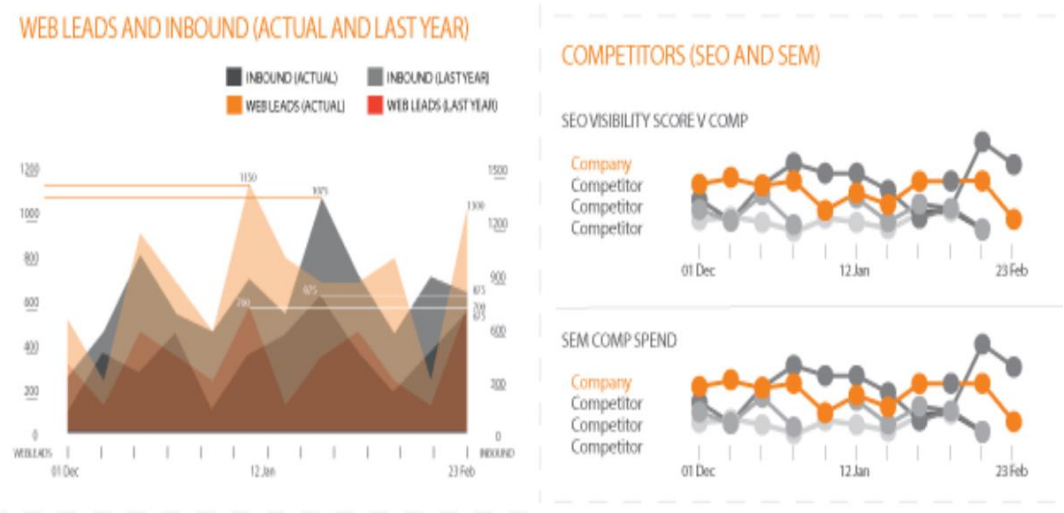
3.3.2 Google Charts

Τα Google Charts συμπεριλαμβάνονται ανάμεσα στις καλύτερες λύσεις για την απεικόνιση μεγάλων δεδομένων, για να μην αναφέρουμε ότι είναι

εντελώς δωρεάν και λαμβάνουν την συνεχή υποστήριξη από την Google.Προσφέρουν μια πληθώρα εργαλείων απεικόνισης, από διαγράμματα πίτας και χρονολογικές σειρές, μέχρι διαδοχικές πολυδιάστατες αλληλεπιδραστικές μήτρες [38]. Οι επιλογές προσαρμογής είναι πολλές και υπάρχει εκτενής ενότητα βοήθειας σε περίπτωση πολύπλοκου customization.

3.3.3 Tableau

Είναι εργαλείο απεικόνισης δεδομένων με κύρια στόχευση την εύκολη δημιουργία διαγραμμάτων, χωρίς να είναι απαραίτητη η γνώση προγραμματισμού [39]. Η μεγαλύτερη καινοτομία έγκειται στη δυνατότητα αξιοποίησης δεδομένων που βρίσκονται στο διαδίκτυο, χωρίς να είναι απαραίτητη η λήψη τους, μέσω μίας διεπαφής που παρέχει η εφαρμογή (βλ. Εικόνα 6). Γενικά, θεωρείται η πιο ευέλικτη και πλούσια σε δυνατότητες εφαρμογή για την απεικόνιση στατιστικών στοιχείων και Big Data.



Εικόνα 6: Αναπαράσταση δεδομένων

Πηγή:www.tableau.org

3.3.4 D3.js

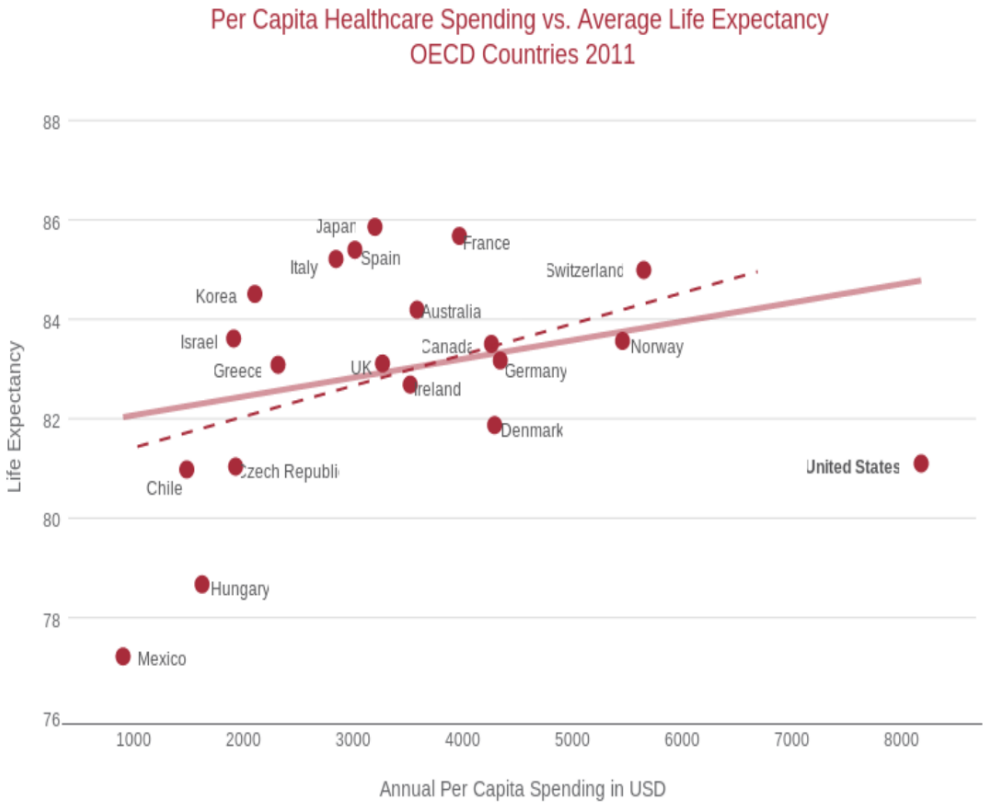
Η πλατφόρμα D3.js αποτελεί εξέλιξη του εργαλείου Protovis, ενώ επιτρέπει την διαχείριση δεδομένων μίας ιστοσελίδας. Είναι υλοποιημένη με JavaScript, χρησιμοποιεί Scalable Vector Graphics (SVG) και βασίζεται στο Document Object Model (DOM) [40]. Ουσιαστικά είναι μία βιβλιοθήκη της JavaScript που μας δίνει την δυνατότητα να διαχειριστούμε Documents βασισμένα σε δεδομένα.

3.3.5 CartoDB

Το CartoDB είναι μία εφαρμογή η οποία προσφέρει απεικόνιση μεγάλου όγκου δεδομένων μέσω της δημιουργίας χαρτών και χρησιμοποιείται κυρίως για αναπαράσταση πληροφορίας με κριτήριο την τοπικότητα των φαινομένων. Ενδείκνυται για μελέτη Big Data, καθώς διαχειρίζεται πολλών ειδών δεδομένα και τύπους αρχείων, παρέχει τη δυνατότητα ενιαίων αποτελεσμάτων και διαθέτει πρότυπα σύνολα δεδομένων τα οποία είναι εύκολα στο χειρισμό τους [41].

3.3.6 Chartio

Το Chartio (βλ. Εικόνα 7) δίνει τη δυνατότητα συνδυασμού των Big Data και της εκτέλεσης ερωτημάτων (queries) στο πρόγραμμα περιήγησης [42]. Το σημαντικότερο πλεονέκτημα της εφαρμογής αυτής είναι η ταχύτητα με την οποία μπορεί να λειτουργήσει, καθώς σε ελάχιστο χρόνο μπορεί να εισάγει δεδομένα από διαφορετικές πηγές και να τα αξιοποιήσει χωρίς να απαιτούνται γνώσεις SQL ή άλλων γλωσσών προγραμματισμού. Χρησιμοποιείται συνήθως για την εξαγωγή απλούστερων διαγραμμάτων.



Εικόνα 7: Τρισδιάστατη απεικόνιση δεδομένων

Πηγή: www.chartio.com

Κεφάλαιο 4 – Γλώσσες προγραμματισμού Big Data

4.1 Προγραμματισμός σε Big Data

Οι γλώσσες προγραμματισμού που χρησιμοποιούνται στην επεξεργασία και στην ανάλυση μεγάλου όγκου δεδομένων είναι οι Java, Scala, Python και R [43].

4.1.1 Java

Η Java είναι αντικειμενοστραφής γλώσσα προγραμματισμού η οποία καλύπτει ένα ευρύ φάσμα λειτουργιών και γνωστικών αντικειμένων της Πληροφορικής. Χρησιμοποιείται ευρέως σήμερα καθώς έχει το πλεονέκτημα ότι είναι συμβατή με οποιοδήποτε λειτουργικό σύστημα και πλατφόρμα.

4.1.2 Scala

Η Scala είναι μια σχετικά νέα γλώσσα προγραμματισμού, που συνδυάζει μοναδικά τον αντικειμενοστραφή (object-oriented) και το συναρτησιακό (functional) προγραμματισμό και εκτελείται στο ευρέως διαδεδομένο περιβάλλον μιας JVM (Java Virtual Machine). Η γλώσσα SCALA έχει σχεδιαστεί για να προσφέρει δυνατότητα επέκτασης του συστήματος καθώς και ενδυνάμωσής του αλλά και ελευθερία έκφρασης στο απεριόριστο. Η γλώσσα σε συνδυασμό με τις βιβλιοθήκες της επιτρέπει στο χρήστη να συγκεντρωθεί στον τομέα του ενδιαφέροντός του χωρίς να τον αποσπούν διάφορα προβλήματα εξαιτίας του χαμηλού επιπέδου δομής.

4.1.3 Python

Η Python είναι μια εύκολη στην εκμάθηση, ισχυρή γλώσσα προγραμματισμού. Έχει αποδοτικές δομές δεδομένων υψηλού επιπέδου και μια απλή αλλά αποτελεσματική προσέγγιση στον αντικειμενοστραφή προγραμματισμό. Η κομψή σύνταξη της Python και οι δυναμικοί τύποι της, μαζί με τη λειτουργία της ως διερμηνευόμενης (αντί μεταγλωττιζόμενης) γλώσσας, την καθιστούν την ιδανική γλώσσα για δημιουργία σεναρίων εντολών και για ταχεία ανάπτυξη εφαρμογών σε πολλούς τομείς και στις περισσότερες πλατφόρμες.

4.1.4 R

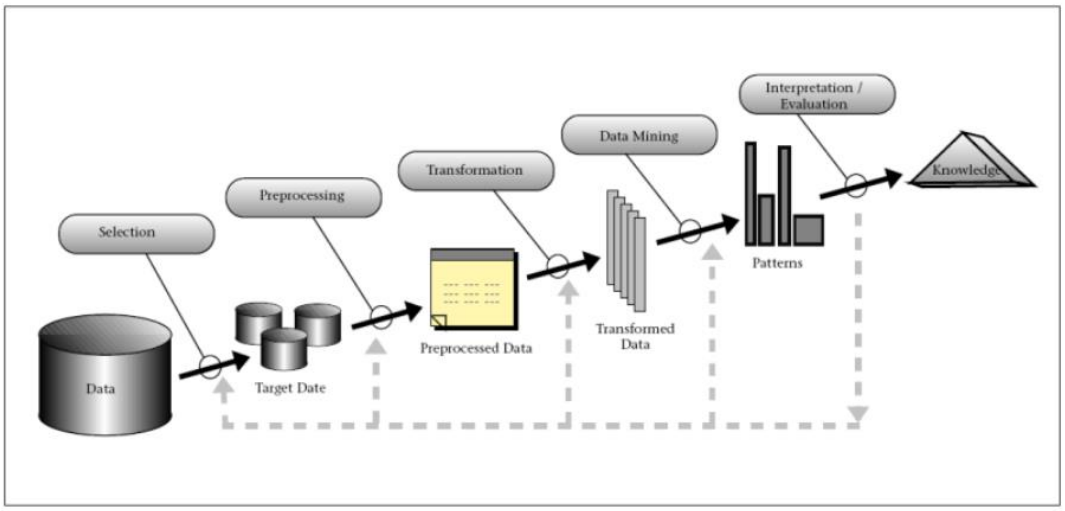
Η R είναι γλώσσα προγραμματισμού και περιβάλλον που παρέχει στον χρήστη τη δυνατότητα να κάνει υπολογιστική στατιστική και γραφήματα. Παρέχει τα απαραίτητα εργαλεία προκειμένου να υλοποιηθεί μια στατιστική ανάλυση, όπως:

- δημιουργία τυχαίων δειγμάτων
- διακριτές και συνεχείς μεταβλητές (Poisson, Gamma, Exponential)
- έλεγχοι υποθέσεων
- στατιστικά τεστ (Kolmogorov-Smirnoff)
- δημιουργία γραφημάτων (ιστόγραμμα, qq plot, pie chart, bar chart)

4.2 Εξόρυξη γνώσης

Εξόρυξη γνώσης είναι η διαδικασία εξαγωγής άγνωστης και ενδεχομένως χρήσιμης γνώσης υπό την μορφή συσχετίσεων προτύπων και τάσεων, μέσω της ανάλυσης και επεξεργασίας βάσεων δεδομένων, συνδυάζοντας και χρησιμοποιώντας τεχνικές από την μηχανική μάθηση, την αναγνώριση προτύπων, την στατιστική, τις βάσεις δεδομένων και την οπτικοποίηση [44]. Παρά το γεγονός ότι υπάρχει μια γενικότερη συμφωνία ότι ο στόχος της εξόρυξης δεδομένων είναι η ανακάλυψη νέας και χρήσιμης πληροφορίας σε βάσεις δεδομένων, τα μέσα για την επίτευξη του στόχου αυτού ποικίλουν σε πολύ υψηλό βαθμό. Τα βασικά στάδια της εξόρυξης γνώσης είναι τα εξής (βλ. Εικόνα 8):

- Συλλογή Δεδομένων (Data Collection)
- Προεπεξεργασία Δεδομένων (Preprocessing Data)
- Μετασχηματισμός Δεδομένων (Transformation)
- Εξόρυξη Δεδομένων (Data Mining)
- Διερμηνεία και Αξιολόγηση (Interpretation/Evaluation)



Εικόνα 8: Τα στάδια της εξόρυξης γνώσης

Πηγή: www.digitaltransformationpro.com

4.2.1 Συλλογή Δεδομένων

Η συλλογή των δεδομένων συνήθως γίνεται είτε αυτόματα με τη χρήση αισθητήρων από κινητές ή οποιουδήποτε άλλου τύπου συσκευές, είτε μη αυτόματα με τη χρήση ερωτηματολογίων και δειγματοληψίας. Τυχόν δυσλειτουργία στους αισθητήρες ή αδυναμία απάντησης κάποιας ερώτησης στα ερωτηματολόγια μπορεί να οδηγήσει σε ελλιπή δεδομένα [45]. Τα συγκεκριμένα προβλήματα, που ενδεχομένως να προκύψουν κατά τη συλλογή δεδομένων, αντιμετωπίζονται στο στάδιο της προεπεξεργασίας των δεδομένων.

4.2.2 Προεπεξεργασία Δεδομένων

Τα δεδομένα ενδέχεται σε αρκετές περιπτώσεις να είναι λανθασμένα ή ελλιπή. Συνεπώς είναι απαραίτητη η διόρθωση ή η απομάκρυνση των λανθασμένων δεδομένων και η απόκτηση ή πρόβλεψη των δεδομένων που είναι ελλιπή [46]. Κατά το στάδιο αυτό αντιμετωπίζονται και δύο πολύ σημαντικά προβλήματα: η ύπαρξη θορύβου και ελλιπών τιμών στα δεδομένα.

4.2.3 Μετασχηματισμός Δεδομένων

Σε αυτό το στάδιο γίνεται η μετατροπή των δεδομένων ώστε να διευκολύνουν την εξόρυξη γνώσης. Χρησιμοποιείται κυρίως για την απομάκρυνση θορύβου, για τη συνάθροιση των δεδομένων, για την κανονικοποίηση τους, δηλαδή την κλιμάκωση των χαρακτηριστικών του

συνόλου δεδομένων σε ένα συγκεκριμένο και περιορισμένο εύρος τιμών, ή τέλος για τη δημιουργία νέων χαρακτηριστικών από τα ήδη υπάρχοντα. Επίσης σε αυτό το στάδιο οι αναλυτές κάνουν το λεγόμενο “track patterns” δηλαδή ψάχνουν για μοτίβα ή ακόμη καλύτερα ψάχνουν για κοινά χαρακτηριστικά ή επαναλαμβανόμενες τιμές στα δεδομένα [47]. Η διαδικασία αυτή είναι πολύ σημαντική για τη μετέπειτα κατηγοριοποίηση των δεδομένων και την πρόβλεψη των τιμών των δεδομένων. Ειδικές μορφές μετασχηματισμού αποτελούν η διακριτοποίηση και η συμπίεση.

4.2.4 Εξόρυξη Δεδομένων

Στο στάδιο αυτό καθορίζεται το είδος της γνώσης που θα αναζητηθεί, ενώ καθορίζεται και ο αλγόριθμος που πρόκειται να χρησιμοποιηθεί. Η εφαρμογή του αλγορίθμου είναι το στάδιο στο οποίο γίνεται η ουσιαστική ανακάλυψη γνώσης από τα δεδομένα [48]. Έχοντας καθαρίσει και μετασχηματίσει τα δεδομένα, είναι έτοιμα να χρησιμοποιηθούν από κάποιον αλγόριθμο, ώστε να δημιουργηθεί κάποιο μοντέλο, συνήθως κατηγοριοποίησης ή πρόβλεψης.

4.2.5 Ερμηνεία και Αξιολόγηση

Στο στάδιο αυτό γίνεται η ερμηνεία και η αξιολόγηση των αποτελεσμάτων με τη βοήθεια γραφικών παραστάσεων ή και των δεδομένων που περιγράφονται. Επίσης κρίνεται απαραίτητη προϋπόθεση και η συμμετοχή ενός ειδικού αναλυτή δεδομένων [49]. Πρέπει να σημειωθεί ότι η διαδικασία μπορεί να περιλαμβάνει συνεχείς επαναλήψεις κι έτσι πολλά βήματα να επαναλαμβάνονται πολλές φορές ή να υπάρχουν βρόγχοι δύο διαδοχικών βημάτων, ωσότου αυτή ολοκληρωθεί επιτυχώς.

4.3 Τεχνικές Εξόρυξης Big Data στον τομέα της υγείας

Οι αλγόριθμοι που χρησιμοποιούνται στην εξόρυξη γνώσης παράγουν δύο ειδών μοντέλα: τα περιγραφικά μοντέλα (descriptive models) και τα μοντέλα πρόβλεψης(predictive models) [50]. Στόχος ενός μοντέλου πρόβλεψης είναι να προβλέψει τιμές για ένα συγκεκριμένο χαρακτηριστικό που παρουσιάζει ενδιαφέρον και που πιθανώς βασίζεται στη συμπεριφορά άλλων χαρακτηριστικών. Για παράδειγμα, η πρόβλεψη μπορεί να βασίζεται στη χρονολογική κατάταξη των δεδομένων. Ένα περιγραφικό μοντέλο

βρίσκει πρότυπα (patterns) ή σχέσεις (relations) που υπάρχουν στα δεδομένα και μελετά τις ιδιότητες τους, ώστε να δοθεί μια αιτιολόγηση της συμπεριφοράς τους.

Στο χώρο της υγείας οι τεχνικές που χρησιμοποιούνται για την εξόρυξη δεδομένων είναι η κατηγοριοποίηση (classification), η συσχέτιση (association) και ι η συσταδοποίηση (clustering). Οι τεχνικές αυτές χρησιμοποιούνται από τα νοσοκομεία και τους οργανισμούς υγείας για να αυξήσουν την πιθανότητα να καταλήξουν σε ασφαλή συμπεράσματα σχετικά με την υγεία των ασθενών.

4.3.1 Κατηγοριοποίηση

Πρόκειται για μια προγνωστική μέθοδο. Περιλαμβάνει την οργάνωση ενός συνόλου αντικειμένων (objects) που περιγράφονται από ένα σύνολο χαρακτηριστικών (attributes), σε μια σειρά από προκαθορισμένες κλάσεις (classes), χρησιμοποιώντας μεθόδους μάθησης με επίβλεψη (supervised learning methods) [51]. Οι τεχνικές της κατηγοριοποίησης χρησιμοποιούν κατά κανόνα ένα σύνολο εκπαίδευσης (training set), όπου όλα τα αντικείμενα είναι ήδη συνδεδεμένα με γνωστές κλάσεις. Ο αλγόριθμος ταξινόμησης μαθαίνει από αυτό το σύνολο, χρησιμοποιώντας τη μάθηση αυτή για την κατασκευή ενός μοντέλου και το μοντέλο αυτό στην συνέχεια ταξινομεί νέα αντικείμενα στις κατάλληλες κλάσεις.

4.3.2 Συσχέτιση

Η συσχέτιση (association) αναφέρεται στη διαδικασία της εξόρυξης γνώσης που δείχνει συσχετίσεις μεταξύ των δεδομένων. Το καλύτερο παράδειγμα αυτού του είδους της εφαρμογής είναι ο προσδιορισμός κανόνων συσχετίσεων. Ένας κανόνας συσχέτισης (association rule) είναι ένα μοντέλο που αναγνωρίζει ειδικούς τύπους συσχέτισης μεταξύ δεδομένων. Χρησιμοποιούνται για να εντοπίσουν ομοιότητες μεταξύ διαφορετικών τύπων δεδομένων και να επιτευχθεί καλύτερη κατανόηση της συμπεριφοράς των χρηστών. Οι κανόνες αυτοί συνδέουν ένα ή περισσότερα (αρχικά μη συσχετιζόμενα) γεγονότα και ανακαλύπτουν σχέσεις που δεν μπορούν εύκολα να προβλεφθούν [52].

Οι κανόνες συσχέτισης είναι ιδιαίτερα σημαντικοί στον τομέα της υγειονομικής περίθαλψης καθώς εξετάζουν τη σχέση μεταξύ μιας ασθένειας,

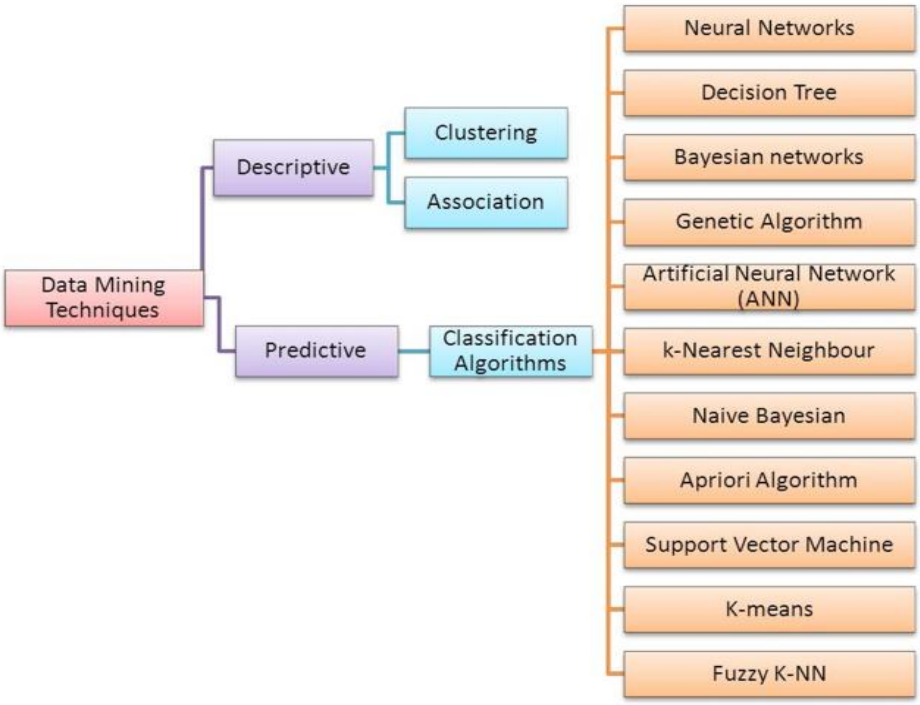
την κατάσταση της υγείας του ασθενούς καθώς και τα συμπτώματα της νόσου που εμφανίζει. Ως πηγή δεδομένων χρησιμοποιείται ο Ηλεκτρονικός Φάκελος Ασθενούς ο οποίος δίνει τη δυνατότητα μιας ενοποιημένης προσέγγισης των μεθόδων της κατηγοριοποίησης και της συσχέτισης για τη μελέτη των δεδομένων υγείας. Αυτή η προσέγγιση έχει μεγάλη επιρροή στον τομέα της υγειονομικής περίθαλψης για τον εντοπισμό των σχέσεων ανάμεσα σε διάφορες ασθένειες, την κατάσταση της ανθρώπινης υγείας και των συμπτωμάτων της νόσου.

4.3.3 Συσταδοποίηση

Η συσταδοποίηση είναι μια περιγραφική μέθοδος, διαφορετική από την ταξινόμηση καθώς δεν έχει προκαθορισμένες κατηγορίες. Μια μεγάλη βάση δεδομένων διαιρείται σε έναν μικρό αριθμό υποομάδων που ονομάζονται συστάδες. Τα δεδομένα ομαδοποιούνται με βάση τα χαρακτηριστικά τους. Ένα ιδιαίτερο χαρακτηριστικό της ομαδοποίησης, σε αντίθεση με την κατηγοριοποίηση, είναι ότι η δομή και το πλήθος των ομάδων είναι καταρχάς άγνωστα και καθορίζονται από τον εκάστοτε αλγόριθμο συσταδιοποίησης [53]. Αυτοί οι αλγόριθμοι βασίζονται στο σύνολό τους στην αρχή της μεγιστοποίησης της ομοιότητας ανάμεσα στα αντικείμενα της ίδιας ομάδας (intra-class similarity) και την ταυτόχρονη αρχή της ελαχιστοποίησης της ομοιότητας μεταξύ των αντικειμένων διαφορετικών ομάδων (inter-class similarity).

4.4 Εφαρμογές Εξόρυξης Big Data στον τομέα της υγείας

Όπως φαίνεται στο παρακάτω σχήμα (βλ. Εικόνα 9) οι διάφοροι αλγόριθμοι κατηγοριοποίησης χρησιμοποιούνται για την ανάλυση ή την πρόβλεψη ασθενειών ή για τη διαχείριση των Big Data στον τομέα της υγείας
.



Εικόνα 9: Τεχνικές εξόρυξης γνώσης

Πηγή: www.semanticscholar.com

4.4.1 Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα αποτελούν κλάδο της Τεχνητής Νοημοσύνης. Είναι υπολογιστικά συστήματα τα οποία προσομοιώνουν τη λειτουργία του ανθρώπινου εγκεφάλου. Είναι απλοποιημένα μοντέλα του ανθρώπινου νευρικού συστήματος που προσπαθούν να βρουν μοτίβα μέσα στα δεδομένα. Είναι κατάλληλα για την εύρεση μη γραμμικών μοτίβων. Τα νευρωνικά δίκτυα αποτελούνται από συνδεδεμένους μεταξύ τους κόμβους. Κάθε ένας από αυτούς δέχεται ως είσοδο ένα σύνολο από δεδομένα, εκτελεί τους απαραίτητους υπολογισμούς και παράγει το αποτέλεσμα [54].

Οι κόμβοι των νευρωνικών δικτύων είναι οι κόμβοι εισόδου, εξόδου και υπολογιστικοί. Οι πρώτοι απλώς υπάρχουν ανάμεσα στο περιβάλλον και τους υπολογιστικούς κόμβους. Οι κόμβοι εξόδου είναι αυτοί που εξάγουν στο περιβάλλον τα τελικά αποτελέσματα του δικτύου. Τέλος οι υπολογιστικοί είναι αυτοί που πολλαπλασιάζουν κάθε στοιχείο εισόδου με ένα βάρος και υπολογίζουν το άθροισμα των γινομένων.

4.4.2 Δέντρα αποφάσεων

Τα δέντρα αποφάσεων είναι στην ουσία διαγράμματα ροής με δενδρική δομή όπου τα φύλλα αναπαριστούν την πρόβλεψη των κλάσεων, τα κλαδιά

δείχνουν το αποτέλεσμα του ελέγχου, οι εσωτερικοί κόμβοι τον έλεγχο σε ένα γνώρισμα και η ρίζα του δένδρου υποδηλώνει την αρχή του δένδρου (data mining). Κάθε εσωτερικός κόμβος υποδηλώνει την επαλήθευση ενός ή περισσοτέρων γνωρισμάτων ενός συνόλου δεδομένων για κάθε δυνατό αποτέλεσμα της δοκιμής που πραγματοποιείται [55]. Η συγκεκριμένη μέθοδος χρησιμοποιεί κάποια δεδομένα εκπαίδευσης από περιπτώσεις με τις οποίες το δέντρο απόφασης δημιουργήθηκε, αρχικά για γενίκευση και αξιολόγηση της αξιοπιστίας των κανόνων που εξάγονται από το δέντρο απόφασης και κατά δεύτερο λόγο για να βελτιώσει τη συλλογή των κανόνων σε έναν ολοκληρωμένο κανόνα ο οποίος θα δώσει ένα αποτέλεσμα.

Τα αποτελέσματα που προκύπτουν από ένα δέντρο απόφασης είναι κατανοητά ακόμη και από έναν που δεν είναι ειδικός στο τομέα αυτό. Γι' αυτό η τεχνική των δέντρων αποφάσεων είναι αρκετά διαδεδομένη στο τομέα της εξόρυξης γνώσης [56]. Υπάρχουν πολλοί αλγόριθμοι για τα δέντρα απόφασης με τους πιο διαδεδομένους να είναι ο C4.5 και ο ID3.

4.4.3 Δίκτυα Bayes

Τα Δίκτυα Bayes ενσωματώνουν αλγόριθμους ταξινόμησης με την υπόθεση της ανεξαρτησίας μεταξύ κάθε ζεύγους χαρακτηριστικών. Μπορούν να εκπαιδευτούν πολύ αποτελεσματικά καθώς μέσα σε ένα μόνο πέρασμα στα δεδομένα εκπαίδευσης, υπολογίζουν τη δεσμευμένη κατανομή πιθανότητας για κάθε χαρακτηριστικό δεδομένης ετικέτας [57].

Ο κατηγοριοποιητής Naive Bayes αποτελεί μια απλουστευμένη εκδοχή των δικτύων Bayes που βασίζονται στην εφαρμογή του θεωρήματος Bayes. Παραμένει μια δημοφιλής βασική μέθοδος κατηγοριοποίησης [58]. Με κατάλληλη προ επεξεργασία των δεδομένων μπορεί να γίνει αρκετά ανταγωνιστικός ακόμα και με πιο ανεπτυγμένες μεθόδους στον τομέα αυτό συμπεριλαμβανομένου και του αλγορίθμου support vector machines. Γενικά υποθέτει ότι η επίδραση ενός γνωρίσματος σε μία κατηγορία είναι ανεξάρτητη από τις τιμές των άλλων γνωρισμάτων. Τα χαρακτηριστικά των Naive Bayes κατηγοριοποιητών είναι τα παρακάτω [59]:

- Εύρωστοι σε απομονωμένα σημεία θορύβου
- Μπορούν να διαχειριστούν ελλιπείς τιμές

- Είναι αρκετά εύρωστοι στην περίπτωση παρουσίας μη σχετικών χαρακτηριστικών.

Αποτελεί μια απλή τεχνική για την κατασκευή ταξινομητών, μοντέλων δηλαδή που ορίζουν ετικέτες κλάσης σε οντότητες προβλημάτων που αναπαρίστανται από διανύσματα με τιμές χαρακτηριστικών, όπου οι ετικέτες αυτές καθορίζονται από ένα πεπερασμένο σύνολο .Δεν είναι ένας μεμονωμένος αλγόριθμος για την εκπαίδευση τέτοιων ταξινομητών, αλλά μια οικογένεια αλγορίθμων που βασίζονται σε ένα κοινό πρότυπο.

4.4.4 Κανόνες Συσχέτισης – Αλγόριθμος Apriori

Οι κανόνες συσχέτισης είναι μία από βασικότερες τεχνικές εξόρυξης μεγάλου όγκου δεδομένων. Σκοπός της είναι η εύρεση των σημαντικών αλληλοεξαρτήσεων μεταξύ των διαφόρων χαρακτηριστικών του συνόλου δεδομένων [60]. Πρόκειται για τους κανόνες βάση των οποίων εκφράζονται συσχετίσεις μεταξύ των αντικειμένων ενός συνόλου δεδομένων και έχουν τη μορφή A->B. Εξετάζεται δηλαδή κατά πόσο η εμφάνιση του συνόλου A θα οδηγήσει στην εμφάνιση ενός συνόλου B. Το συμπέρασμα που εξάγεται είναι ότι τα δύο σύνολα συσχετίζονται μεταξύ τους καθώς η ύπαρξη του ενός οδηγεί στην ύπαρξη του άλλου.

Μέσω του data mining προκύπτει ένας τεράστιος όγκος από κανόνες συσχέτισης. Είναι σημαντικό να οριστεί, ποιοι κανόνες από αυτούς είναι σημαντικοί. Ένας κανόνας συσχέτισης θεωρείται ικανοποιητικός όταν προσφέρει σημαντική γνώση στον ερευνητή. Για να καθοριστεί ποιο κανόνες είναι σημαντικοί και προσφέρουν γνώση, χρησιμοποιούνται τα μέτρα ενδιαφέροντος (interesting measures).

Στον τομέα της υγείας οι κανόνες συσχέτισης χρησιμοποιούνται για εξόρυξη και ανακάλυψη νέας γνώσης παρά για την πρόβλεψη αποτελεσμάτων. Συγκεκριμένα χρησιμοποιούνται για [61]:

- Την αναζήτηση και τον εντοπισμό συχνά εμφανιζόμενων patterns στο DNA.
- Την αναζήτηση αλληλουχιών πρωτεϊνών στην ανάλυση καρκινικών δεδομένων.
- Την ανακάλυψη πιθανής οικονομικής απάτης και υπέρογκων χρεώσεων σε σχέση με ιατρικές και ασφαλιστικές δαπάνες.

Τα δομικά στοιχεία ενός κανόνα συσχέτισης είναι τα παρακάτω:

- I={i1,i2,….,in} είναι το σύνολο από διακριτά στοιχεία (Items).
- Το Στοιχειοσύνολο (Itemset) είναι ένα υποσύνολο του I και k-itemset είναι ένα Στοιχειοσύνολο με k στοιχεία.
- T={t1,t2,….,tn} είναι ένα σύνολο από συναλλαγές (transactions) όπου κάθε tn αποτελεί ένα Itemset και ισχύει ότι T⊆I.

ΟρισμόςX→Y:

Ένας κανόνας συσχέτισης εκφράζεται με τη σχέση X→Y όπου X και Y είναι Στοιχειοσύνολα και ισχύει ότι X⊆I, Y⊆I και X∩Y=∅.
Το X ονομάζεται LHS (Left Hand Side) ή αλλιώς «προηγούμενο» του κανόνα
Το Y ονομάζεται RHS (Right Hand Side) ή αλλιώς «επακόλουθο» του κανόνα.

Τα χαρακτηριστικά γνωρίσματα που καθιστούν έναν κανόνα σημαντικό είναι:

- Ο κανόνας που προκύπτει να έχει ικανοποιητικό βαθμό βεβαιότητας.
- Ο κανόνας που προκύπτει να είναι χρήσιμος.
- Ο κανόνας που προκύπτει να προσδίδει επιπλέον πληροφορία.
- Ο κανόνας που προκύπτει να είναι εύκολα κατανοητός από τον άνθρωπο.

Τα μέτρα ενδιαφέροντος καθορίζουν το πόσο σημαντικός και ενδιαφέρον είναι ένας κανόνας που προκύπτει από την εξόρυξη δεδομένων. Η σημαντικότητα ενός κανόνα καθορίζεται μέσω των στατιστικών μεταβλητών Support και Confidence.

Η μεταβλητή Support μετράει πόσο συχνά εμφανίζεται ένα Itemset στα δεδομένα

$$\text{Support(X)} = \frac{count(X)}{N}$$

όπου N είναι το σύνολο των δοσοληψιών.

Confidence (Εμπιστοσύνη) ενός κανόνα συσχέτισης X→Y, αποτελεί μια πιθανότητα υπό συνθήκη, P(Y|X), δηλαδή την πιθανότητα μια συναλλαγή που περιέχει το X να περιέχει επίσης και το Y.

$$Confidence = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Με τον όρο Συχνό Στοιχειοσύνολο (Frequent Itemset) χαρακτηρίζονται σύνολα από Στοιχεία (Items) που εμφανίζονται συχνά μαζί σε ένα σύνολο συναλλαγών. Σε ένα Frequent Itemset το support είναι μεγαλύτερο από ένα ελάχιστο κατώφλι (min support threshold) το οποίο μεταβάλλεται ανάλογα με τη φύση του προβλήματος.

Ερμηνεία μέτρων ενδιαφέροντος

- Μικρή τιμή του μέτρου support σημαίνει ότι ο κανόνας έχει μικρό ενδιαφέρον, καθώς αφορά σε ένα μικρό αριθμό συναλλαγών.
- Με τη βοήθεια του μέτρου support μπορεί να εξαιρεθούν κανόνες με μικρό ενδιαφέρον.
- Ένας κανόνας με μικρό support, υπάρχει πιθανότητα να εμφανίζεται τυχαία.
- Το μέτρο confidence μετρά την αξιοπιστία.
- Όσο μεγαλύτερο είναι το μέτρο του confidence, τόσο μεγαλύτερη είναι η πιθανότητα εμφάνισης του στοιχειοσυνόλου Υ σε κανόνα που περιέχει το στοιχειοσύνολο Χ.
- Κανόνες που προέρχονται από το ίδιο σύνολο, έχουν το ίδιο support.

Οι υποκατηγορίες των κανόνων συσχέτισης είναι οι εξής:

- Ποσοτικοί κανόνες συσχέτισης: Περιγράφουν συσχετίσεις μεταξύ ποσοτικών αντικειμένων.
- Boolean κανόνες συσχέτισης: Αναφέρεται στην ύπαρξη ή μη ενός αντικειμένου σε έναν κανόνα συσχέτισης.
- Κανόνες συσχέτισης επιπέδου: Προκύπτουν από την ύπαρξη ιεραρχικών επιπέδων ενός Item.
- Κανόνες συσχέτισης μονής ή πολλών διαστάσεων: Διαχωρίζονται με βάση τον αριθμό των ιδιοτήτων που περιλαμβάνουν.

Το πρόβλημα της εύρεσης κανόνων συσχέτισης εστιάζεται στην εύρεση όλων των κανόνων που έχουν μία καθορισμένη από το χρήστη ελάχιστη τιμή support και confidence. Χρησιμοποιώντας για είσοδο ένα σύνολο από Τ συναλλαγές, λαμβάνονται σαν έξοδος όλοι οι κανόνες που έχουν support μεγαλύτερο από ένα κατώφλι (min_support) και confidence support μεγαλύτερο από ένα κατώφλι (min_confidence) [62]. Οι τιμές των κατωφλίων έχουν οριστεί εκ των προτέρων. Για την εύρεση κανόνων συσχέτισης ακολουθούνται τα εξής βήματα:

- Παράγονται όλοι οι πιθανοί κανόνες συσχέτισης.
- Υπολογίζεται το support και το confidence για κάθε έναν κανόνα που έχει παραχθεί.
- Εξαιρούνται οι κανόνες με μικρότερο support και confidence από τα κατώφλια min_support και min_confidence που έχουν οριστεί.

Αν υπάρχουν n διαφορετικά στοιχεία (Items) τότε ισχύει ότι:

- Ο συνολικός αριθμός στοιχειοσυνόλων θα είναι 2^n.
- Ο συνολικός αριθμός των πιθανών κανόνων συσχέτισης είναι: 3n - 2n+1+1.

Ο αλγόριθμος Apriori ανήκει στην κατηγορία αλγορίθμων εύρεσης Boolean κανόνων συσχέτισης μέσω της εξόρυξης στοιχειοσυνόλων με μεγάλη συχνότητα εμφάνισης. Βασική αρχή του Apriori είναι ότι όλα τα υποσύνολα ενός Συχνού Στοιχειοσυνόλου (Frequent Itemset) είναι επίσης συχνά. Για παράδειγμα αν ένα στοιχειοσύνολο {AB} είναι συχνό τότε τα υποσύνολα {A} και {B} είναι επίσης συχνά [63].

Για να εξαχθούν κανόνες συσχέτισης αρχικά εντοπίζονται τα συχνά Itemsets Στη συνέχεια καθορίζεται ένα διάστημα ή κατώφλι υποστήριξης(min_sup) και ένα κατώφλι εμπιστοσύνης (min_conf) ώστε να εξαχθούν κανόνες συσχέτισης που είναι σημαντικοί.

Αρχικά, διατρέχοντας τη βάση δεδομένων δημιουργείται το σύνολο των υποψηφίων Itemsets, C1. Όσα από αυτά ικανοποιούν τον περιορισμό του min_sup αποτελούν το σύνολο L1 των συχνών Itemsets. Η διαδικασία αυτή επαναλαμβάνεται έως ότου καθοριστεί ένα σύνολο Lk με συχνά Itemsets μεγέθους k (k-Itemsets). Η επαναληπτική αυτή διαδικασία σταματά στο Lk,

αν δεν υπάρχει κάποιο k+1-Itemset που να ικανοποιεί τον περιορισμό min_sup.

Ο αλγόριθμος Apriori βασίζεται σε δύο βήματα που εφαρμόζονται σε κάθε επανάληψη, το βήμα Συνένωσης (Join step) και το βήμα Κλαδέματος (Pruning step). Κατ' επέκταση το Lk προκύπτει αφού εφαρμοστεί αρχικά το βήμα συνένωσης στα σύνολα Lk-1 και Lk-1 και στη συνέχεια το βήμα κλαδέματος στο σύνολο που θα προκύψει [64].

- Βήμα συνένωσης: Στο βήμα αυτό προκύπτει το σύνολο Ck των υποψήφιων Itemsets, από τη συνένωση του Lk-1 με τον εαυτό του.
- Βήμα Κλαδέματος: Από το σύνολο Ck υπολογίζεται το support του κάθε k-στοιχειοσυνόλου και απορρίπτονται τα k στοιχειοσύνολα των οποίων το support δεν ικανοποιεί τον περιορισμό του min_sup (δηλαδή support < min_sup). Έτσι προκύπτει το σύνολο Lk όλων των frequent itemset.

4.4.5 Λογιστική Παλινδρόμηση (Logistic regression)

Το μοντέλο της Λογιστικής παλινδρόμησης (logistic regression) αποτελεί ειδική περίπτωση των γενικευμένων γραμμικών μοντέλων [65]. Η Λογιστική Παλινδρόμηση είναι μία τεχνική σχεδιασμένη για την πραγματοποίηση ανάλυσης δεδομένων που αφορούν την μελέτη και την πρόβλεψη τιμών κάποιας κατηγορικής εξαρτημένης μεταβλητής και χρησιμοποιεί ποσοτικές και ποιοτικές ανεξάρτητες μεταβλητές. Είναι μια γενίκευση της απλής γραμμικής παλινδρόμησης για την περίπτωση όπου η εξαρτημένη μεταβλητή(Υ)είναι δίτιμη (δηλαδή παίρνει την τιμή 0 όταν απουσιάζει το χαρακτηριστικό ή την τιμή 1 όταν υπάρχει το χαρακτηριστικό) [66].

Χρησιμοποιείται σε αναλύσεις στους παρακάτω τομείς:

- ✓ της υγείας για την μελέτη της θεραπείας ή όχι των ασθενών,
- ✓ του Marketing για την αγορά ή όχι κάποιων προϊόντων,
- ✓ της παιδείας για την επιτυχία ή όχι των μαθητών στις εξετάσεις.

Η εξίσωση της Λογιστικής Παλινδρόμησης

Η εξίσωση της Λογιστικής Παλινδρόμησης είναι $y=a+b_1x_1+b_2x_2....+b_kx_k$ όπου:

- b_1, b2, Bk: οι συντελεστές των ανεξάρτητων μεταβλητών στην εξίσωση της παλινδρόμησης.
- x_1,x_2,x_k:οι ανεξάρτητες μεταβλητές
- y: η εξαρτημένη μεταβλητή.

Βήματα εκτέλεσης Λογιστικής Παλινδρόμησης

Τα βήματα για την εκτέλεση του μοντέλου της Λογιστικής Παλινδρόμησης είναι τα εξής [67]:

- Προσδιορισμός των ανεξάρτητων και της εξαρτημένης μεταβλητής.
- Διερεύνηση των δεδομένων για πιθανή ύπαρξη ακραίων ή ελλειπουσών τιμών.
- Έλεγχος της ικανοποίησης των υποθέσεων για την σωστή εφαρμογή της Λογιστικής Παλινδρόμησης.
- Δημιουργία της εξίσωσης παλινδρόμησης.
- Μελέτη της επίδρασης κάθε ανεξάρτητης μεταβλητής στο μοντέλο.
- Εξέταση της ικανοποίησης των υποθέσεων του μοντέλου παλινδρόμησης και διερεύνηση της πιθανότητας κάποια συγκεκριμένη τιμή να επηρεάζει υπερβολικά τα αποτελέσματα.

4.4.6 Αλγόριθμος Κ κοντινότερων γειτόνων (K Nearest Neighbors – KNN)

Αποτελεί μία γνωστή και συχνά χρησιμοποιούμενη τεχνική κατηγοριοποίησης που στηρίζεται στη χρήση μέτρων βασισμένων στην απόσταση. Η λειτουργία του είναι η εξής: η τιμή της συνάρτησης για ένα νέο στιγμιότυπο βασίζεται αποκλειστικά και μόνο στις αντίστοιχες τιμές των k πιο κοντινών στιγμιότυπων εκπαίδευσης, τα οποία και αποτελούν τους γείτονές του [70]. Για την παραμετροποίηση του αλγορίθμου θα πρέπει να αποσαφηνιστούν τα εξής:

- Ο ορισμός της απόστασης μεταξύ δύο στιγμιότυπων, δηλαδή

μιας τιμής πάνω στο χώρο των στιγμιότυπων, που θα εκφράζει την εγγύτητα, ή αλλιώς την ομοιότητα μεταξύ των στιγμιότυπων.

- Η τιμή του k.

Για το πρώτο ζήτημα, υπάρχουν πολλές εναλλακτικές επιλογές. Η απόφαση εξαρτάται από τα ειδικά χαρακτηριστικά του χώρου στιγμιότυπων του προβλήματος. Ιδιαίτερη σημασία έχει αν στην αναπαράσταση των στιγμιότυπων περιλαμβάνονται αριθμητικά ή συμβολικά χαρακτηριστικά [71].

Για να προσδιοριστεί ποιο σύνολο εκπαίδευσης είναι πλησιέστερο σε ένα νέο στιγμιότυπο χρησιμοποιείται κατά κύριο λόγο η Ευκλείδεια απόσταση. Πιο συγκεκραμένα, αν τα στιγμιότυπα αναπαρίστανται ως διανύσματα από χαρακτηριστικά που παίρνουν τιμές πραγματικούς αριθμούς, δηλαδή το στιγμιότυπο x αναπαρίσταται από το διάνυσμα: <a1(x), a2(x),…,an(x)>, όπου ar(x) δηλώνει την τιμή του r-οστού χαρακτηριστικού του x, τότε η απόσταση d(xi,xj) μεταξύ δύο στιγμιότυπων xi και xj ορίζεται ως: d(xi,xj)= $\sqrt{\sum_{n}^{r} = 1(a_r\,(x_i)- a_r(x_j))^2}$.

Μια εναλλακτική της Ευκλείδειας απόστασης είναι η city-block μετρική και η Manhattan [72]. Η διαφορά μεταξύ των χαρακτηριστικών τιμών δεν είναι τετράγωνο, αλλά μόλις προστεθούν (μετά τη λήψη της απόλυτης τιμής).

Διαφορετικά χαρακτηριστικά γνωρίσματα μετρώνται σε διαφορετικές κλίμακες, έτσι ώστε αν ο τύπος της Ευκλείδειας απόστασης χρησιμοποιήθηκε κατευθείαν, τα αποτελέσματα από μερικά χαρακτηριστικά ίσως είναι εντελώς επισκιασμένα από άλλα που είχαν μεγαλύτερες κλίμακες μέτρησης. Άρα είναι σύνηθες να ομαλοποιούνται όλες τις τιμές των χαρακτηριστικών και να βρίσκονται στο διάστημα 0 και 1.

Όσον αφορά στα αριθμητικά χαρακτηριστικά η διαφορά ανάμεσα σε δύο τιμές είναι ακριβώς η αριθμητική διαφορά μεταξύ τους, και αυτή είναι η διαφορά που τετραγωνίζεται και προστίθεται η απόδοση της συνάρτησης. Για ονομαστικά χαρακτηριστικά που λαμβάνουν τιμές που είναι σύμβολα, η διαφορά μεταξύ δύο τιμών που δεν είναι η ίδια συχνά θεωρείται ότι είναι ένα, ενώ αν οι τιμές είναι οι ίδιες, η διαφορά είναι μηδέν. Δεν απαιτείται

κανονικοποίηση σε αυτή την περίπτωση, επειδή χρησιμοποιούνται μόνο οι τιμές 0 και 1 [73].

Ο αλγόριθμος KNN χειρίζεται τις ελλείπεις τιμές ως εξής [74]:

- για τα ονομαστικά δεδομένα, θεωρεί ότι ένα χαρακτηριστικό που λείπει είναι μέγιστα διαφορετικό από οποιαδήποτε άλλη χαρακτηριστική τιμή. Έτσι, αν μία ή και οι δύο τιμές λείπουν, ή αν οι τιμές είναι διαφορετικές, η διαφορά μεταξύ τους λαμβάνεται ως ένα. Η διαφορά είναι μηδέν μόνο αν δεν λείπουν και είναι και οι δύο το ίδιο.
- για αριθμητικά χαρακτηριστικά, η διαφορά μεταξύ δύο τιμών που λείπουν λαμβάνεται επίσης ως ένα. Ωστόσο, αν λείπει μία μόνο τιμή, η διαφορά λαμβάνεται συχνά είτε ως μέγεθος της άλλης τιμής ή ένα μικρότερο από αυτό μέγεθος, οποιοδήποτε είναι μεγαλύτερο. Αυτό σημαίνει ότι αν οι δύο τιμές λείπουν, η διαφορά τους είναι τόσο μεγάλη όσο μπορεί πιθανώς να είναι.

Η επιλογή του k είναι σημαντική για τον τρόπο λειτουργίας του συγκεκριμένου αλγορίθμου. Η επιλογή του k μπορεί να θεωρηθεί ως ένας από τους πιο σημαντικούς παράγοντες του μοντέλου ο οποίος μπορεί να επηρεάσει τόσο πολύ την ποιότητα των προβλέψεων [75]. Ένας κατάλληλος τρόπος για να "δούμε" τον αριθμό των κοντινότερων γειτονιών k είναι να σκεφτούμε τον αριθμό αυτό σαν μια παράμετρο ομαλότητας (smoothing parameter). Σε κάθε περίπτωση, μία πιθανή ή μικρή τιμή για το k οδηγεί σε πολύ μεγάλη διακύμανση όσον αφορά τις προβλέψεις. Αντίθετα αν δώσουμε στο k μεγάλη τιμή τότε οδηγούμαστε σε ένα μοντέλο με μεγάλη μεροληψία. Από τα παραπάνω προκύπτει πως το k θα πρέπει να είναι αρκετά μεγάλο ώστε να ελαχιστοποιήσει την πιθανότητα λάθους κατάταξης αλλά και αρκετά μικρό (με σεβασμό πάντα στον αριθμό των παρατηρήσεων που περιλαμβάνει το δείγμα) ώστε οι k κοντινές παρατηρήσεις να είναι αρκετά κοντά στο άγνωστο σημείο [76]. Έτσι λοιπόν και όπως με κάθε παράμετρο ομαλότητας (smoothing parameter) υπάρχει μια βέλτιστη τιμή για το k η οποία καταφέρνει να φέρει την ισορροπία μεταξύ μεροληψίας και διακύμανσης στο μοντέλο.

4.4.7 Αλγόριθμος SVM (Support Vector Machine)

Ο αλγόριθμος Support Vector Machine (SVM) είναι ένα σύνολο μεθόδων εκμάθησης που χρησιμοποιούνται για προβλήματα ταξινόμησης και παλινδρομήσης [77]. Η κύρια ιδέα του SVM είναι να κατασκευαστεί ένα υπερεπίπεδο, έτσι ώστε η απόσταση διαχωρισμού μεταξύ των θετικών και αρνητικών παραδειγμάτων να μεγιστοποιείται. Τα διανύσματα των πιο κοντινών στοιχείων στο υπερεπίπεδο αυτό είναι τα υποστηρικτικά διανύσματα (support vectors) το οποίο επιτυγχάνεται ακολουθώντας την αρχή της ελαχιστοποίησης δομικού ρίσκου (structural risk minimization). Η ιδέα της ελαχιστοποίησης του δομικού ρίσκου είναι να βρεθεί μια υπόθεση για την οποία μπορούμε να εγγυηθούμε το χαμηλότερο πραγματικό σφάλμα. Το πραγματικό σφάλμα της υπόθεσης είναι η πιθανότητα της υπόθεσης να κάνει λάθος σε ένα τυχαία επιλεγμένο παράδειγμα το οποίο δεν έχει εξεταστεί στο παρελθόν. Το πλεονέκτημα της τεχνικής αυτής είναι ότι επιτυγχάνονται καλές επιδόσεις στα προβλήματα ταξινόμησης χωρίς να ενσωματώνεται γνώση από τον τομέα του προβλήματος [78].

Αντιλαμβανόμενος τα δεδομένα εισόδου σαν δύο σύνολα διανυσμάτων σε ένα ν-διάστατο χώρο, ο αλγόριθμος SVM θα κατασκευάσει ένα ξεχωριστό διαχωριστικό επίπεδο σε αυτόν το χώρο, που θα μεγιστοποιεί την απόσταση μεταξύ των δύο συνόλων. Για τον υπολογισμό της απόστασης αυτής, κατασκευάζονται δύο παράλληλα υπερεπίπεδα, ένα σε κάθε πλευρά του διαχωριστικού υπερεπιπέδου, τα οποία “σπρώχνονται” πάνω στα δύο σύνολα δεδομένων [79]. Ιδανικά ένας καλός διαχωρισμός επιτυγχάνεται από το υπερεπίπεδο που έχει τη μεγαλύτερη απόσταση από τα γειτονικά σημεία δεδομένων και των δύο συνόλων, δεδομένου ότι σε γενικές γραμμές όσο μεγαλύτερη είναι η απόσταση τόσο καλύτερο είναι το λάθος γενίκευσης του ταξινομητή.

Ο ταξινομητής SVM ανήκει στην κατηγορία των γενικευμένων γραμμικών ταξινομητών. Ένα πλεονέκτημά τους είναι ότι μπορούν παράλληλα να ελαχιστοποιούν το εμπειρικό σφάλμα ταξινόμησης και να μεγιστοποιούν τη γεωμετρική απόσταση. Ένα ακόμη πλεονέκτημα του SVM είναι η ικανότητά του να μαθαίνει ανεξάρτητα από τις διαστάσεις του χώρου χαρακτηριστικών [80]. Ο SVM μετράει την πολυπλοκότητα των υποθέσεων με βάση την απόσταση που μπορεί να διαχωρίσουν τα στοιχεία, και όχι με

βάση τον αριθμό των χαρακτηριστικών. Αυτό σημαίνει ότι μπορούμε να γενικεύσουμε ακόμη και με την παρουσία πάρα πολλών χαρακτηριστικών, αν τα στοιχεία μας μπορούν να διαχωριστούν με ένα ευρύ περιθώριο χρησιμοποιώντας συναρτήσεις από το χώρο υποθέσεων.

Επίσης μπορεί να εφαρμοστεί αποτελεσματικά σε ένα ευρύ φάσμα προβλημάτων ταξινόμησης καθώς κλιμακώνεται σε τεράστια σύνολα δεδομένων και είναι ανεξάρτητος του τομέα του προβλήματος. Επιπλέον, μπορεί να αναπτυχθούν αποτελεσματικές συναρτήσεις πυρήνα για κάθε συγκεκριμένο πρόβλημα, προκειμένου να επιτευχθούν ακόμα καλύτερα αποτελέσματα. Ο SVM έχει πολλές επιτυχημένες εφαρμογές στον τομέα της βιοπληροφορικής, της ανίχνευσης προσώπου και αναγνώρισης χειρογράφου κειμένου καθώς και στην κατηγοριοποίηση κειμένων [81].

4.4.8 Αλγόριθμος Random Forest

Ο αλγόριθμος Random Forest είναι μια τεχνική εκμάθησης για ταξινόμηση που δημιουργήθηκε από τον Breimann και στην πράξη κατασκευάζει μια πληθώρα δέντρων αποφάσεων και έχει ως απόκριση την κλάση που παρουσιάζεται συχνότερα ως απόκριση των επί μέρους δέντρων αποφάσεων [82]. Ουσιαστικά είναι ένας ταξινομητής που αποτελείται από μια συλλογή ταξινομητών μορφής δέντρων αποφάσεων. Κάθε δέντρο απόφασης αναπτύσσεται σε σχέση με ένα τυχαίο διάνυσμα όπου κ=1….L είναι ανεξάρτητα και ισόνομα μεταξύ τους. Κάθε δέντρο “ψηφίζει” για την πιο δημοφιλή κλάση εισόδου x.

Ένα από τα πιο επιτυχημένα ευρήματα του random forest είναι η τυχαία επιλογή εισόδου (input). Ενώ επιλέγουμε τυχαία bootstrap δείγματα από το σύνολο των δεδομένων εκπαίδευσης, η τυχαία επιλογή λαμβάνει χώρα σε κάθε κόμβο του δέντρου [83]. Διαλέγουμε τυχαία ένα subset S με M χαρακτηριστικά από το αρχικό σετ v χαρακτηριστικών και ψάχνουμε από το S το καλύτερο χαρακτηριστικό για να διαχωρίσουμε τον κόμβο. Βασικό πλεονέκτημα του ταξινομητή Random Forest είναι η ικανότητά του να διαχειρίζεται αποτελεσματικά μεγάλο αριθμό ανεξάρτητων μεταβλητών και ο μικρός απαιτούμενος χρόνος εκτέλεσής του [84].

4.5 Τεχνικές για τη μέτρηση της απόδοσης των μοντέλων κατηγοριοποίησης

Για την αξιολόγηση της απόδοσης των μοντέλων κατηγοριοποίησης χρησιμοποιούνται διάφορες τεχνικές. Πριν την ανάλυσή τους ακλουθούν κάποιες κάποιοι σημαντικές έννοιες οι οποίες χρησιμοποιούνται στις μετρικές αξιολόγησης. Ειδικότερα [85]:

- True Positive (TP): είναι ο συνολικός αριθμός των θετικών κλάσεων που προβλέφθηκαν σωστά.
- True Negative (TN):είναι ο συνολικός αριθμός των αρνητικών κλάσεων που προβλέφθηκαν σωστά.
- False Positive (FP) είναι ο συνολικός αριθμός των θετικών κλάσεων που προβλέφθηκαν λανθασμένα
- False Negative (FN) είναι ο συνολικός αριθμός των αρνητικών κλάσεων που προβλέφθηκαν λανθασμένα.

Στις παρακάτω ενότητες παρουσιάζονται κάποιες οι τεχνικές για τη μέτρηση της απόδοσης των μοντέλων κατηγοριοποίησης.

4.5.1 Ορθότητα(Accuracy)

Προσδιορίζει την αναλογία του αριθμού των σωστών προβλέψεων σε σχέση με το σύνολο των περιπτώσεων [86]. Υπολογίζεται με τον τύπο:

$$Accuracy = \frac{TP + TN}{N}$$

4.5.2 Ακρίβεια (Precision)

Η ακρίβεια εκτιμά την ορθότητα των αποτελεσμάτων. Είναι η αναλογία των θετικών κλάσεων που προβλέφθηκαν σωστά προς τον συνολικό αριθμό των περιπτώσεων που κατηγοριοποιήθηκαν ως θετικές [86]. Υπολογίζεται ως εξής:

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

4.5.3 Ανάκληση (Recall)

Η Ανάκληση είναι η αναλογία των θετικών περιπτώσεων που κατηγοριοποιήθηκαν σωστά, προς τον αριθμό όλων των θετικών περιπτώσεων [86]. Υπολογίζεται με τον τύπο:

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

4.5.4 F-measure

Παρέχει μία συνολική εκτίμηση των μοντέλων, καθώς συνδυάζει δύο άλλες μετρικές, την ανάκληση και την ακρίβεια [86]. Η μετρική F-Measure στην ουσία είναι ο αρμονικός μέσος όρος της ανάκλησης και της ακρίβειας, και υπολογίζεται ως εξής:

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

4.5.5 Πίνακας σύγχυσης (confusion-matrix)

Αποτελεί ένα εργαλείο για την μέτρηση της ποιότητας των αλγορίθμων εξόρυξης γνώσης. Κάθε γραμμή και στήλη του πίνακα αντιστοιχεί σε μία κλάσης ταξινόμησης [87]. Οι γραμμές του πίνακα αντιστοιχούν στις τιμές που θα έπρεπε να προκύψουν από την εφαρμογή του αλγορίθμου εξόρυξης γνώσης ενώ οι στήλες αντιστοιχούν στις πραγματικές τιμές. Αποτελείται από δύο σειρές και δύο στήλες όπου εκεί αναφέρεται ο αριθμός των αληθώς θετικών, ψευδώς θετικών, αληθώς αρνητικών και ψευδώς αρνητικών [87]. Έτσι με αυτόν τον τρόπο έχουμε μια πιο λεπτομερή ανάλυση από την απλή αναλογία σωστών προβλέψεων. Η γενική μορφή ενός τέτοιου πίνακα φαίνεται παρακάτω.

		Predicted	
Actual Values		Negative(0)	Positive(1)
	Negative(0)	TN	FP
	Positive(1)	FN	TP

4.5.6 Διάγραμμα ROC

Ένας ακόμη τρόπος για την αξιολόγηση της αποτελεσματικότητας ενός μοντέλου ταξινόμησης είναι το διάγραμμα ROC (Receiver Operating Characteristics) όπου κριτήριο για την αποτελεσματικότητα ενός μοντέλου λοιπόν αποτελεί η μορφή της καμπύλης ROC και συγκεκριμένα το εμβαδόν της περιοχής κάτω από την καμπύλη (AUC, Area Under Curve) [88]. Έχει αποδειχθεί ότι η AUC αποτελεί καλύτερο μέτρο αξιολόγησης των αλγορίθμων κατηγοριοποίησης σε σχέση με την ακρίβεια. Μια τέλεια δοκιμασία έχει AUC= 1, ενώ μια κακή δοκιμασία έχει AUC<0.5.

Η τιμή AUC=1 επιτυγχάνεται εάν ο ταξινομητής ταξινομεί όλα τα θετικά δείγματα πάνω από όλα τα αρνητικά, ενώ AUC=0 όταν συμβαίνει το αντίθετο.Η τιμή AUC=0.5 επιτυγχάνεται σε μία σειρά από διαφορετικά σενάρια:

- Ο ταξινομητής αναθέτει την ίδια βαθμολογία-score σε όλα τα παραδείγματα δοκιμής, θετικά και αρνητικά. Σε αυτήν την περίπτωση η καμπύλη είναι η αύξουσα διαγώνιος.
- Οι κατανομές των υποδειγμάτων κάθε κλάσης είναι παρόμοιες, το οποίο έχει ως αποτέλεσμα η καμπύλη να είναι πολύ κοντά (αλλά όχι ακριβώς) στην αύξουσα διαγώνιο.
- Ο ταξινομητής δίνει στα μισά υποδείγματα μίας κλάσης την υψηλότερη κατάταξη ενώ στα άλλα μισά την χαμηλότερη.

Κεφάλαιο 5 - Μεγάλα Δεδομένα στο Εθνικό Σύστημα Υγείας

5.1 Big Data στο Εθνικό Σύστημα Υγείας

Η αποτελεσματικότητα των ιατρικών και διαχειριστικών πληροφοριών στο Εθνικό Σύστημα Υγείας μπορεί να συμβάλει θετικά στην ενίσχυση της θέσης του ασθενή, στην υποστήριξη των γιατρών και άλλων υπηρεσιών και στην ανάπτυξη των δυνατοτήτων κλινικής και διαχειριστικής διακυβέρνησης μέσα σε ένα σύνθετο και πολύπλοκο περιβάλλον όπως είναι το σύστημα υγείας [89]. Έτσι τα τελευταία χρόνια, σε όλο τον κόσμο αλλά και στη χώρα μας δίνεται ιδιαίτερη έμφαση στον τομέα της διαχείρισης και του ελέγχου των πληροφοριών του συστήματος υγείας.

Όσον αφορά στην περίπτωση της Ελλάδας, έρευνες έδειξαν ότι η συλλογή των δεδομένων είναι ικανοποιητική και ότι η επεξεργασία ιατρικών δεδομένων είναι πενιχρή από τους οργανισμούς και τους φορείς του δημόσιου τομέα. Τα δεδομένα αυτά αναφέρονται κυρίως:

- στη συνταγογράφηση φαρμακευτικών σκευασμάτων και παρακλινικών και απεικονιστικών εξετάσεων
- στη χρήση αγαθών και υπηρεσιών υγείας
- κατανάλωση φαρμάκων και ιατροτεχνολογικών προϊόντων
- ιατρικά και νοσηλευτικά δεδομένα που συνδέονται με τη νοσηρότητα
- οικονομικά και διαχειριστικά δεδομένα των μονάδων υγείας

Η μεγάλη ευκαιρία της Ελλάδας για την έναρξη χρήσης των Big Data είναι μέσω του συστήματος της ηλεκτρονικής συνταγογράφησης του οποίου η εκκίνηση για τον έλεγχο της συνταγογραφικής συμπεριφοράς του ιατρικού σώματος μετετράπη σε απαρχή συγκρότησης βάσης δεδομένων Big Data με δυνητικά θετικές επιπτώσεις στην υγεία και την ιατρική περίθαλψη. Κάτω από αυτές τις συνθήκες, είναι αναγκαίο να εξετασθεί το οργανωτικό και το διοικητικό πλαίσιο προς την κατεύθυνση ενός συστήματος διαχείρισης και ελέγχου στη συλλογή πληροφοριών και εμπειρογνωμοσύνης στην επεξεργασία, την ανάλυση, την ταξινόμηση και τη μετάδοση της έγκυρης πληροφορίας στον τομέα της υγείας. Αυτό το εγχείρημα μπορεί να επιτευχθεί

με τη συμμετοχή και συμβολή των βασικών φορέων συλλογής «παραγωγής των πληροφοριών (ΗΔΙΚΑ, ΕΟΠΥΥ, ΕΟΦ, ΕΟΦ, ΚΕΕΛΠΝΟ, ΕΣΑΝ, ΕΛΣΤΑΤ) [90].

5.2 Ευρωπαϊκή Εμπειρία

Εδώ και μία δεκαετία περίπου οι φαρμακευτικές εταιρείες καταχωρούν δεδομένα τα οποία μπορούν να χρησιμοποιήσουν ανά πάσα στιγμή τα συστήματα υγείας. Επίσης όλοι οι εμπλεκόμενοι φορείς έχουν πρόσβαση σε ένα νέο και με πολλές δυνατότητες πεδίο γνώσης. Επιπλέον πολλές εταιρείες δημιουργούν εφαρμογές που βοηθούν ασθενείς, τους ιατρούς και άλλους ενδιαφερόμενους στην εξόρυξη νέας γνώσης [91]. Ειδικότερα:

- Σκανδιναβία: διαθέτει βάσεις δεδομένων και αρχεία ασθενών για κάθε νόσημα σχεδόν (Σουηδία, Φινλανδία, Νορβηγία).
- Ηνωμένο Βασίλειο: η συνδρομή για τη βάση των δεδομένων των γενικών γιατρών είναι 500000 λίρες ετησίως για κάθε εταιρεία που θέλει να αναλύει δεδομένα για διάφορα νοσήματα
- Ουγγαρία: το ταμείο ασφάλισης υγείας διαχειρίζεται και εισπράττει τέλη συντήρησης και ανάλυσης της βάσης δεδομένων 10 εκατομμυρίων ασθενών.
- Βέλγιο: ένα δίκτυο 22 νοσοκομείων και Κέντρων Υγείας συνδέεται για ανάλυση δεδομένων, κλινικές δοκιμές κ.α.
- Δανία: Έχει πρόγραμμα Big Data analytics για την βελτίωση της ποιότητας και της ολοκληρωμένης φροντίδας ασθενών με χρόνιες παθήσεις όπως το διαβήτη και τις καρδιαγγειακές νόσους.
- Εσθονία: Οι εθνικές βάσεις δεδομένων είναι αποκεντρωμένες & κάθε κυβερνητική υπηρεσία ή επιχείρηση είναι σε θέση να επιλέξει τη πληροφορία που χρειάζεται.

5.3 Ευρωπαϊκό Πλαίσιο Ιατρικής Πληροφορίας – EMIF

Το Ευρωπαϊκό Πλαίσιο Ιατρικής Πληροφορίας, το οποίο αναφέρεται και ως EMIF (European Medical Information Framework), δημιουργήθηκε τον Ιανουάριο του 2013, με σκοπό την αξιοποίηση των Big Data για τη βελτίωση

της υγείας [92]. Βασικός του στόχος είναι η δημιουργία ενός περιβάλλοντος το οποίο να επιτρέπει την αποδοτική επαναχρησιμοποίηση των διαθέσιμων ιατρικών δεδομένων, μέσω μίας πλατφόρμας, της EMIF-Platform. Χρηματοδοτείται από την Πρωτοβουλία Ιατρικής Καινοτομίας (Innovative Medicines Initiative ή IMI) Προκειμένου να διασφαλιστεί η άμεση εφαρμογή της δράσης, παράλληλα με την εκκίνηση του προγράμματος, ορίστηκαν δύο εξειδικευμένες ερευνητικές δράσεις με σκοπό να συμβάλλουν στην ανάπτυξη του Πλαισίου:

- **EMIF-AD:** Έρευνα που συμβάλλει στην αναγνώριση και επιβεβαίωση των παραγόντων που επισπεύδουν την εμφάνιση της νόσου Alzheimer (EMIF-Alzheimer Disease).
- **EMIF-Metabolic:** Έρευνα που στοχεύει στον εντοπισμό των μεταβολικών επιπλοκών της παχυσαρκίας.

Η κυριότερη λειτουργία της EMIF-Platform είναι η διαχείριση της επαναχρησιμοποίησης των ιατρικών Big Data. Δεδομένης της ποικιλίας των πηγών δεδομένων, τα οποία ενδέχεται να έχουν χρησιμότητα, η πλατφόρμα επιτρέπει την αναγνώριση, αξιολόγηση και επιλογή των κατάλληλων πηγών δεδομένων, από τον κατάλογο δεδομένων που διαθέτει (EMIF data catalogue).

Επειδή τα δεδομένα προέρχονται από πολλές διαφορετικές πηγές και εντοπίζονται σε διαφορετικές μορφές και δομές, η πλατφόρμα δίνει έμφαση στη εναρμόνισή τους σύμφωνα με αυστηρά οριοθετημένες προδιαγραφές, ώστε να επιτρέπει την αποδοτική χρήση τους στην έρευνα. Γενικότερα, η ανάπτυξη της EMIF-Platform έγινε με στόχο να παρέχει δυνατότητες πρόσβασης στα ιατρικά δεδομένα, ανάλυσης και οπτικής απεικόνισης. Τα προγράμματα EMIF-AD και EMIF-Metabolic χρησιμοποιήθηκαν ως σενάρια χρήσης και τα αντίστοιχα αποτελέσματα διαμόρφωσαν τον τρόπο με τον οποίο πραγματοποιούνται οι επιθυμητές λειτουργίες μέσω της πλατφόρμας [93].

Έπειτα από τρία έτη λειτουργίας, πραγματοποιήθηκε αξιολόγηση της EMIF-Platform και εκδόθηκε η τρίτη έκδοση καταλόγου (EMIF Catalogue v3). Η λειτουργία των EMIF-AD και EMIF-Metabolic αποδείχθηκε καθοριστική για την αποτίμηση του καταλόγου και τη διαμόρφωσή του, ενώ προστέθηκαν νέες δοκιμές ελέγχου ώστε να απαντηθούν ερευνητικές ερωτήσεις μέσω της

εξόρυξης των κατάλληλων δεδομένων. Το αρχικό λογισμικό έχει αναβαθμιστεί με νέες προσθήκες και εργαλεία βελτιστοποίησης της ροής δεδομένων, δημιουργώντας ένα εργαλείο διαχείρισης (TASKA). Για την αποτίμηση της EMIF-Platform χρησιμοποιήθηκε το OMOP Common Model (Observational Medical Outcomes Partnership).

5.4 Open PHACTS

Το πρόγραμμα Open PHACTS (Open Pharmacological Concept Triple Store) εγκαινιάστηκε το 2013 και αποτελεί μία Ευρωπαϊκή πρωτοβουλία σύμπραξης δημόσιου και ιδιωτικού τομέα μεταξύ ερευνητών, ακαδημαϊκού χώρου, επιχειρήσεων, φαρμακοβιομηχανιών και άλλων οργανισμών με στόχο την αποδοτικότερη, οικονομικότερη και ταχύτερη ανακάλυψη νέων φαρμάκων. Στα πλαίσια του προγράμματος υπάρχει συνεργασία με 27 Πανεπιστήμια σε όλη την Ευρώπη, 6 φαρμακευτικές εταιρίες και 4 εταιρίες που δραστηριοποιούνται στο χώρο των Big Data [94]. Βασικός σκοπός είναι η εξάλειψη των εμποδίων για τη φαρμακευτική έρευνα, ενώ τα αποτελέσματα της χρήσης είναι ελεύθερα και διαθέσιμα στην πλατφόρμα GitHub.Η αξιοποίηση των δεδομένων γίνεται μέσω της εφαρμογής Open PHACTS Discovery Platform, η οποία διατίθεται δωρεάν και ενσωματώνει φαρμακευτικά δεδομένα από μεγάλο πλήθος πηγών, ενώ παρέχει εργαλεία και υπηρεσίες εξόρυξης δεδομένων, μέσω μίας φιλικής προς το χρήστη διεπαφής.

Ο μεγάλος αριθμός των διαθέσιμων βάσεων δεδομένων στον τομέα της ανάπτυξης φαρμάκων δημιουργεί την ανάγκη προσδιορισμού προτεραιοτήτων και μεθόδων για την επιλογή των κατάλληλων πληροφοριών σε ένα τεράστιο σύνολο από Big Data . Στο πλαίσιο αυτό, η πρωτοβουλία Open PHACTS υλοποιεί την αναζήτηση βάσει της σημασιολογικής βαρύτητας των ερευνητικών ερωτημάτων που πραγματοποιούνται στα πλαίσια της φαρμακευτικής έρευνας. Ο κατάλογος με τα συνηθέστερα και σημαντικότερα ερωτήματα ξεκίνησε από επιστήμονες ευρωπαϊκών φαρμακευτικών εταιρειών και στη συνέχεια η λίστα επεκτάθηκε και βελτιώθηκε συμπεριλαμβάνοντας ακαδημαϊκές έρευνες. Αρχικά, ορίστηκε ένα σύνολο 83 ερωτήσεων, οι οποία ήταν ομαδοποιημένες κατά τομέα και κατά προτεραιότητα. Όλες οι ερωτήσεις γίνονται σε φυσική γλώσσα και

απαιτούν την ενσωμάτωση τουλάχιστον δύο διαφορετικών πηγών δεδομένων.

Το πρόγραμμα Open PHACTS έχει σαφή αντίκτυπο με διάφορους τρόπους. Η πιο σημαντική συνεισφορά είναι η χρήση του συστήματος στην επιστημονική έρευνα. Αρκετές επιστημονικές δημοσιεύσεις προέρχονται από την εκτεταμένη χρήση του συστήματος, το οποίο επιτρέπει την ανάλυση δεδομένων που ήταν πολύ δύσκολο να επιτευχθεί στο παρελθόν. Πολλές φαρμακευτικές εταιρείες έχουν ενσωματώσει τα εσωτερικά τους δεδομένα μέσω του Open PHACTS, ώστε να μπορούν εύκολα να πραγματοποιήσουν ερωτήματα σε όλες τις πληροφορίες που είναι στη διάθεσή τους, τόσο στις δημόσιες όσο και στις ιδιωτικές [94].

Μια ακόμα συνεισφορά προέρχεται από την διαπίστωση ότι μεγάλες ποσότητες ποικίλων σημασιολογικών φαρμακευτικών δεδομένων μπορούν να αναλυθούν με αποδοτικό τρόπο, κάτι το οποίο επιβεβαιώνεται από σημαντικούς φορείς, όπως το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (European Bioinformatics Institute ή EBI) και εμπορικούς παρόχους όπως η Thomson-Reuters. Η επιτυχία του προγράμματος Open PHACTS έχει αποδείξει την πρακτικότητα της χρήσης των Data στη βιοϊατρική έρευνα. Μάλιστα, το γεγονός ότι οι πάροχοι επέλεξαν να προσφέρουν τα δεδομένα τους, ενισχύει την αξία της δράσης και βοηθά τα μέγιστα στην διατήρηση του συστήματος Open PHACTS.

5.5 Πρόγραμμα φροντίδας πρόωρα γεννημένων βρεφών με χρήση Big Data Analytics

Παρά το μικρό του μέγεθος, κάθε βρέφος παράγει μεγαλύτερο όγκο δεδομένων από αυτό που μπορεί να επεξεργαστεί και να αποκρυπτογραφήσει ένα νοσοκομειακό πληροφοριακό σύστημα. Στο πλαίσιο αυτό εγκαινιάστηκε το 2013 στον Καναδά το πρόγραμμα Artemis, μία ερευνητική σύμπραξη μεταξύ του Πανεπιστημίου University of Ontario Institute of Technology, της IBM, και της Παιδιατρικής Κλινικής του Τορόντο [95] Έπειτα, έλαβε τη στήριξη πολλών οργανισμών μεταξύ των οποίων του Canada Foundation for Innovation και του Canadian Institutes for Health Research. Μάλιστα το εγχείρημα έλαβε χρηματοδότηση 3 εκατομμυρίων δολαρίων ώστε να δημιουργηθεί μία εμπορική πλατφόρμα η οποία θα

διατεθεί στην αγορά για την καλύτερη φροντίδα των πρόωρα γεννημένων βρεφών.

Στα πλαίσια της προσπάθειας αυτής παρακολουθούνται επί του παρόντος περισσότερα από 1.000 πρόωρα γεννημένα βρέφη. Το περιβάλλον του Artemis, που είναι χτισμένο σε πλατφόρμα της IBM, συνδέεται με τα νοσοκομεία Women & Infants Hospital in Providence, R.I. και Children's Hospital of Fudan University in Shanghai στην Κίνα. Για τα νοσοκομεία που φροντίζουν πρόωρα βρέφη, η συνεισφορά από το πρόγραμμα Artemis είναι μεγάλη. Το εγχείρημα αυτό έχει αποδείξει μέσω της ιατρικής έρευνας ότι [95]:

- Μπορεί να συνδυάσει πληροφορίες από τους καρδιακούς παλμούς και την αναπνοή με άλλα φυσιολογικά δεδομένα για τη μείωση των ψευδών θετικών ενδείξεων (false positive) σε ελέγχους ανίχνευσης σηψαιμίας σε σύγκριση με την αποκλειστική χρήση πληροφοριών καρδιακών παλμών.
- Μπορεί να εντοπίσει αυτόματα το είδος της άπνοιας ενός βρέφους με ακρίβεια μεγαλύτερη από 98%.
- Μπορεί αυτόματα να διαχωρίζει τις καταστάσεις ύπνου και αφύπνισης στα νεογνά για να βοηθήσει τους γιατρούς να εκτιμήσουν τον τρόπο ανάπτυξης του εγκεφάλου ενός βρέφους.
- Μπορεί να παρέχει πληροφορίες σχετικά με τα επίπεδα οξυγόνου που είναι σε θέση να χρησιμοποιηθούν ώστε τα βρέφη να μην λαμβάνουν υπερβολικές ποσότητες οξυγόνου, γεγονός που μπορεί να οδηγήσει σε μόνιμη οφθαλμική βλάβη.

Το έργο χρησιμοποιεί τεχνικές εξόρυξης, κατοχυρωμένες με δίπλωμα ευρεσιτεχνίας οι οποίες έχουν σχεδιαστεί για την εξαγωγή μη τετριμμένων και δυνητικά χρήσιμων αφηρημένων πληροφοριών από μεγάλες συλλογές δεδομένων, στην περίπτωση των οποίων τα αριθμητικά δεδομένα παράγονται από συσκευές παρακολούθησης. Το πρόγραμμα Artemis χρησιμοποιεί τρία συστήματα ιατρικής συνδεσιμότητας από τα κλινικά κέντρα Capsule Tech, ExcelMedical και True Process, για την τροφοδοσία δεδομένων σε πραγματικό χρόνο σε μια βάση δεδομένων με βάση το

υπολογιστικό νέφος (cloud computing) και ένα περιβάλλον ανάλυσης που βασίζεται στην πλατφόρμα InfoSphere και στη σχεσιακή βάση δεδομένων DB2, αμφότερα προϊόντα της IBM.

5.6 IBM Watson

Ο IBM Watson είναι ένας υπερυπολογιστής (supercomputer) τον οποίο κατασκεύασε η εταιρία IBM. Πρόκειται για ένα σύνθετο υπολογιστικό σύστημα, το οποίο είναι σε θέση να δίνει απαντήσεις σε ερωτήματα διατυπωμένα σε φυσική γλώσσα. Τον Φεβρουάριο του 2011, ανακοινώθηκε ότι η IBM θα συνεργάζεται με την εταιρία Nuan Communications στα πλαίσια ενός ερευνητικού προγράμματος με στόχο την ανάπτυξη ενός εμπορικού προϊόντος εκμετάλλευσης των δυνατοτήτων κλινικής υποστήριξης αποφάσεων του Watson, προβλεπόμενη διάρκεια 18 έως 24 μήνες [96]. Οι ιατροί στο Πανεπιστήμιο Columbia θα προσπαθούσαν να εντοπίσουν τα κρίσιμα ζητήματα στην άσκηση της Ιατρικής στα οποία η τεχνολογία του Watson είναι σε θέση να συμβάλλει, ενώ οι ιατροί στο Πανεπιστήμιο του Maryland θα επιχειρούσαν προσδιορίσουν τον αποδοτικότερο τρόπο με τον οποίο η τεχνολογία του Watson θα μπορούσε αλληλεπιδράσει με τους ιατρούς ούτως ώστε να παρέχεται η καλύτερη δυνατή φροντίδα στους ασθενείς.

Ο Watson χαίρει πλέον της ευρείας αποδοχής του ιατρικού προσωπικού. Σύμφωνα με δηλώσεις του Manoj Saxena, πρώην επικεφαλής του τμήματος επιχειρήσεων υγείας της IBM, το 90% των νοσηλευτών που χρησιμοποιούν τον Watson, πλέον ακολουθούν πιστά τις προτάσεις του.

Το ερώτημα που γεννάται είναι ποιες λειτουργίες πραγματοποιεί ο Watson και με ποιο τρόπο καταφέρνει τα εκπληκτικά αυτά αποτελέσματα, αξιοποιώντας τα Big Data με τα οποία τροφοδοτείται. Η συνήθης διαδικασία που ακολουθείται είναι η ακόλουθη: Το ιατρικό προσωπικό διατυπώνει σε φυσική γλώσσα το πρόβλημα που αντιμετωπίζει, περιγράφοντας τα συμπτώματα και άλλους σχετικούς παράγοντες. Έπειτα, ο Watson εκτελεί τα εξής βήματα [96]:

- ανατρέχει στα δεδομένα του ασθενούς

- διαχωρίζει τις λέξεις που ειπώθηκαν
- επιχειρεί να εκτιμήσει τις σημαντικότερες λέξεις-κλειδιά
- αναζητά κοινά στοιχεία με το διαθέσιμο αρχείο ιατρικών δεδομένων σχηματίζει υποθέσεις
- διατυπώνει λίστα με τις πιθανές αιτίες δίνοντας μάλιστα αντίστοιχη βαρύτητα σε κάθε εκτίμηση

Οι πηγές των Big Data στις οποίες ανατρέχει ο Watson μπορεί να είναι προτεινόμενες μέθοδοι θεραπείας από τη βιβλιογραφία, σημειώσεις και καταγραφές ιατρών και νοσηλευτών, ηλεκτρονικά ιατρικά αρχεία ασθενών, κλινικές δοκιμές και έρευνες, επιστημονικά άρθρα, καθώς επίσης και πληροφορίες που παρέχονται από τους ίδιους τους ασθενείς. Αν και αναπτύχθηκε και διαφημίστηκε ως σύμβουλος διάγνωσης και θεραπείας, στην πραγματικότητα, ο Watson έχει χρησιμοποιηθεί κυρίως στην θεραπεία ασθενών που έχουν ήδη διαγνωσθεί με κάποια ασθένεια, προτείνοντας τρόπους αντιμετώπισης.

Κεφάλαιο 6 – Big Data και Προσωπικά Δεδομένα

6.1 Προστασία προσωπικών δεδομένων

Η ανάπτυξη των τεχνολογιών πληροφορικής και επικοινωνιών, η χρήση πλέον της πληροφορίας σε κάθε έργο καθώς και η απομάκρυνση από το φορέα της και τον αρχικό σκοπό της συλλογής και της επεξεργασίας της, εγείρουν σημαντικά ζητήματα σε σχέση με την προστασία της [97]. Έτσι διαμορφώνεται το αίτημα για την προστασία προσωπικών δεδομένων. Αποτελεί ένα αίτημα αναπόσπαστα συνδεδεμένο με την τεχνολογική εξέλιξη, καθώς οι υφιστάμενες ρυθμίσεις δεν προσφέρουν επαρκή προστασία έναντι των κινδύνων που συνεχώς εμφανίζονται. Η προστασία προσωπικών δεδομένων δεν περιορίζεται στην προστασία της πληροφορίας που το ίδιο το άτομο θεωρεί ως ιδιωτική, αλλά αφορά κάθε πληροφορία που αναφέρεται σε ένα φυσικό πρόσωπο [98].

Στην προστασία των προσωπικών δεδομένων δεν υφίσταται διάκριση μεταξύ των πληροφοριών σε ιδιωτικές ή απόρρητες. Έτσι η προστασία των δεδομένων είναι ευρύτερη της ιδιωτικότητας και παράλληλα στενότερη, δεδομένου ότι περιλαμβάνει και άλλα στοιχεία πέραν των προσωπικών δεδομένων [99]. Το πλαίσιο της προστασίας προσωπικών δεδομένων έχει μεγάλη ευρύτητα και δεν περιορίζεται στην εξουσιοδοτημένη ή μη χρήση της πληροφορίας. Είναι ένα σύνολο κανόνων, προϋποθέσεων, όρων, εξουσιών και απαγορεύσεων σε σχέση με τη συλλογή και επεξεργασία των προσωπικών δεδομένων.

6.2 Προστασία προσωπικών δεδομένων και Big Data

Η εκτεταμένη συλλογή και περαιτέρω επεξεργασία από τα συστήματα διαχείρισης Big Data, στα οποία συνήθως περιέχονται και προσωπικά δεδομένα, έχει προκαλέσει σοβαρές ανησυχίες και προβληματισμούς, αναδεικνύοντας το ζήτημα της προστασίας της «Ιδιωτικότητας» (Privacy) σε μείζον θέμα [100]. Ειδικότερα, οι ανησυχίες αυτές εστιάζουν σε ζητήματα που σχετίζονται με την ευρείας κλίμακας ηλεκτρονική επιτήρηση, την αποκάλυψη προσωπικών και ευαίσθητων προσωπικών δεδομένων σε

τρίτους και την κατηγοριοποίηση και αντιμετώπιση των ανθρώπων με βάση το ηλεκτρονικό τους προφίλ [101].

Προκειμένου λοιπόν να εκμεταλλευόμαστε τα πλεονεκτήματα που προκύπτουν από την χρήση των Big Data και παράλληλα να έχουμε τον μέγιστο δυνατό βαθμό προστασίας της Ιδιωτικότητας, καθίσταται επιτακτικό, αφενός να περιοριστούν τα προβλήματα που προκύπτουν από την χρήση των συστημάτων αυτών και αφετέρου να ενσωματωθούν σε αυτά οι κατάλληλοι μηχανισμοί και δικλίδες ασφαλείας, που θα εγγυώνται την προστασία της ιδιωτικής ζωής του ατόμου [102].

Οι χρήστες για να αποκτήσουν ισχυρότερα δικαιώματα έχουν ανάγκη από την πρόσβαση, τη διόρθωση, τη διαγραφή ή και την άρνηση στην επεξεργασία των δεδομένων τους σε offline και online κατάσταση. Η συναίνεση είναι ένα άλλο εργαλείο για τα άτομα να ελέγξουν ή να επηρεάσουν την συλλογή και χρήση των προσωπικών τους δεδομένων. Ωστόσο, η συναίνεση είναι ευάλωτη σε κακή χρήση, ιδίως σε απευθείας σύνδεση [102].

Οι διαχειριστές big data από την πλευρά τους, θα πρέπει να επενδύσουν στην καλύτερη προστασία της ιδιωτικής ζωής και της προστασίας των δεδομένων και ειδικότερα κατά τη φάση σχεδιασμού των έργων τους έτσι ώστε να καθορίζονται οι κανόνες και το νομικό πλαίσιο σχετικά με την προστασία των δεδομένων. Είναι ακόμη σαφές ότι οι τρέχουσες προθεσμίες διατήρησης των δεδομένων μπορούν εύκολα να αμφισβητηθούν, καθώς όλο και περισσότερα δεδομένα δημιουργούνται που θα μπορούσαν εύκολα να διατηρηθούν πάνω από τα επιτρεπτά σημερινά χρονικά όρια [103].

Επίσης, πέρα από τις γενικές αρχές θα πρέπει να θεσπιστούν και συγκεκριμένοι κανόνες οι οποίοι θα θέτουν περιορισμούς και θα αποτρέπουν οποιαδήποτε παράνομη χρήση των Big Data. Η Ευρωπαϊκή Ένωση έχει θεσπίσει νομοθεσία και απαιτεί συγκεκριμένα τη συγκατάθεση για τη χρήση ορισμένων cookies ώστε να ελέγχει τη χρήση των data της κίνησης (traffic) ή της θέσης (location). Χορηγούνται ακόμη ειδικά πρόσθετα δικαιώματα

προστασίας που παρέχουν στους πολίτες το δικαίωμα να αντιταχθούν σε διάφορες μορφές άμεσου marketing.

6.3 Big Data και GDPR στην υγεία

Είναι αδιαμφισβήτητο γεγονός η εισαγωγή των νέων τεχνολογιών σε όλους τους τομείς και απολύτως ορατές οι αλλαγές που έχει επιφέρει στον τρόπο ζωής όλων. Ο όγκος και το εύρος των προσωπικών δεδομένων που έχουν καταχωρηθεί σε μυριάδες βάσεις δεδομένων αντιστοίχως πολλών διαφορετικών εταιριών και ο στόχος της διαφύλαξης και αξιοποίησής τους προς όφελος του κοινού οδήγησε τους αρμόδιους ευρωπαϊκούς φορείς στην ψήφιση και θέση σε ισχύ του Ευρωπαϊκού Κανονισμού Προστασίας Προσωπικών Δεδομένων (GDPR) [104]. Ο νέος κανονισμός τέθηκε σε εφαρμογή στις 25 Μαΐου 2018 και όλες οι εταιρίες οι οποίες επεξεργάζονται προσωπικά δεδομένα Ευρωπαίων πολιτών θα πρέπει συμμορφωθούν με τις διατάξεις του νέου Κανονισμού.

Στον κλάδο της υγείας τα πράγματα είναι ιδιαιτέρως ευαίσθητα, καθώς η προστασία των ευαίσθητων προσωπικών δεδομένων και οι πληροφορίες όπως το ιατρικό ιστορικό, ο τρόπος ζωής και διατροφής συνθέτουν την εικόνα του ατόμου που δυνητικά θα αποτελέσει ασθενή σε κάποια φάση της ζωής του. Η καταγραφή, συγκέντρωση και επεξεργασία των στοιχείων αυτών σε μεγάλο μέρος του πληθυσμού, εξυπηρετεί με το βέλτιστο τρόπο στην πρόληψη επιδημιών, στη θεραπεία ασθενειών, στη βελτίωση της ζωής, στην πρόληψη, ακόμα και στην εξατομικευμένη φαρμακευτική αγωγή κάθε ασθενούς.

Οι ραγδαίες τεχνολογικές εξελίξεις, η ευρεία πρόσβαση στο διαδίκτυο, οι υπηρεσίες cloud computing δημιούργησαν κενά στον προηγούμενο κανονισμό, καθιστώντας τον ξεπερασμένο και αναποτελεσματικό. Στην εποχή των big data, ο κανονισμός στοχεύει στην μεγίστη προστασία των δεδομένων προσωπικού χαρακτήρα από την όποια επεξεργασία στην οποία υποβάλλονται [104].

6.3.1 Αρχή του GDPR

Τα φυσικά πρόσωπα θα πρέπει να έχουν τον έλεγχο των δικών τους δεδομένων προσωπικού χαρακτήρα. Θα πρέπει να ενισχυθούν η ασφάλεια δικαίου και η πρακτική ασφάλεια για τα φυσικά πρόσωπα, τους οικονομικούς παράγοντες και τις δημόσιες αρχές [105]. Ειδικότερα:

- Κάθε πρόσωπο έχει δικαίωμα στην προστασία των δεδομένων που το αφορά.
- Η επεξεργασία των δεδομένων θα πρέπει να προορίζεται για να εξυπηρετεί τον άνθρωπο.
- Η επίτευξη ενός ομοιογενούς χώρου ελευθερίας ασφάλειας και δικαιοσύνης, στον οποίο θα διακινούνται ελεύθερα τα προσωπικά δεδομένα, ανάμεσα στα κράτη μέλη.

6.3.2 Δεδομένα προσωπικού χαρακτήρα

- Γενετικά δεδομένα: τα δεδομένα προσωπικού χαρακτήρα που αφορούν τα γενετικά χαρακτηριστικά φυσικού προσώπου που κληρονομήθηκαν ή αποκτήθηκαν, όπως προκύπτουν, ιδίως, από ανάλυση βιολογικού δείγματος του εν λόγω φυσικού προσώπου και τα οποία παρέχουν μοναδικές πληροφορίες σχετικά με την φυσιολογία ή την υγεία του εν λόγω φυσικού προσώπου.

- Βιομετρικά δεδομένα: δεδομένα προσωπικού χαρακτήρα τα οποία προκύπτουν από ειδική τεχνική επεξεργασία συνδεόμενη με φυσικά, βιολογικά ή συμπεριφορικά χαρακτηριστικά φυσικού προσώπου και τα οποία επιτρέπουν ή επιβεβαιώνουν την αδιαμφισβήτητη ταυτοποίηση του εν λόγω φυσικού προσώπου, όπως εικόνες προσώπου ή δακτυλοσκοπικά δεδομένα.

- Δεδομένα που αφορούν την υγεία: δεδομένα προσωπικού χαρακτήρα τα οποία σχετίζονται με τη σωματική ή ψυχική υγεία ενός φυσικού προσώπου, περιλαμβανομένης της παροχής υπηρεσιών υγειονομικής φροντίδας, και τα οποία αποκαλύπτουν πληροφορίες σχετικά με την κατάσταση της υγείας του.

6.3.3 Κριτήρια νομιμότητας

Ο GDPR ορίζει έξι προϋποθέσεις “νομιμότητας” της επεξεργασίας και απαιτεί να ισχύει τουλάχιστον μία (Άρθρο 6) [106]:

- Το υποκείμενο των δεδομένων έχει συναινέσει στην επεξεργασία
- Η επεξεργασία είναι απαραίτητη για εκτέλεση σύμβασης (όπου το υποκείμενο είναι συμβαλλόμενος) ή για να ληφθούν μέτρα κατ’ αίτηση του υποκειμένου πριν τη σύναψη σύμβασης.
- Η επεξεργασία είναι απαραίτητη για τη συμμόρφωση με έννομη υποχρέωση του υπευθύνου επεξεργασίας.
- Η επεξεργασία είναι απαραίτητη για τη διαφύλαξη ζωτικού συμφέροντος (του υποκειμένου ή άλλου φυσικού προσώπου).
- Η επεξεργασία είναι απαραίτητη για τους σκοπούς των έννομων συμφερόντων που επιδιώκει ο υπεύθυνος επεξεργασίας ή τρίτος. Εξαιρούνται οι δημόσιες αρχές κατά την άσκηση των καθηκόντων τους.

6.3.4 Κριτήρια επιτυχίας του GDPR.

Για την επιτυχή εφαρμογή του νόμου GDPR στον τομέα της Υγείας είναι απαραίτητο να υπάρχει ένας πιστοποιημένος Υπεύθυνος Προστασίας Δεδομένων. Κύριες αρμοδιότητές του θα είναι:

- Η ενημέρωση των εργαζομένων και των επαγγελματιών υγείας σχετικά με τις διατάξεις του νόμου.
- Η εκπαίδευση των εργαζομένων και των επαγγελματιών υγείας.
- Η αξιολόγηση των προγραμμάτων εκπαίδευσης και των εκπαιδευόμενων.
- Ο βαθμός ετοιμότητας του νοσοκομείου, του οργανισμού ή του δημόσιου φορέα να εφαρμόσει τη νέα νομοθεσία.
- Η ανάληψη δράσεων και πρωτοβουλιών για τη νέα νομοθεσία.

Β.ΕΙΔΙΚΟ ΜΕΡΟΣ

Σκοπός της Διατριβής

Σκοπός της παρούσας διατριβής είναι η διερεύνηση της άποψης και της γνώσης των επιστημόνων της Πληροφορικής της Υγείας καθώς και των επαγγελματιών υγείας σχετικά την τεχνολογία των Big Data. Η συγκεκριμένη έρευνα στοχεύει στη διερεύνηση της αντίληψης που έχουν οι επιστήμονες της Πληροφορικής Υγείας και οι επαγγελματίες υγείας (ιατροί και νοσηλευτές) για την τεχνολογία των Μεγάλων Δεδομένων (Big Data). Επιπλέον, σκοπός της παρούσας διατριβής είναι η μελέτη τεχνικών εξόρυξης γνώσης για δεδομένα μεγάλου όγκου που αφορούν το πεδίο της Υγείας. Παράλληλα σκοπός της έρευνας είναι η μελέτη στατιστικών και υπολογιστικών αλγορίθμων ανάλυσης μεγάλου όγκου δεδομένων υγείας που έχουν ως αποτέλεσμα την παραγωγή νέας γνώσης, την εξαγωγή στατιστικά σημαντικής πληροφορίας για τους επαγγελματίες υγείας. Συγκεκριμένα θα γίνει η επιλογή του κατάλληλου συνόλου δεδομένων (datasets) έτσι ώστε να εφαρμοστούν τα κατάλληλα μοντέλα πρόβλεψης ή κατηγοριοποίησης. Στη συνέχεια χρησιμοποιώντας τους κατάλληλους αλγορίθμους θα γίνει συγκριτική ανάλυση και παρουσίαση των αποτελεσμάτων.

Κεφάλαιο 7

Εμπειρική διερεύνηση για την καταγραφή της άποψης των επιστημόνων της Πληροφορικής Υγείας σχετικά με την Τεχνολογία των Big Data (Μεγάλα Δεδομένα)

7.1 Σκοπός

Σκοπός της έρευνας που παρουσιάζεται στο παρόν κεφάλαιο είναι η καταγραφή της άποψης των επιστημόνων της Πληροφορικής Υγείας σχετικά την τεχνολογία των Big Data. Η συγκεκριμένη έρευνα στοχεύει στη διερεύνηση της αντίληψης που έχουν οι επιστήμονες της Πληροφορικής Υγείας για την τεχνολογία των Μεγάλων Δεδομένων (Big Data) και αν θεωρούν εφικτή την εφαρμογής της στο χώρο της υγείας.

Για την εμπειρική διερεύνηση της άποψης των επιστημόνων της Πληροφορικής Υγείας σχετικά με την τεχνολογία των Big Data τέθηκαν τα παρακάτω ερευνητικά ερωτήματα:

- Γνωρίζουν οι επιστήμονες της Πληροφορικής Υγείας τι είναι τα Big Data;
- Ποια είναι η πηγή ενημέρωσης των επιστημόνων της Πληροφορικής Υγείας για την τεχνολογία των Big Data;
- Ποια είναι, κατά την άποψη των επιστημόνων της Πληροφορικής Υγείας, η μορφή των Big Data;
- Γνωρίζουν περιπτώσεις χρήσης της τεχνολογίας των Big Data στην Ελλάδα ή στο εξωτερικό στον τομέα της Υγείας;
- Ποιες θεωρούν οι επιστήμονες της Πληροφορικής Υγείας ότι είναι οι κύριες πηγές για τη συλλογή μεγάλου όγκου δεδομένων που αφορούν τον τομέα της υγείας;
- Οι επιστήμονες της Πληροφορικής Υγείας θεωρούν την τεχνολογία των Big Data χρήσιμη για τον τομέα της Υγείας;
- Οι επιστήμονες της Πληροφορικής Υγείας πιστεύουν ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα μπορούσε να αυξήσει την αποτελεσματικότητα των παρεχόμενων Υπηρεσιών Υγείας;

- Οι επιστήμονες της Πληροφορικής Υγείας πιστεύουν ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα βοηθήσει τους Επαγγελματίες Υγείας στη διαδικασία λήψης αποφάσεων;
- Οι επιστήμονες της Πληροφορικής Υγείας θεωρούν τι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα δώσει τη δυνατότητα παροχής υπηρεσιών υγείας προσαρμοσμένων στις ανάγκες των ασθενών;
- Οι επιστήμονες της Πληροφορικής Υγείας θεωρούν ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας προσφέρει αποτελεσματικότερη πρόληψη στους πληθυσμούς υψηλού κινδύνου.

7.2 Μέθοδος

Από το Μάιο μέχρι το Δεκέμβριο του 2017 διενεργήθηκε μία εμπειρική έρευνα για να καταγράψει τις γνώσεις και τις απόψεις των επιστημόνων της Πληροφορικής Υγείας σχετικά με τα Big Data. Οι επιστήμονες της Πληροφορικής Υγείας ήταν απόφοιτοι και εν ενεργεία φοιτητές του Διιδρυματικού Μεταπτυχιακού Προγράμματος «Πληροφορική της Υγείας». Για να εξασφαλιστεί ότι στην έρευνα συμμετείχαν τα εν λόγω άτομα, έγινε αποστολή του ερωτηματολογίου στα email των ατόμων ζητώντας του να συνδράμουν στην έρευνα.

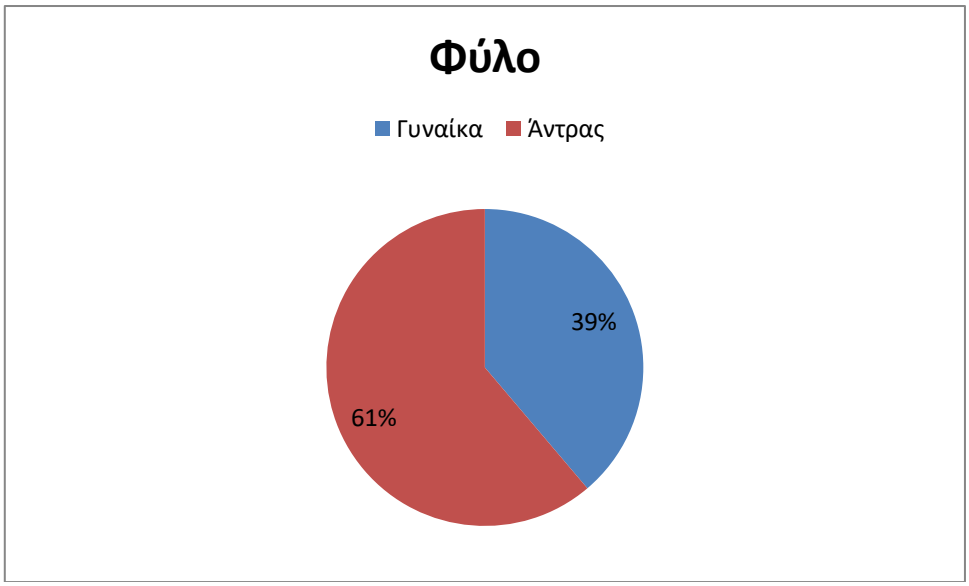
Η μέθοδος συλλογής των στοιχείων έγινε μέσω ηλεκτρονικού ερωτηματολογίου το οποίο δημιουργήθηκε μέσω του Google Forms. Το ερωτηματολόγιο βασίστηκε στην έρευνα των Mathisen, Wienhofen και Roman [107]. Μετάφραση του ερωτηματολογίου στα ελληνικά έγινε κατά τέτοιο τρόπο ώστε να είναι κατανοητή η ορολογία στο δείγμα καθώς και να ανταποκρίνεται στην ελληνική πραγματικότητα. Από το παραπάνω ερωτηματολόγιο οι ερωτήσεις 1-10 χρησιμοποιήθηκαν αυτούσιες ενώ οι υπόλοιπες δεν χρησιμοποιήθηκαν καθώς αφορούσαν τεχνικές εξόρυξης γνώσης και οι οποίες δεν συνάδουν με τους σκοπούς της παρούσας έρευνας. Επιπλέον προστέθηκαν ερωτήσεις για την καταγραφή της άποψης των ατόμων σχετικά με τα Big Data.

Το μέγεθος του δείγματος ήταν 69 άτομα. Το ερωτηματολόγιο αποτελούνταν από δύο μέρη. Το πρώτο μέρος περιελάμβανε ερωτήσεις που αφορούσαν τα δημογραφικά χαρακτηριστικά του δείγματος ενώ το δεύτερο μέρος περιελάμβανε ερωτήσεις που αφορούσαν την καταγραφή της γνώσης και των απόψεων των ατόμων σχετικά με τα Big Data.

Η επεξεργασία και η στατιστική ανάλυση των εμπειρικών δεδομένων, έγινε με τη χρήση του λογισμικού πακέτου SPSS 22. Η εκτίμηση της αξιοπιστίας του ερωτηματολογίου έγινε με τον συντελεστή αξιοπιστίας Cronbach alpha. Για την καταγραφή των απόψεων και τον έλεγχο συσχετίσεων υπολογίστηκαν τα συνολικά σκορ και εφαρμόστηκαν οι κατάλληλοι παραμετρικοί και μη παραμετρικοί έλεγχοι για τον έλεγχο συσχετίσεων. Το χρησιμοποιούμενο επίπεδο στατιστικής σημαντικότητας σε όλες τις στατιστικές δοκιμασίες, ορίστηκε στο 0,05.

7.3 Αποτελέσματα

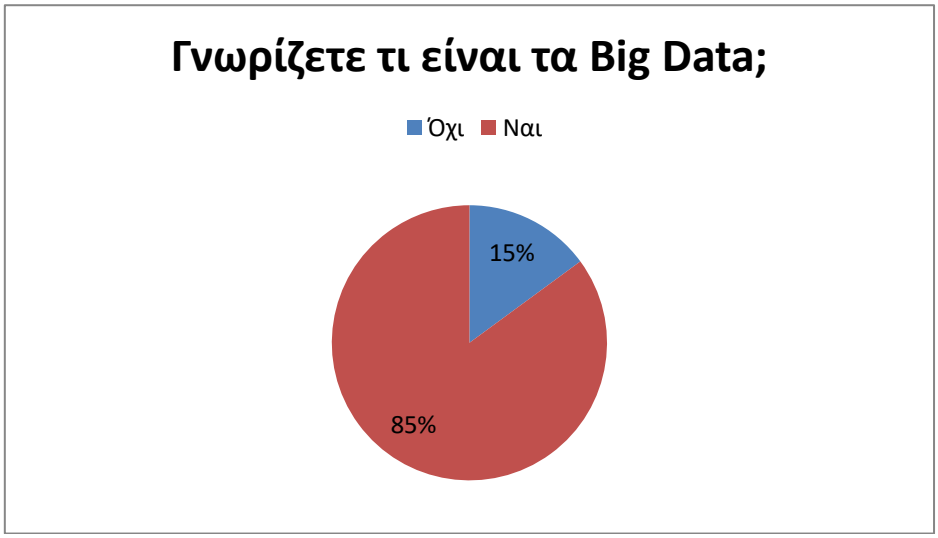
7.3.1 Περιγραφική στατιστική



Γράφημα 1: Φύλο του δείγματος

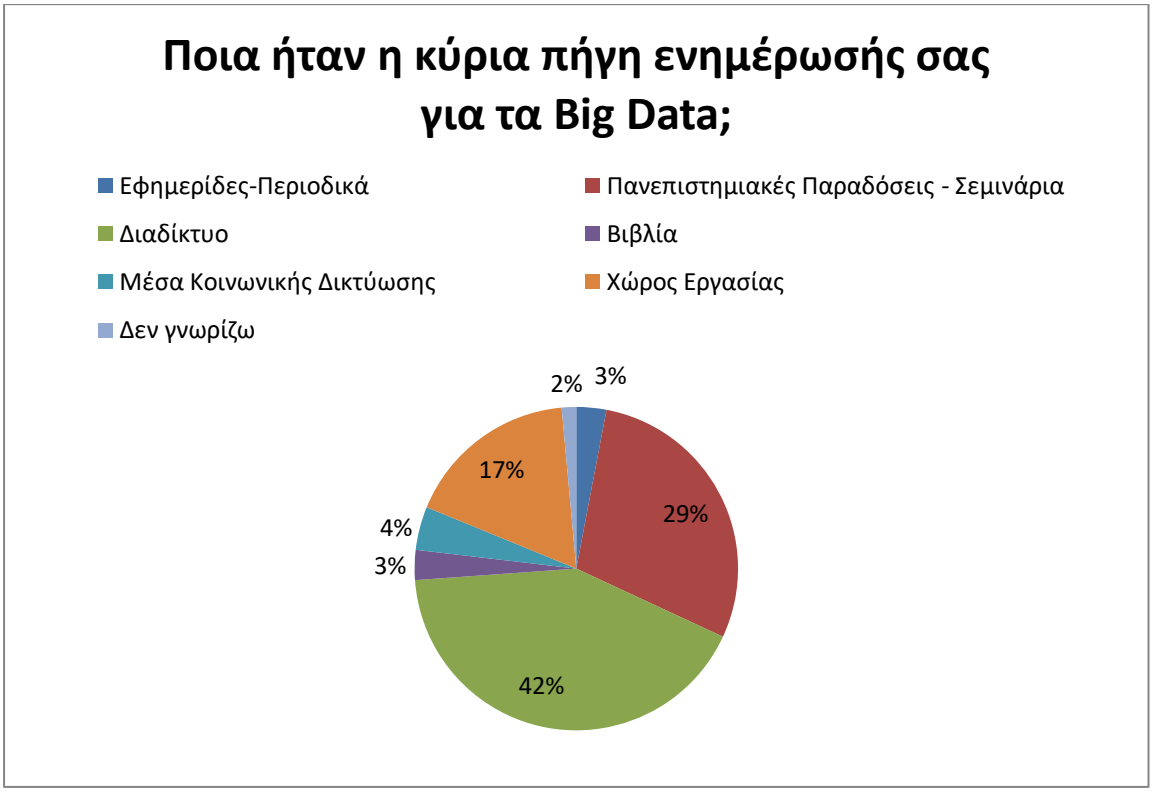
Το δείγμα είναι 69 άτομα. Όσον αφορά στο φύλο το 39% είναι γυναίκες και το 61% άνδρες (Γράφημα 1).

Από τον Πίνακα 1 (Παράρτημα Α) παρατηρούμε ότι η μέγιστη ηλικία είναι τα 63 έτη και η μικρότερη τα 18 έτη. Η μέση τιμή της ηλικίας του δείγματος είναι 28,69 έτη.



Γράφημα 2: Γνώση του δείγματος σχετικά με τα Big Data

Το 85% του δείγματος (Γράφημα 2) απάντησε ότι γνωρίζει την ορολογία των Μεγάλων Δεδομένων.



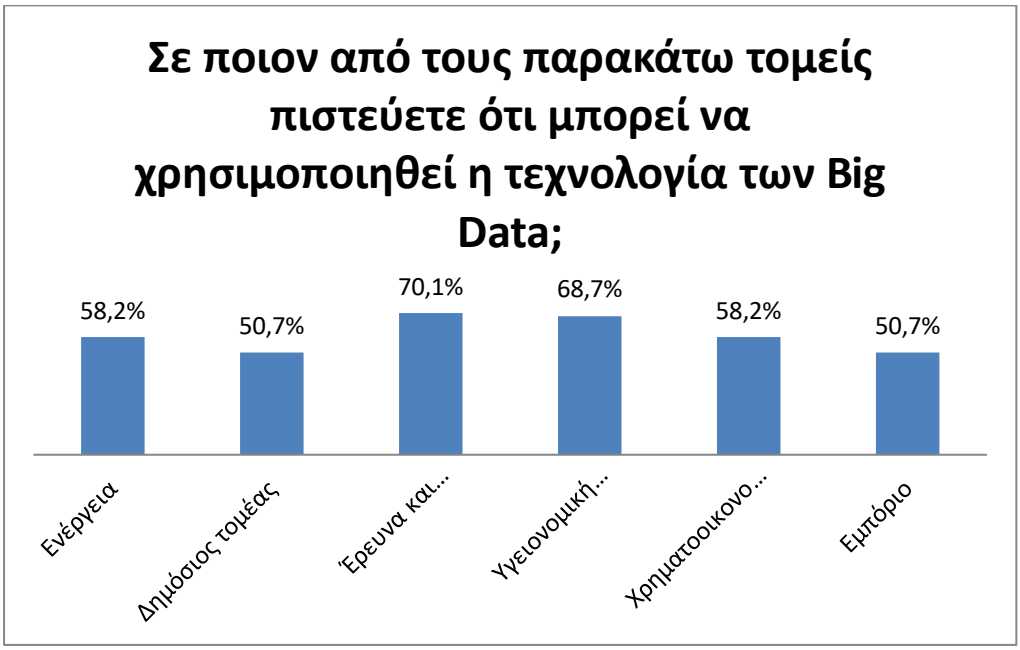
Γράφημα 3: Κύρια πηγή ενημέρωσης του δείγματος σχετικά με τα Big Data

Ως κύρια πηγή ενημέρωσης για τα Big Data το 42% του δείγματος ανέφερε το Διαδίκτυο, το 29% τις Πανεπιστημιακές Παραδόσεις- Σεμινάρια και το 17% το Χώρο εργασίας. Επίσης το 4% των ατόμων ανέφεραν τα Μέσα

Κοινωνικής Δικτύωσης, το 3% Εφημερίδες-Περιοδικά και Βιβλία ενώ το 2% ανέφερε ότι δεν γνώριζε κάτι (Γράφημα 3).

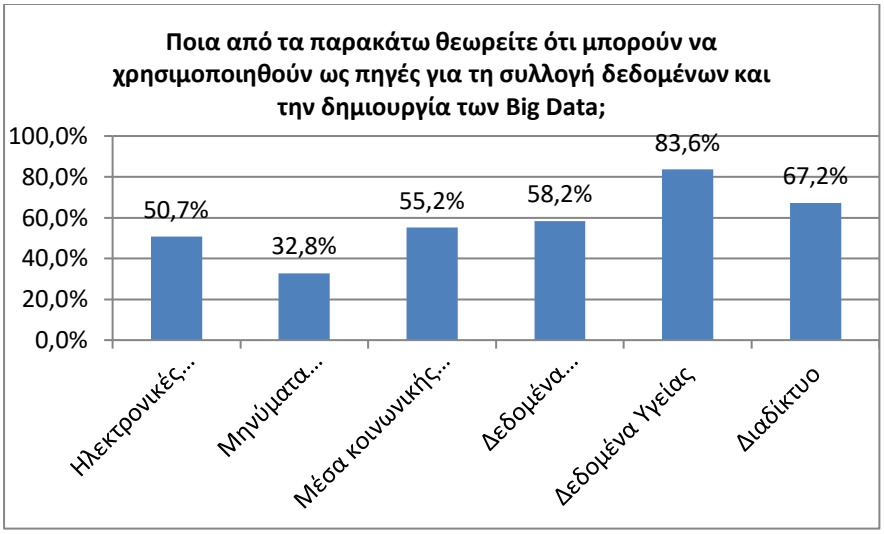
Γράφημα 4: Γνώση του δείγματος σχετικά με την τεχνολογία των Big Data

Το 75% του δείγματος (Γράφημα 4) απάντησε ότι γνωρίζει την την τεχνολογία των Big Data.



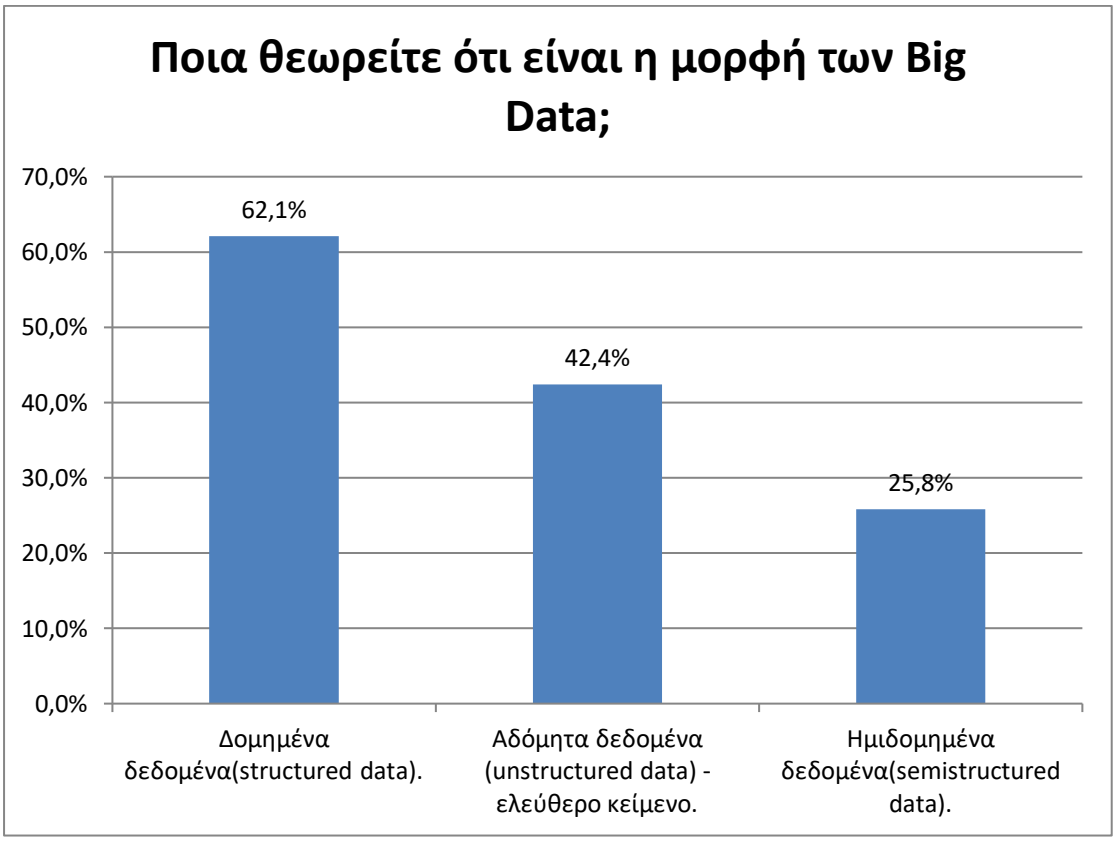
Γράφημα 5: Πεδίο εφαρμογής της τεχνολογίας των Big Data

Το Γράφημα 5 δείχνει ότι τα υψηλότερα ποσοστά συγκέντρωσε η Έρευνα και Εκπαίδευση (70,1%). Ακολουθεί η Υγειονομική Περίθαλψη με 68,7% και στη συνέχεια η Ενέργεια και οι Χρηματοοικονομικές υπηρεσίες με ποσοστό 58,25 η καθεμία.



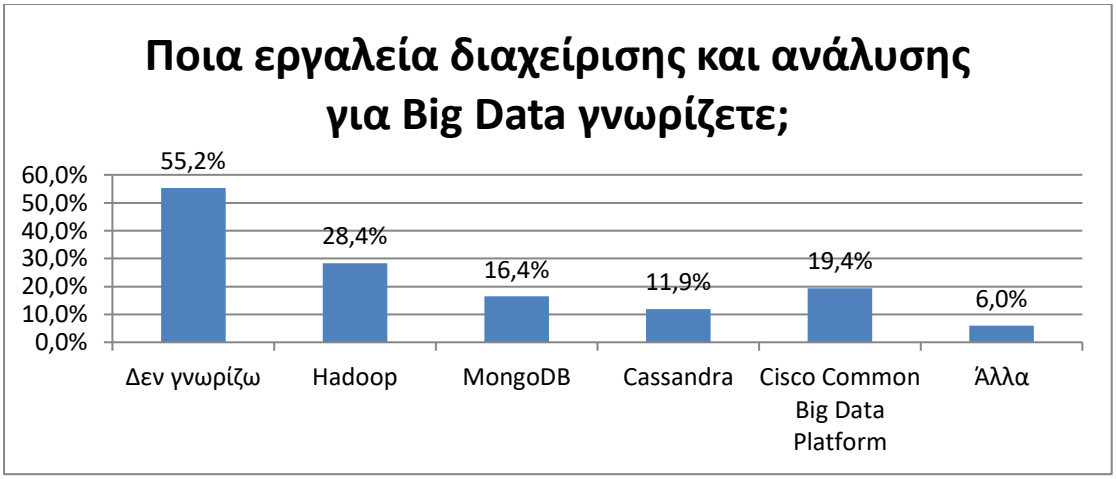
Γράφημα 6: Συλλογή δεδομένων των Big Data

Το Γράφημα 6 δείχνει ότι τα υψηλότερα ποσοστά συγκέντρωσαν τα Δεδομένα υγείας (83,6%). Ακολουθεί το Διαδίκτυο με 67,2%, τα Δεδομένα Αισθητήρων με 58,2% και τα Μέσα Κοινωνικής δικτύωσης με 55,2%.



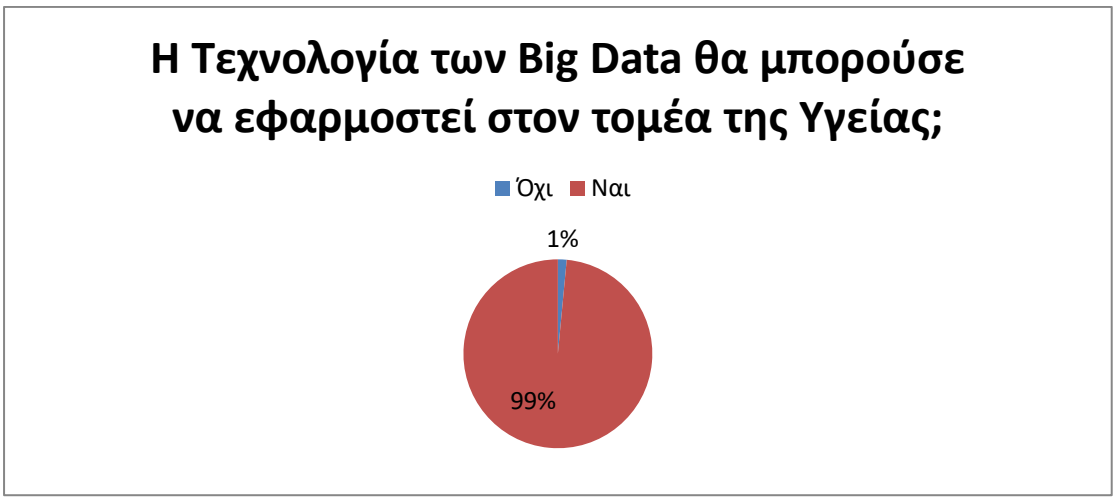
Γράφημα 7: Μορφή των Big Data

Από το Γράφημα 7 παρατηρούμε ότι οι ερωτηθέντες επέλεξαν τα δομημένα δεδομένα ως μορφή των Big Data με ποσοστό 62,1%. Ακολουθούν τα αδόμητα δεδομένα με 42,4% και τα ημιδομημένα δεδομένα με 25,8%.



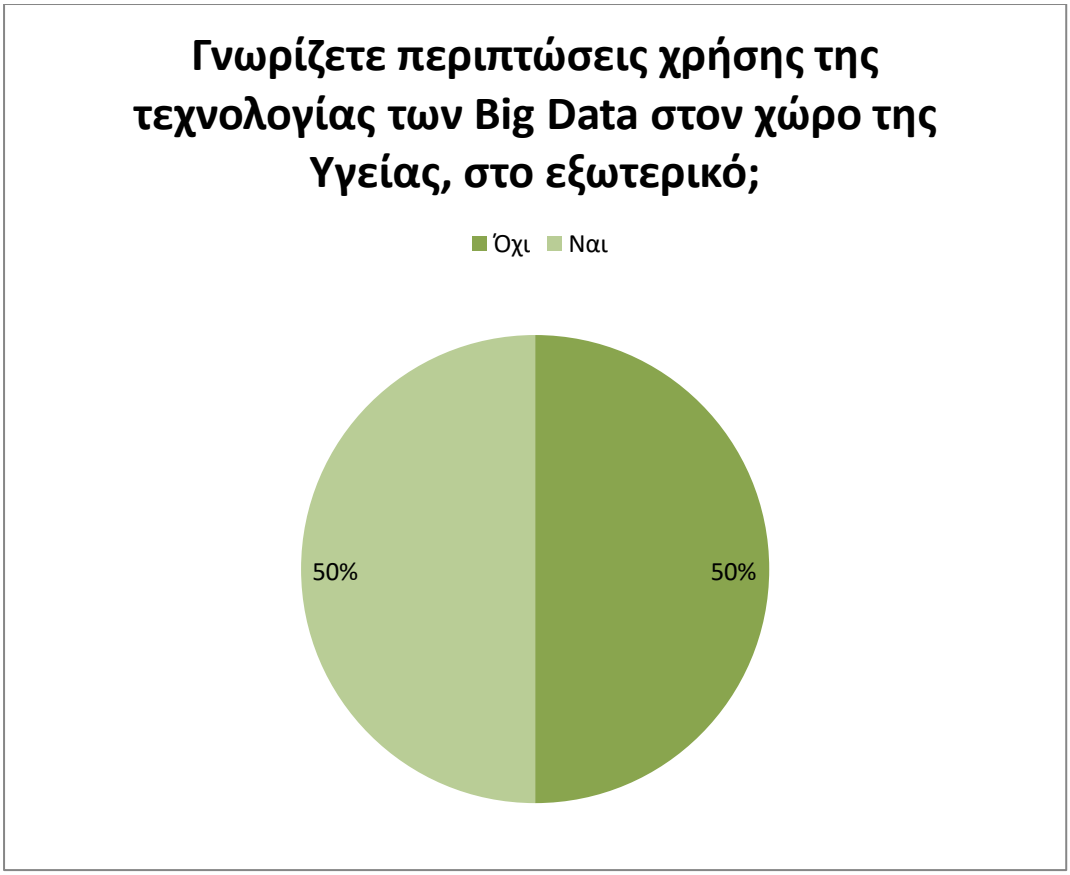
Γράφημα 8: Εργαλεία διαχείρισης και ανάλυσης για Big Data

Από το Γράφημα 8 παρατηρούμε ότι τα περισσότερα άτομα ανέφεραν ότι δεν γνωρίζουν εργαλεία ανάλυσης και διαχείρισης για Big Data σε ποσοστό 55,2%. Επίσης οι επιστήμονες της Πληροφορικής Υγείας ανέφεραν το Hadoop σε ποσοστό 28,4% και το Cisco Common Big Data Platform σε ποσοστό 19,4%.



Γράφημα 9: Εφαρμογή των Big Data στον τομέα της Υγείας

Σε ποσοστό 99% οι επιστήμονες της Πληροφορικής Υγείας πιστεύουν ότι η τεχνολογία των Big Data θα μπορούσε να εφαρμοστεί στον τομέα της Υγείας (Γράφημα 9).



Γράφημα 10: Περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας στο εξωτερικό

Όπως φαίνεται στο Γράφημα 10 μοιρασμένα είναι τα ποσοστό (50%) των ατόμων που γνωρίζουν περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας στο εξωτερικό.

Γράφημα 11: Περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας στην Ελλάδα

Το 83% (Γράφημα 11) των ατόμων δεν γνωρίζει περιπτώσεις χρήσης των Big data στην Ελλάδα.

Γράφημα 12: Συλλογή δεδομένων των Big Data στον τομέα της Υγείας

Το Γράφημα 12 δείχνει ότι τα υψηλότερα ποσοστά συγκέντρωσε ο Ηλεκτρονικός Φάκελος Ασθενή (86,4%). Ακολουθεί η Ηλεκτρονική Συνταγογράφηση με 69,7%,τα Συστήματα Υποστήριξης Κλινικών Αποφάσεων με 51,5% και τα Επιστημονικά Περιοδικά με 24,2%.

Για την καταγραφή της άποψης των επαγγελματιών υγείας σχετικά την τεχνολογία των Big Data χρησιμοποιήθηκαν ερωτήσεις σε κλίμακα Likert με 7 διαβαθμίσεις όπως φαίνεται παρακάτω:

Διαφωνώ Απόλυτα 1 2 3 4 5 6 7 Συμφωνώ Απόλυτα

Για την κάθε ερώτηση υπολογίστηκε ο μέσος όρος όπως φαίνεται στον Πίνακα 2 (Παράρτημα Α). Τα αποτελέσματα του πίνακα δείχνουν ότι:

- οι ερωτηθέντες συμφωνούν με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας είναι χρήσιμη.
- οι ερωτηθέντες συμφωνούν με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα μπορούσε να αυξήσει την αποτελεσματικότητα των παρεχόμενων Υπηρεσιών Υγείας.

- οι ερωτηθέντες συμφωνούν μερικώς με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας προσφέρει αποτελεσματικότερη πρόληψη στους πληθυσμούς υψηλού κινδύνου
- οι ερωτηθέντες συμφωνούν με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα δώσει τη δυνατότητα παροχής υπηρεσιών υγείας προσαρμοσμένων στις ανάγκες των ασθενών.
- οι ερωτηθέντες συμφωνούν μερικώς με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα δώσει τη δυνατότητα παροχής υπηρεσιών υγείας προσαρμοσμένων στις ανάγκες των ασθενών.

7.3.2 Επαγωγική στατιστική

Για τον έλεγχο της αξιοπιστίας του ερωτηματολογίου υπολογίστηκε ο δείκτης a του Cronbach. Από τον Πίνακα 3 (Παράρτημα Α) βλέπουμε ότι a= 0,814 που δείχνει υψηλή αξιοπιστία.

7.3.2.1 Έλεγχοι Κανονικότητας ποσοτικών μεταβλητών

Από τον Πίνακα 4 (Παράρτημα Α) παρατηρούμε ότι οι ποσοτικές μεταβλητές δεν ακολουθούν την κανονική κατανομή αφού έχουν p-value<0,05.

7.3.2.2 Συσχέτιση Ηλικίας -Γνώσης Big Data

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στη Γνώση των Big Data.
- H1: Υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στη Γνώση των Big Data.

p-value=0,066> 0,05 (Πίνακας 5-Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ της ηλικίας και γνώσης των Big Data.

7.3.2.3 Συσχέτιση Φύλου -Γνώσης Big Data

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στο Φύλο και στη Γνώση των Big Data.

- H1: Υπάρχει συσχέτιση ανάμεσα στο Φύλο και στη Γνώση των Big Data.

p-value=0,536>0,05 (Πίνακας 6 - Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ του φύλου και γνώσης των Big Data.

7.3.2.4 Συσχέτιση Φύλου – Μορφή των Big Data

- Δομημένα δεδομένα

  - Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.
  - H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

p-value=0,603>0,05 (Πίνακας 7 - Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

- Αδόμητα δεδομένα

  - Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.
  - H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

p-value=0,017<0,05 (Πίνακας 8 - Παράρτημα Α). Επομένως απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές. Επίσης r=0,012 (Πίνακας 9 - Παράρτημα Α) που σημαίνει ότι η συσχέτιση μεταξύ των μεταβλητών είναι ασθενής. Επίσης παρατηρούμε ότι είναι περισσότεροι οι άνδρες που πιστεύουν ότι τα Big Data αφορούν αδόμητα δεδομένα σε σχέση με τις γυναίκες.

- Ημιδομημένα δεδομένα

  - Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.
  - H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

p-value=0,102>0,05 (Πίνακας 10 - Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

7.3.2.5 Έλεγχος Συσχέτισης Φύλου –Σκορ Χρησιμότητας των Big Data

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Χρησιμότητας.
- Η1: Υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Χρησιμότητας.

p-value=0,271>0,05 (Πίνακας 11 -Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ του Φύλου και Σκορ Χρησιμότητας.

7.3.2.6 Έλεγχος Συσχέτισης Φύλου –Σκορ Αποτελεσματικότητας Υπηρεσιών Υγείας

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Αποτελεσματικότητας.
- Η1: Υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Αποτελεσματικότητας.

p-value= 0,428>0,05 (Πίνακας 12 -Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ του Φύλου και Σκορ Αποτελεσματικότητας.

7.3.2.7 Έλεγχος Συσχέτισης Φύλου –Σκορ Αποφάσεων

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Αποφάσεων.
- Η1: Υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Αποφάσεων.

p-value=0,671>0,05 (Πίνακας 13 -Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ του Φύλου και Σκορ Αποφάσεων.

7.3.2.8 Έλεγχος Συσχέτισης Φύλου –Σκορ Παροχής Υπηρεσιών Υγείας

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Παροχής Υπηρεσιών Υγείας.
- Η1: Υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Παροχής Υπηρεσιών Υγείας.

p-value=0,059>0,05 (Πίνακας 14 -Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ του Φύλου και Σκορ Παροχής Υπηρεσιών Υγείας.

7.3.2.9 Έλεγχος Συσχέτισης Φύλου –Σκορ Πρόληψης

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Πρόληψης.
- Η1: Υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Πρόληψης.

p-value=0,197>0,05 (Πίνακας 15 -Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ του Φύλου και του Σκορ Πρόληψης.

7.3.2.10 Έλεγχος Συσχέτισης Ηλικίας –Σκορ Χρησιμότητας των Big Data

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στο Σκορ Χρησιμότητας
- Η1: Υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στο Σκορ Χρησιμότητας

p-value=0,904>0,05 (Πίνακας 16-Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ των δύο μεταβλητών.

7.3.2.11 Έλεγχος Συσχέτισης Ηλικίας –Σκορ Αποτελεσματικότητας Υπηρεσιών Υγείας

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στο Σκορ Αποτελεσματικότητας.
- Η1: Υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στο Σκορ Αποτελεσματικότητας

p-value=0,905>0,05 (Πίνακας 16-Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν συσχέτιση μεταξύ των δύο μεταβλητών.

7.3.2.12 Έλεγχος Συσχέτισης Ηλικίας –Σκορ Αποφάσεων

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στο Σκορ Αποφάσεων.
- Η1: Υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στο Σκορ Αποφάσεων.

p-value= 0,924>0,05 (Πίνακας 16 -Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ των δύο μεταβλητών.

7.3.2.13 Έλεγχος Συσχέτισης Ηλικίας –Σκορ Παροχής Υπηρεσιών Υγείας

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές
- H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές

p-value=0,909>0,05 (Πίνακας 16-Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ των δύο μεταβλητών.

7.3.2.14 Έλεγχος Συσχέτισης Ηλικίας –Σκορ Πρόληψης

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.
- H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

p-value=0,927>0,05 (Πίνακας 16-Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ των δύο μεταβλητών.

7.3.2.15 Έλεγχος Συσχέτισης Φύλου –Χρήση των Big Data στον τομέα της Υγείας

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές
- H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές

p-value=0,422>0,05 (Πίνακας 17-Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

7.3.2.16 Έλεγχος Συσχέτισης Ηλικίας –Χρήση των Big Data στον τομέα της Υγείας

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές
- H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές

p-value=0,242>0,05 (Πίνακας 18-Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

7.3.2.17 Έλεγχος Συσχέτισης Φύλου –Περιπτώσεις Χρήσεις Big Data εξωτερικό

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές
- H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές

βp-value=0,044<0,05 (Πίνακας 19-Παράρτημα Α). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει ασθενής συσχέτιση ανάμεσα στις δύο μεταβλητές καθώς r=0,242 (Πίνακας 20-Παράρτημα Α).Επίσης οι γυναίκες γνωρίζουν περισσότερο από ότι οι άνδρες περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της υγείας στο εξωτερικό

7.3.2.18 Έλεγχος Συσχέτισης Ηλικίας –Περιπτώσεις Χρήσεις Big Data εξωτερικό

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές
- Η1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές

p-value=0,058>0,05 (Πίνακας 21-Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

7.3.2.19 Έλεγχος Συσχέτισης Φύλου –Περιπτώσεις Χρήσεις Big Data στην Ελλάδα

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές
- Η1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές

p-value=0,822>0,05 (Πίνακας 22-Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

7.3.2.20 Έλεγχος Συσχέτισης Ηλικίας –Περιπτώσεις Χρήσεις Big Data στην Ελλάδα

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές
- Η1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές

p-value=0,8253>0,05 (Πίνακας 23-Παράρτημα Α). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

7.4 Συζήτηση

Τα αποτελέσματα της έρευνας δείχνουν ότι η πλειονότητα των ατόμων που ερωτήθηκαν ήταν άνδρες. Το δείγμα στελεχώθηκε από άτομα όλων των ηλικιών με την πλειονότητα αυτών να είναι άνδρες. Επίσης η πλειονότητα των επιστημόνων της Πληροφορικής υγείας, απάντησε ότι γνωρίζει για τα Big

Data. Όσον αφορά στη μορφή των μεγάλων δεδομένων οι απαντήσεις έδειξαν ότι τα άτομα του δείγματος θεωρούν ότι τα Big Data περιλαμβάνουν κατά κύριο λόγο δομημένα δεδομένα και στη συνέχεια αδόμητα δεδομένα.

Σαν κύριες πηγές ενημέρωσης για τα Big Data η πλειονότητα των επιστημόνων της Πληροφορικής Υγείας ανέφερε το Διαδίκτυο και τις Πανεπιστημιακές Παραδόσεις-Σεμινάρια. Αξιοσημείωτο είναι το γεγονός ότι αρκετοί ανέφεραν σαν πηγή ενημέρωσης το χώρο εργασίας τους. Ως πεδίο εφαρμογής των Big Data η πλειονότητα των ατόμων απάντησαν την Έρευνα και Εκπαίδευση και την Υγειονομική. Όσον αφορά στο ποια δεδομένα θα πρέπει να συλλεχθούν για τα Big Data η πλειονότητα των ερωτηθέντων επέλεξε τα δεδομένα υγείας ακολουθούμενα από τα δεδομένα του Διαδικτύου. Επίσης, η πλειοψηφία των ατόμων δείχνει να μην γνωρίζει κάποιο εργαλείο διαχείρισης και ανάλυσης Big Data σε ποσοστό. Οι υπόλοιποι ανέφεραν το Hadoop και το Cisco Common Big Data Platform.

Στην ερώτηση για τον αν η τεχνολογία των Data θα μπορούσε να εφαρμοστεί στον τομέα της Υγείας τα αποτελέσματα ήταν υπέρ του Ναι. Ωστόσο τα μισά άτομα δεν γνωρίζουν περιπτώσεις χρήσης των Big Data στο εξωτερικό. Αντίστοιχα για την Ελλάδα, η πλειονότητα των επιστημόνων της Πληροφορικής Υγείας δεν γνωρίζει περιπτώσεις χρήσης των Big Data στη χώρα μας.

Επιπλέον τα αποτελέσματα της έρευνας έδειξαν ότι επιστήμονες της Πληροφορικής Υγείας συμφωνούν με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας είναι χρήσιμη. Ακόμη συμφωνούν με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα μπορούσε να αυξήσει την αποτελεσματικότητα των παρεχόμενων Υπηρεσιών Υγείας. Η πλειονότητα συμφωνεί με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα βοηθήσει τους Επαγγελματίες Υγείας στη διαδικασία λήψης αποφάσεων καθώς και στο ότι θα δώσει τη δυνατότητα παροχής υπηρεσιών υγείας προσαρμοσμένων στις ανάγκες των ασθενών. Ακόμη οι ερωτηθέντες συμφωνούν με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα δώσει τη δυνατότητα παροχής υπηρεσιών υγείας προσαρμοσμένων στις ανάγκες των ασθενών.

Σύμφωνα με τα παραπάνω αποτελέσματα οι επιστήμονες της Πληροφορικής Υγείας δείχνουν να γνωρίζουν κάποια πράγματα σχετικά με τα Big Data. Επιπλέον θεωρούν χρήσιμη την συγκεκριμένη τεχνολογία στο χώρο της υγείας καθώς θεωρούν ότι θα αυξήσει την αποτελεσματικότητα και την ποιότητα των παρεχόμενων υπηρεσιών. Επίσης, αρκετοί από αυτούς εξέφρασαν στα σχόλια του ερωτηματολογίου την επιφύλαξή τους σχετικά με το κατά πόσο είναι εφικτή η χρήση της συγκεκριμένης τεχνολογίας στο Εθνικό Σύστημα Υγείας. Ακόμη, κάποιοι σχολίασαν κατά πόσο μπορεί να επιτευχθεί η ασφάλεια, η ακεραιότητα και η εμπιστευτικότητα των δεδομένων ενώ κάποιο ανέφεραν ότι η τεχνολογία των Big Data θα δημιουργούσε ηθικά διλήμματα σχετικά με την πρόσβαση και την επεξεργασία ευαίσθητων δεδομένων.

Θα πρέπει να τονιστεί ότι μετά την αναζήτηση στη διεθνή βιβλιογραφία δεν βρέθηκαν αντίστοιχες μελέτες οι οποίες θα μπορούσαν να συμπεριληφθούν στη διατριβή και να γίνει συγκριτική ανάλυση των αποτελεσμάτων.

Ένας από τους περιορισμούς της μελέτης είναι ότι οι απόψεις του δείγματος μπορεί να είναι θετικά επηρεασμένες σχετικά με τη Big Data,μιας που οι περισσότεροι είναι ενημερωμένοι για τη συγκεκριμένη τεχνολογία, τις δυνατότητες που αυτή έχει αλλάζοντας τον τρόπο χειρισμού του τεράστιου όγκου των δεδομένων καθώς και τις λύσεις που μπορεί να προσφέρει στον χώρο της υγείας. Επίσης ένας ακόμη περιορισμός της μελέτης είναι ότι τα άτομα δεν έχουν χρησιμοποιήσει ή δεν έχουν παρακολουθήσει κάποιο σεμινάριο ή workshop όπου θα μπορούσαν να έρθουν σε επαφή με εφαρμογές χειρισμού μεγάλου όγκου δεδομένων και εξαγωγής συμπερασμάτων σε πραγματικό χρόνο.

7.5 Συμπεράσματα

Η μελέτη που παρουσιάζεται στο παρόν Κεφάλαιο στοχεύει στη διερεύνηση της αντίληψης που έχουν οι επιστήμονες της Πληροφορικής Υγείας για την τεχνολογία των Μεγάλων Δεδομένων (Big Data). Τα αποτελέσματα της παραπάνω έρευνας δείχνουν ότι ένα μεγάλο ποσοστό του δείγματος ανέφερε ότι γνωρίζει την τεχνολογία των Big Data. Αξιοσημείωτο είναι το γεγονός ότι οι περισσότεροι επιστήμονες της Πληροφορικής Υγείας

αναφέρουν ότι έχουν ακούσει για τη συγκεκριμένη τεχνολογία και στο χώρο εργασίας τους.

Ανάμεσα στα ευρήματα της παρούσας έρευνας, αναδεικνύεται ότι οι επιστήμονες της Πληροφορικής Υγείας θεωρούν χρήσιμη την τεχνολογία των Big Data και πιστεύουν ότι θα συμβάλει στην πρόληψη της υγείας του πληθυσμού αλλά και ότι θα βελτιώσει την αποτελεσματικότητα των παρεχόμενων υπηρεσιών υγείας.

Οι μελλοντικές εργασίες πάνω σε αυτό το πεδίο θα μπορούσαν να είναι η επανάληψη της παρούσας έρευνας στα ίδια άτομα αφού προηγουμένως είχαν παρακολουθήσει κάποιο σεμινάριο σχετικό με εφαρμογές χειρισμού μεγάλου όγκου δεδομένων καθώς και εφαρμογές Data Mining σχετικές με το χώρο της υγείας έτσι ώστε να έχουν μια ολοκληρωμένη εικόνα για την τεχνολογία των Big Data.

Κεφάλαιο 8

Εμπειρική διερεύνηση για την καταγραφή της άποψης των επαγγελματιών Υγείας σχετικά με την Τεχνολογία των Big Data (Μεγάλα Δεδομένα)

8.1 Σκοπός

Σκοπός της έρευνας που παρουσιάζεται στο παρόν κεφάλαιο είναι η καταγραφή της άποψης των επαγγελματιών Υγείας σχετικά την τεχνολογία των Big Data. Η συγκεκριμένη έρευνα στοχεύει στη διερεύνηση της αντίληψης που έχουν οι επαγγελματίες υγείας (ιατροί και νοσηλευτές) για την τεχνολογία των Μεγάλων Δεδομένων (Big Data) και αν θεωρούν εφικτή την εφαρμογής της στο χώρο της υγείας.

Για την εμπειρική διερεύνηση της άποψης των επαγγελματιών υγείας σχετικά με την τεχνολογία των Big Data τέθηκαν τα παρακάτω ερευνητικά ερωτήματα:

- Γνωρίζουν οι επαγγελματίες Υγείας τι είναι τα Big Data;
- Ποια είναι η πηγή ενημέρωσης των επαγγελματιών Υγείας για την τεχνολογία των Big Data;
- Ποια είναι, κατά την άποψη των επαγγελματιών Υγείας, η μορφή των Big Data;
- Γνωρίζουν περιπτώσεις χρήσης της τεχνολογίας των Big Data στην Ελλάδα ή στο εξωτερικό στον τομέα της Υγείας;
- Ποιες θεωρούν οι επαγγελματίες Υγείας ότι είναι οι κύριες πηγές για τη συλλογή μεγάλου όγκου δεδομένων που αφορούν τον τομέα της υγείας;
- Οι επαγγελματίες Υγείας θεωρούν την τεχνολογία των Big Data χρήσιμη για τον τομέα της Υγείας;
- Οι επαγγελματίες Υγείας πιστεύουν ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα μπορούσε να αυξήσει την αποτελεσματικότητα των παρεχόμενων Υπηρεσιών Υγείας;

- Οι επαγγελματίες Υγείας πιστεύουν ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα βοηθήσει τους Επαγγελματίες Υγείας στη διαδικασία λήψης αποφάσεων;
- Οι επαγγελματίες Υγείας θεωρούν τι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα δώσει τη δυνατότητα παροχής υπηρεσιών υγείας προσαρμοσμένων στις ανάγκες των ασθενών;
- Οι επαγγελματίες Υγείας θεωρούν ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας προσφέρει αποτελεσματικότερη πρόληψη στους πληθυσμούς υψηλού κινδύνου.

8.2 Μέθοδος

Έχοντας σαν βάση την προηγούμενη έρευνα τον Δεκέμβριο του 2018 διενεργήθηκε μία εμπειρική έρευνα για να καταγράψει τις γνώσεις και τις απόψεις των επαγγελματιών υγείας σχετικά με τα Big Data. Η μέθοδος συλλογής των στοιχείων έγινε μέσω ηλεκτρονικού ερωτηματολογίου.

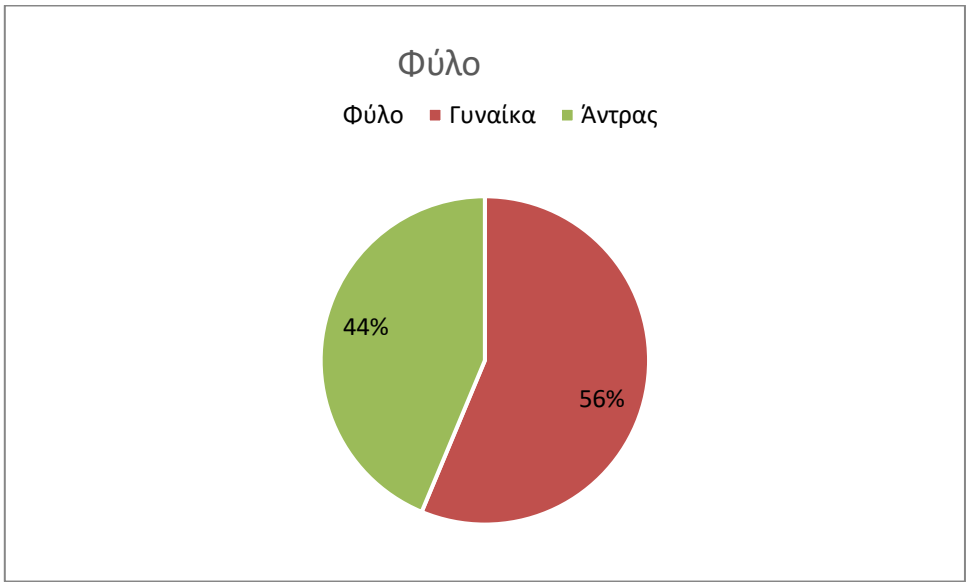
Το ερωτηματολόγιο βασίστηκε στην έρευνα των Mathisen, Wienhofen και Roman [107] και δημιουργήθηκε μέσω του Google Forms. Η μετάφραση του ερωτηματολογίου στα ελληνικά έγινε κατά τέτοιο τρόπο ώστε να είναι κατανοητή η ορολογία στο δείγμα καθώς και να ανταποκρίνεται στην ελληνική πραγματικότητα. Στο ερωτηματολόγιο έγιναν κάποιες προσθαφαιρέσεις σε ερωτήματα λαμβάνοντας υπόψη το μορφωτικό επίπεδο και την επαγγελματική κατάρτιση των ερωτηθέντων. Συγκεκριμένα προστέθηκαν ερωτήσεις σχετικά με το επάγγελμα των ερωτηθέντων και την επαγγελματική τους εμπειρία. Επίσης προστέθηκαν ερωτήσεις προσαρμοσμένες στη χρήση των Big Data στον τομέα της Υγείας(ερώτησες 8-12). Τέλος προστέθηκαν και ερωτήσεις για την καταγραφή της άποψης των επαγγελματιών Υγείας σχετικά με τη χρήση των Big Data Υγεία. Ο διαμοιρασμός του ερωτηματολογίου έγινε ηλεκτρονικά μέσω μηνυμάτων ηλεκτρονικού ταχυδρομείου.

Το μέγεθος του δείγματος ήταν 151 ιατροί και νοσηλευτές Το ερωτηματολόγιο αποτελούνταν από δύο μέρη. Το πρώτο μέρος περιελάμβανε ερωτήσεις που αφορούσαν τα δημογραφικά χαρακτηριστικά του δείγματος

ενώ το δεύτερο μέρος περιελάμβανε ερωτήσεις που αφορούσαν την καταγραφή της γνώσης και των απόψεων των ατόμων σχετικά με τα Big Data.

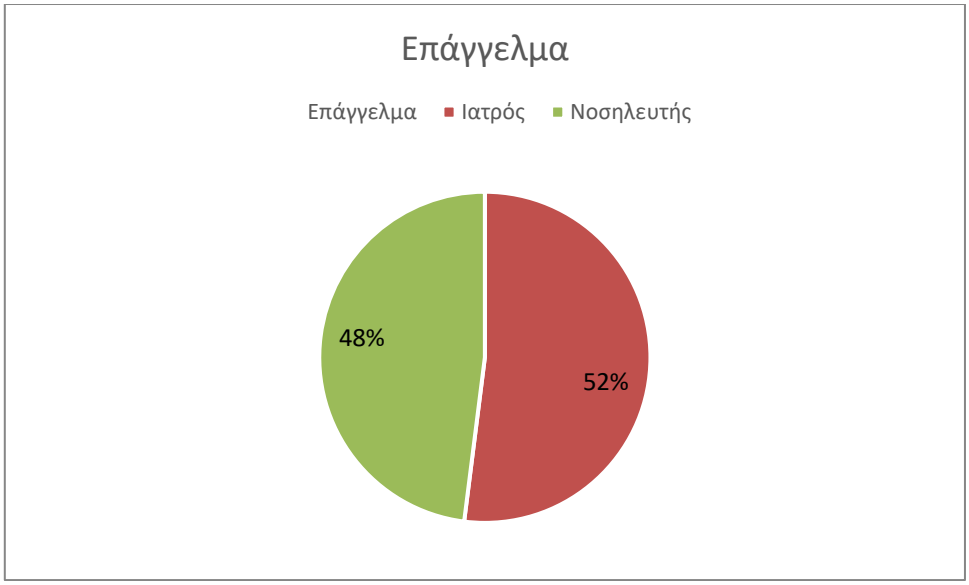
Η επεξεργασία και η στατιστική ανάλυση των εμπειρικών δεδομένων, έγινε με τη χρήση του λογισμικού πακέτου SPSS 25.0. Η εκτίμηση της αξιοπιστίας του ερωτηματολογίου έγινε με τον συντελεστή αξιοπιστίας Cronbach alpha. Για την καταγραφή των απόψεων και τον έλεγχο συσχετίσεων υπολογίστηκαν τα συνολικά σκορ και εφαρμόστηκαν οι κατάλληλοι παραμετρικοί και μη παραμετρικοί έλεγχοι για τον έλεγχο συσχετίσεων. Το χρησιμοποιούμενο επίπεδο στατιστικής σημαντικότητας σε όλες τις στατιστικές δοκιμασίες, ορίστηκε στο 0,05.

8.3.1 Περιγραφική στατιστική



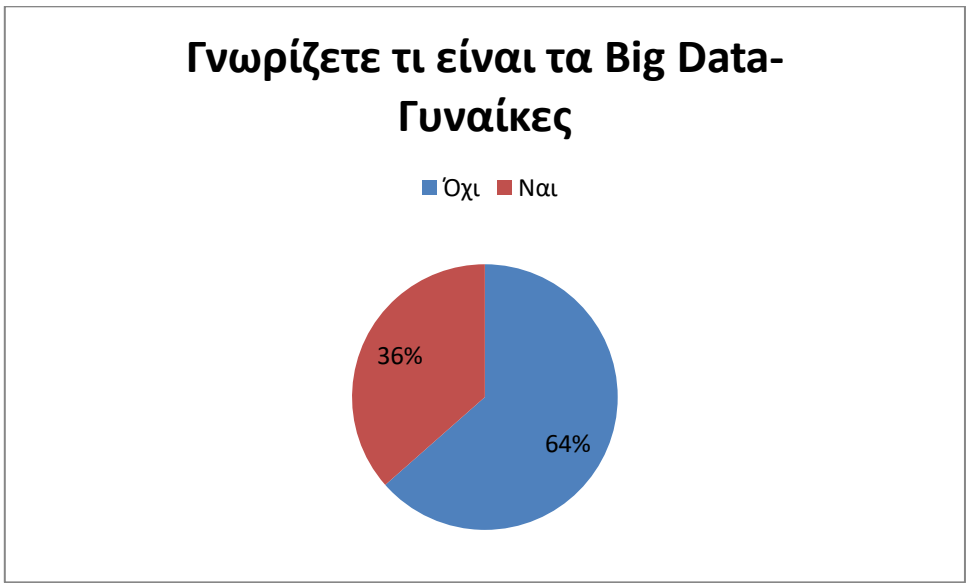
Γράφημα 13: Φύλο του δείγματος

Το δείγμα είναι 151 άτομα. Η μέση ηλικία του δείγματος είναι 38,9 έτη (Πίνακας 1 – Παράρτημα Β). Όσον αφορά στο φύλο το 56% είναι γυναίκες και το 44% άνδρες (Γράφημα 13).



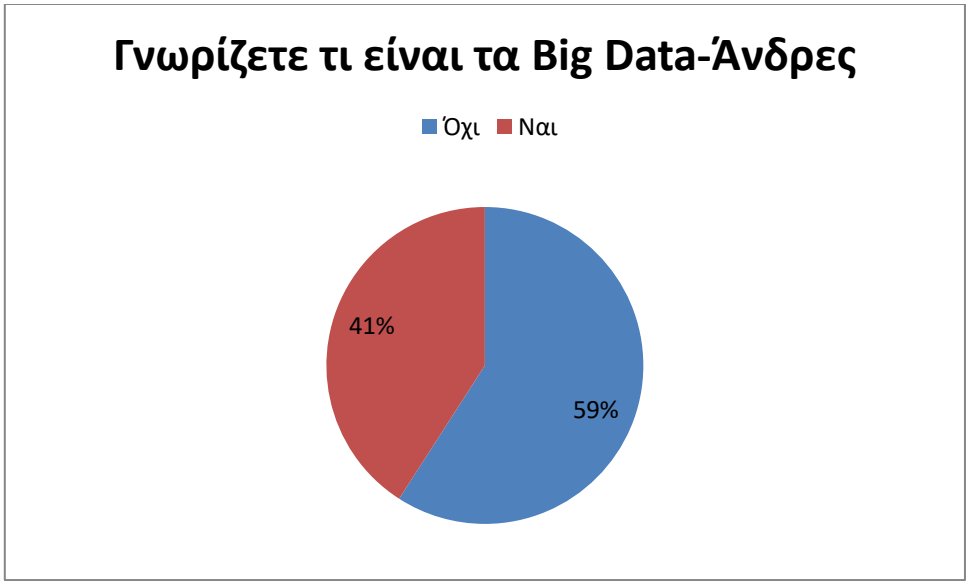
Γράφημα 14: Κατανομή των επαγγελματιών υγείας

Επίσης το 52% είναι Ιατροί και το 48% Νοσηλευτές (Γράφημα 14).



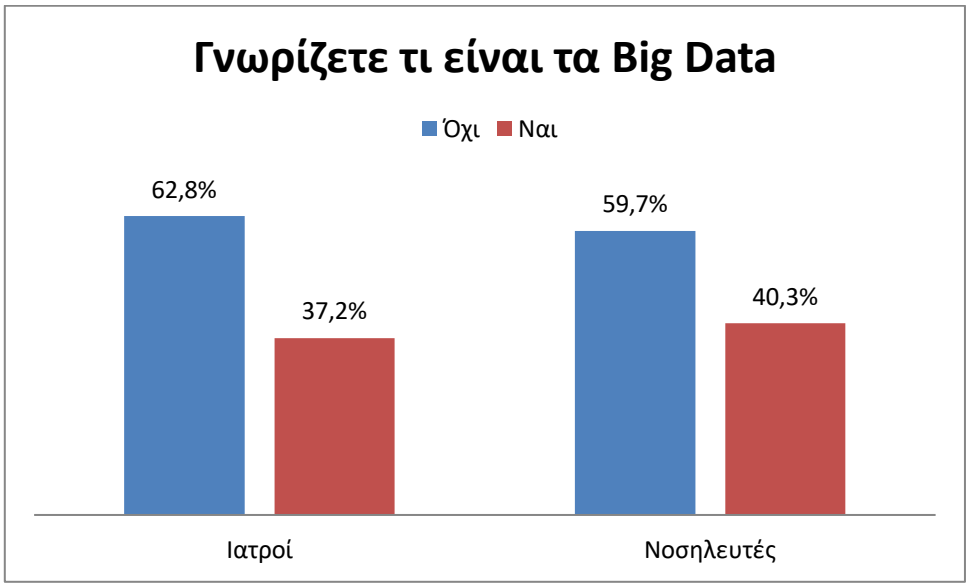
Γράφημα 15: Γνώση των γυναικών για τα Big Data

Στην ερώτηση για το αν γνωρίζουνε τι είναι τα Big Data οι γυναίκες απάντησαν ότι δεν γνωρίζουν για τη συγκεκριμένη τεχνολογία σε ποσοστό 63% (Γράφημα 15).



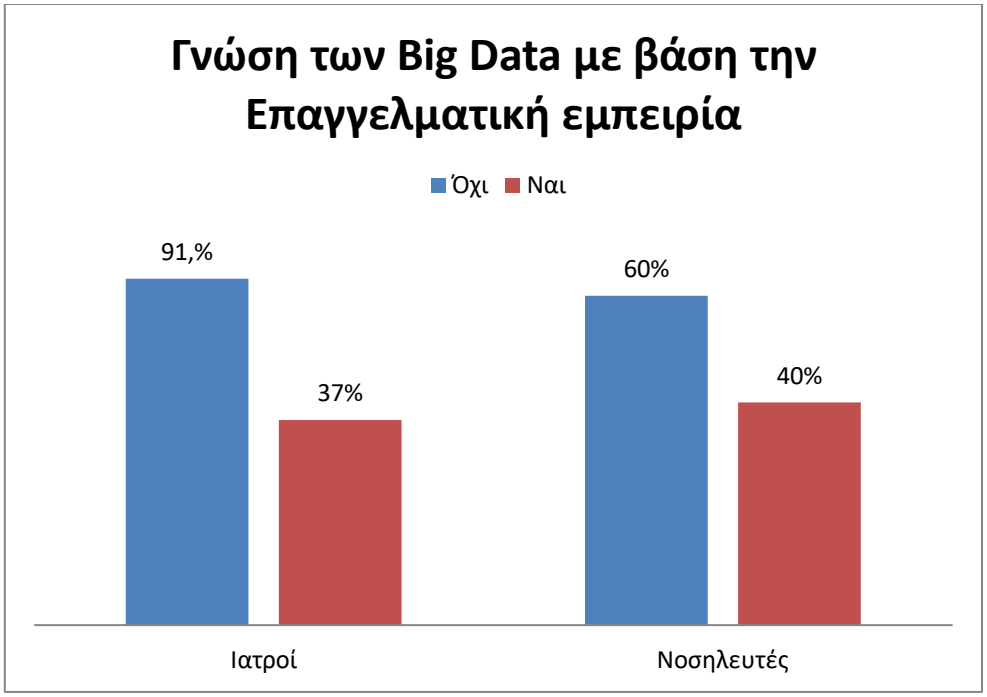
Γράφημα 16: Γνώση των ανδρών για τα Big Data

Αντίστοιχα οι άνδρες δεν γνωρίζουν τη συγκεκριμένη τεχνολογία σε ποσοστό 59% (Γράφημα 16).



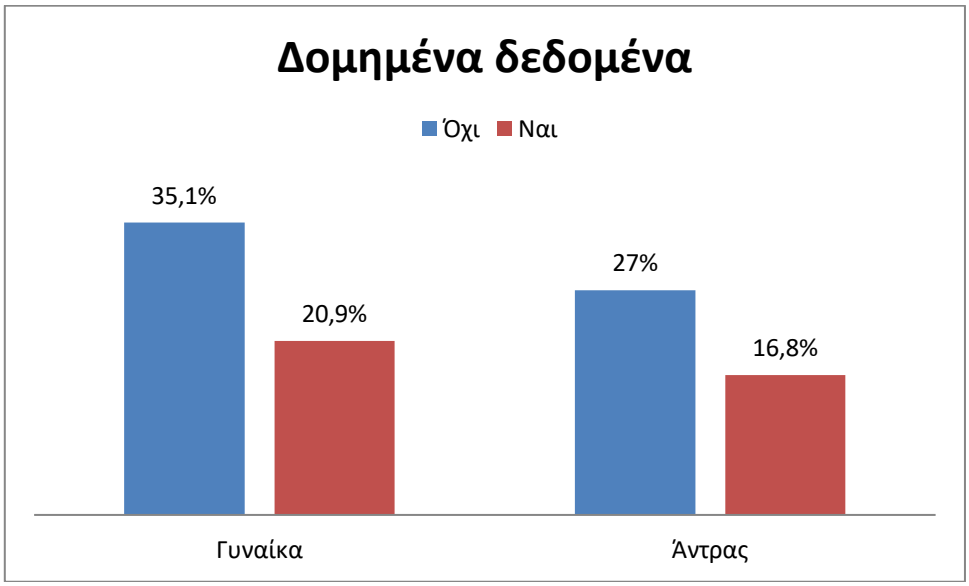
Γράφημα 17: Γνώση των επαγγελματιών υγείας για τα Big Data

Το Γράφημα 17 δείχνει ότι το 62,8% των ιατρών δεν γνωρίζει τι είναι τα Big Data. Επίσης οι νοσηλευτές δεν γνωρίζουν για τα Big Data σε ποσοστό 59,7%. Τέλος παρατηρούμε ότι είναι περισσότεροι οι νοσηλευτές(40,3%) που γνωρίζουν για τα Big Data σε σχέση με τους ιατρούς (37,2%).



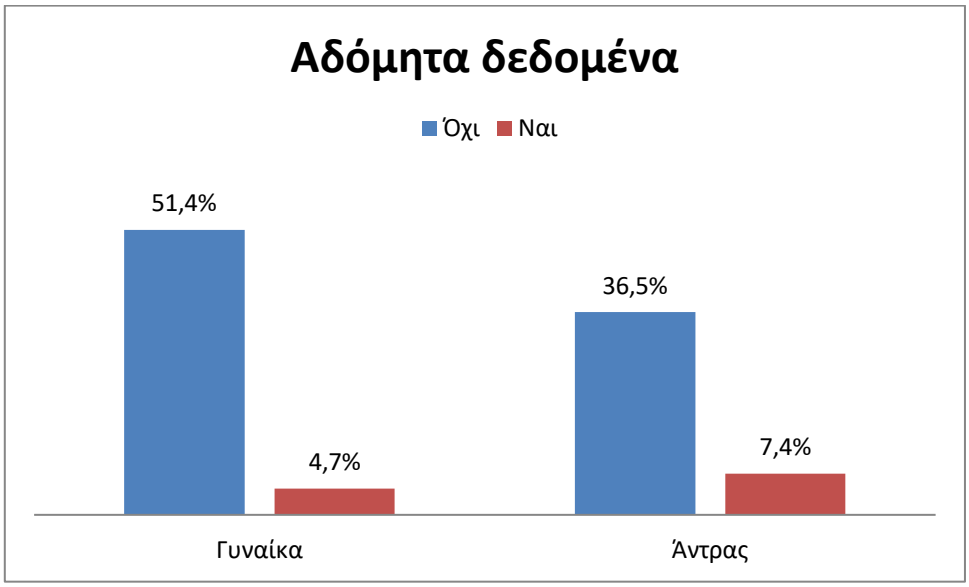
Γράφημα 18: Γνώση των επαγγελματιών υγείας για τα Big Data με βάση την επαγγελματική τους εμπειρία

Το 91% των ιατρών με επαγγελματική εμπειρία δεν γνωρίζουν για τα Big Data. Αντίστοιχα το 60% των νοσηλευτών με επαγγελματική εμπειρία δεν γνωρίζουν για τα Big Data. Επίσης το Γράφημα 18 δείχνει ότι είναι περισσότεροι οι νοσηλευτές με επαγγελματική εμπειρία που γνωρίζουν για τα Big Data σε σχέση με τους ιατρούς με ποσοστό 40% έναντι 37%.



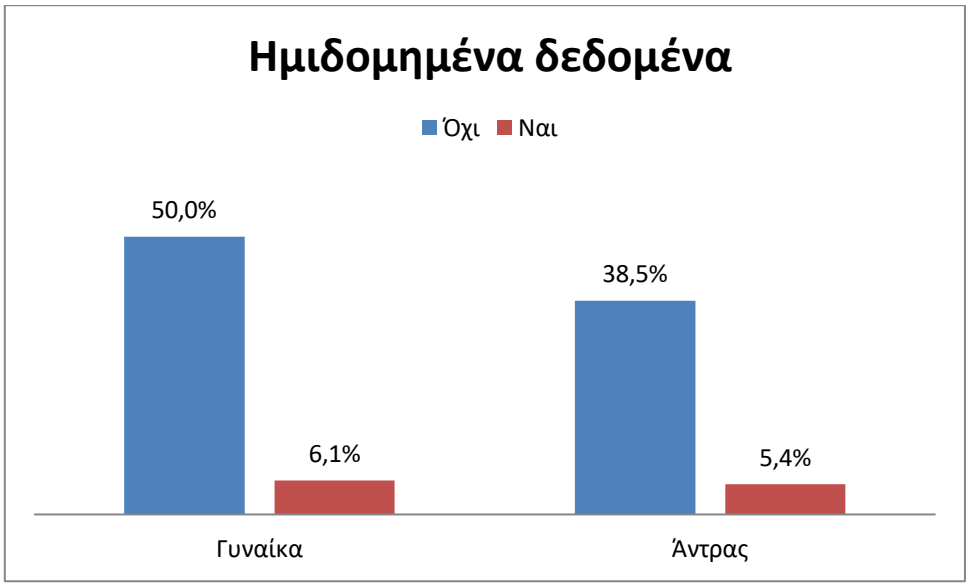
Γράφημα 19: Μορφή των Big Data με βάση το φύλο

Το 16,8% των ανδρών θεωρούν ότι τα Big Data είναι δομημένα δεδομένα. Το 20,9% των γυναικών θεωρούν ότι τα Big Data είναι δομημένα δεδομένα (Γράφημα 19).



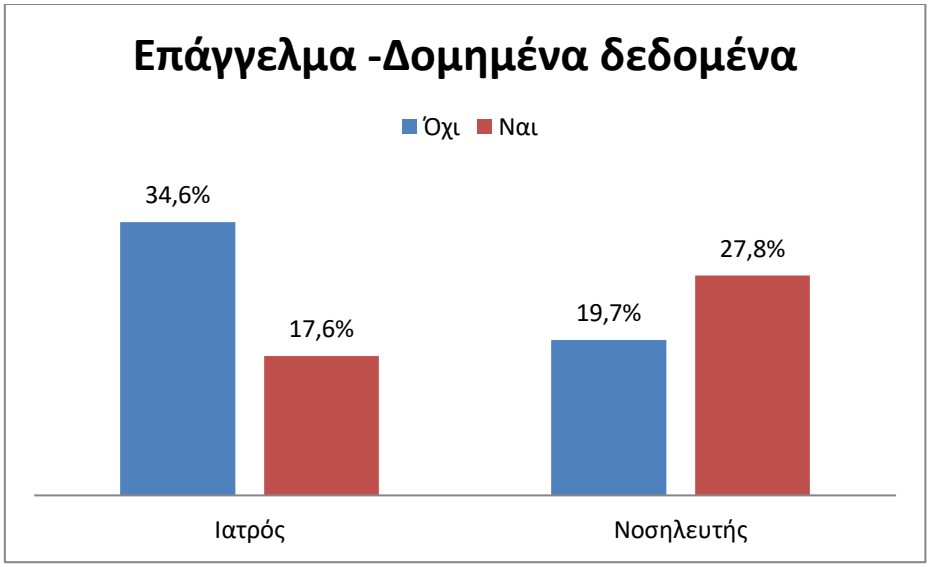
Γράφημα 20: Big Data ως αδόμητα δεδομένα με βάση το φύλο

Το 7,4% των ανδρών και το 4,7% των γυναικών θεωρούν ότι τα Big Data δεν περιέχουν αδόμητα δεδομένα (Γράφημα 20).



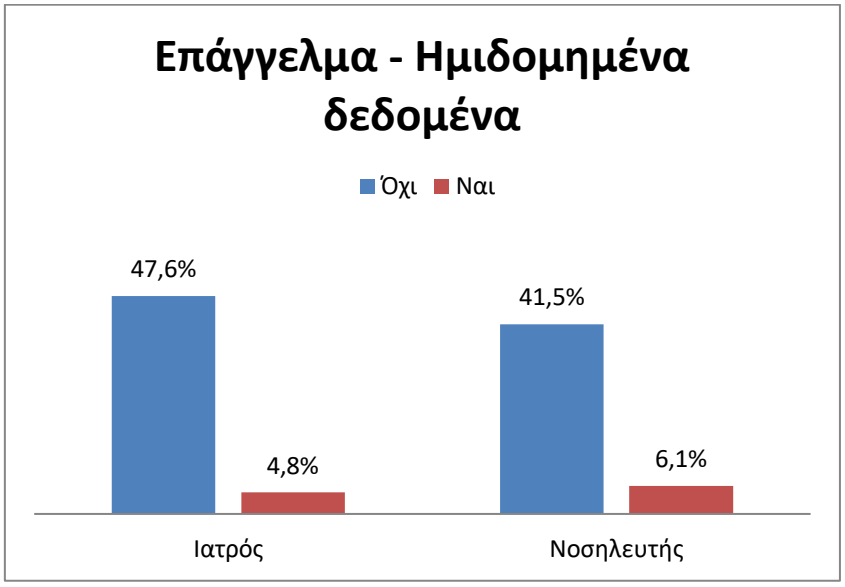
Γράφημα 21: Big Data ως ημιδομημένα δεδομένα με βάση το φύλο

Το 50% των γυναικών ανέφερε ότι τα Big data δεν περιέχουν ημιδομημένα δεδομένα. Αντίστοιχη είναι η άποψη των ανδρών σε ποσοστό 38,5% (Γράφημα 21).



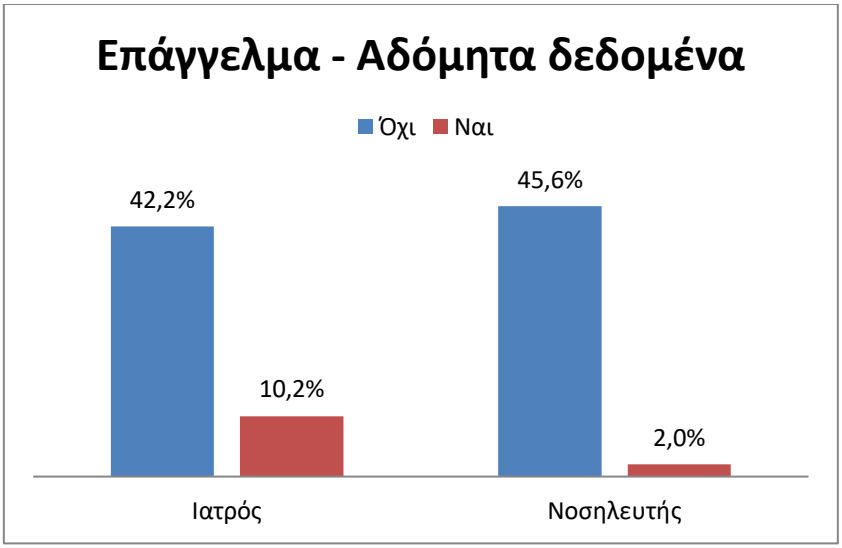
Γράφημα 22: Big Data ως δομημένα δεδομένα με βάση το επάγγελμα

Το 34,6% των ιατρών θεωρεί ότι τα Big Data δεν έχουν τη μορφή δομημένων δεδομένων. Αντίθετα το 27,8% των νοσηλευτών θεωρεί ότι τα Big Data έχουν τη μορφή δομημένων δεδομένων (Γράφημα 22).



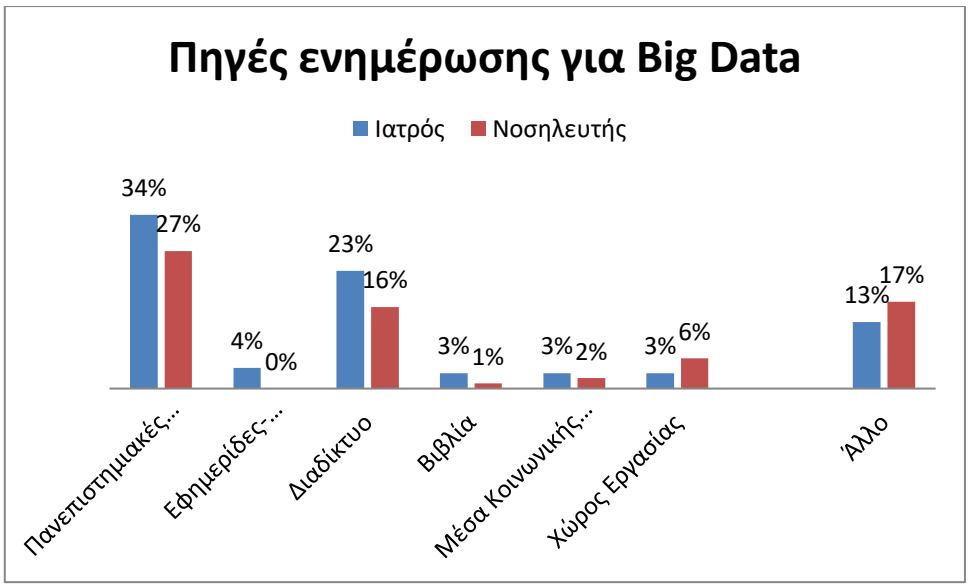
Γράφημα 23: Big Data ως ημιδομημένα δεδομένα με βάση το επάγγελμα

Το Γράφημα 23 δείχνει ότι το 6,15 των νοσηλευτών θεωρεί ότι τα Big data περιέχουν ημιδομημένα δεδομένα . Αντίστοιχη είναι και η των ιατρών σε ποσοστό 4,8.



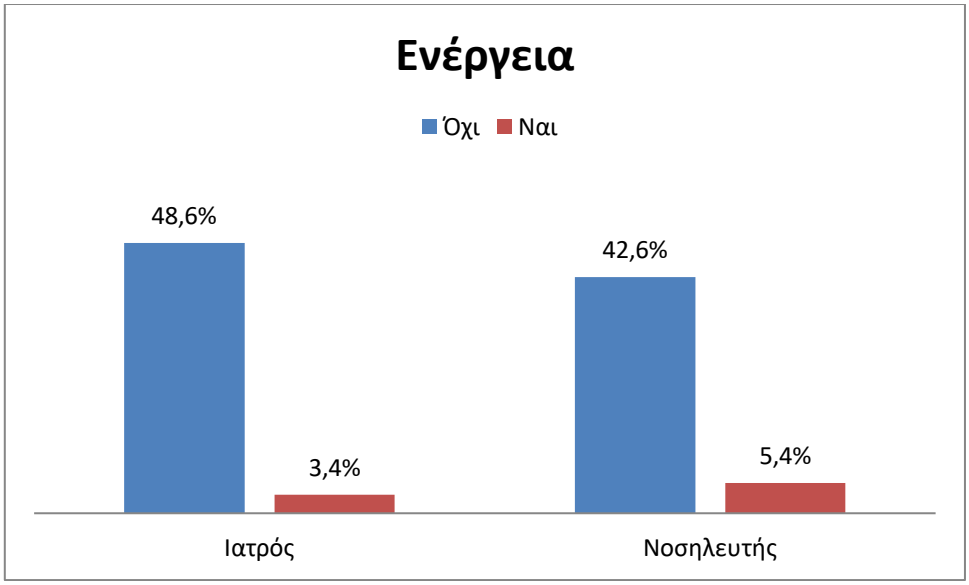
Γράφημα 24: Big Data ως αδόμητα δεδομένα με βάση το επάγγελμα

Το 10.2% των ιατρών και το 2% των νοσηλευτών θεωρούν ότι τα Big Data περιέχουν αδόμητα δεδομένα (Γράφημα 24).



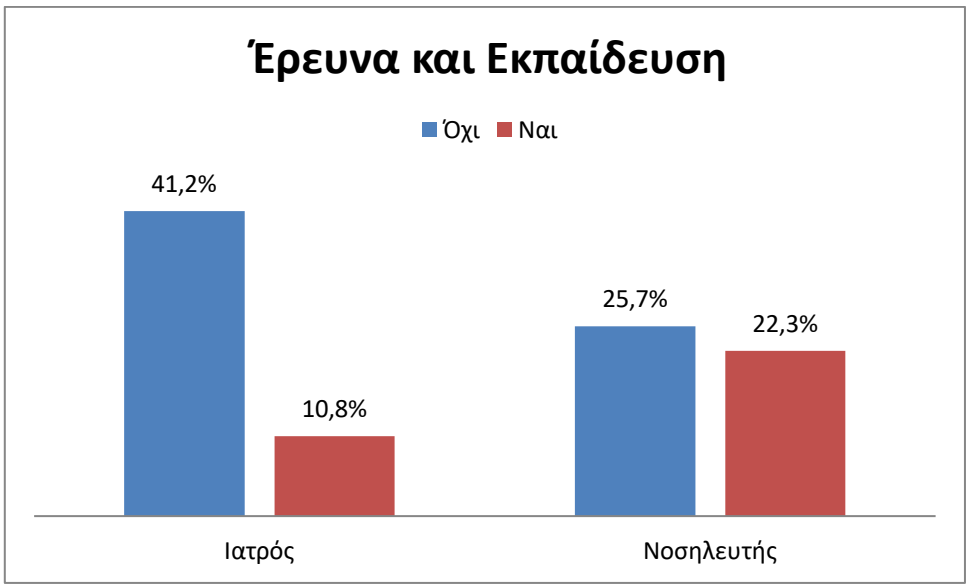
Γράφημα 25: Πηγές ενημέρωσης για Big Data

Από το Γράφημα 25 παρατηρούμε ότι το 34% των ιατρών και το 27% των νοσηλευτών έμαθαν για τα Big Data από Πανεπιστημιακές Παραδόσεις – Σεμινάρια. Στη συνέχεια οι ιατροί ανέφεραν το Διαδίκτυο σε ποσοστό 23%. Το Διαδίκτυο ανέφεραν και οι νοσηλευτές σε ποσοστό 16%. Τέλος το 13% των ιατρών και το 17% των νοσηλευτών ανέφεραν άλλη πηγή ενημέρωσης για την τεχνολογία των Big Data.



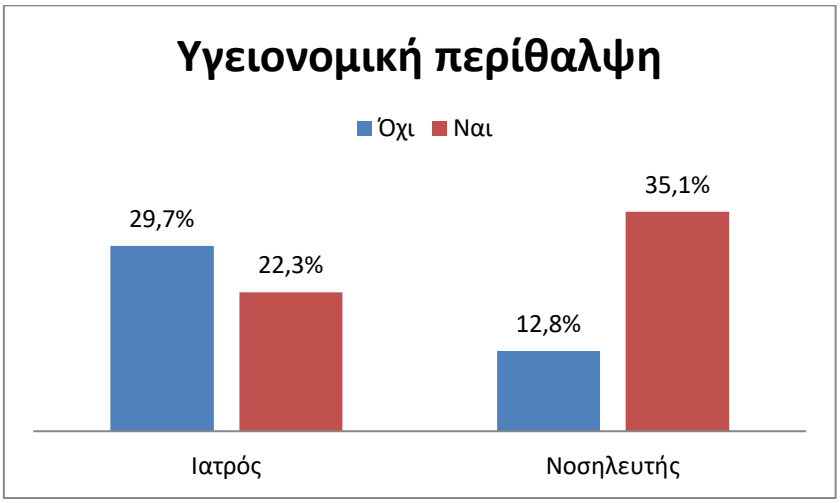
Γράφημα 26:Χρήση των Big Data στον τομέα της ενέργειας

Από το Γράφημα 26 παρατηρούμε ότι το 3,4% των ιατρών και το 5,4% των νοσηλευτών θεωρούν ότι η τεχνολογία των Big Data μπορεί να χρησιμοποιηθεί στον τομέα της ενέργειας.



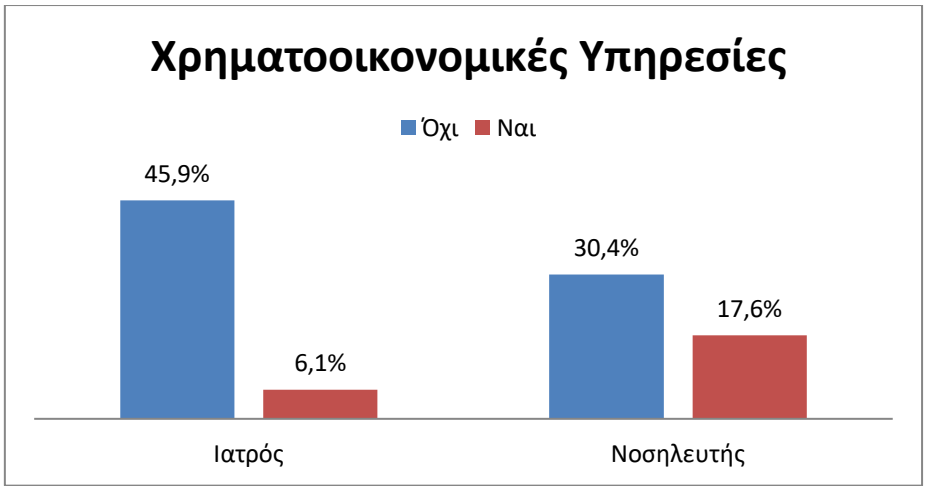
Γράφημα 27: Χρήση των Big Data στην Έρευνα και Εκπαίδευση

Το Γράφημα 27 δείχνει ότι το 10,8% των ιατρών θεωρεί ότι η τεχνολογία των Big Data μπορεί να χρησιμοποιηθεί στην Έρευνα και Εκπαίδευση. Υψηλότερο είναι το ποσοστό των νοσηλευτών που έχουν την ίδια άποψη. Πιο συγκεκριμένα το 22,3% θεωρεί ότι η τεχνολογία των Big Data μπορεί να χρησιμοποιηθεί στην Έρευνα και Εκπαίδευση.



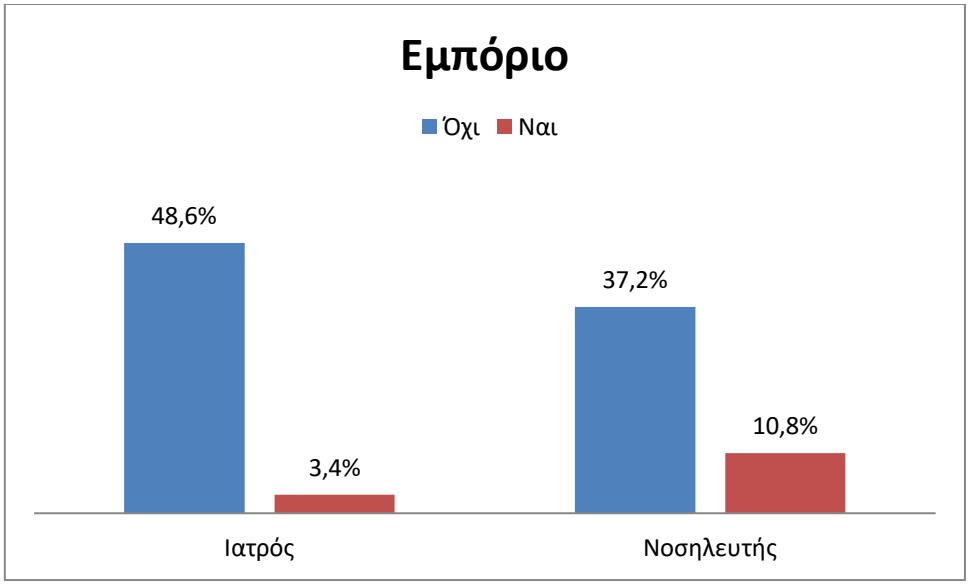
Γράφημα 28: Χρήση των Big Data στην Υγειονομική Περίθαλψη

Το Γράφημα 28 δείχνει ότι το 22,3% των ιατρών θεωρεί ότι η τεχνολογία των Big Data μπορεί να χρησιμοποιηθεί στην Υγειονομική Περίθαλψη. Το 35,1% των νοσηλευτών θεωρεί ότι η τεχνολογία των Big Data μπορεί να χρησιμοποιηθεί στην Υγειονομική Περίθαλψη, ποσοστό σημαντικά υψηλότερο από το αντίστοιχο των ιατρών.



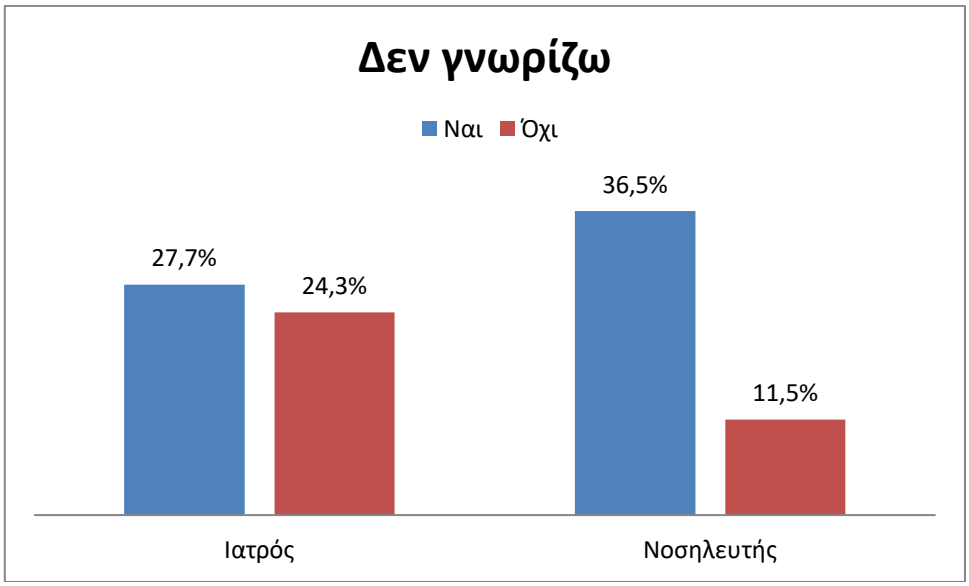
Γράφημα 29: Χρήση των Big Data στα χρηματοοικονομικά.

Το Γράφημα 29 δείχνει ότι μόλις το 6,1% των ιατρών θεωρεί ότι η τεχνολογία των Big Data μπορεί να χρησιμοποιηθεί σε Χρηματοοικονομικές Υπηρεσίες. Το 17,6% των νοσηλευτών θεωρεί ότι η τεχνολογία των Big Data μπορεί να χρησιμοποιηθεί στις Χρηματοοικονομικές Υπηρεσίες.



Γράφημα 30: Χρήση των Big Data στο Εμπόριο.

Το Γράφημα 30 δείχνει ότι το 48,6% των ιατρών θεωρεί ότι η τεχνολογία των Big Data δεν μπορεί να χρησιμοποιηθεί στο Εμπόριο. Το ίδιο πιστεύουν και οι νοσηλευτές σε ποσοστό 37,2%.

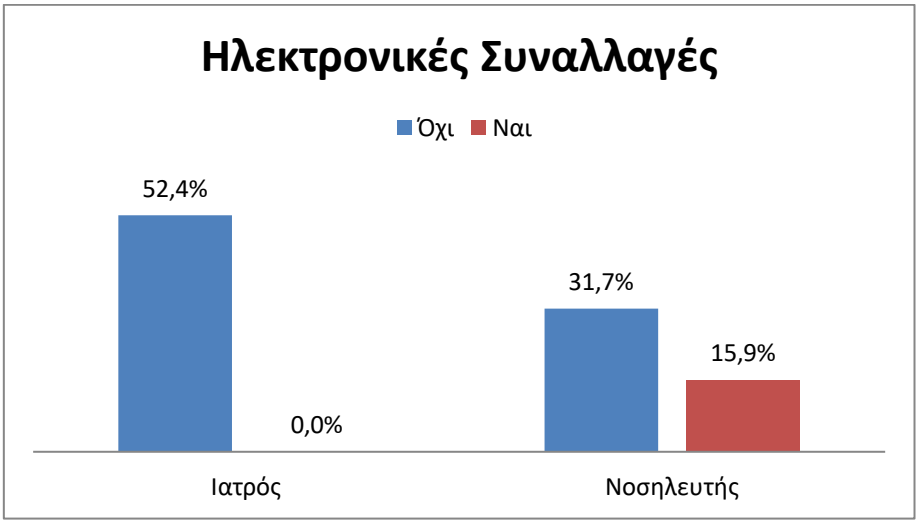


Γράφημα 31: Αποτελέσματα της απάντησης «Δεν γνωρίζω»

Από τα Γραφήματα 26,27,28,29,30 βλέπουμε ότι η πλειοψηφία των ιατρών δεν θεωρεί ότι η τεχνολογία των Big Data μπορεί να χρησιμοποιηθεί σε κάποιον από τους παραπάνω τομείς. Το ίδιο ισχύει και για τους νοσηλευτές με εξαίρεση τον τομέα της Υγειονομικής περίθαλψης. Σημαντικά υψηλότερο είναι και το ποσοστό των ιατρών (22,3%) που θεωρούν ότι η

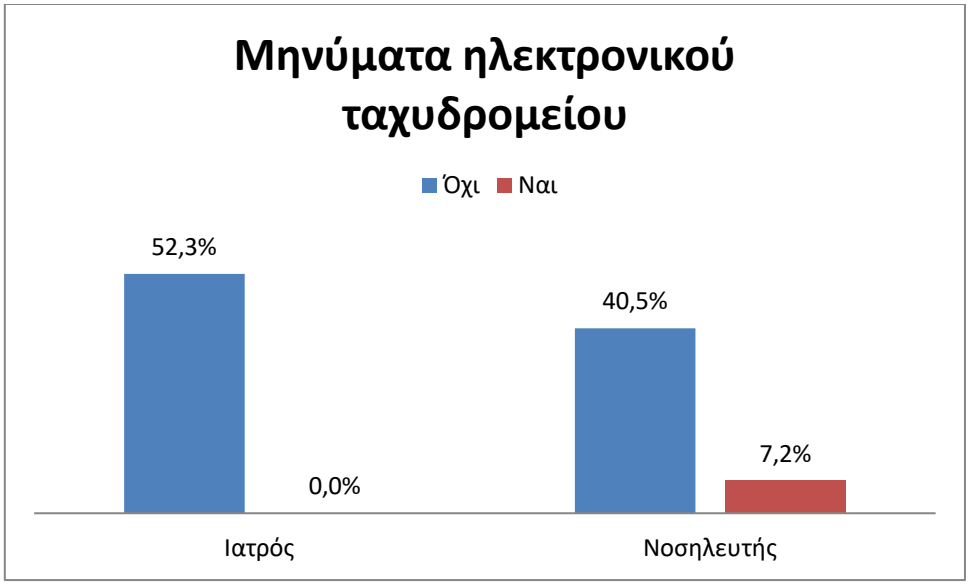
τεχνολογία των Big Data μπορεί να χρησιμοποιηθεί στον τομέα της Υγειονομικής περίθαλψης. Τέλος το Γράφημα 31 δείχνει ότι το 27,75 των ιατρών και το 36,5% των νοσηλευτών απάντησαν ότι δεν γνωρίζουν αν μπορεί η τεχνολογία των Big Data μπορεί να χρησιμοποιηθεί σε κάποιον από τους παραπάνω τομείς.

Στη συνέχεια οι ερωτηθέντες έπρεπε να απαντήσουν στη ερώτηση για το ποια από δεδομένα θεωρούν ότι μπορούν να χρησιμοποιηθούν ως πηγές για τη συλλογή και την δημιουργία των Big Data.



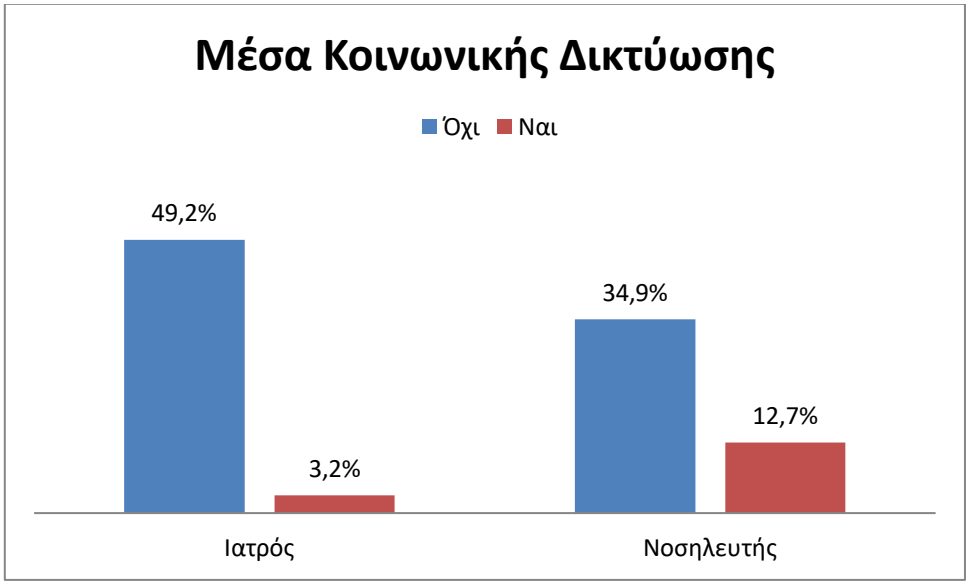
Γράφημα 32: Ηλεκτρονικές Συναλλαγές ως πηγή δεδομένων Big Data

Το Γράφημα 32 δείχνει ότι το 52,4% των ιατρών και το 31,7% των νοσηλευτών θεωρούν ότι δεδομένα ηλεκτρονικών συναλλαγών δεν μπορούν να χρησιμοποιηθούν ως πηγές δεδομένων για data.



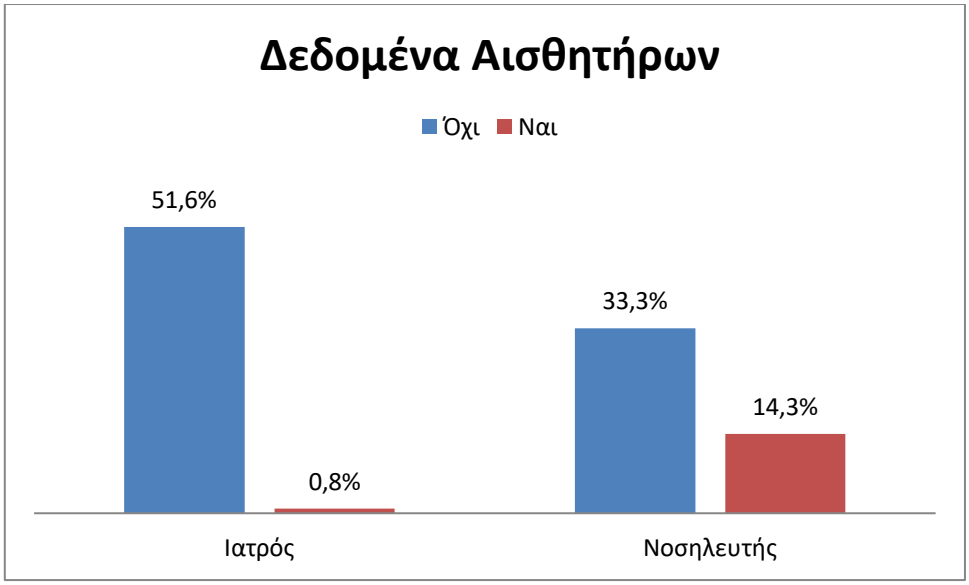
Γράφημα 33: Μηνύματα Ηλεκτρονικού ταχυδρομείου ως πηγή δεδομένων Big Data

Το Γράφημα 33 δείχνει ότι το 52,3% των ιατρών και το 40,5% των νοσηλευτών θεωρούν ότι τα δεδομένα από μηνύματα ηλεκτρονικού ταχυδρομείου δεν μπορούν να χρησιμοποιηθούν ως πηγές δεδομένων για data.



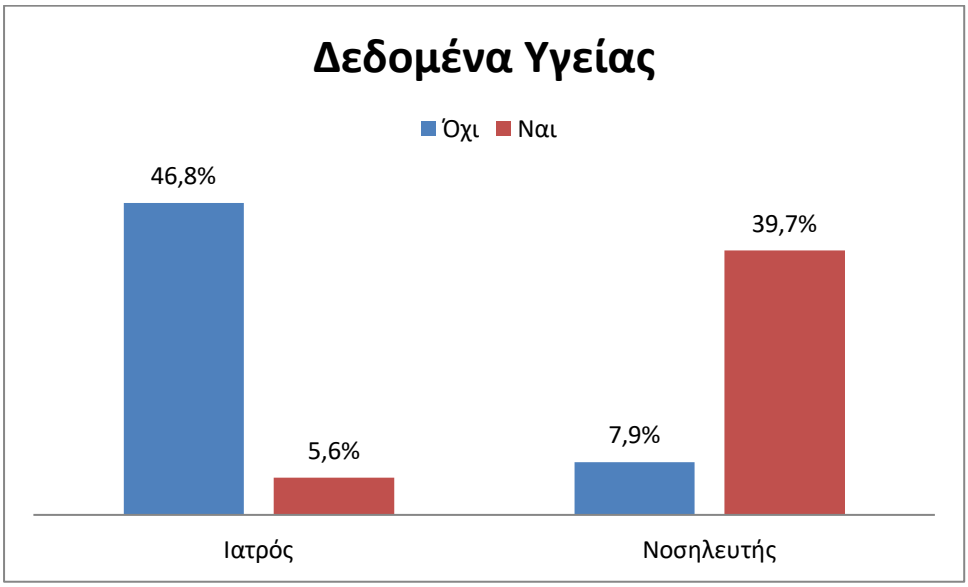
Γράφημα 34: Μέσα κοινωνικής δικτύωσης ως πηγή δεδομένων Big Data

Το Γράφημα 34 δείχνει ότι το 3,2% των ιατρών και το 12,7% των νοσηλευτών θεωρούν ότι δεδομένα των μέσων κοινωνικής δικτύωσης μπορούν να χρησιμοποιηθούν ως πηγές δεδομένων για data.



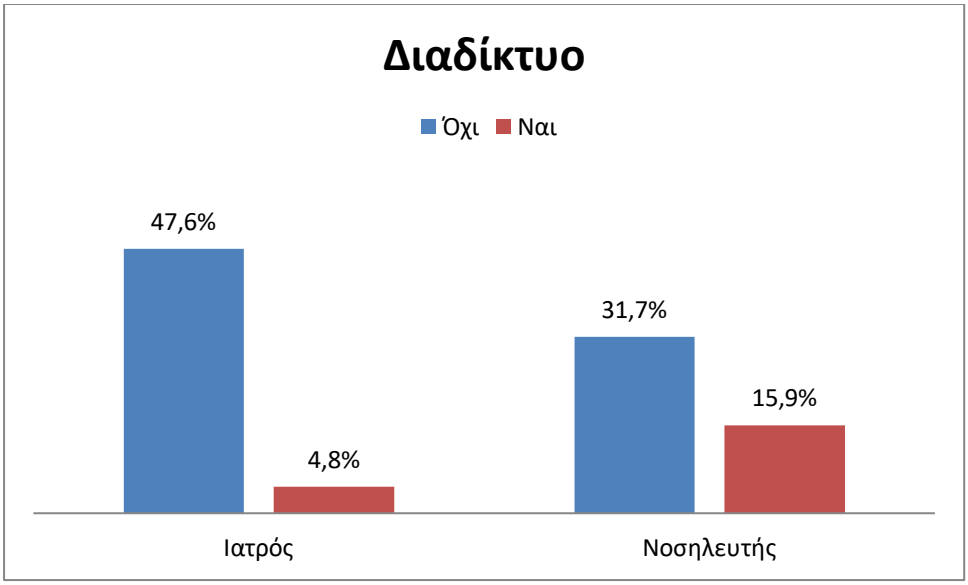
Γράφημα 35: Δεδομένα αισθητήρων ως πηγή δεδομένων Big Data

Το 0,8% (Γράφημα 35) των ιατρών πιστεύει ότι δεδομένα αισθητήρων μπορούν να χρησιμοποιηθούν ως πηγή δεδομένων για Big data. Αντίστοιχη άποψη εκφράζουν και οι νοσηλευτές σε ποσοστό 14,3%.



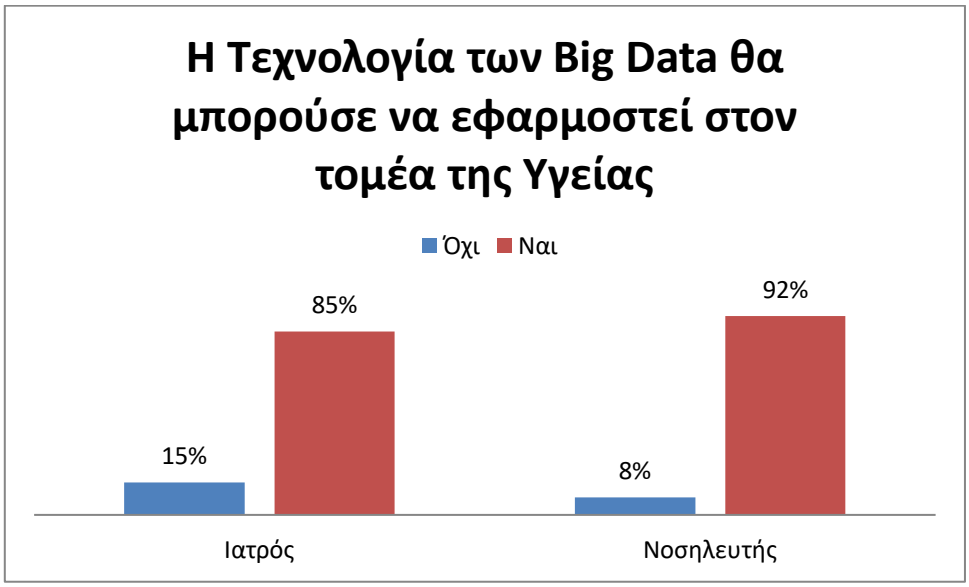
Γράφημα 36: Δεδομένα υγείας ως πηγή δεδομένων Big Data

Το 5,6% (Γράφημα 36) των ιατρών πιστεύει ότι τα δεδομένα υγείας μπορούν να χρησιμοποιηθούν ως πηγή δεδομένων για Big data. Το ποσοστό των νοσηλευτών που έχουν την ίδια άποψη είναι σημαντικά μεγαλύτερο από αυτό των ιατρών κατά 34,1%.



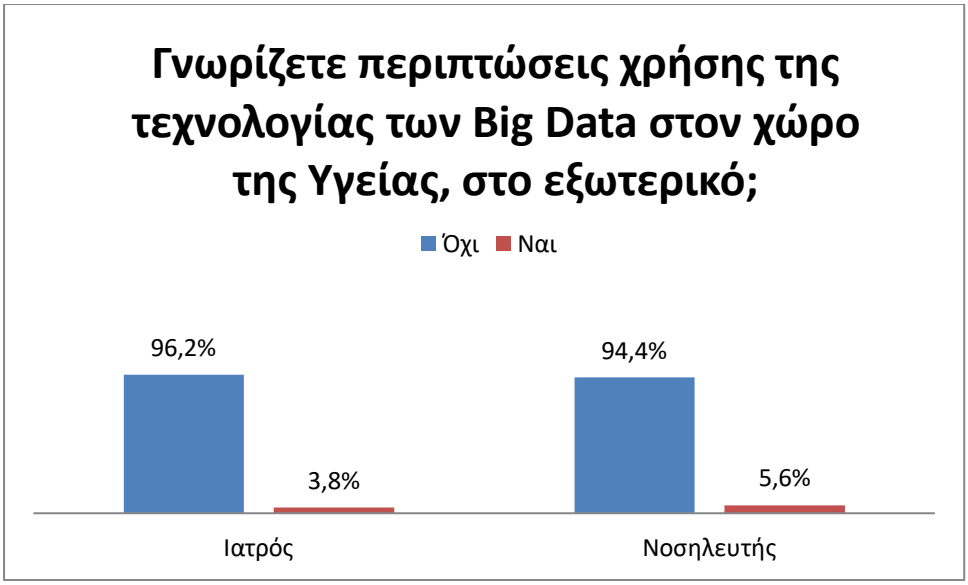
Γράφημα 37: Διαδίκτυο ως πηγή δεδομένων Big Data

Το 4,8% (Γράφημα 37) των ιατρών πιστεύει ότι δεδομένα από το διαδίκτυο μπορούν να χρησιμοποιηθούν ως πηγή δεδομένων για Big data. Αντίστοιχη άποψη εκφράζουν και οι νοσηλευτές σε ποσοστό 15.9%.



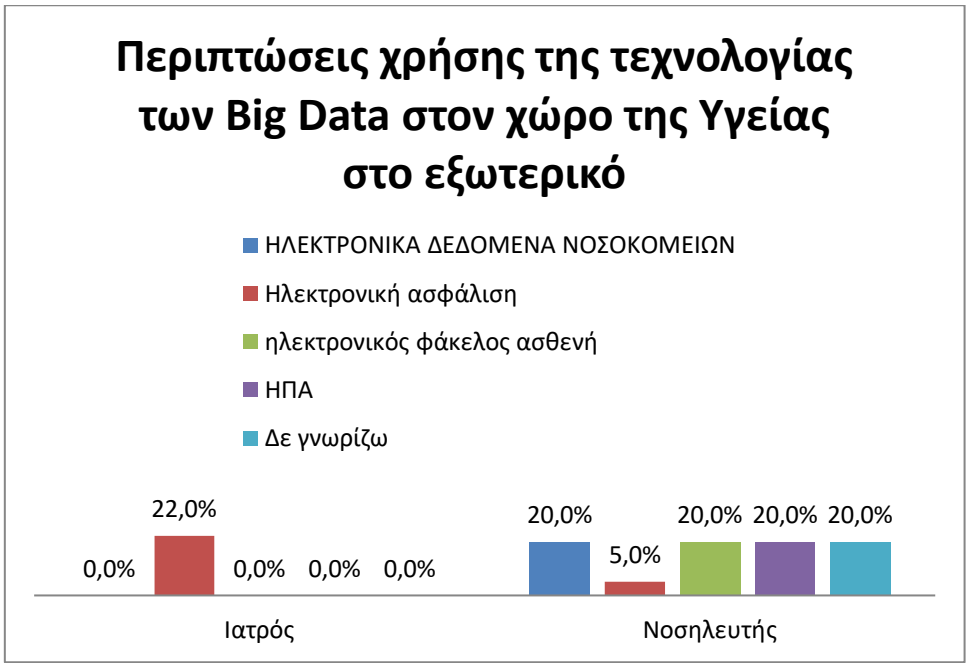
Γράφημα 38: Η Τεχνολογία των Big Data θα μπορούσε να εφαρμοστεί στον τομέα της Υγείας

Το Γράφημα 38 δείχνει ότι το 85% των ιατρών αλλά και το 92% των νοσηλευτών πιστεύει ότι η τεχνολογία των Big Data θα μπορούσε να εφαρμοστεί στον τομέα της Υγείας.



Γράφημα 39: Γνωρίζετε περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας, στο εξωτερικό;

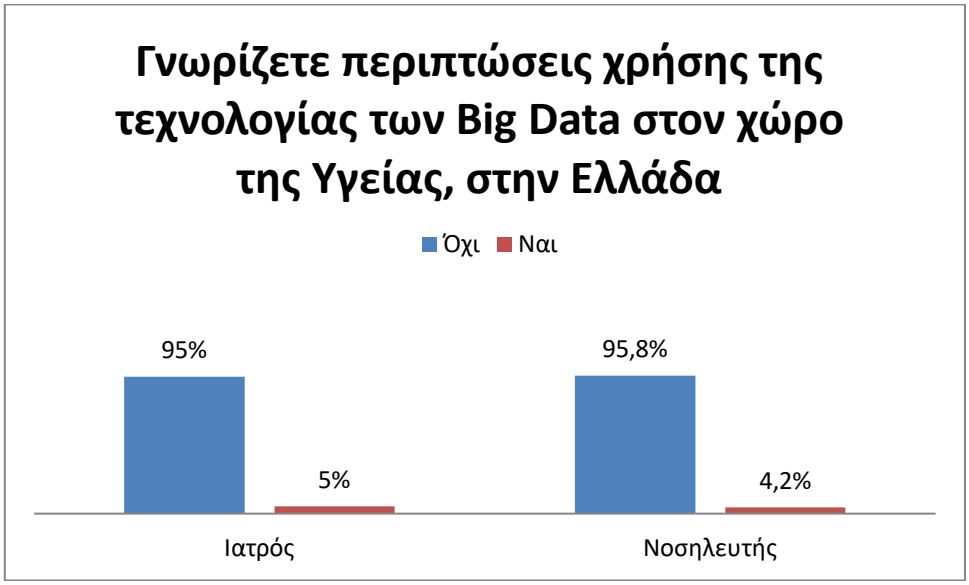
Το 96,2% των ιατρών και το 94,4% των νοσηλευτών αντίστοιχα δεν γνωρίζει περιπτώσεις χρήσης των Big Data στο εξωτερικό (Γράφημα 39).



Γράφημα 40: Περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας στο εξωτερικό

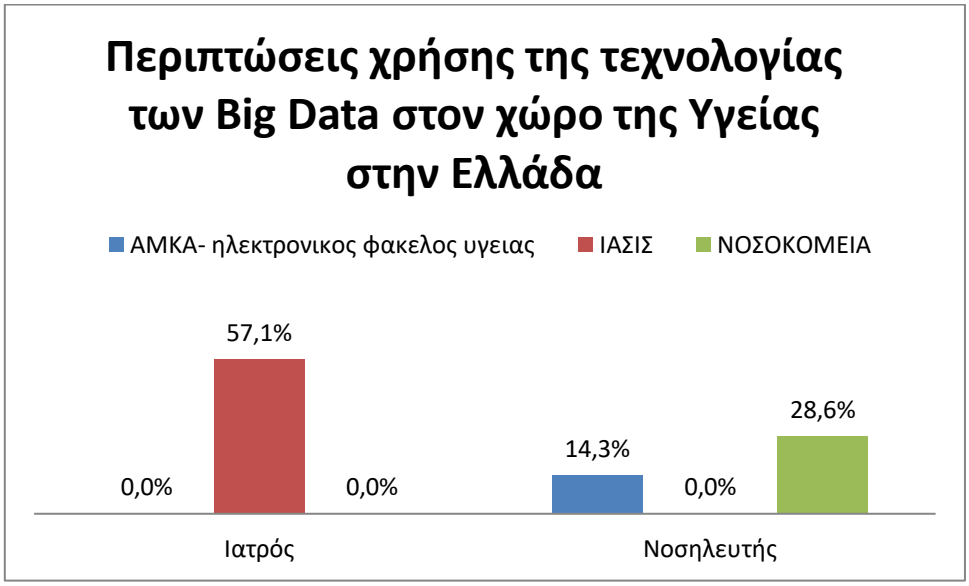
Οι ιατροί ανέφεραν σε ποσοστό 22% τον τομέα της ηλεκτρονικής ασφάλισης ως περίπτωση χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας στο εξωτερικό (Γράφημα 40). Αντίστοιχα οι νοσηλευτές ανέφεραν σε ποσοστό 20% τα ηλεκτρονικά δεδομένα νοσοκομείων, 20% την ηλεκτρονική ασφάλιση, 20% την ηλεκτρονικό φάκελο ασθενή, 20% τις ΗΠΑ

και 5% την ηλεκτρονική ασφάλιση. Επίσης ένα 20% των νοσηλευτών απάντησε ότι δεν γνωρίζει.



Γράφημα 41: Γνωρίζετε περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας, στην Ελλάδα

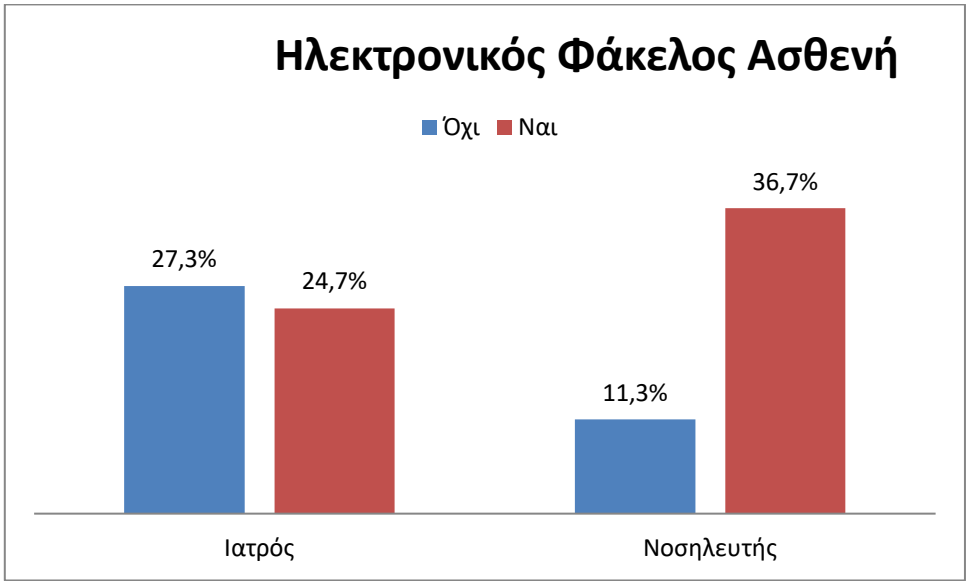
Το 95% (Γράφημα 41) των ιατρών και το 95,8% των νοσηλευτών δεν γνωρίζει περιπτώσεις χρήσης Big data στην Ελλάδα. Όσοι γνώριζαν ανέφεραν τα παρακάτω.



Γράφημα 42: Περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας στην Ελλάδα

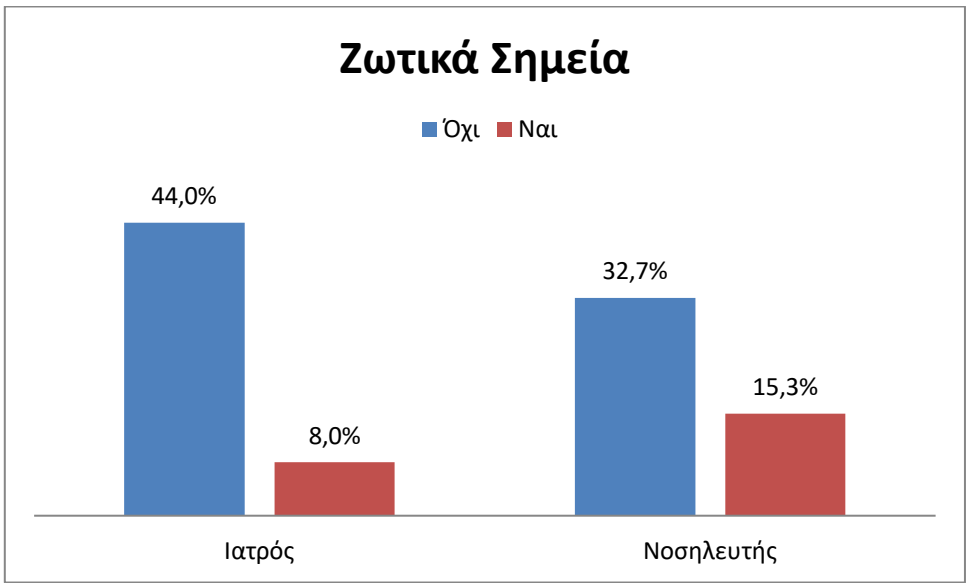
Για τις περιπτώσεις χρήσης των Big Data στην Ελλάδα οι νοσηλευτές σε ποσοστό 14,3 % ανέφεραν τον ΑΜΚΑ, τον Ηλεκτρονικό φάκελο υγείας σε

ποσοστό 28,6% (Γράφημα 42). Αντίστοιχα οι ιατροί ανέφεραν μόνο το ΙΑΣΥΣ σε ποσοστό 57,1%.



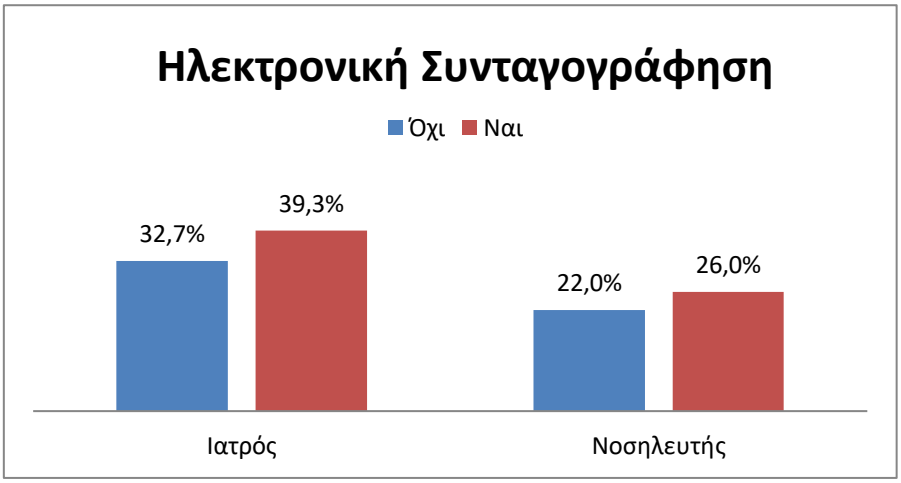
Γράφημα 43: Ηλεκτρονικός Φάκελος Ασθενή

Παρατηρούμε στο Γράφημα 43 ότι το 24,7% των ιατρών και το 36,7% των νοσηλευτών πιστεύουν ότι ο Ηλεκτρονικός Φάκελος Ασθενή μπορούσε να χρησιμοποιηθεί ως πηγή συλλογής δεδομένων για Big Data στον τομέα της υγείας.



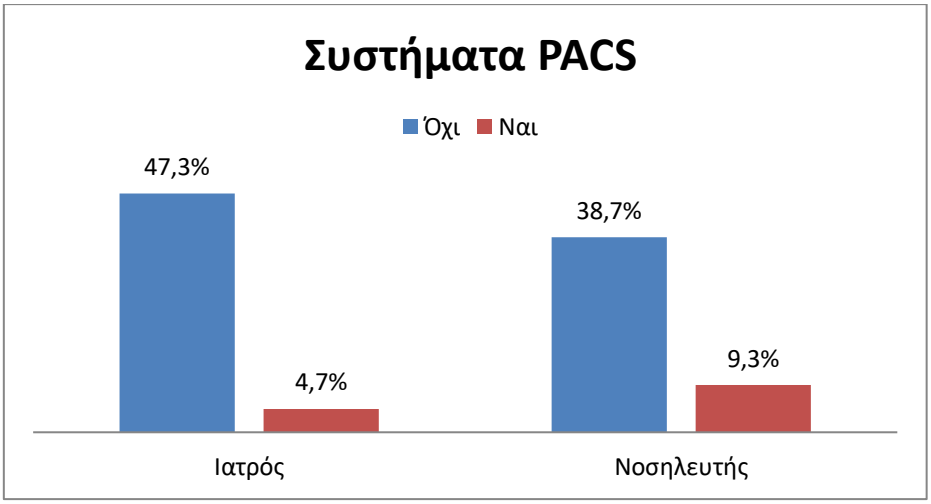
Γράφημα 44: Ζωτικά σημεία

Το Γράφημα 44 δείχνει ότι το 8% των ιατρών και το 15,3% των νοσηλευτών πιστεύουν ότι δεδομένα από ζωτικά σημεία θα μπορούσαν να χρησιμοποιηθούν ως πηγή συλλογής δεδομένων για Big Data στον τομέα της υγείας.



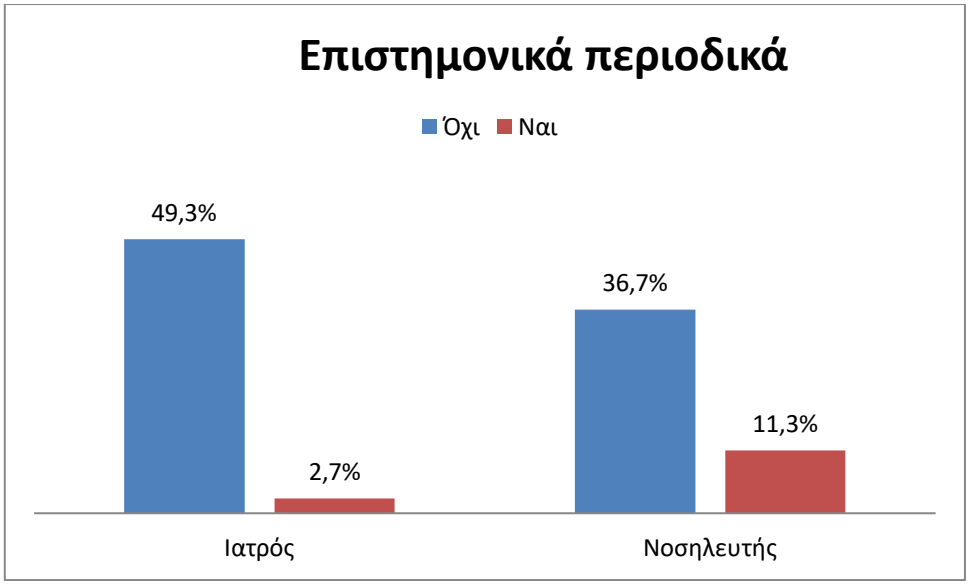
Γράφημα 45: Ηλεκτρονική Συνταγογράφηση

Το Γράφημα 45 δείχνει ότι το 39,3% των ιατρών και το 26% των νοσηλευτών πιστεύουν ότι δεδομένα της ηλεκτρονικής συνταγογράφησης θα μπορούσαν να χρησιμοποιηθούν ως πηγή συλλογής δεδομένων για Big Data στον τομέα της υγείας.



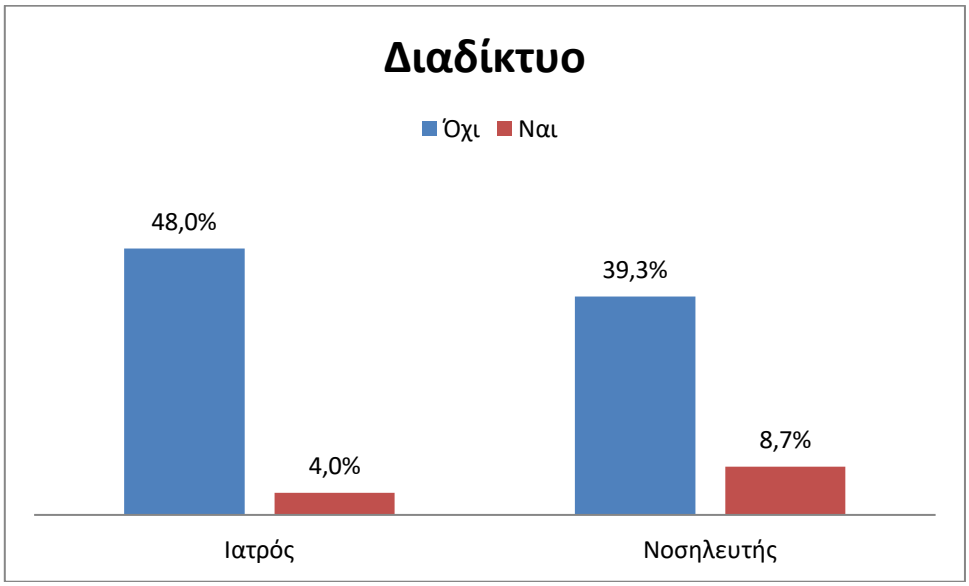
Γράφημα 46: Συστήματα PACS

Το Γράφημα 46 δείχνει ότι το 4,7% των ιατρών και το 9,3% των νοσηλευτών πιστεύουν ότι δεδομένα από συστήματα PACS θα μπορούσαν να χρησιμοποιηθούν ως πηγή συλλογής δεδομένων για Big Data στον τομέα της υγείας.



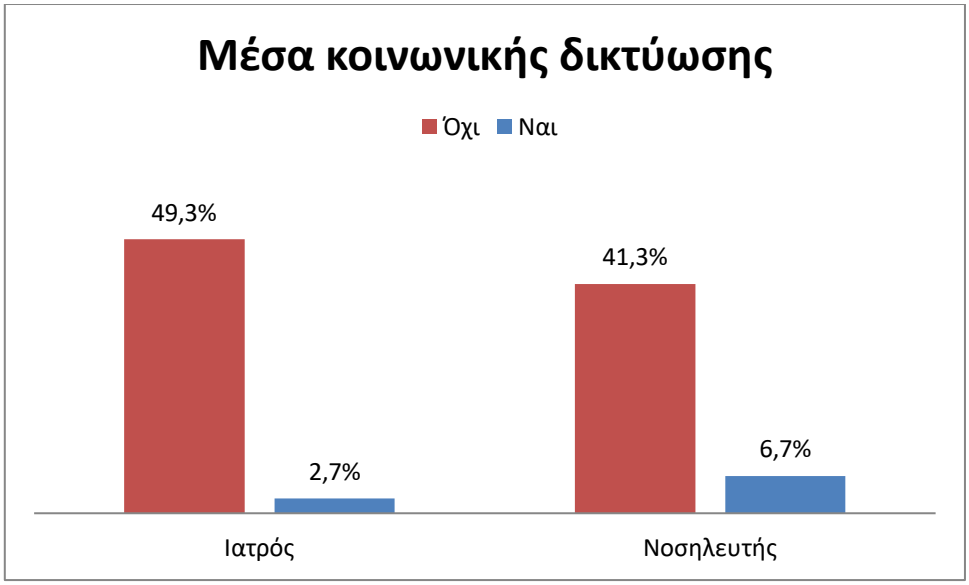
Γράφημα 47: Επιστημονικά περιοδικά

Το Γράφημα 47 δείχνει ότι το 2,7% των ιατρών και το 11,3% των νοσηλευτών πιστεύουν ότι δεδομένα από επιστημονικά περιοδικά θα μπορούσαν να χρησιμοποιηθούν ως πηγή συλλογής δεδομένων για Big Data στον τομέα της υγείας.



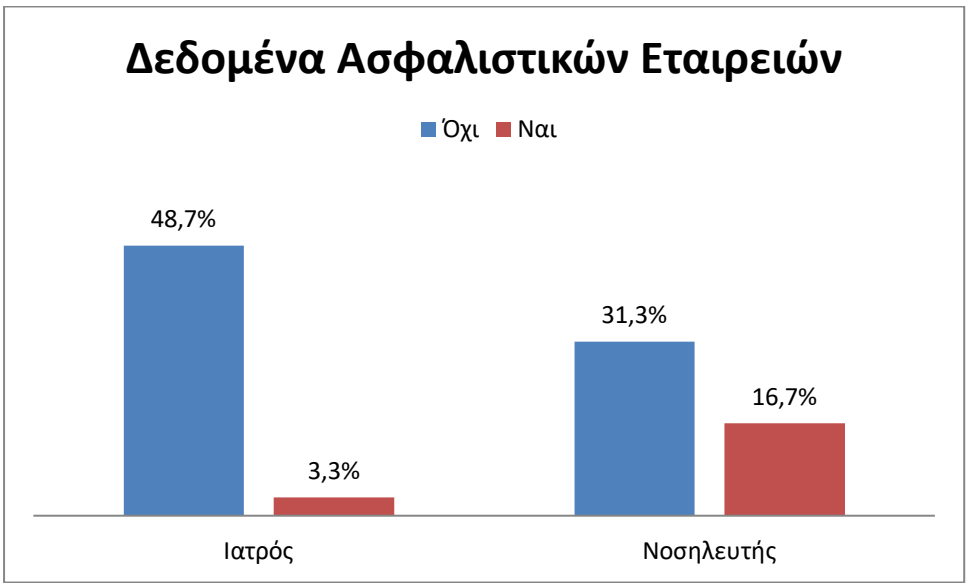
Γράφημα 48: Διαδίκτυο

Το Γράφημα 48 δείχνει ότι το 4% των ιατρών και το 8,7% των νοσηλευτών πιστεύουν ότι δεδομένα διαδικτύου θα μπορούσαν να χρησιμοποιηθούν ως πηγή συλλογής δεδομένων για Big Data στον τομέα της υγείας.



Γράφημα 49: Μέσα κοινωνικής δικτύωσης

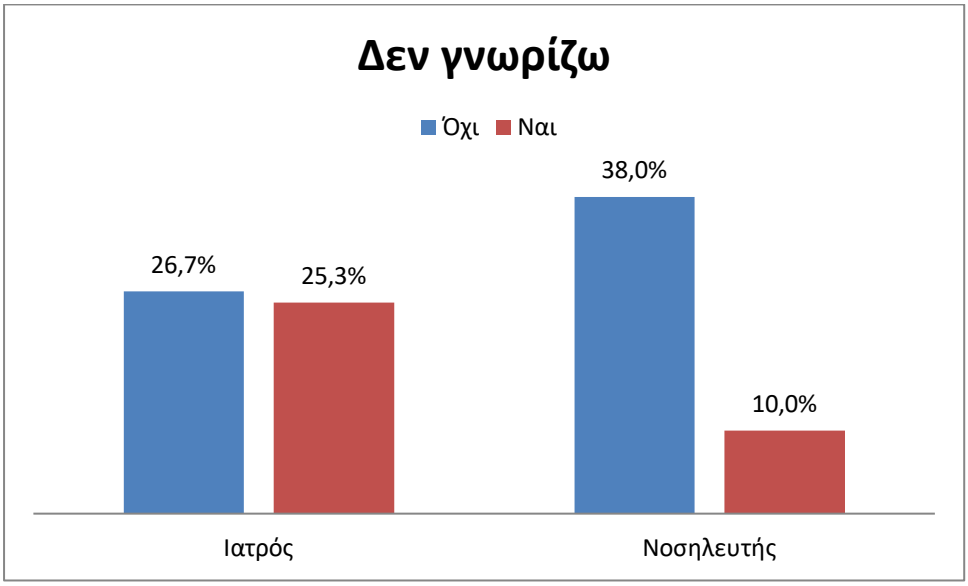
Το Γράφημα 49 δείχνει ότι το 2,7% των ιατρών και το 6,7% των νοσηλευτών πιστεύουν ότι δεδομένα από τα μέσα κοινωνικής δικτύωσης θα μπορούσαν να χρησιμοποιηθούν ως πηγή συλλογής δεδομένων για Big Data στον τομέα της υγείας.



Γράφημα 50: Δεδομένα Ασφαλιστικών Εταιρειών

Το Γράφημα 50 δείχνει ότι το 3,3% των ιατρών και το 26,7% των νοσηλευτών πιστεύουν ότι δεδομένα ασφαλιστικών εταιρειών θα μπορούσαν

να χρησιμοποιηθούν ως πηγή συλλογής δεδομένων για Big Data στον τομέα της υγείας.



Γράφημα 51:Αποτελέσματα της απάντησης «Δεν γνωρίζω»

Το Γράφημα 51 δείχνει ότι το 25,3% των ιατρών και το 10,0% των νοσηλευτών πιστεύουν ότι δεδομένα ασφαλιστικών εταιρειών θα μπορούσαν να χρησιμοποιηθούν ως πηγή συλλογής δεδομένων για Big Data στον τομέα της υγείας.

Από τα παραπάνω γραφήματα βλέπουμε ότι η πλειοψηφία των ιατρών δεν γνωρίζει ποια δεδομένα θα μπορούσαν να χρησιμοποιηθούν ως πηγές για τη συλλογή δεδομένων και την δημιουργία των Big Data . Το ίδιο ισχύει και για τους νοσηλευτές με εξαίρεση την ηλεκτρονική συνταγογράφηση και τον ηλεκτρονικό φάκελο ασθενή.

Για την καταγραφή της άποψης των επαγγελματιών υγείας σχετικά την τεχνολογία των Big Data χρησιμοποιήθηκαν ερωτήσεις σε κλίμακα Likert με 7 διαβαθμίσεις όπως φαίνεται παρακάτω:

Διαφωνώ Απόλυτα 1 2 3 4 5 6 7 Συμφωνώ Απόλυτα

Για την κάθε ερώτηση υπολογίστηκε ο μέσος όρος όπως φαίνεται στον Πίνακα 2 (Παράρτημα Β). Τα αποτελέσματα του πίνακα δείχνουν ότι:

- οι ερωτηθέντες συμφωνούν με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας είναι χρήσιμη.
- οι ερωτηθέντες συμφωνούν με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα μπορούσε να αυξήσει την αποτελεσματικότητα των παρεχόμενων Υπηρεσιών Υγείας.
- οι ερωτηθέντες ούτε συμφωνούν ούτε διαφωνούν με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα βοηθήσει τους Επαγγελματίες Υγείας στη διαδικασία λήψης αποφάσεων.
- οι ερωτηθέντες συμφωνούν μερικώς με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα δώσει τη δυνατότητα παροχής υπηρεσιών υγείας προσαρμοσμένων στις ανάγκες των ασθενών.
- οι ερωτηθέντες συμφωνούν με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα δώσει τη δυνατότητα παροχής υπηρεσιών υγείας προσαρμοσμένων στις ανάγκες των ασθενών.

8.3.2 Επαγωγική στατιστική

Για τον έλεγχο της αξιοπιστίας του ερωτηματολογίου υπολογίστηκε ο δείκτης a του Cronbach. Από τον Πίνακα 3 (Παράρτημα Β) βλέπουμε ότι a= 0,888 που δείχνει υψηλή αξιοπιστία.

8.3.2.1 Έλεγχος Κανονικότητας για Ηλικία

- Ho: Η κατανομή της Ηλικίας είναι κανονική.
- H1: Η κατανομή της Ηλικίας δεν είναι κανονική.

p-value=0,000<0,05 (Πίνακας 4-Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς η μεταβλητή Ηλικία δεν ακολουθεί την κανονική κατανομή.

8.3.2.2 Συσχέτιση Ηλικίας -Γνώσης Big Data

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στη Γνώση των Big Data.
- H1: Υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στη Γνώση των Big Data.

p-value=0,018< 0,05 (Πίνακας 5-Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει συσχέτιση μεταξύ της ηλικίας και γνώσης των Big Data. Η τιμή του συντελεστή συσχέτισης είναι -0,229 (Πίνακας 6-Παράρτημα Β) που δείχνει αρνητική συσχέτιση. Δηλαδή όσο μικρότερη είναι η ηλικία τόσο καλύτερη η γνώση των Big Data.

8.3.2.3 Συσχέτιση Φύλου - Γνώσης Big Data

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στο Φύλο και στη Γνώση των Big Data.
- H1: Υπάρχει συσχέτιση ανάμεσα στο Φύλο και στη Γνώση των Big Data.

p-value=0,578>0,05 (Πίνακας 7 - Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ του φύλου και γνώσης των Big Data.

8.3.2.4 Συσχέτιση Επαγγέλματος -Γνώσης Big Data

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στο Επάγγελμα και στη Γνώση των Big Data.
- H1: Υπάρχει συσχέτιση ανάμεσα στο Επάγγελμα και στη Γνώση των Big Data.

p-value=0,697>0,05 (Πίνακας 8 - Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ του Επαγγέλματος και Γνώσης των Big Data.

8.3.2.5 Συσχέτιση Επαγγελματικής Εμπειρίας -Γνώσης Big Data

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στην Επαγγελματική Εμπειρία και στη Γνώση των Big Data.
- H1: Υπάρχει συσχέτιση ανάμεσα στην Επαγγελματική Εμπειρία και στη Γνώση των Big Data.

p-value = 0,714>0,05 (Πίνακας 9 - Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ της Επαγγελματικής Εμπειρίας και Γνώσης των Big Data.

8.3.2.6 Συσχέτιση Διάρκειας Επαγγελματικής Εμπειρίας -Γνώσης Big Data

- Έλεγχος Κανονικότητας για Διάρκεια Επαγγελματικής Εμπειρίας

  - Ho: Η κατανομή της μεταβλητής είναι κανονική.
  - H1: Η κατανομή της μεταβλητής δεν είναι κανονική.

p-value=0,000<0,05 (Πίνακας 10 - Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς η μεταβλητή Ηλικία δεν ακολουθεί την κανονική κατανομή.

  - Ho: Δεν υπάρχει συσχέτιση ανάμεσα στη Διάρκεια Επαγγελματικής Εμπειρίας και στη Γνώση των Big Data.
  - H1: Υπάρχει συσχέτιση ανάμεσα στη Διάρκεια Επαγγελματικής Εμπειρίας και στη Γνώση των Big Data.

p-value=0,002<0,05 (Πίνακας 11 - Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει ασθενής αρνητική (Πίνακας 12 - Παράρτημα Β) συσχέτιση μεταξύ της Διάρκειας Επαγγελματικής Εμπειρίας και Γνώσης των Big Data αφού r = -0,278. Ειδικότερα όσο μεγαλύτερη είναι η επαγγελματική εμπειρία τόσο μικρότερη είναι η γνώση για τα Big Data.

8.3.2.7 Συσχέτιση Φύλου – Μορφή των Big Data

- Δομημένα δεδομένα

  - Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.
  - H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

p-value=0,89>0,05 (Πίνακας 13 - Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

- Αδόμητα δεδομένα

  - Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.
  - H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

p-value=0,117>0,05 (Πίνακας 14 - Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

- Ημιδομημένα δεδομένα

  - Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.
  - H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

p-value=0,782>0,05 (Πίνακας 15 - Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

8.3.2.8 Συσχέτιση Ηλικίας– Μορφή των Big Data

- Δομημένα δεδομένα
  - Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.
  - H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

p-value=0,010<0,05 (Πίνακας 16 -Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές. Συγκεκριμένα r = -0,239 (Πίνακας 17 -Παράρτημα Β) που σημαίνει ότι άτομα μικρότερης ηλικίας πιστεύουν ότι τα Big Data αφορούν κυρίως δομημένα δεδομένα.

- Αδόμητα δεδομένα

  - Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.
  - H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

p-value=0,78 0>0,05 (Πίνακας 18 -Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

- Ημιδομημένα δεδομένα

  - Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.
  - H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

p-value=0,004<0,05 (Πίνακας 19 -Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές. Συγκεκριμένα r = -0,213 (Πίνακας 20 -Παράρτημα Β) που σημαίνει ότι άτομα μικρότερης ηλικίας πιστεύουν ότι τα Big Data αφορούν κυρίως ημιδομημένα δεδομένα.

8.3.2.9 Συσχέτιση Επαγγέλματος– Μορφή των Big Data

- Δομημένα δεδομένα

  - Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.
  - H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

p-value=0,338>0,05 (Πίνακας 21-Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

- Αδόμητα δεδομένα
  - Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.
  - H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

p-value =0,05=0,05 (Πίνακας 22 -Παράρτημα Β). Οριακά απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές. Συγκεκριμένα r = -0,05 (Πίνακας 23 -Παράρτημα Β) που σημαίνει ότι οι ιατροί ή οι νοσηλευτές πιστεύουν ότι τα Big Data αφορούν κυρίως ημιδομημένα δεδομένα.

- Ημιδομημένα δεδομένα

  - Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.
  - H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

p-value=0,464>0,05 (Πίνακας 24 -Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

8.3.2.10 Έλεγχος Συσχέτισης Φύλου –Σκορ Χρησιμότητας των Big Data

- Έλεγχος Κανονικότητας για Σκορ Χρησιμότητας των Big Data

- Ho: Η κατανομή του Σκορ Χρησιμότητας των Big Data είναι κανονική.
- Η1: Η κατανομή Σκορ Χρησιμότητας των Big Data δεν είναι κανονική.

p-value=0,000<0,05 (Πίνακας 25 -Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς η μεταβλητή Σκορ Χρησιμότητας δεν ακολουθεί την κανονική κατανομή.

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Χρησιμότητας.
- Η1: Υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Χρησιμότητας.

p-value=0,045<0,05 (Πίνακας 26 -Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει ασθενής αρνητική συσχέτιση μεταξύ του Φύλου και Σκορ Χρησιμότητας αφού r= -0,177 (Πίνακας 27 -Παράρτημα Β).

8.3.2.11 Έλεγχος Συσχέτισης Φύλου –Σκορ Αποτελεσματικότητας Υπηρεσιών Υγείας

- Έλεγχος Κανονικότητας για Σκορ Αποτελεσματικότητας Υπηρεσιών Υγείας

- Ho: Η κατανομή του Σκορ Αποτελεσματικότητας είναι κανονική.
- Η1: Η κατανομή του Σκορ Αποτελεσματικότητας των Big Data δεν είναι κανονική.

p-value=0,000<0,05 (Πίνακας 28 -Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς η μεταβλητή Σκορ Χρησιμότητας δεν ακολουθεί την κανονική κατανομή.

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Αποτελεσματικότητας.
- Η1: Υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Αποτελεσματικότητας.

p-value= 0,009<0,05 (Πίνακας 29 -Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει ασθενής αρνητική συσχέτιση μεταξύ του Φύλου και Σκορ Αποτελεσματικότητας αφού r= -0,214 (Πίνακας 30 -Παράρτημα Β) σημαίνει ότι οι άνδρες θεωρούν σε μεγαλύτερο βαθμό από τις

γυναίκες ότι η χρήση των Big Data θα αυξήσει την αποτελεσματικότητα των παρεχόμενων υπηρεσιών υγείας.

8.3.2.12 Έλεγχος Συσχέτισης Φύλου –Σκορ Αποφάσεων

- Έλεγχος Κανονικότητας για Σκορ Λήψης Αποφάσεων

  - Ho: Η κατανομή του Σκορ Αποφάσεων είναι κανονική.
  - H1: Η κατανομή του Σκορ Αποφάσεων δεν είναι κανονική.

p-value=0,000<0,05 (Πίνακας 31 -Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς η μεταβλητή Σκορ Αποφάσεων δεν ακολουθεί την κανονική κατανομή.

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Αποφάσεων.
- H1: Υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Αποφάσεων.

p-value=0,394>0,05 (Πίνακας 32 -Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ του Φύλου και Σκορ Αποφάσεων.

8.3.2.13 Έλεγχος Συσχέτισης Φύλου –Σκορ Παροχής Υπηρεσιών Υγείας

- Έλεγχος Κανονικότητας για Σκορ Παροχής Υπηρεσιών Υγείας

  - Ho: Η κατανομή του Σκορ Παροχής Υπηρεσιών Υγείας είναι κανονική.
  - H1: Η κατανομή του Σκορ Παροχής Υπηρεσιών Υγείας δεν είναι κανονική.

p-value=0,000<0,05 (Πίνακας 33 -Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς η μεταβλητή Σκορ Παροχής Υπηρεσιών Υγείας δεν ακολουθεί την κανονική κατανομή.

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Παροχής Υπηρεσιών Υγείας.
- H1: Υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Παροχής Υπηρεσιών Υγείας.

p-value=0,924>0,05 (Πίνακας 34-Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ του Φύλου και Σκορ Παροχής Υπηρεσιών Υγείας.

8.3.2.14 Έλεγχος Συσχέτισης Φύλου –Σκορ Πρόληψης

- Έλεγχος Κανονικότητας για Σκορ Πρόληψης

  - Ho: Η κατανομή του Σκορ Πρόληψης είναι κανονική.
  - H1: Η κατανομή του Σκορ Πρόληψης δεν είναι κανονική.

p-value=0,000<0,05 (Πίνακας 35 -Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς η μεταβλητή Σκορ Πρόληψης δεν ακολουθεί την κανονική κατανομή.

  - Ho: Δεν υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Πρόληψης.
  - H1: Υπάρχει συσχέτιση ανάμεσα στο Φύλο και στο Σκορ Πρόληψης.

p-value=0,235>0,05 (Πίνακας 36-Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ του Φύλου και του Σκορ Πρόληψης.

8.3.2.15 Έλεγχος Συσχέτισης Ηλικίας –Σκορ Χρησιμότητας των Big Data

- Έλεγχος Κανονικότητας για Σκορ Χρησιμότητας των Big Data
  - Ho: Η κατανομή του Σκορ Χρησιμότητας των Big Data είναι κανονική.
  - H1: Η κατανομή Σκορ Χρησιμότητας των Big Data δεν είναι κανονική.

p-value=0,000<0,05 (Πίνακας 37-Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς η μεταβλητή Σκορ Χρησιμότητας δεν ακολουθεί την κανονική κατανομή.

  - Ho: Δεν υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στο Σκορ Χρησιμότητας
  - H1: Υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στο Σκορ Χρησιμότητας

p-value=0,000<0,05 (Πίνακας 38-Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει αρνητική συσχέτιση μεταξύ των δύο μεταβλητών καθώς r= -0,519. Συγκεκριμένα τα άτομα μεγαλύτερη ηλικίας θεωρούν ότι η τεχνολογία των Big data στον τομέα της υγείας δεν είναι χρήσιμη σε αντίθεση με άτομα νεότερης ηλικίας.

8.3.2.16 Έλεγχος Συσχέτισης Ηλικίας –Σκορ Αποτελεσματικότητας Υπηρεσιών Υγείας

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στο Σκορ Αποτελεσματικότητας.
- Η1: Υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στο Σκορ Αποτελεσματικότητας

p-value=0,000<0,05 (Πίνακας 39-Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει αρνητική συσχέτιση μεταξύ των δύο μεταβλητών καθώς r= -0,516. Συγκεκριμένα τα άτομα μεγαλύτερη ηλικίας θεωρούν ότι η τεχνολογία των Big Data στον τομέα της υγείας δεν θα αυξήσει την αποτελεσματικότητα των παρεχόμενων υπηρεσιών υγείας σε αντίθεση με τα άτομα νεότερης ηλικίας.

8.3.2.17 Έλεγχος Συσχέτισης Ηλικίας –Σκορ Αποφάσεων

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στο Σκορ Αποφάσεων.
- Η1: Υπάρχει συσχέτιση ανάμεσα στην Ηλικία και στο Σκορ Αποφάσεων.

p-value= 0,08>0,05 (Πίνακας 40 -Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ των δύο μεταβλητών.

8.3.2.18 Έλεγχος Συσχέτισης Ηλικίας –Σκορ Παροχής Υπηρεσιών Υγείας

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές
- Η1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές

p-value=0,034<0,05 (Πίνακας 41-Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει αρνητική συσχέτιση μεταξύ των δύο μεταβλητών καθώς r= -0,187. Συγκεκριμένα τα άτομα μεγαλύτερη ηλικίας θεωρούν ότι η τεχνολογία των Big Data στον τομέα της υγείας δεν θα δώσει

τη δυνατότητα παροχής υπηρεσιών υγείας προσαρμοσμένων στις ανάγκες των ασθενών.

8.3.2.19 Έλεγχος Συσχέτισης Ηλικίας –Σκορ Πρόληψης

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.
- H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

p-value=0,000<0,05 (Πίνακας 42-Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει αρνητική συσχέτιση μεταξύ των δύο μεταβλητών καθώς r= -0,396. Συγκεκριμένα τα άτομα μεγαλύτερη ηλικίας θεωρούν ότι η τεχνολογία των Big Data στον τομέα της υγείας δεν συμβάλει στην αποτελεσματικότερη πρόληψη στους πληθυσμούς υψηλού κινδύνου.

8.3.2.20 Έλεγχος Συσχέτισης Επαγγέλματος –Σκορ Χρησιμότητας των Big Data

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στο Επάγγελμα και στο Σκορ Χρησιμότητας
- H1: Υπάρχει συσχέτιση ανάμεσα στο Επάγγελμα και στο Σκορ Χρησιμότητας

p-value=0,000<0,05 (Πίνακας 43-Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει θετική συσχέτιση μεταξύ του Επαγγέλματος και Σκορ Χρησιμότητας αφού r= 0,342 (Πίνακας 44-Παράρτημα Β). Ειδικότερα τόσο οι ιατροί όσο και οι νοσηλευτές θεωρούν χρήσιμη την τεχνολογία των Big Data στον τομέα της Υγείας.

8.3.2.21 Έλεγχος Συσχέτισης Επαγγέλματος – Σκορ Αποτελεσματικότητας Υπηρεσιών Υγείας

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στο Επάγγελμα και στο Σκορ Αποτελεσματικότητας
- H1: Υπάρχει συσχέτιση ανάμεσα στο Επάγγελμα και στο Σκορ Αποτελεσματικότητας

p-value=0,001<0,05 (Πίνακας 45 - Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει ασθενής θετική συσχέτιση μεταξύ του Επαγγέλματος και του Σκορ Αποτελεσματικότητας αφού r= 0,291 (Πίνακας

46-Παράρτημα Β). Ειδικότερα τόσο οι ιατροί όσο και οι νοσηλευτές θεωρούν ότι η τεχνολογία των Big Data στον τομέα της Υγείας θα αυξήσει την ποιότητα των παρεχόμενων υπηρεσιών υγείας.

8.3.2.22 Έλεγχος Συσχέτισης Επαγγέλματος – Σκορ Αποφάσεων

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στο Επάγγελμα και στο Σκορ Αποφάσεων
- Η1: Υπάρχει συσχέτιση ανάμεσα στο Επάγγελμα και στο Σκορ Αποφάσεων

p-value=0,630>0,05 (Πίνακας 47-Παράρτημα Β).Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ του Επαγγέλματος και Σκορ Αποφάσεων.

8.3.2.23 Έλεγχος Συσχέτισης Επαγγέλματος – Σκορ Παροχής Υπηρεσιών Υγείας

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στο Επάγγελμα και στο Σκορ Παροχής Υπηρεσιών Υγείας
- Η1: Υπάρχει συσχέτιση ανάμεσα στο Επάγγελμα και στο Σκορ Παροχής Υπηρεσιών Υγείας

p-value=0,642>0,05 (Πίνακας 48-Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση μεταξύ του Επαγγέλματος και Σκορ Παροχής Υπηρεσιών Υγείας

8.3.2.24 Έλεγχος Συσχέτισης Επαγγέλματος – Σκορ Πρόληψης

- Ηο: Δεν υπάρχει συσχέτιση ανάμεσα στο Επάγγελμα και στο Σκορ Πρόληψης
- Η1: Υπάρχει συσχέτιση ανάμεσα στο Επάγγελμα και στο Σκορ Πρόληψης

p-value=0,029<0,05 (Πίνακας 49-Παράρτημα Β). Απορρίπτουμε τη μηδενική υπόθεση και συνεπώς υπάρχει ασθενής θετική συσχέτιση μεταξύ του Επαγγέλματος και Σκορ Πρόληψης αφού r=0,99 (Πίνακας 50-Παράρτημα Β).

Ειδικότερα τόσο οι ιατροί όσο και οι νοσηλευτές θεωρούν ότι η τεχνολογία των Big Data στον τομέα της Υγείας θα προσφέρει αποτελεσματικότερη πρόληψη στους πληθυσμούς υψηλού κινδύνου.

8.3.2.25 Έλεγχος Συσχέτισης Επαγγέλματος –Περιπτώσεις Χρήσεις Big Data εξωτερικό

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές
- H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές

p-value=0,620>0,05 (Πίνακας 51-Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

8.3.2.26 Έλεγχος Συσχέτισης Επαγγέλματος –Περιπτώσεις Χρήσεις Big Data στην Ελλάδα

- Ho: Δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές
- H1: Υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές

p-value=0,780>0,05 (Πίνακας 51-Παράρτημα Β). Δεν απορρίπτουμε τη μηδενική υπόθεση και συνεπώς δεν υπάρχει συσχέτιση ανάμεσα στις δύο μεταβλητές.

8.4 Συζήτηση

Τα προαναφερθέντα αποτελέσματα δείχνουν ότι η πλειονότητα των ατόμων που ερωτήθηκαν ήταν γυναίκες. Το δείγμα στελεχώθηκε από άτομα όλων των ηλικιών με την πλειονότητα αυτών να είναι νέοι. Επιπλέον η πλειονότητα των ατόμων είναι ιατροί. Επίσης η πλειονότητα των ανδρών και των γυναικών απάντησε ότι δεν γνωρίζει για τα Big Data. Τα αποτελέσματα της έρευνας έδειξαν ότι περισσότεροι επαγγελματίες δεν έχουν ακούσει για τη συγκεκριμένη τεχνολογία. Όσον αφορά στη μορφή των μεγάλων δεδομένων οι απαντήσεις έδειξαν ότι τόσο οι ιατροί όσο και οι νοσηλευτές θεωρούν ότι τα Big Data περιλαμβάνουν κατά κύριο λόγο δομημένα δεδομένα.

Αξιοσημείωτο είναι το γεγονός ότι το ποσοστό των επαγγελματιών υγείας που ανέφεραν σαν πηγή ενημέρωσης το χώρο εργασίας τους είναι πολύ χαμηλό. Ως πεδίο εφαρμογής των Big Data η πλειονότητα των ιατρών και νοσηλευτών απάντησαν την υγειονομική περίθαλψη. Στη συγκεκριμένη

περίπτωση το ποσοστό των νοσηλευτών είναι μεγαλύτερο από αυτό των ιατρών. Όσον αφορά στο ποια δεδομένα θα πρέπει να συλλεχθούν για τα Big Data η πλειονότητα των ερωτηθέντων επέλεξε τα δεδομένα υγείας ακολουθούμενα από τα δεδομένα του Διαδικτύου.

Στην ερώτηση για τον αν η τεχνολογία των Data θα μπορούσε να εφαρμοστεί στον τομέα της Υγείας τα αποτελέσματα ήταν υπέρ του Ναι. Επίσης η πλειονότητα των επαγγελματιών υγείας δεν γνωρίζει περιπτώσεις χρήσης των Big Data στο εξωτερικό. Όσοι απάντησαν θετικά ανέφεραν ως περιπτώσεις χρήσης τα ηλεκτρονικά δεδομένα των νοσοκομείων, την ηλεκτρονική ασφάλιση, τον ηλεκτρονικό φάκελο ασθενή και τις ΗΠΑ. Αντίστοιχα για την Ελλάδα, η πλειονότητα των επαγγελματιών υγείας δεν γνωρίζει περιπτώσεις χρήσης των Big Data στη χώρα μας. Όσοι απάντησαν θετικά ανέφεραν ως περιπτώσεις χρήσης τον ΑΜΚΑ, τον ηλεκτρονικό φάκελο υγείας ,το πρόγραμμα ΙΑΣΥΣ και τα νοσοκομεία.

Όσον αφορά στη συλλογή δεδομένων για τα Big Data στην υγεία η πλειονότητα των επαγγελματιών υγείας ανέφερε τον ηλεκτρονικό φάκελο ασθενή και την ηλεκτρονική συνταγογράφηση ως πηγές δεδομένων .Επιπλέον τα αποτελέσματα της έρευνας έδειξαν ότι οι ιατροί και οι νοσηλευτές συμφωνούν με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας είναι χρήσιμη. Ακόμη συμφωνούν με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα μπορούσε να αυξήσει την αποτελεσματικότητα των παρεχόμενων Υπηρεσιών Υγείας. Η πλειονότητα ούτε συμφωνεί αλλά ούτε και διαφωνεί με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα βοηθήσει τους Επαγγελματίες Υγείας στη διαδικασία λήψης αποφάσεων καθώς και στο ότι θα δώσει τη δυνατότητα παροχής υπηρεσιών υγείας προσαρμοσμένων στις ανάγκες των ασθενών. Ακόμη οι ερωτηθέντες συμφωνούν μερικώς με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα δώσει τη δυνατότητα παροχής υπηρεσιών υγείας προσαρμοσμένων στις ανάγκες των ασθενών. Τέλος οι έλεγχοι συσχετίσεων έδειξαν ότι υπάρχει συσχέτιση μεταξύ:

- της ηλικίας και γνώσης των Big Data (τα αποτελέσματα έδειξαν ότι άτομα μεγαλύτερης ηλικίας γνώριζαν λιγότερα για τα Big Data σε σχέση με τα άτομα μικρότερης ηλικίας).
- της διάρκειας επαγγελματικής εμπειρίας και γνώσης των Big Data (τα αποτελέσματα έδειξαν ότι όσο μεγαλύτερη είναι η επαγγελματική εμπειρία τόσο μικρότερη είναι η γνώση για τα Big Data).
- της ηλικίας και στην μορφή των Big Data(τα αποτελέσματα έδειξαν ότι άτομα μικρότερης ηλικίας πιστεύουν ότι τα Big Data αφορούν κυρίως δομημένα δεδομένα και ημιδομημένα δεδομένα.
- του επαγγέλματος και της μορφής των Big Data που σημαίνει ότι οι ιατροί ή οι νοσηλευτές πιστεύουν ότι τα Big Data αφορούν κυρίως ημιδομημένα δεδομένα.
- του φύλου και της άποψης των επαγγελματιών υγείας ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας είναι χρήσιμη που σημαίνει ότι οι γυναίκες (ή οι άνδρες) θεωρούν

περισσότερο από τους άνδρες (ή τις γυναίκες) ότι η εφαρμογή της τεχνολογίας των Big Data είναι χρήσιμη.

- του φύλου και της άποψης των επαγγελματιών υγείας ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα μπορούσε να αυξήσει την αποτελεσματικότητα των παρεχόμενων Υπηρεσιών Υγείας που σημαίνει ότι οι γυναίκες (ή οι άνδρες) θεωρούν περισσότερο από τους άνδρες (ή τις γυναίκες) ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα μπορούσε να αυξήσει την αποτελεσματικότητα των παρεχόμενων Υπηρεσιών Υγείας
- της ηλικίας και της άποψης των επαγγελματιών υγείας ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας είναι χρήσιμη. Συγκεκριμένα τα άτομα μεγαλύτερη ηλικίας θεωρούν ότι η τεχνολογία των Big data στον τομέα της υγείας δεν είναι χρήσιμη σε αντίθεση με άτομα νεότερης ηλικίας.
- της ηλικίας και της άποψης των επαγγελματιών υγείας ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα μπορούσε να αυξήσει την αποτελεσματικότητα των παρεχόμενων Υπηρεσιών Υγείας. Συγκεκριμένα τα άτομα μεγαλύτερη ηλικίας θεωρούν ότι η τεχνολογία των Big Data στον τομέα της υγείας δεν θα αυξήσει την αποτελεσματικότητα των παρεχόμενων υπηρεσιών υγείας σε αντίθεση με τα άτομα νεότερης ηλικίας.
- της ηλικίας και της άποψης των επαγγελματιών υγείας ότι η τεχνολογία των Big Data θα δώσει τη δυνατότητα παροχής υπηρεσιών υγείας προσαρμοσμένων στις ανάγκες των ασθενών. Συγκεκριμένα τα άτομα μεγαλύτερη ηλικίας θεωρούν ότι η τεχνολογία των Big Data στον τομέα της υγείας δεν θα δώσει τη δυνατότητα παροχής υπηρεσιών υγείας προσαρμοσμένων στις ανάγκες των ασθενών.
- της ηλικίας και της άποψης των επαγγελματιών υγείας ότι η τεχνολογία των Big Data προσφέρει αποτελεσματικότερη πρόληψη στους πληθυσμούς υψηλού κινδύνου. Ειδικότερα, τα άτομα μεγαλύτερη ηλικίας θεωρούν ότι η τεχνολογία των Big Data στον τομέα της υγείας δεν συμβάλει στην

αποτελεσματικότερη πρόληψη στους πληθυσμούς υψηλού κινδύνου.

- του επαγγέλματος και της άποψης των επαγγελματιών υγείας ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας είναι χρήσιμη. Ειδικότερα οι ιατροί ή οι νοσηλευτές θεωρούν χρήσιμη την τεχνολογία των Big Data στον τομέα της Υγείας.
- του επαγγέλματος και της άποψης των επαγγελματιών υγείας ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα μπορούσε να αυξήσει την αποτελεσματικότητα των παρεχόμενων Υπηρεσιών Υγείας. Ειδικότερα οι ιατροί ή οι νοσηλευτές θεωρούν ότι η τεχνολογία των Big Data στον τομέα της Υγείας θα αυξήσει την ποιότητα των παρεχόμενων υπηρεσιών υγείας.
- του επαγγέλματος και της άποψης των επαγγελματιών υγείας ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας προσφέρει αποτελεσματικότερη πρόληψη στους πληθυσμούς υψηλού κινδύνου. Ειδικότερα οι ιατροί ή οι νοσηλευτές θεωρούν ότι η τεχνολογία των Big Data στον τομέα της Υγείας θα προσφέρει αποτελεσματικότερη πρόληψη στους πληθυσμούς υψηλού κινδύνου.

Τα ευρήματα της έρευνας δείχνουν ότι υπάρχει άγνοια από την πλευρά των επαγγελματιών υγείας σχετικά με την τεχνολογία των Big Data. Ωστόσο η πλειοψηφία αυτών θεωρεί ότι η συγκριμένη τεχνολογία είναι χρήσιμη στον τομέα της Υγείας. Είναι χαρακτηριστικό το γεγονός ότι οι περισσότεροι επαγγελματίες υγείας, σύμφωνα με τα σχόλια του ερωτηματολογίου, θεωρούν τα Big Data ως μία εφαρμογή ηλεκτρονικού φακέλου ασθενούς που θα αντικαταστήσει τον χειρόγραφο και θα βελτιώσει την πρόσβαση στην ιατρική πληροφορία.

Επομένως γίνεται αντιληπτό ότι έχουν μια λανθασμένη εντύπωση σχετικά με την τεχνολογία των Big Data και τη φύση της εργασίας της το οποίο είναι και ένας από τους περιορισμούς της έρευνας δεδομένου ότι δεν υπάρχει και το ανάλογο γνωστικό και τεχνολογικό υπόβαθρο του δείγματος. Ένας ακόμη περιορισμός είναι ότι κατά την αναζήτηση στη διεθνή

βιβλιογραφία δεν βρέθηκαν αντίστοιχες μελέτες οι οποίες θα μπορούσαν να συμπεριληφθούν στη διατριβή και να γίνει συγκριτική ανάλυση των αποτελεσμάτων.

Μελλοντικές επεκτάσεις της παρούσας έρευνας θα μπορούσαν να είναι η επανάληψη της παρούσας έρευνας στα ίδια άτομα αφού προηγουμένως είχαν παρακολουθήσει κάποιο σεμινάριο σχετικό με εφαρμογές χειρισμού μεγάλου όγκου δεδομένων καθώς και εφαρμογές Data Mining σχετικές με το χώρο της υγείας έτσι ώστε να έχουν μια ολοκληρωμένη εικόνα για την τεχνολογία των Big Data. Με αυτό τον τρόπο, οι επαγγελματίες υγείας θα κατανοήσουν την συγκεκριμένη τεχνολογία και τα οφέλη από τη χρήση της στον τομέα της Υγείας.

8.5 Συμπεράσματα

Η μελέτη που παρουσιάζεται στο παρόν Κεφάλαιο στοχεύει στη διερεύνηση της αντίληψης που έχουν οι Επαγγελματίες Υγείας για την τεχνολογία των Μεγάλων Δεδομένων (Big Data). Τα αποτελέσματα της παραπάνω έρευνας δείχνουν ότι ένα πολύ μικρό ποσοστό του δείγματος ανέφερε ότι γνωρίζει την τεχνολογία των Big Data. Είναι ενδεικτικό ότι οι περισσότεροι επαγγελματίες αναφέρουν ότι δεν έχουν ακούσει για τη συγκεκριμένη τεχνολογία ούτε στο χώρο εργασίας γεγονός που σημαίνει ότι ίσως και να μην έχουν γίνει ενέργειες για την υιοθέτηση της συγκεκριμένης τεχνολογίας από τους αρμόδιους φορείς της υγείας. Επιπλέον, είναι αξιοσημείωτο ότι αρκετοί ερωτηθέντες στα σχόλια του ερωτηματολογίου αναφέρουν ότι μία τέτοια τεχνολογία είναι δύσκολο να εφαρμοστεί στα Νοσοκομεία της χώρας. Σημαντικό είναι επίσης το γεγονός ότι υπάρχει συσχέτιση μεταξύ της ηλικίας και της γνώσης των Data. Τα αποτελέσματα έδειξαν ότι άτομα μικρότερης ηλικίας διαθέτουν περισσότερες γνώσεις για τα Big Data. Επίσης, κάποιοι ερωτηθέντες σχολίασαν ότι δεν γνωρίζουν για την τεχνολογία των Big Data ωστόσο τους ακούγεται ως κάτι θετικό. Παρότι οι επαγγελματίες υγείας δείχνουν θετική στάση ως προς την τεχνολογία των Big Data, κάτι τέτοιο δεν μπορεί να οδηγήσει σε ασφαλή συμπεράσματα καθώς δεν γνωρίζουν πολλά για τη συγκεκριμένη τεχνολογία ενώ θεωρούν ότι δεν μπορεί να εφαρμοστεί στο Ελληνικό Σύστημα Υγείας.

Τι θα πρέπει να γίνει για τη βελτίωση της υπάρχουσας κατάστασης; Πρώτα από όλα θα πρέπει να υπάρξει βούληση από την κυβέρνηση καθώς και από τους φορείς του ΕΣΥ για τη βελτίωση της τεχνολογικής υποδομής των νοσοκομείων. Θα πρέπει να γίνει άμεση εφαρμογή του συστήματος Πληροφοριακών Συστημάτων Νοσοκομείων και Ηλεκτρονικού Φακέλου Ασθενών τα οποία θα παρέχουν άμεση και αποκεντρωμένη πρόσβαση στους επαγγελματίες υγείας. Επίσης, το Εθνικό Σύστημα Υγείας θα πρέπει να αξιοποιήσει τις λεγόμενες Αποθήκες Δεδομένων(Data Warehouses) οι οποίες θα συμβάλουν στην ενοποίηση (intergration) ετερογενών συστημάτων και στην χρήση συστημάτων επιχειρηματικής ευφυΐας. Τέλος, χρειάζεται εθνικό σχέδιο ,όραμα στρατηγική και επενδύσεις σχετικά με την ψηφιοποίηση του χώρου της υγείας καθώς και την εκμετάλλευση των δεδομένων των ασθενών για την εξόρυξη νέας γνώσης.

Όσον αφορά τους επαγγελματίες υγείας, στο νέο οικοσύστημα της υγείας όπως διαμορφώνεται από τις εξελίξεις στην τεχνολογία και τα μεγάλα δεδομένα, χρειάζεται μια νέα παγκόσμια γλώσσα, νέες δεξιότητες και κυρίως, νέος τρόπο σκέψης. Πρώτον, επαγγελματίες υγείας πρέπει να κατανοήσουν τις ξεχωριστές διαφορές μεταξύ Big Data και άλλων ηλεκτρονικών συνόλων δεδομένων. Θα πρέπει να γίνουν ενημερωτικά σεμινάρια, οργανωμένα από τους φορείς του ΕΣΥ, τα οποία θα επιμορφώνουν τους ιατρούς και τους νοσηλευτές και θα γίνεται χρήση της τεχνολογίας των Big Data σε πραγματικά χρόνο μέσα από μελέτες περίπτωσης. Δεύτερον, αυτές οι διαφορές σημαίνουν ότι αυτός ο αναδυόμενος τομέας απαιτεί έναν νέο τρόπο σκέψης και εργασίας, τον οποίο πρέπει να συνηθίσουν οι επαγγελματίες του τομέα της υγείας, αν πρόκειται να χρησιμοποιήσουν τα μεγάλα δεδομένα για τη βελτίωση της περίθαλψης των ασθενών Τέλος, οι επαγγελματίες υγείας πρέπει να γνωρίζουν ποια δυνητικά οφέλη προσφέρει η τεχνολογία Big Data και πώς μπορούν να συμβάλουν στο κίνημα της επιστήμης των δεδομένων.

Κεφάλαιο 9

9.1 Μέτρηση της απόδοσης των μοντέλων κατηγοριοποίησης για την πρόβλεψη καρδιαγγειακών νοσημάτων.

Σε αυτή την ενότητα της διατριβής θα γίνει χρήση τεχνικών κατηγοριοποίησης με σκοπό την πρόβλεψη καρδιαγγειακών νοσημάτων με χρήση ενός συνόλου δεδομένων. Η κατηγοριοποίηση είναι μια από τις βασικότερες εργασίες Εξόρυξης Δεδομένων. Είναι εργασία επιβλεπόμενης μάθησης, που στόχο έχει την ανακάλυψη της σχέσης ανάμεσα σε ένα γνώρισμα στόχο με ονομαστικές τιμές και σε ένα σύνολο άλλων γνωρισμάτων. Στην κατηγοριοποίηση εφαρμόζεται ένας επαγωγικός αλγόριθμος και κατασκευάζεται ένα μοντέλο. Η διαδικασία της κατηγοριοποίησης περιλαμβάνει τρία στάδια [108].

- ✓ Στο πρώτο στάδιο ο αλγόριθμος επεξεργάζεται τα δεδομένα του συνόλου εκπαίδευσης και κατασκευάζει ένα μοντέλο.
- ✓ Στο δεύτερο στάδιο ελέγχεται η ικανότητα του μοντέλου να προβλέπει την κλάση άγνωστων παρατηρήσεων.
- ✓ Εάν η επίδοση του μοντέλου κριθεί ικανοποιητική, τότε ακολουθεί το τρίτο στάδιο, το οποίο συνίσταται στη χρήση του μοντέλου για τη διατύπωση προβλέψεων.

Κατά την εκπαίδευση πρέπει να αποφευχθεί η υπερπροσαρμογή του μοντέλου, η απομνημόνευση δηλαδή του συγκεκριμένου συνόλου εκπαίδευσης. Αποτέλεσμα της υπερπροσαρμογής είναι η πτώση της απόδοσης έναντι άγνωστων παρατηρήσεων.

Οι αλγόριθμοι κατηγοριοποίησης που θα χρησιμοποιηθούν είναι: Λογιστική Παλινδρόμηση, Naive Bayes Classifier, Δένδρα αποφάσεων, Αλγόριθμος Κ κοντινότερων γειτόνων, Αλγόριθμος SVM (Support Vector Machine) και Random Forest.

9.2 Εφαρμογή για τη μέτρηση της απόδοσης των μοντέλων κατηγοριοποίησης στην πρόβλεψη καρδιαγγειακών νοσημάτων

Η επιλογή των δεδομένων έγινε από το ηλεκτρονικό αποθετήριο του Πανεπιστημίου της Καλιφόρνια (University of California Irvine) που βρίσκεται

στο σύνδεσμο: https://archive.ics.uci.edu/ml/datasets.php. Η ιστοσελίδα περιέχει datasets τα οποία μπορεί κάποιος να τα επεξεργαστεί χρησιμοποιώντας τεχνικές εξόρυξης γνώσης. Το dataset που χρησιμοποιήθηκε περιέχει πρωτογενή δεδομένα από το εργαστήριο Cleveland Heartlab, το οποίο παρέχει υπηρεσίες υγείας που αφορούν στην πρόληψη και διάγνωση καρδιαγγειακών νοσημάτων και είναι το Εθνικό Κέντρο Αριστείας σε καρδιολογικά νοσήματα [109].

Για την εφαρμογή των αλγορίθμων κατηγοριοποίησης έγινε η σχετική προεπεξεργασία των δεδομένων (έλεγχος ελλιπών τιμών, κωδικοποίηση κατηγορικών μεταβλητών). Η ανάλυση των δεδομένων έγινε μέσω του Jupyter Notebook που περιλαμβάνεται στο λογισμικό Anaconda 3. Η συγγραφή του σχετικού κώδικα έγινε με χρήση της γλώσσας προγραμματισμού Python έκδοση 3.8. Για την εφαρμογή των αλγορίθμων κατηγοριοποίησης καθώς και για τη δημιουργία των κατάλληλων γραφημάτων έγινε εγκατάσταση και χρήση των κατάλληλων πακέτων που ενσωματώνει και υποστηρίζει η Python τα οποία είναι τα παρακάτω [110]:

- Pandas: είναι μια βιβλιοθήκη που προσφέρει υψηλής απόδοσης εργαλεία για ανάλυση δεδομένων στην Python και δομές δεδομένων που είναι εύκολες στη χρήση.
- Matplotlib: είναι βιβλιοθήκη σχεδίασης δισδιάστατων γραφημάτων καθώς δημιουργεί γραφήματα όπως ιστογράμματα, ραβδογράμματα, κυκλικά διαγράμματα κ.ά. Επιπρόσθετα πακέτα εργαλείων είναι το mplot3d για την δημιουργία 3D διαγραμμάτων και η διεπαφή seaborn 2 για απεικόνιση στατιστικών δεδομένων.
- Imbalanced-learn: είναι ένα πακέτο της Python, το οποίο προσφέρει τεχνικές αναδειγματοληψίας που χρησιμοποιούνται συνήθως σε σύνολα δεδομένων για να αντιμετωπιστεί η ανισορροπία μεταξύ δύο ή περισσότερων κλάσεων. Περιέχει αλγορίθμους οι οποίοι προσαρμόζουν την κατανομή των κλάσεων σε ένα σύνολο δεδομένων, όπως για παράδειγμα μέθοδοι undersampling, oversampling, ή και συνδυασμούς των δύο αυτών μεθόδων.
- NumPy: είναι η κύρια βιβλιοθήκη της Python για επιστημονικούς υπολογισμούς. Υποστηρίζει πολυδιάστατους πίνακες, μεθόδους

γραμμικής άλγεβρας και άλλες υψηλού επιπέδου μαθηματικές λειτουργίες.

- SciPy: είναι μια βιβλιοθήκη ανοιχτού κώδικα που χρησιμοποιείται από επιστήμονες, αναλυτές και μηχανικούς για επιστημονικούς και τεχνικούς υπολογισμούς. Έχει δημιουργηθεί για να υποστηρίζει πίνακες NumPy και υπολογίζει αποδοτικά αριθμητικές λειτουργίες.
- Scikit-learn: περιέχει ένα σύνολο από επιβλεπόμενους και μη-επιβλεπόμενους αλγορίθμους μηχανικής μάθησης και έχει σχεδιαστεί να λειτουργεί με τις βιβλιοθήκες NumPy και SciPy.

- **Επιλογή των δεδομένων**

Τα δεδομένα είναι αποθηκευμένα σε ένα CSV (Comma Separated Values) αρχείο. Το αρχείο αποτελείται από 16 διαφορετικές μεταβλητές (ή χαρακτηριστικά) από τις οποίες η καθεμία περιέχει 4270 εγγραφές ή δεδομένα όπως φαίνεται και στην Εικόνα 10.

male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	Disease
1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0

Εικόνα 10: Εγγραφές του dataset

Ακολουθούν οι μεταβλητές του συνόλου δεδομένων καθώς και ο τύπος της καθεμίας (βλ. Εικόνα 11).

```
Out[4]: male                int64
        age                 int64
        education         float64
        currentSmoker       int64
        cigsPerDay        float64
        BPMeds            float64
        prevalentStroke     int64
        prevalentHyp        int64
        diabetes            int64
        totChol           float64
        sysBP             float64
        diaBP             float64
        BMI               float64
        heartRate         float64
        glucose           float64
        Disease             int64
        dtype: object
```

Εικόνα 11:Μεταβλητές dataset

Οι 15 πρώτες μεταβλητές είναι ουσιαστικά οι ανεξάρτητες μεταβλητές ενώ η μεταβλητή με την ονομασία Disease είναι η εξαρτημένη μεταβλητή. Συγκεκριμένα το αρχείο δεδομένων περιλαμβάνει τις παρακάτω μεταβλητές

Μεταβλητή	Επεξήγηση	Είδος
male	φύλο	Nominal (1=male,0=female)
age	ηλικία	Ποσοτική-Συνεχής
education	εκπαίδευση	Κατηγορική (1 = Some High School; 2 = High School or GED; 3 = Some College or Vocational School; 4 = college
currentSmoker	καπνιστής	Κατηγορική (0 = nonsmoker;

		1 = smoker)
cigsPerDay	τσιγάρα ανά ημέρα	Ποσοτική-Συνεχής
BPMeds	Θεραπεία για αρτηριακή πίεση	Κατηγορική (0 = Not on Blood Pressure medications; 1 = Is on Blood Pressure medications)
prevalentStroke	Περίπτωση εγκεφαλικού επεισοδίου	Κατηγορική (0 = No; 1 = Yes)
prevalentHyp	Αρτηριακή υπέρταση	Κατηγορική (0 = No; 1 = Yes)
diabetes	Διαβήτης	Κατηγορική (0 = No; 1 = Yes)
totChol	Επίπεδα Χοληστερόλης	Ποσοτική-Συνεχής
sysBP	Συστολική Πίεση	Ποσοτική-Συνεχής
diaBP	Διαστολική Πίεση	Ποσοτική-Συνεχής
BMI	Δείκτης Μάζας Σώματος	Ποσοτική-Συνεχής
heartRate	Καρδιακοί παλμοί	Ποσοτική-Συνεχής
glucose	Επίπεδα Γλυκόζης	Ποσοτική-Συνεχής
Disease	Ύπαρξη ασθένειας	Κατηγορική (0 = No; 1 = Yes)

- **Επεξεργασία των δεδομένων**

Για να μπορέσουμε να εφαρμόσουμε τους κατάλληλους αλγόριθμους θα πρέπει να κάνουμε αρχικά ένα έλεγχο για την ύπαρξη ελλιπών τιμών .

```
Out[5]: male               0
        age                0
        education        105
        currentSmoker      0
        cigsPerDay        29
        BPMeds            53
        prevalentStroke    0
        prevalentHyp       0
        diabetes           0
        totChol           50
        sysBP              0
        diaBP              0
        BMI               19
        heartRate          1
        glucose          388
        Disease            0
        dtype: int64
```

Εικόνα 12:Δεδομένα με ελλιπείς τιμές

Παρατηρούμε στην Εικόνα 12 ότι υπάρχουν ελλιπή δεδομένα. Επομένως, διαγράφουμε εκείνες τις εγγραφές των οποίων οι μεταβλητές περιέχουν ελλιπή δεδομένα (βλ. Εικόνες 13 & 14).

```
Out[16]: male               0
         age                0
         education          0
         currentSmoker      0
         cigsPerDay         0
         BPMeds             0
         prevalentStroke    0
         prevalentHyp       0
         diabetes           0
         totChol            0
         sysBP              0
         diaBP              0
         BMI                0
         heartRate          0
         glucose            0
         Disease            0
         dtype: int64
```

Εικόνα 13:Απαλοιφή ελλιπών τιμών

```
(3658, 16)
['male', 'age', 'education', 'currentSmoker', 'cigsPerDay', 'BPMeds', 'prevalentStroke', 'prevalentHyp', 'diabetes', 'totCh
ol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose', 'Disease']
```

Εικόνα 14:Το νέο dataset

Στη συνέχεια θα εφαρμόσουμε τους διάφορους αλγόριθμους κατηγοριοποίησης. Για να το πετύχουμε αυτό θα πρέπει να διαχωρίσουμε τις ανεξάρτητες μεταβλητές από την εξαρτημένη μεταβλητή. Επίσης για να εφαρμόσουμε τις τεχνικές κατηγοριοποίησης θα πρέπει να χωρίσουμε τα δεδομένα μας σε training set και test set. Με τη χρήση των κατάλληλων εντολών το 75% των δεδομένων θα χρησιμοποιηθεί για την εκπαίδευση του μοντέλου μας και το υπόλοιπο 25% για τη δοκιμή του . Σύμφωνα με τη βιβλιογραφία η συνηθισμένη αναλογία για το διαχωρισμό των δεδομένων είναι είτε 75%-25% είτε 80%- 20%. Και στις δύο περιπτώσεις ο διαχωρισμός γίνεται με αυτές τις παραπάνω αναλογίες έτσι ώστε το test set να είναι επαρκές για την εξαγωγή στατιστικά σημαντικών συμπερασμάτων.

9.2.1 Λογιστική Παλινδρόμηση

Εφαρμόζοντας το μοντέλο της λογιστικής παλινδρόμησης με χρήση του πακέτου Scikit-learn παίρνουμε τα παρακάτω αποτελέσματα (βλ. Εικόνα 15).

Logit Regression Results

Dep. Variable:	Disease	No. Observations:	3658
Model:	Logit	Df Residuals:	3643
Method:	MLE	Df Model:	14
Date:	Tue, 14 May 2019	Pseudo R-squ.:	0.07063
Time:	10:30:53	Log-Likelihood:	-1450.4
converged:	True	LL-Null:	-1560.6
		LLR p-value:	3.572e-39

	coef	std err	z	P>|z|	[0.025	0.975]
male	0.4015	0.105	3.834	0.000	0.196	0.607
age	0.0289	0.006	4.949	0.000	0.017	0.040
education	-0.1973	0.048	-4.100	0.000	-0.292	-0.103
currentSmoker	-0.2111	0.153	-1.380	0.168	-0.511	0.089
cigsPerDay	0.0206	0.006	3.326	0.001	0.008	0.033
BPMeds	0.3657	0.230	1.587	0.113	-0.086	0.817
prevalentStroke	0.7057	0.485	1.456	0.146	-0.245	1.656
prevalentHyp	0.9162	0.124	7.365	0.000	0.672	1.160
diabetes	0.7305	0.298	2.452	0.014	0.147	1.314
totChol	-0.0010	0.001	-0.925	0.355	-0.003	0.001
sysBP	0.0127	0.004	3.345	0.001	0.005	0.020
diaBP	-0.0259	0.006	-4.232	0.000	-0.038	-0.014
BMI	-0.0483	0.012	-3.928	0.000	-0.072	-0.024
heartRate	-0.0224	0.004	-5.647	0.000	-0.030	-0.015
glucose	0.0020	0.002	0.967	0.334	-0.002	0.006

Εικόνα 15:Αποτελέσματα Λογιστικής Παλινδρόμησης

Η Εικόνα 15 μας δίνει κάποιες χρήσιμες πληροφορίες όπως το coef είναι οι συντελεστές παλινδρόμησης καθώς και το κριτήριο Z το οποίο χρησιμοποιείται για να ελεγχθεί κατά πόσο κάθε μία από τις μεταβλητές του μοντέλου της λογιστικής παλινδρόμησης είναι στατιστικά σημαντική [111]. Η κάθε τιμή του κριτηρίου Z συγκρίνεται με το P ή αλλιώς p-value. Αν το P<0,05 για μία μεταβλητή τότε η συγκεκριμένη μεταβλητή προσφέρει στατιστικά σημαντική πληροφορία στην πρόβλεψη της τιμής της εξαρτημένης μεταβλητής

Από τον πίνακα φαίνεται ότι οι μεταβλητές currentSmoker, BPMeds, prevalentStroke, totChol, glucose έχουν p-value>0,05 και επομένως δεν συμβάλουν στατιστικά σημαντικά στην πρόβλεψη καρδιαγγειακών νοσημάτων. Επομένως δεν είναι στατιστικά σημαντικές και αφαιρούνται από το μοντέλο της λογιστικής παλινδρόμησης όπως δείχνει ο παρακάτω πίνακας.

Logit Regression Results

Dep. Variable:	Disease	No. Observations:	3658
Model:	Logit	Df Residuals:	3648
Method:	MLE	Df Model:	9
Date:	Tue, 14 May 2019	Pseudo R-squ.:	0.06785
Time:	10:38:30	Log-Likelihood:	-1454.7
converged:	True	LL-Null:	-1560.6
		LLR p-value:	1.114e-40

	coef	std err	z	P>|z|	[0.025	0.975]
male	0.4034	0.104	3.889	0.000	0.200	0.607
age	0.0280	0.006	5.029	0.000	0.017	0.039
education	-0.2031	0.048	-4.270	0.000	-0.296	-0.110
cigsPerDay	0.0138	0.004	3.315	0.001	0.006	0.022
prevalentHyp	0.9581	0.121	7.931	0.000	0.721	1.195
diabetes	0.9244	0.231	4.007	0.000	0.472	1.377
sysBP	0.0133	0.004	3.550	0.000	0.006	0.021
diaBP	-0.0267	0.006	-4.383	0.000	-0.039	-0.015
BMI	-0.0477	0.012	-3.939	0.000	-0.071	-0.024
heartRate	-0.0237	0.004	-6.193	0.000	-0.031	-0.016

Εικόνα 16:Αποτελέσματα Λογιστικής Παλινδρόμησης

Με βάση τις τιμές των coef (coefficients) και του κριτηρίου Z (βλ. Εικόνα 16) παρατηρούμε τον συντελεστή παλινδρόμησης του συγκεκριμένου μοντέλου. Βλέπουμε για παράδειγμα ότι μία αύξησης στην τιμή της αρτηριακής υπέρτασης αυξάνει τον κίνδυνο εμφάνισης καρδιαγγειακών

νοσημάτων. Ομοίως και στην περίπτωση του διαβήτη. Αντίθετα βλέπουμε ότι μία αύξηση της τιμής του δείκτη μάζας σώματος μειώνει τον κίνδυνο εμφάνισης καρδιαγγειακών νοσημάτων.

Στη συνέχεια κατασκευάζουμε το confusion matrix του μοντέλου.

Actual Values	Predicted		
		Negative(0)	Positive(1)
	Negative(0)	TN=763	FP=5
	Positive(1)	FN=138	TP=9

Παρατηρούμε ότι από 915 δεδομένα , τα 772 είναι σωστές προβλέψεις και οι 143 είναι λάθος. Μία καλή προσέγγιση για την αξιολόγηση των αποτελεσμάτων είναι να μετρήσουμε τα true/false positive (tp / fp) και τα true/false negative (tn / fn). Τα αποτελέσματα του πίνακα δείχνουν ότι από τα 901 δεδομένα , στην πραγματικότητα τα 763 παρατηρήθηκαν ως αληθώς αρνητικά. Από τα 14 που παρατηρήθηκαν ως θετικά τα 9 είναι αληθώς θετικά. Με βάση τα αποτελέσματα του Confusion Matrix θα υπολογίσουμε και τις υπόλοιπες μετρικές του μοντέλου μας. .

Στη συνέχεια θα υπολογίσουμε την ακρίβεια του μοντέλου μας. Η μέτρηση της απόδοσης για τα μοντέλα κατηγοριοποίησης γίνεται με τη μέθοδο cross validation. Αποτελεί μια πολύ χρήσιμη τεχνική για την αξιολόγηση της απόδοσης των μοντέλων μηχανικής μάθησης καθώς υπολογίζει πόσο ακριβείς είναι οι προβλέψεις που θα δώσει ο κατηγοριοποιητής στην πράξη. Ο λόγος που χρησιμοποιείται η συγκεκριμένη τεχνική είναι ότι ικανοποιείται το αίτημα της ανεξαρτησίας μεταξύ των παραδειγμάτων εκπαίδευσης και επικύρωσης και αμβλύνεται η διάσταση μεταξύ των τιμών των μέτρων αποτελεσματικότητας για διαφορετικά σύνολα επικύρωσης, καθώς η έξοδος της μεθόδους είναι ο μέσος όρος τους.

Προκειμένου τα αποτελέσματα που προκύπτουν να είναι αξιόπιστα, έγινε σε όλα τα μοντέλα διασταυρωτική επαλήθευση με την μέθοδο cross-validation. Με αυτόν τον τρόπο, το αρχικό δείγμα είναι χωρίζεται τυχαία σε k επιμέρους δείγματα [112]. Από τα επιμέρους δείγματα k, ένα ενιαίο υπο-δείγμα διατηρείται ως σύνολο δεδομένων επικύρωσης για τη δοκιμή του

μοντέλου, και τα υπόλοιπα k-1 επιμέρους δείγματα χρησιμοποιούνται ως δεδομένα εκπαίδευσης [112]. Η διαδικασία cross-validation συνέχεια επαναλαμβάνεται k φορές, με το καθένα από τα επιμέρους δείγματα k να χρησιμοποιείται ακριβώς μία φορά ως δεδομένα επικύρωσης. Από τα αποτελέσματα υπολογίζεται ο μέσος όρος που χρησιμοποιείται για να παραχθεί μια ενιαία εκτίμηση.

```
Mean Accuracy 0.8515575351236336
```

Εικόνα 17:Μέτρηση της ακρίβειας του μοντέλου

Παρατηρούμε (βλ. Εικόνα 17) ότι η εκτέλεση του πειράματός μας με τη λογιστική παλινδρόμηση και με χρήση της διαδικασίας cross-validation επιστρέφει αποτελέσματα με μέση ακρίβεια 85%.

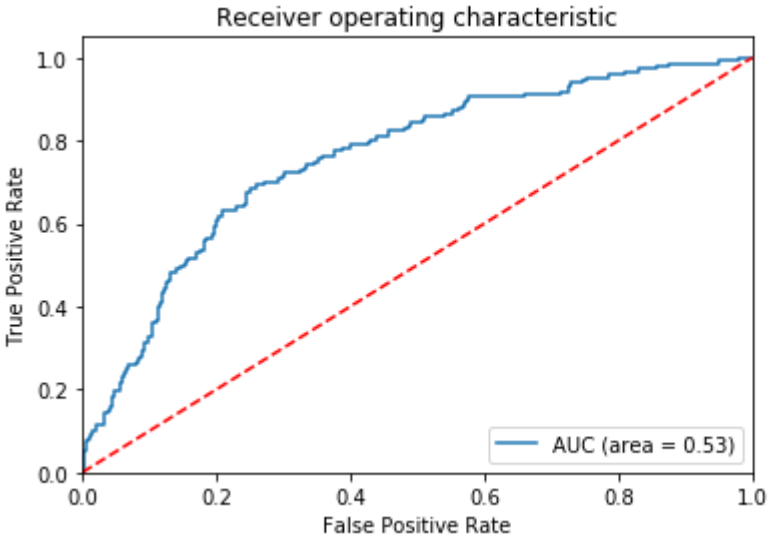
Στη συνέχεια ακολουθεί ο υπολογισμός των μετρικών precision , recall και f1 measure (βλ. Εικόνα 18).

```
precision= 0.6428571428571429 recall= 0.061224489795918366
f1= 0.11180124223602485
```

Εικόνα 18: Υπολογισμός των μετρικών precision, recall και f1 measure

Παρατηρούμε ότι η εκτέλεση του πειράματός μας με τη λογιστική παλινδρόμηση δίνει Precision=64,3%,Recall=61,2% και F1-score=11,2%.

Ένας ακόμη τρόπος για την αξιολόγηση της αποτελεσματικότητας ενός μοντέλου ταξινόμησης είναι το διάγραμμα ROC (Receiver Operating Characteristics) όπου κριτήριο για την αποτελεσματικότητα ενός μοντέλου λοιπόν αποτελεί η μορφή της καμπύλης ROC και συγκεκριμένα το εμβαδόν της περιοχής κάτω από την καμπύλη (AUC, Area Under Curve).



Γράφημα 52:Αποτελέσματα AUC

Το παραπάνω γράφημα δείχνει ότι η περιοχή κάτω από την ROC καμπύλη είναι 0.53.Επομένως διαφέρει οριακά στατιστικά σημαντικά από την τιμή0.5 που υποδηλώνει μηδενική ικανότητα του μοντέλου να προβλέπει καρδιαγγειακά νοσήματα. Η τιμής της AUC δείχνει ότι το μοντέλο δεν έχει καλή προγνωστική ικανότητα.

9.2.2 Παρατηρήσεις

Εκτελώντας τη λογιστική παλινδρόμηση στο παραπάνω μοντέλο βλέπουμε ότι ενώ έχει αρκετά καλή μέση ακρίβεια ωστόσο η τιμή της AUC, που είναι καλύτερο μέτρο αξιολόγησης των αλγορίθμων κατηγοριοποίησης σε σχέση με την ακρίβεια δείχνει την κακή προγνωστική ικανότητα του μοντέλου. Επομένως θα πρέπει να βρούμε τρόπους να βελτιώσουμε το μοντέλο της λογιστικής παλινδρόμησης που εφαρμόσαμε προηγουμένως.

Μέσα από αναζήτηση στη διεθνή βιβλιογραφία ένα σύνηθες πρόβλημα που αντιμετωπίζουν οι αλγόριθμοι κατηγοριοποίησης είναι αυτό της ανισορροπίας των κλάσεων(class imbalance problem).

9.2.3 Ανισορροπία των κλάσεων

Το πρόβλημα της ανισορροπίας των κλάσεων εμφανίζεται όταν μια κατηγορία του συνόλου αντιπροσωπεύεται από μεγάλο αριθμό εγγραφών σε σχέση με το άλλο ή τα άλλα, τα οποία αντιπροσωπεύονται από λιγότερα[30].

Αυτό αποτελεί πρόβλημα, διότι οι περισσότεροι αλγόριθμοι μηχανικής μάθησης λειτουργούν καλύτερα όταν ο αριθμός των εγγραφών σε κάθε κλάση είναι περίπου ο ίδιος. Όταν ο αριθμός των εγγραφών σε μία τάξη υπερβαίνει κατά πολύ την άλλη ή τις άλλες, προκύπτουν προβλήματα.

Τα προβλήματα αυτά έχουν να κάνουν με την ορθότητα (accuracy) του μοντέλου η οποία έχει συνήθως υψηλή τιμή με αποτέλεσμα ο κατηγοριοποιητής να προβλέπει σχεδόν πάντα την κλάση με τις περισσότερες εγγραφές μη λαμβάνοντας υπόψη την άλλη [113]. Αυτό έχει σαν συνέπεια ο κατηγοριοποιητής να μην θεωρείται κατάλληλος για τα δεδομένα, παρά την υψηλή τιμής της ορθότητας που παρουσιάζει. Για αυτόν τον λόγο είναι σημαντικό να αξιολογούμε την επίδοση του μοντέλου χρησιμοποιώντας και άλλες μετρικές, όπως ο πίνακας σύγχυσης (confusion matrix), η ακρίβεια (precision), η ανάκληση (recall), το f-score και οι καμπύλες ROC

Η προτεινόμενη λύση για το πρόβλημα της ανισορροπίας των κλάσεων είναι η αναδειγματοληψία (resampling) η οποία διακρίνεται σε υπερδειγματοληψία (oversampling) και υποδειγματοληψία (undersampling) [114].

9.2.4 Υπερδειγματοληψία (oversampling)

Η υπερδειγματοληψία αυξάνει τον αριθμό των εγγραφών στην κλάση μειοψηφίας (minority class) με τυχαία αναπαραγωγή τους προκειμένου να παρουσιάσουν μια υψηλότερη αναπαράσταση της τάξης των μειονοτήτων στο δείγμα. Τα πλεονεκτήματα της μεθόδου είναι τα εξής [115]:

- Σε αντίθεση με την υποδειγματοληψία η μέθοδος αυτή δεν οδηγεί σε απώλεια δεδομένων
- Παρουσιάζει καλύτερη απόδοση από την υποδειγματοληψία

Ένα μειονέκτημα της μεθόδου είναι ότι αυξάνει την υπερπροσαρμογή καθώς αναπαράγει τις εγγραφές της κλάσης μειοψηφίας.

9.2.5 Υποδειγματοληψία (undersampling)

Η υποδειγματοληψία εξισορροπεί το σύνολο δεδομένων μειώνοντας το μέγεθος της κλάσης πλειοψηφίας. Αυτή η μέθοδος χρησιμοποιείται όταν υπάρχει επαρκής ποσότητα δεδομένων [115]. Διατηρώντας όλα τα δείγματα

στη κλάση μειοψηφίας και επιλέγοντας τυχαία ίσο αριθμό δειγμάτων στην κλάση πλειοψηφίας, μπορεί να ανακτηθεί ένα νέο ισορροπημένο σύνολο δεδομένων για περαιτέρω μοντελοποίηση. Τα πλεονεκτήματα της μεθόδου είναι:

- Μπορεί να βοηθήσει στην βελτίωση του χρόνου εκτέλεσης και των προβλημάτων αποθήκευσης μειώνοντας τον αριθμό των δειγμάτων εκπαίδευσης όταν το σύνολο δεδομένων εκπαίδευσης είναι τεράστιο.

Τα μειονεκτήματα της μεθόδου είναι τα εξής [115]:

- Μπορεί να απορρίψει δυνητικά χρήσιμες πληροφορίες που θα μπορούσαν να είναι σημαντικές για τα μοντέλα κατηγοριοποίησης
- Σε περίπτωση που το δείγμα επιλέγεται με τυχαία δειγματοληψία, η μέθοδος αυτή οδηγεί σε ανακριβή αποτελέσματα

9.2.6 SMOTE (Synthetic Minority Over-Sampling Technique)

Είναι μία από τις ποιο εξελιγμένες τεχνικές αντιμετώπισης προβλημάτων ανισορροπίας [116]. Το βασικό χαρακτηριστικό της SMOTE είναι ότι στην κλάση μειοψηφίας γίνεται oversampling δημιουργώντας συνθετικά δείγματα και όχι αναπαράγοντας ήδη υπάρχοντα. Στην κλάση μειοψηφίας γίνεται oversampling παίρνοντας κάθε δείγμα από την κλάση και δημιουργώντας συνθετικά δείγματα, τα οποία χρησιμοποιούν τους k Nearest Neighbors των δειγμάτων της κλάσης. Τα πλεονεκτήματα της μεθόδου είναι τα εξής:

- Μειώνει το πρόβλημα της υπερπροσαρμογής που προκαλείται από τυχαία υπερδειγματοληψία, καθώς παράγονται συνθετικά δείγματα και όχι τυχαία αναπαραγωγή στιγμιότυπων της κλάσης.
- Δεν υπάρχει απώλεια χρήσιμων πληροφοριών.

Τα μειονεκτήματα της μεθόδου είναι τα εξής:

- Ενώ παράγει συνθετικά δείγματα, η SMOTE δεν λαμβάνει υπόψη γειτονικά δείγματα από άλλες κατηγορίες. Αυτό μπορεί να οδηγήσει σε επικάλυψη των κλάσεων και μπορεί να εισαγάγει επιπλέον θόρυβο.

Οι μέθοδοι αυτοί αναδειγματοληψίας, χρησιμοποιούν ευριστικές τεχνικές, οι οποίες προσπαθούν να προσεγγίσουν τη βέλτιστη κατανομή των

δειγμάτων, ώστε να τα επεξεργαστούμε και να βγάλουμε ασφαλή συμπεράσματα για τα δεδομένα.

Εκτελώντας τον κατάλληλο κώδικα, το πρόγραμμα επιστρέφει τα παρακάτω αποτελέσματα όπως φαίνονται στην Εικόνα 19:

```
Total percent without cardiovascular disease 84.77310005467469
Total percent with cardiovascular disease 15.226899945325314
```

Εικόνα 19: Ποσοστό ατόμων με η χωρίς καρδιαγγειακά νοσήματα

Είναι εμφανές το πρόβλημα της ανισορροπίας των δεδομένων καθώς παρατηρούμε ότι το ποσοστό των ατόμων χωρίς καρδιαγγειακά νοσήματα είναι πολύ μεγαλύτερο από των αντίστοιχων με καρδιαγγειακά νοσήματα. Συνεπώς πρέπει να επιλύσουμε το παραπάνω πρόβλημα.

Για να αντιμετωπίσουμε το πρόβλημα της ανισορροπίας των κλάσεων στη δικιά μας περίπτωση αποφασίσαμε να χρησιμοποιήσουμε την τεχνική SMOTE καθώς αποτελεί μία εξελιγμένη τεχνική υπερδειγματοληψίας. Αυτό σημαίνει ότι έχει όλα τα πλεονεκτήματα της υπερδειγματοληψίας που αναπτύχθηκαν προηγουμένως ενώ αντιμετωπίζει αποτελεσματικά και τα μειονεκτήματά της. Επίσης πλεονεκτεί σε σχέση με την υποδειγματοληψία στο ότι δεν παρουσιάζει απώλεια δεδομένων και δίνει πιο ακριβή αποτελέσματα αφού δεν χρησιμοποιεί τυχαία δειγματοληψία.

Εφαρμόζοντας την SMOTE παίρνουμε τα παρακάτω αποτελέσματα.

```
length of oversampled data is  4666
Number of no cardiovascular disease in oversampled data 2333
Number of no cardiovascular disease  2333
Proportion of no cardiovascular disease  data in oversampled data is  0.5
Proportion of cardiovascular disease data in oversampled data is  0.5
```

Εικόνα 20: Αποτελέσματα της υπερδειγματοληψίας SMOTE

Όπως φαίνεται στην Εικόνα 20 επιλύθηκε το πρόβλημα της ανισορροπίας των δεδομένων. Στη συνέχεια εφαρμόζουμε τον κώδικα για τη λογιστική παλινδρόμηση τα αποτελέσματα της οποίας φαίνονται στον παρακάτω πίνακα.

```
Optimization terminated successfully.
         Current function value: 0.396494
         Iterations 6
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.071
Dependent Variable: Disease          AIC:              2930.7498
Date:               2019-05-13 21:31 BIC:              3023.8199
No. Observations:   3658             Log-Likelihood:   -1450.4
Df Model:           14               LL-Null:          -1560.6
Df Residuals:       3643             LLR p-value:      3.5723e-39
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
-----------------------------------------------------------------
                   Coef.  Std.Err.    z    P>|z|   [0.025  0.975]
-----------------------------------------------------------------
male                0.4015   0.1047  3.8337 0.0001  0.1962  0.6067
age                 0.0289   0.0058  4.9494 0.0000  0.0175  0.0404
education          -0.1973   0.0481 -4.1000 0.0000 -0.2916 -0.1030
currentSmoker      -0.2111   0.1530 -1.3797 0.1677 -0.5110  0.0888
cigsPerDay          0.0206   0.0062  3.3260 0.0009  0.0085  0.0328
BPMeds              0.3657   0.2304  1.5870 0.1125 -0.0859  0.8173
prevalentStroke     0.7057   0.4848  1.4556 0.1455 -0.2445  1.6559
prevalentHyp        0.9162   0.1244  7.3649 0.0000  0.6724  1.1600
diabetes            0.7305   0.2980  2.4518 0.0142  0.1465  1.3145
totChol            -0.0010   0.0011 -0.9252 0.3549 -0.0032  0.0011
sysBP               0.0127   0.0038  3.3450 0.0008  0.0053  0.0201
diaBP              -0.0259   0.0061 -4.2321 0.0000 -0.0380 -0.0139
BMI                -0.0483   0.0123 -3.9281 0.0001 -0.0725 -0.0242
heartRate          -0.0224   0.0040 -5.6473 0.0000 -0.0302 -0.0147
glucose             0.0020   0.0021  0.9670 0.3336 -0.0021  0.0061
=================================================================
```

Εικόνα 21: Αποτελέσματα λογιστικής παλινδρόμησης

Από την Εικόνα 21 φαίνεται ότι οι μεταβλητές currentSmoker, BPMeds, prevalentStroke, totChol, glucose έχουν p-value>0,05 και επομένως δεν συμβάλουν στατιστικά σημαντικά στην πρόβλεψη καρδιαγγειακών νοσημάτων. Επομένως δεν είναι στατιστικά σημαντικές και αφαιρούνται από το μοντέλο της λογιστικής παλινδρόμησης όπως δείχνει η Εικόνα 22

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                Disease   No. Observations:                 3658
Model:                          Logit   Df Residuals:                     3648
Method:                           MLE   Df Model:                            9
Date:                Tue, 14 May 2019   Pseudo R-squ.:                0.06785
Time:                        10:53:24   Log-Likelihood:                -1454.7
converged:                       True   LL-Null:                       -1560.6
                                        LLR p-value:                1.114e-40
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
male            0.4034      0.104      3.889      0.000       0.200       0.607
age             0.0280      0.006      5.029      0.000       0.017       0.039
education      -0.2031      0.048     -4.270      0.000      -0.296      -0.110
cigsPerDay      0.0138      0.004      3.315      0.001       0.006       0.022
prevalentHyp    0.9581      0.121      7.931      0.000       0.721       1.195
diabetes        0.9244      0.231      4.007      0.000       0.472       1.377
sysBP           0.0133      0.004      3.550      0.000       0.006       0.021
diaBP          -0.0267      0.006     -4.383      0.000      -0.039      -0.015
BMI            -0.0477      0.012     -3.939      0.000      -0.071      -0.024
heartRate      -0.0237      0.004     -6.193      0.000      -0.031      -0.016
==============================================================================
```

Εικόνα 22: Τιμές συντελεστών παλινδρόμησης

Με βάση τις τιμές των coef(coefficients) και του κριτηρίου Z παρατηρούμε τις τιμές των συντελεστών παλινδρόμησης του συγκεκριμένου

μοντέλου (βλ. Εικόνα 22). Βλέπουμε για παράδειγμα ότι μία αύξησης στην τιμή της αρτηριακής υπέρτασης αυξάνει τον κίνδυνο εμφάνισης καρδιαγγειακών νοσημάτων. Ομοίως και στην περίπτωση του διαβήτη. Αντίθετα βλέπουμε ότι μία αύξηση της τιμής του δείκτη μάζας σώματος μειώνει τον κίνδυνο εμφάνισης καρδιαγγειακών νοσημάτων.

Στη συνέχεια κατασκευάζουμε το confusion matrixτου μοντέλου.

Actual Values	Predicted		
		Negative(0)	Positive(1)
	Negative(0)	TN=402	FP=189
	Positive(1)	FN=201	TP=357

Παρατηρούμε ότι πρόβλημα της ανισορροπίας των δεδομένων έχει άμεσο αντίκτυπο στα αποτελέσματα του confusion matrix. Συγκεκριμένα βλέπουμε ότι η απώλεια των δεδομένων μικρότερη καθώς έχουμε 1149 εγγραφές από 915 που είχαμε προηγουμένως. Επίσης, οι 759 αφορούν σωστές προβλέψεις έναντι 390 λανθασμένων. Μία καλή προσέγγιση για την αξιολόγηση των αποτελεσμάτων είναι να μετρήσουμε τα true/false positive (tp / fp) και τα true/false negative (tn / fn). Τα αποτελέσματα του πίνακα δείχνουν ότι από τα 603 δεδομένα , στην πραγματικότητα τα 402 παρατηρήθηκαν ως αληθώς αρνητικά. Από τα 546 που παρατηρήθηκαν ως θετικά τα 357 είναι αληθώς θετικά.

Στη συνέχεια ακολουθεί ο υπολογισμός των μετρικών precision , recall και f1 measure (βλ. Εικόνα 23).

```
precision= 0.6648936170212766
recall= 0.6510416666666666
f1= 0.6578947368421052
```

Εικόνα 23: Υπολογισμός των μετρικών precision, recall και f1 measure

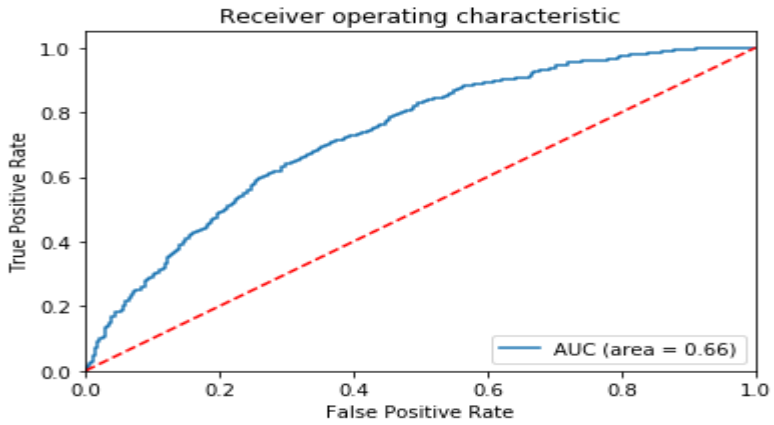
Το μοντέλο της λογιστικής παλινδρόμησης δίνει Precision=64,5%,Recall=65,1% και F1-score=65,7%.

```
Mean Accuracy 0.6624527713583508
```

Εικόνα 24:Μέτρηση της ακρίβειας του μοντέλου

Παρατηρούμε ότι η εκτέλεση του πειράματός μας με τη λογιστική παλινδρόμηση και με χρήση της διαδικασίας cross-validation επιστρέφει αποτελέσματα με μέση ακρίβεια 66% (βλ. Εικόνα 24) μικρότερη από την ακρίβεια του προηγούμενου μοντέλου λόγω του ότι της SMOTE υπερδειγματοληψίας.

Ένας ακόμη τρόπος για την αξιολόγηση της αποτελεσματικότητας ενός μοντέλου ταξινόμησης είναι το διάγραμμα ROC (Receiver Operating Characteristics) όπου κριτήριο για την αποτελεσματικότητα ενός μοντέλου λοιπόν αποτελεί η μορφή της καμπύλης ROC και συγκεκριμένα το εμβαδόν της περιοχής κάτω από την καμπύλη (AUC, Area Under Curve).



Γράφημα 53:Αποτελέσματα AUC

Το Γράφημα 53 δείχνει ότι η περιοχή κάτω από την ROC καμπύλη είναι 0.68.Επομένως διαφέρει στατιστικά σημαντικά από την τιμή0.5 που υποδηλώνει μηδενική ικανότητα του μοντέλου να προβλέπει καρδιαγγειακά νοσήματα.

Παρατηρούμε ότι λύνοντας το πρόβλημα της ανισορροπίας των δεδομένων το μοντέλο δίνει μικρότερη ακρίβεια αλλά μεγαλύτερη τιμή της AUC σε σχέση με το αρχικό γεγονός που δείχνει ότι είναι καλύτερο στην πρόβλεψη των καρδιαγγειακών νοσημάτων.

9.2.7 Βελτίωση του μοντέλου αφαιρώντας τη μεταβλητή εκπαίδευση

Θέλοντας να βελτιώσουμε την απόδοση του μοντέλου της λογιστικής παλινδρόμησης, θεωρήσαμε ότι η μεταβλητή educationθα μπορούσε να διαγραφεί καθώς δεν συνιστά, από ιατρικής άποψης, έναν παράγοντα σχετικό

με την εμφάνιση καρδιαγγειακών νοσημάτων, επομένως δεν έχει και λογική να βρίσκεται στο σύνολο των δεδομένων μας.

Επομένως μέσω της Python διαγράφουμε τη μεταβλητή education όπως φαίνεται και στην Εικόνα 25:

```
male               0
age                0
currentSmoker      0
cigsPerDay        29
BPMeds            53
prevalentStroke    0
prevalentHyp       0
diabetes           0
totChol           50
sysBP              0
diaBP              0
BMI               19
heartRate          1
glucose          388
Disease            0
dtype: int64
```

Εικόνα 25: Μεταβλητές του συνόλου δεδομένων

Η παραπάνω εικόνα δείχνει επίσης ποιες μεταβλητές έχουν ελλιπείς τιμές τις οποίες και απαλείφουμε (βλ Εικόνα 26).

```
male               0
age                0
currentSmoker      0
cigsPerDay         0
BPMeds             0
prevalentStroke    0
prevalentHyp       0
diabetes           0
totChol            0
sysBP              0
diaBP              0
BMI                0
heartRate          0
glucose            0
Disease            0
dtype: int64
```

Εικόνα 26: Αφαίρεση ελλιπών τιμών

```
percentage without cardiovascular disease 84.75073313782991
percentage with cardiovascular disease 15.249266862170089
```

Εικόνα 26: Ανισορροπία των κλάσεων

Επειδή εμφανίζεται το πρόβλημα της ανισορροπίας των κλάσεων στην Εικόνα 27 χρησιμοποιούμε το αλγόριθμο SMOTE για να επιλύσουμε το πρόβλημα.

```
length of oversampled data is  4762
Number of no subscription in oversampled data 2381
Number of subscription 2381
Proportion of no cardiovascular disease  data in oversampled data is  0.5
Proportion of cardiovascular disease data in oversampled data is  0.5
```

Εικόνα 27: Αποτελέσματα της υπερδειγματοληψίας SMOTE

Στη συνέχεια εφαρμόζουμε τη λογιστική παλινδρόμηση

Logit Regression Results

Dep. Variable:	Disease	No. Observations:	3751
Model:	Logit	Df Residuals:	3737
Method:	MLE	Df Model:	13
Date:	Wed, 15 May 2019	Pseudo R-squ.:	0.06313
Time:	10:27:54	Log-Likelihood:	-1500.6
converged:	True	LL-Null:	-1601.7
		LLR p-value:	4.723e-36

	coef	std err	z	P>|z|	[0.025	0.975]
male	0.4089	0.103	3.976	0.000	0.207	0.610
age	0.0286	0.006	4.989	0.000	0.017	0.040
currentSmoker	-0.2367	0.150	-1.574	0.115	-0.531	0.058
cigsPerDay	0.0214	0.006	3.528	0.000	0.010	0.033
BPMeds	0.3241	0.227	1.426	0.154	-0.121	0.770
prevalentStroke	0.7838	0.483	1.623	0.105	-0.163	1.731
prevalentHyp	0.9718	0.122	7.964	0.000	0.733	1.211
diabetes	0.7825	0.291	2.686	0.007	0.212	1.354
totChol	-0.0015	0.001	-1.386	0.166	-0.004	0.001
sysBP	0.0133	0.004	3.548	0.000	0.006	0.021
diaBP	-0.0301	0.006	-5.027	0.000	-0.042	-0.018
BMI	-0.0465	0.012	-3.804	0.000	-0.070	-0.023
heartRate	-0.0228	0.004	-5.817	0.000	-0.031	-0.015
glucose	0.0019	0.002	0.899	0.369	-0.002	0.006

Εικόνα 28: Αποτελέσματα λογιστικής παλινδρόμησης

Στην Εικόνα 28 φαίνεται ότι οι μεταβλητές currentSmoker,BPMeds, prevalentStroke,totChol,glucose έχουν p-value>0,05 και επομένως δεν συμβάλουν στατιστικά σημαντικά στην πρόβλεψη καρδιαγγειακών νοσημάτων και επομένως τις αφαιρούμε από το μοντέλο μας. Τα αποτελέσματα είναι τα παρακάτω (βλ. Εικόνα 29):

Logit Regression Results

Dep. Variable:	Disease	No. Observations:	4762
Model:	Logit	Df Residuals:	4753
Method:	MLE	Df Model:	8
Date:	Wed, 15 May 2019	Pseudo R-squ.:	0.1074
Time:	10:31:56	Log-Likelihood:	-2946.4
converged:	True	LL-Null:	-3300.8
		LLR p-value:	9.086e-148

	coef	std err	z	P>|z|	[0.025	0.975]
male	0.4986	0.073	6.864	0.000	0.356	0.641
age	0.0442	0.004	11.244	0.000	0.037	0.052
cigsPerDay	0.0209	0.003	7.218	0.000	0.015	0.027
prevalentHyp	0.9508	0.086	11.031	0.000	0.782	1.120
diabetes	1.0525	0.187	5.618	0.000	0.685	1.420
sysBP	0.0147	0.003	5.167	0.000	0.009	0.020
diaBP	-0.0276	0.005	-6.025	0.000	-0.037	-0.019
BMI	-0.0374	0.008	-4.484	0.000	-0.054	-0.021
heartRate	-0.0237	0.003	-9.117	0.000	-0.029	-0.019

Εικόνα 29: Τιμές συντελεστών παλινδρόμησης

Στη συνέχεια κατασκευάζουμε το confusion matrixτου μοντέλου.

Actual Values	Predicted		
		Negative(0)	Positive(1)
	Negative(0)	TN=409	FP=183
	Positive(1)	FN=189	TP=410

Από τις 1191 εγγραφές από 915 που είχαμε προηγουμένως. Επίσης , οι 819 αφορούν σωστές προβλέψεις έναντι 372 λανθασμένων. Μία καλή προσέγγιση για την αξιολόγηση των αποτελεσμάτων είναι να μετρήσουμε τα true/false positive (tp / fp) και τα true/false negative (tn / fn). Τα αποτελέσματα του πίνακα δείχνουν ότι από τα 598 δεδομένα , στην πραγματικότητα τα 409 παρατηρήθηκαν ως αληθώς αρνητικά. Από τα 593 που παρατηρήθηκαν ως θετικά τα 410 είναι αληθώς θετικά.

Στη συνέχεια ακολουθεί ο υπολογισμός των μετρικών precision , recall και f1 measure.

```
precision= 0.6913996627318718
recall= 0.6844741235392321
f1= 0.6879194630872484
```

Εικόνα 30: Υπολογισμός των μετρικών precision, recall και f1 measure

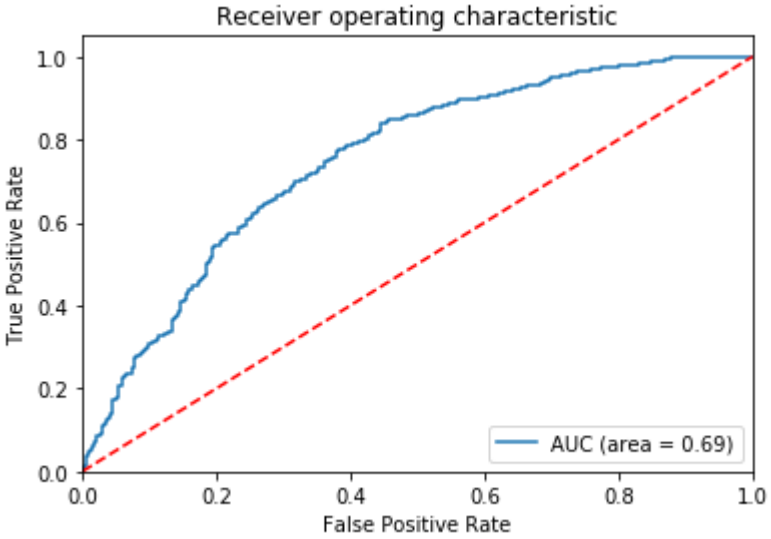
Το μοντέλο της λογιστικής παλινδρόμησης δίνει Precision=69,1%, Recall=68,4% και F1-score=68,7% (βλ. Εικόνα 30).

```
Mean Accuracy 0.6885745930171232
```

Εικόνα 31:Μέτρηση της ακρίβειας του μοντέλου

Παρατηρούμε ότι η εκτέλεση του πειράματός μας με τη λογιστική παλινδρόμηση και με χρήση της διαδικασίας cross-validation επιστρέφει αποτελέσματα με μέση ακρίβεια περίπου 69% η οποία είναι μεγαλύτερη από την ακρίβεια του προηγούμενου μοντέλου (βλ. Εικόνα 31).

Ένας ακόμη τρόπος για την αξιολόγηση της αποτελεσματικότητας ενός μοντέλου ταξινόμησης είναι το διάγραμμα ROC (Receiver Operating Characteristics) όπου κριτήριο για την αποτελεσματικότητα ενός μοντέλου λοιπόν αποτελεί η μορφή της καμπύλης ROC και συγκεκριμένα το εμβαδόν της περιοχής κάτω από την καμπύλη (AUC, Area Under Curve).



Γράφημα 54:Αποτελέσματα AUC

Παρατηρούμε ότι επιλύοντας το πρόβλημα της ανισορροπίας των κλάσεων και διαγράφοντας την μεταβλητή education η ακρίβεια του μοντέλου είναι ελαφρώς μεγαλύτερη ενώ μεγαλύτερη είναι και η τιμή της AUC (βλ. Γράφημα 54). Επομένως αυτή τη μεθοδολογία θα εφαρμόσουμε για την μελέτη της απόδοσης και των υπόλοιπων αλγόριθμων κατηγοριοποίησης.

9.2.8 Αλγόριθμος NAIVE BAYES

Για την εφαρμογή του Naive Bayes Classifier έχουν γίνει τα εξής βήματα:

- Διαγραφή της μεταβλητής education
- Αφαίρεση των ελλιπών τιμών
- Εφαρμογή της SMOTE μεθοδολογίας για το επίλυση του προβλήματος της ανισορροπίας των κλάσεων (βλ. Εικόνα 32).

```
length of oversampled data is  4762
Number of no cardiovascular disease in oversampled data 2381
Number of no cardiovascular disease  2381
Proportion of no cardiovascular disease  data in oversampled data is  0.5
Proportion of cardiovascular disease data in oversampled data is  0.5
```

Εικόνα 32: Αποτελέσματα της υπερδειγματοληψίας

Εφαρμόζοντας τον αλγόριθμο του Bayes παίρνουμε τα παρακάτω αποτελέσματα.

συνέχεια κατασκευάζουμε το confusion matrix του μοντέλου.

Actual Values	Predicted

		Negative(0)	Positive(1)
	Negative(0)	TN=496	FP=96
	Positive(1)	FN=353	TP=246

Από τις 1191 εγγραφές, οι 742 αφορούν σωστές προβλέψεις έναντι 449 λανθασμένων. Μία καλή προσέγγιση για την αξιολόγηση των αποτελεσμάτων είναι να μετρήσουμε τα true/false positive (tp / fp) και τα true/false negative (tn / fn). Τα αποτελέσματα του πίνακα δείχνουν ότι από τα 849 δεδομένα , στην πραγματικότητα τα 496 παρατηρήθηκαν ως αληθώς αρνητικά. Από τα 342 που παρατηρήθηκαν ως θετικά τα 246 είναι αληθώς θετικά.

Στη συνέχεια ακολουθεί ο υπολογισμός των μετρικών precision , recall και f1 measure.

```
precision= 0.7192982456140351
recall= 0.41068447412353926
f1= 0.5228480340063761
```

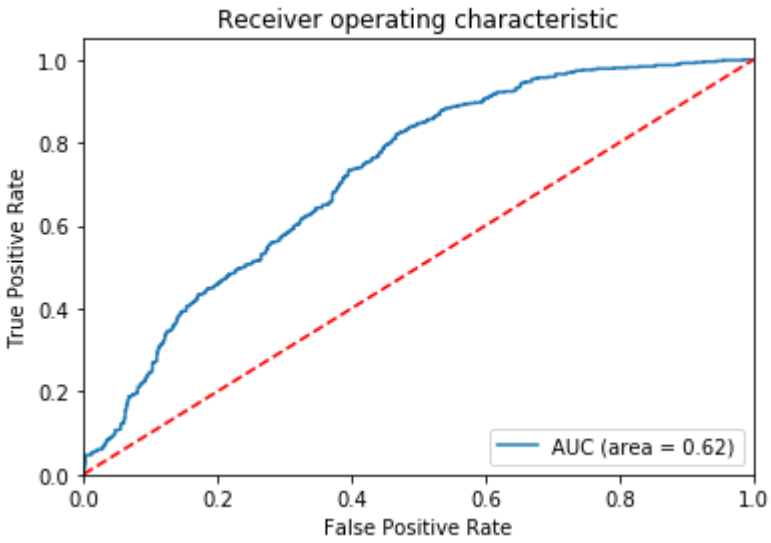
Εικόνα 33: Υπολογισμός των μετρικών precision, recall και f1 measure

Το μοντέλο του Naive Bayes Classifier δίνει Precision=72%, Recall=41 % και F1-score=52,2% (βλ. Εικόνα 33).

```
precision= 0.7192982456140351
```

Εικόνα 34:Μέτρηση της ακρίβειας του μοντέλου

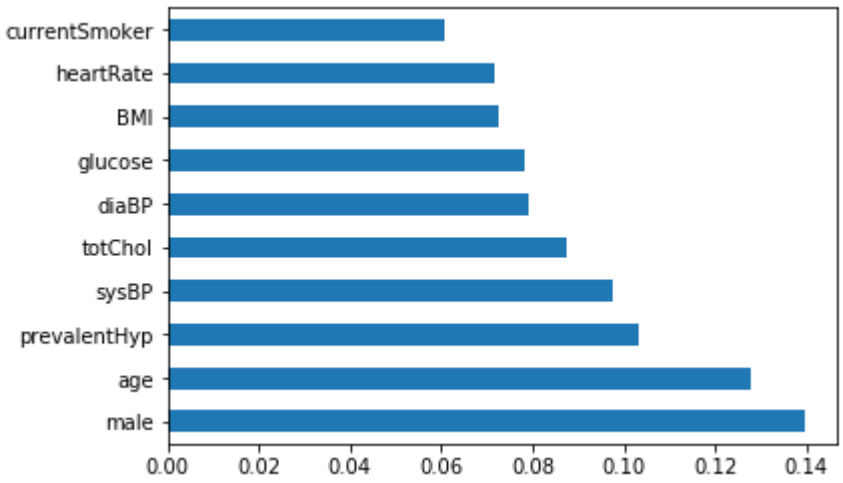
Η εφαρμογή του παραπάνω μοντέλου και με χρήση της διαδικασίας cross-validation επιστρέφει αποτελέσματα με μέση ακρίβεια ~72 % (βλ. Εικόνα 34). Στη συνέχεια θα ελέγξουμε την προγνωστική ικανότητα του μοντέλου μέσω της καμπύλης ROC εξετάζοντας την τιμή της AUC.



Γράφημα 55: Αποτελέσματα AUC

Από το Γράφημα 55 βλέπουμε ότι η τιμή της AUC είναι 0.62. Επομένως διαφέρει στατιστικά σημαντικά από την τιμή 0.5 και υποδηλώνει ότι το μοντέλο μας μπορεί να προβλέπει καρδιαγγειακά νοσήματα σε σχετικά ικανοποιητικό βαθμό.

Λαμβάνοντας υπόψη τα παραπάνω αποτελέσματα παρατηρούμε ότι το μοντέλο προσφέρει σχετικά ικανοποιητική απόδοση. Η βελτίωση του μοντέλου μπορεί να γίνει μέσω της κατάλληλης επιλογής χαρακτηριστικών. Το παρακάτω γράφημα μας δείχνει ποια χαρακτηριστικά(features) μπορούμε να κρατήσουμε ώστε να βελτιώσουμε την απόδοση του μοντέλου μας.



Γράφημα 56:Επιλογή χαρακτηριστικών

Το Γράφημα 56 μας δείχνει ποια χαρακτηριστικά(features) μπορούμε να κρατήσουμε ώστε να βελτιώσουμε την απόδοση του μοντέλου μας.

9.2.9 Αλγόριθμος DECISION TREE

Για την εφαρμογή του αλγορίθμου Decision Tree έχουν γίνει τα εξής βήματα:

- Διαγραφή της μεταβλητής education
- Αφαίρεση των ελλιπών τιμών
- Εφαρμογή της SMOTE μεθοδολογίας για το επίλυση του προβλήματος της ανισορροπίας των κλάσεων(βλ. Εικόνα 35).

```
length of oversampled data is  4762
Number of no cardiovascular disease in oversampled data 2381
Number of no cardiovascular disease  2381
Proportion of no cardiovascular disease  data in oversampled data is  0.5
Proportion of cardiovascular disease data in oversampled data is  0.5
```

Εικόνα 35: Αποτελέσματα της υπερδειγματοληψίας

Εφαρμόζοντας τον αλγόριθμο Decision Tree παίρνουμε τα παρακάτω αποτελέσματα.

Actual Values	Predicted		
		Negative(0)	Positive(1)
	Negative(0)	TN=464	FP=128
	Positive(1)	FN=104	TP=495

Από τις 1191 εγγραφές παρατηρούμε ότι οι 959 αφορούν σωστές προβλέψεις έναντι 232 λανθασμένων. Μία καλή προσέγγιση για την αξιολόγηση των αποτελεσμάτων είναι να μετρήσουμε τα true/false positive (tp / fp) και τα true/false negative (tn / fn). Τα αποτελέσματα του πίνακα δείχνουν ότι από τα 568 δεδομένα , στην πραγματικότητα τα 464 παρατηρήθηκαν ως αληθώς αρνητικά. Από τα 623 που παρατηρήθηκαν ως θετικά τα 495 είναι αληθώς θετικά.

Στη συνέχεια ακολουθεί ο υπολογισμός των μετρικών precision , recall και f1 measure (βλ. Εικόνα 36).

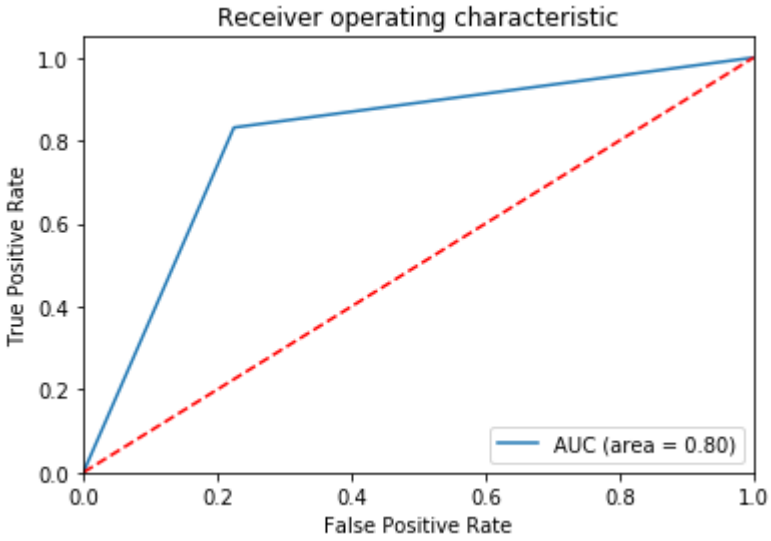
```
precision= 0.7945425361155698
recall= 0.8263772954924875
f1= 0.8101472995090017
```

Εικόνα 36: Υπολογισμός των μετρικών precision, recall και f1 measure

```
Mean Accuracy 0.8436060616715306
```

Εικόνα 37:Μέτρηση της ακρίβειας του μοντέλου

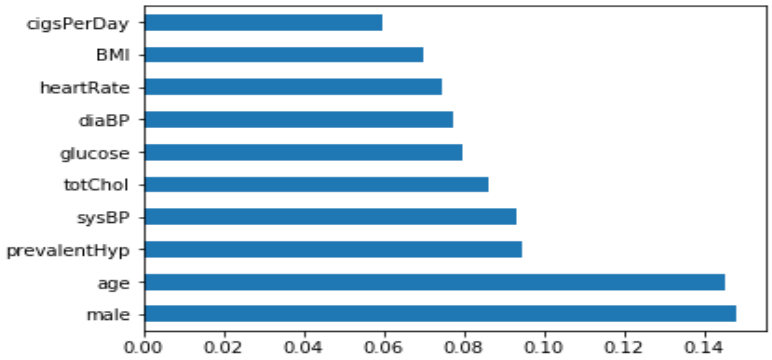
Η εφαρμογή του παραπάνω μοντέλου και με χρήση της διαδικασίας cross-validation επιστρέφει αποτελέσματα με μέση ακρίβεια 84,3% (βλ. Εικόνα 37). Στη συνέχεια θα ελέγξουμε την προγνωστική ικανότητα του μοντέλου μέσω της καμπύλης ROC εξετάζοντας την τιμή της AUC.



Γράφημα 57: Αποτελέσματα AUC

Από το Γράφημα 57 βλέπουμε ότι η τιμή της AUC είναι 0.8. Επομένως διαφέρει στατιστικά σημαντικά από την τιμή 0.5.Επίσης η τιμή της AUC υποδηλώνει ότι το μοντέλο μας προσφέρει αρκετά καλή απόδοση στην πρόγνωση καρδιαγγειακών νοσημάτων.

Για την περίπτωση που επιθυμούμε τη βελτίωση του μοντέλου μία λύση είναι μέσω της κατάλληλης επιλογής χαρακτηριστικών. Το παρακάτω γράφημα μας δείχνει ποια χαρακτηριστικά(features) μπορούμε να κρατήσουμε ώστε να βελτιώσουμε την απόδοση του μοντέλου μας (βλ. Γράφημα 58).



Γράφημα 58: Επιλογή Χαρακτηριστικών

9.2.10 Αλγόριθμος k-Nearest Neighbor

Για την εφαρμογή του αλγορίθμου KNN έχουν γίνει τα εξής βήματα:

- Διαγραφή της μεταβλητής education
- Αφαίρεση των ελλιπών τιμών
- Εφαρμογή της SMOTE μεθοδολογίας για το επίλυση του προβλήματος της ανισορροπίας των κλάσεων (βλ. Εικόνα 38).

```
length of oversampled data is  4762
Number of no cardiovascular disease in oversampled data 2381
Number of no cardiovascular disease  2381
Proportion of no cardiovascular disease  data in oversampled data is  0.5
Proportion of cardiovascular disease data in oversampled data is  0.5
```

Εικόνα 38: Αποτελέσματα της υπερδειγματοληψίας SMOTE

Εφαρμόζοντας τον αλγόριθμο k-κοντινότερου γείτονα παίρνουμε τα παρακάτω αποτελέσματα.

Actual Values	Predicted		
		Negative(0)	Positive(1)
	Negative(0)	TN=765	FP=33
	Positive(1)	FN=126	TP=14

Παρατηρούμε ότι στα 938 δεδομένα, τα 779 είναι σωστές προβλέψεις και οι 159 είναι λάθος. Μία καλή προσέγγιση για την αξιολόγηση των αποτελεσμάτων είναι να μετρήσουμε τα true/false positive (tp / fp) και τα true/false negative (tn / fn). Τα αποτελέσματα του πίνακα δείχνουν ότι από τα 891 που παρατηρήθηκαν ως αρνητικά, στην πραγματικότητα αυτά είναι 765. Από τα 55 που παρατηρήθηκαν αρνητικά τα πραγματικά είναι τα 14.

Στη συνέχεια ακολουθεί ο υπολογισμός των μετρικών precision , recall και f1 measure (βλ. Εικόνα 39).

```
precision= 0.2978723404255319
recall= 0.1
f1= 0.14973262032085563
```

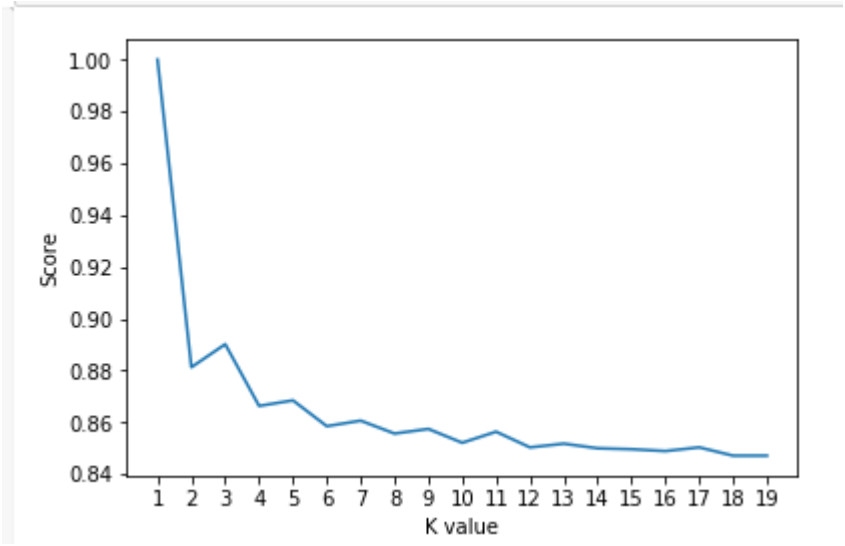
Εικόνα 39: Υπολογισμός των μετρικών precision, recall και f1 measure

Το μοντέλο KNN δίνει Precision=29,7%, Recall=0,1% και F1-score=14,9%.

```
Mean Accuracy 0.8284316655532505
```

Εικόνα 40:Μέτρηση της ακρίβειας του μοντέλου

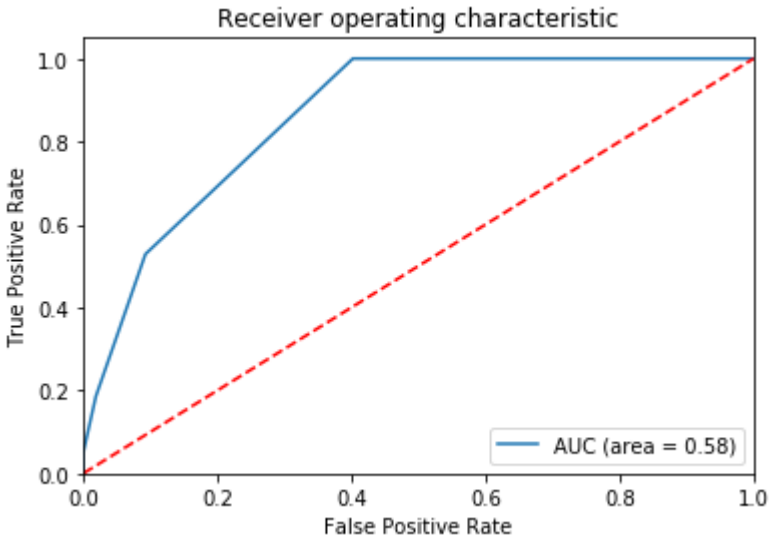
Η εφαρμογή του παραπάνω μοντέλου και με χρήση της διαδικασίας cross-validation επιστρέφει αποτελέσματα με μέση ακρίβεια ~83% (βλ. Εικόνα 40).



Γράφημα 59: Απόδοση KNN

Επίσης παρατηρούμε ότι ο αλγόριθμος φτάνει στη μέγιστη απόδοση με Score=89% όταν k=3 (βλ. Γράφημα 59).

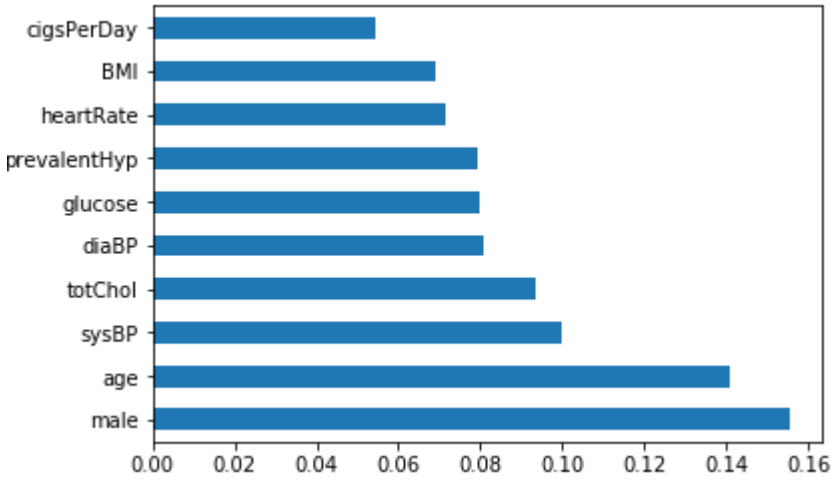
Στη συνέχεια θα ελέγξουμε την προγνωστική ικανότητα του μοντέλου μέσω της καμπύλης ROC εξετάζοντας την τιμή της AUC.



Γράφημα 60: Αποτελέσματα AUC

Από το Γράφημα 60 βλέπουμε ότι η τιμή της AUC είναι 0.51. Επομένως δεν διαφέρει στατιστικά σημαντικά από την τιμή 0.5.Επίσης η τιμή της AUC υποδηλώνει ότι το μοντέλο μας προσφέρει κακή απόδοση στην πρόγνωση καρδιαγγειακών νοσημάτων.

Για την περίπτωση που επιθυμούμε τη βελτίωση του μοντέλου μία λύση είναι μέσω της κατάλληλης επιλογής χαρακτηριστικών. Το παρακάτω γράφημα μας δείχνει ποια χαρακτηριστικά (features) μπορούμε να κρατήσουμε ώστε να βελτιώσουμε την απόδοση του μοντέλου μας (βλ. Γράφημα 61).



Γράφημα 61: Επιλογή Χαρακτηριστικών

9.2.11 Αλγόριθμος SVM

Για την εφαρμογή του αλγορίθμου SVM έχουν γίνει τα εξής βήματα:

- Διαγραφή της μεταβλητής education
- Αφαίρεση των ελλιπών τιμών
- Εφαρμογή της SMOTE μεθοδολογίας για το επίλυση του προβλήματος της ανισορροπίας των κλάσεων (βλ. Εικόνα 41).

```
length of oversampled data is  4762
Number of no cardiovascular disease in oversampled data 2381
Number of no cardiovascular disease  2381
Proportion of no cardiovascular disease  data in oversampled data is  0.5
Proportion of cardiovascular disease data in oversampled data is  0.5
```

Εικόνα 62: Αποτελέσματα της υπερδειγματοληψίας SMOTE

Εφαρμόζοντας τον αλγόριθμο SVM παίρνουμε τα παρακάτω αποτελέσματα.

Actual Values	Predicted		
		Negative(0)	Positive(1)
	Negative(0)	TN=798	FP=0
	Positive(1)	FN=140	TP=0

Παρατηρούμε ότι στα 938 δεδομένα, τα 798 είναι σωστές προβλέψεις και οι 140 είναι λάθος. Μία καλή προσέγγιση για την αξιολόγηση των αποτελεσμάτων είναι να μετρήσουμε τα true/false positive (tp / fp) και τα true/false negative (tn / fn). Τα αποτελέσματα του πίνακα δείχνουν ότι από τα 938 που παρατηρήθηκαν ως αρνητικά, στην πραγματικότητα αυτά είναι 798, ενώ δεν υπάρχουν δεδομένα που να κατηγοριοποιήθηκαν ως αληθώς ή ψευδώς θετικά.

Στη συνέχεια ακολουθεί ο υπολογισμός των μετρικών precision , recall και f1 measure (βλ. Εικόνα 42).

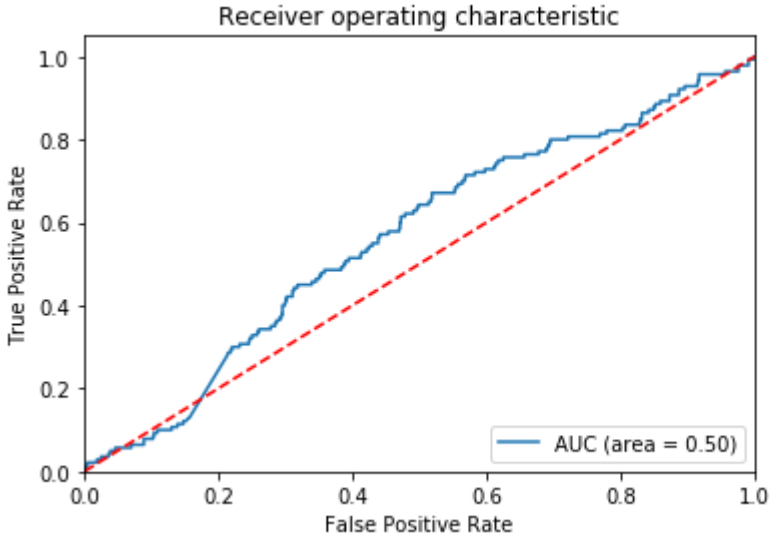
```
precision= 0.0
recall= 0.0
f1= 0.0
```

Εικόνα 42: Υπολογισμός των μετρικών precision, recall και f1 measure

```
Mean Accuracy 0.8301009106571499
```

Εικόνα 43:Μέτρηση της ακρίβειας του μοντέλου

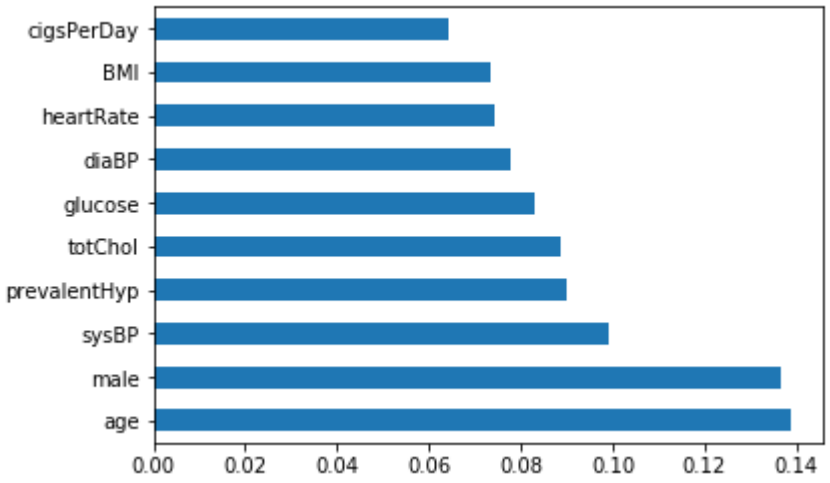
Η εφαρμογή του παραπάνω μοντέλου και με χρήση της διαδικασίας cross-validation επιστρέφει αποτελέσματα με μέση ακρίβεια 83% (βλ. Εικόνα 43).



Γράφημα 62: Αποτελέσματα AUC

Από το Γράφημα 62 βλέπουμε ότι η τιμή της AUCείναι 0.50. Επομένως δεν διαφέρει στατιστικά σημαντικά από την τιμή 0, 5. Έτσι οδηγούμαστε στο συμπέρασμα ότι για το συγκεκριμένο σύνολο δεδομένων ο αλγόριθμος Support Vector Machine δεν είναι κατάλληλος για την πρόγνωση καρδιαγγειακών νοσημάτων.

Για την περίπτωση που επιθυμούμε τη βελτίωση του μοντέλου μία λύση είναι μέσω της κατάλληλης επιλογής χαρακτηριστικών. Το παρακάτω γράφημα μας δείχνει ποια χαρακτηριστικά (features) μπορούμε να κρατήσουμε ώστε να βελτιώσουμε την απόδοση του μοντέλου μας (βλ. Γράφημα 63).



Γράφημα 63: Επιλογή χαρακτηριστικών

9.2.12 Αλγόριθμος Random Forest

Για την εφαρμογή του αλγορίθμου Random Forest έχουν γίνει τα εξής βήματα:

- Διαγραφή της μεταβλητής education
- Αφαίρεση των ελλιπών τιμών
- Εφαρμογή της SMOTE μεθοδολογίας για το επίλυση του προβλήματος της ανισορροπίας των κλάσεων (βλ. Εικόνα 44).

```
length of oversampled data is  4762
Number of no cardiovascular disease in oversampled data 2381
Number of no cardiovascular disease  2381
Proportion of no cardiovascular disease  data in oversampled data is  0.5
Proportion of cardiovascular disease data in oversampled data is  0.5
```

Εικόνα 44: Επίλυση του προβλήματος της ανισορροπίας των κλάσεων

Εφαρμόζοντας τον αλγόριθμο Random Forest παίρνουμε τα παρακάτω αποτελέσματα.

		Predicted	
Actual Values		Negative(0)	Positive(1)
	Negative(0)	TN=793	FP=5
	Positive(1)	FN=131	TP=9

Παρατηρούμε ότι στα 938 δεδομένα, τα 802 είναι σωστές προβλέψεις και οι 136 είναι λάθος. Μία καλή προσέγγιση για την αξιολόγηση των

αποτελεσμάτων είναι να μετρήσουμε τα true/false positive (tp / fp) και τα true/false negative (tn / fn). Τα αποτελέσματα του πίνακα δείχνουν ότι από τα 824 που παρατηρήθηκαν ως αρνητικά στην πραγματικότητα αυτά είναι 793 αληθώς αρνητικά. Από τα 14 που παρατηρήθηκαν θετικά τα 9 είναι αληθώς θετικά.

Στη συνέχεια (βλ. Εικόνα 45) ακολουθεί ο υπολογισμός των μετρικών precision , recall και f1 measure.

```
precision= 0.6428571428571429
recall= 0.06428571428571428
f1= 0.11688311688311687
```

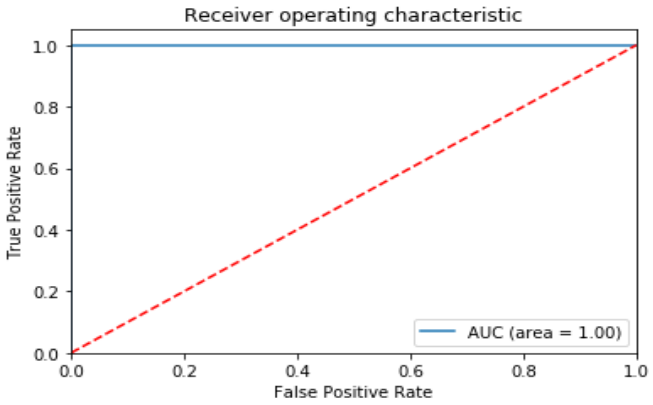
Εικόνα 45: Υπολογισμός των μετρικών precision, recall και f1 measure

Για τον αλγόριθμο Random Forest έχουμε Precision=64,2%,Recall=6,4% και F1-score=11,6% (βλ. Εικόνα 46).

```
Mean Accuracy 0.9110298512710523
```

Εικόνα 46:Μέτρηση της ακρίβειας του μοντέλου

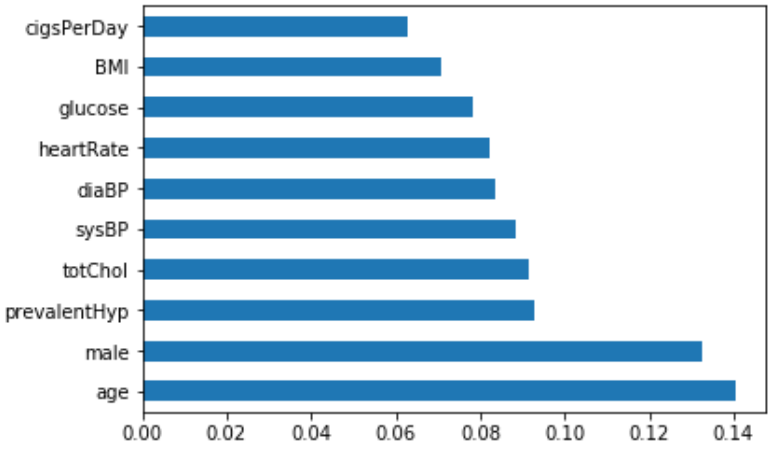
Η εφαρμογή του παραπάνω μοντέλου και με χρήση της διαδικασίας cross-validation επιστρέφει αποτελέσματα με μέση ακρίβεια 91%.



Γράφημα 64: Αποτελέσματα AUC

Από το Γράφημα 64 βλέπουμε ότι η τιμή της AUC είναι 1 . Επομένως διαφέρει στατιστικά σημαντικά από την τιμή 0.5. Έτσι οδηγούμαστε στο συμπέρασμα ότι για το συγκεκριμένο σύνολο δεδομένων ο αλγόριθμος Random Forest είναι ο πλέον κατάλληλος για την πρόγνωση καρδιαγγειακών νοσημάτων καθώς για AUC=1 έχει την μέγιστη προγνωστική ικανότητα..

Για την περίπτωση που επιθυμούμε τη βελτίωση του μοντέλου μία λύση είναι μέσω της κατάλληλης επιλογής χαρακτηριστικών. Το παρακάτω γράφημα μας δείχνει ποια χαρακτηριστικά(features) μπορούμε να κρατήσουμε ώστε να βελτιώσουμε την απόδοση του μοντέλου μας (βλ. Γράφημα 65).



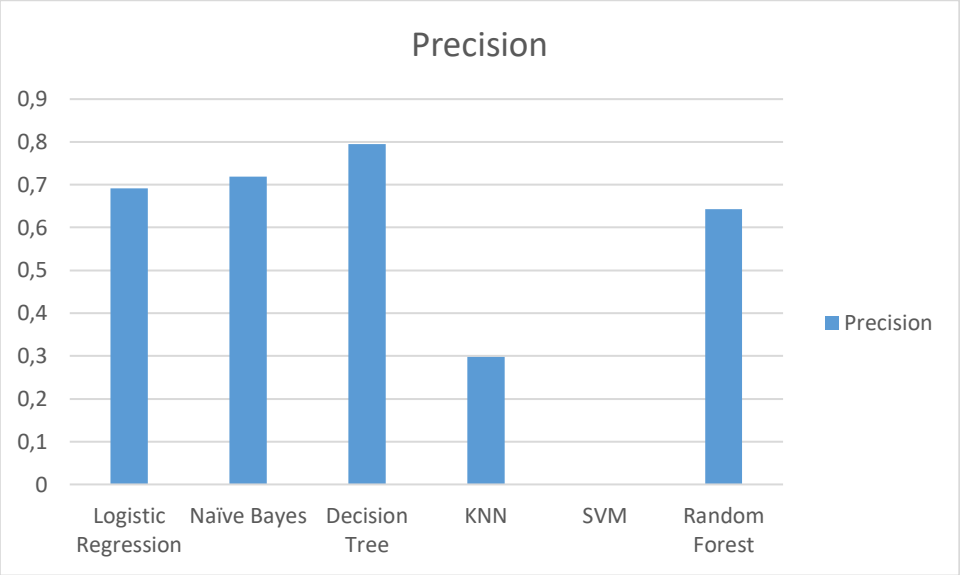
Γράφημα 65: Αποτελέσματα AUC

9.3 Συζήτηση

Για τους αλγόριθμους κατηγοριοποίησης της παρούσας διατριβής τα αποτελέσματα ως προς τις μετρικές που χρησιμοποιήθηκαν φαίνονται παρακάτω.

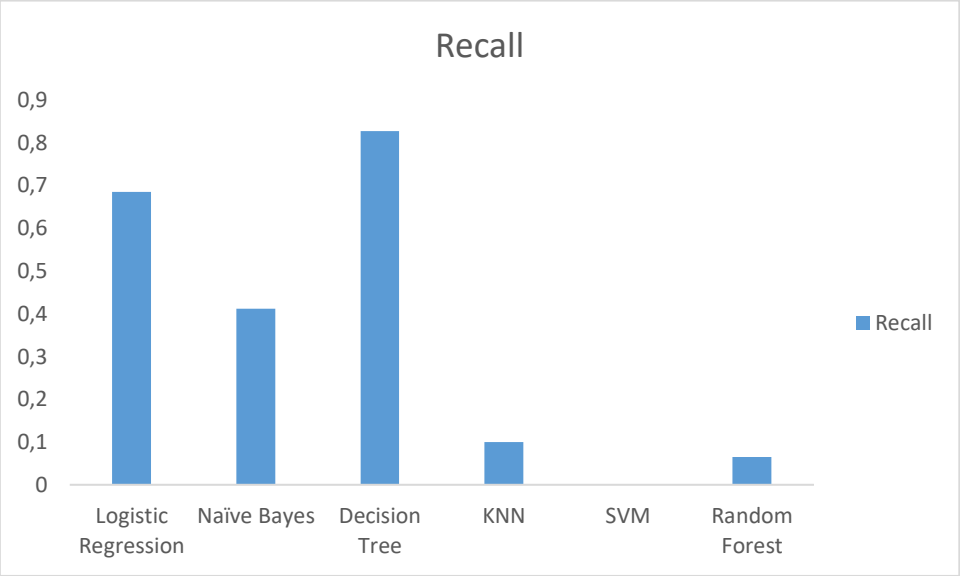
	Precision	Recall	F1-score	Accuracy	AUC
Logistic Regression	0,69139	0,68447	0,68791	0,68857	0,69
Naïve Bayes	0,71929	0,41068	0,52284	0,71929	0,62
Decision Tree	0,79454	0,82637	0,81014	0,8436	0,8
KNN	0,29787	0,1	0,14973	0,82843	0,51
SVM	0	0	0	0,8301	0,5
Random Forest	0,64285	0,06428	0,11688	0,91102	1

Εικόνα 47:Αποτελέσματα κατηγοριοποιητών



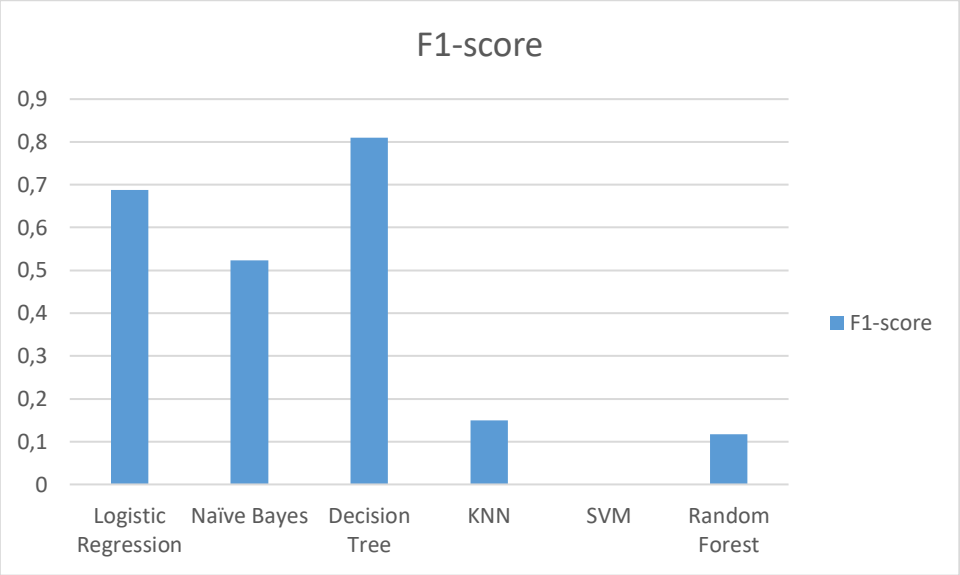
Γράφημα 666: Αποτελέσματα Precision

Παρατηρούμε ότι μεγαλύτερο Precision έχει ο αλγόριθμος Decision tree ακολουθούμενος από το Naive Bayes (βλ. Γράφημα 66). Το χαμηλότερο Precision το παρουσιάζει ο SVM (μηδέν) και στη συνέχεια ο αλγόριθμος KNN.



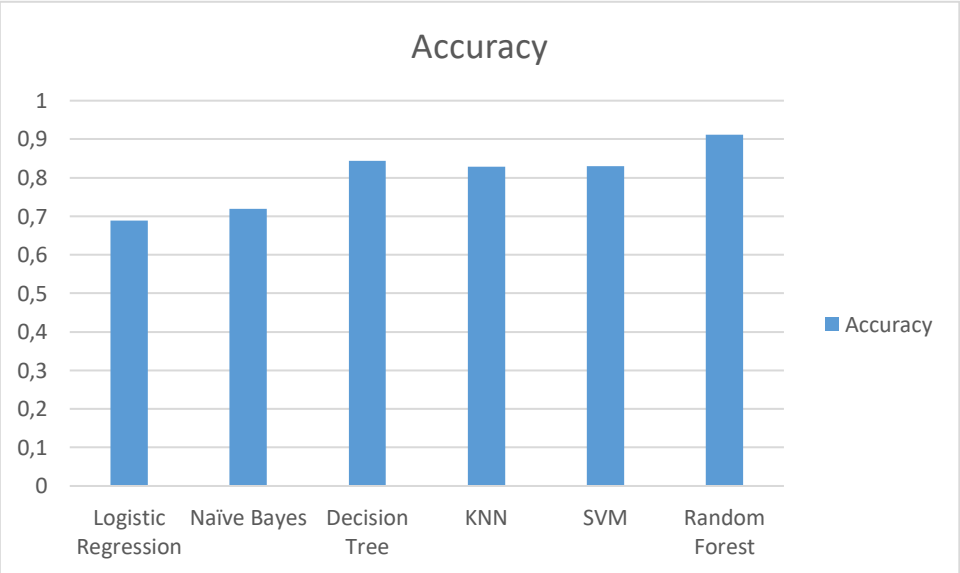
Γράφημα 67: Αποτελέσματα Recall

Ως προς την ανάκληση το μεγαλύτερο σκορ το επιτυγχάνει ο Decision tree και ακολουθεί η Λογιστική Παλινδρόμηση. Ο αλγόριθμος SVM έχει μηδενικό σκορ (βλ. Γράφημα 67). Ο αλγόριθμος Random Forest επιτυγχάνει το χαμηλότερο σκορ ως προς την ανάκληση.



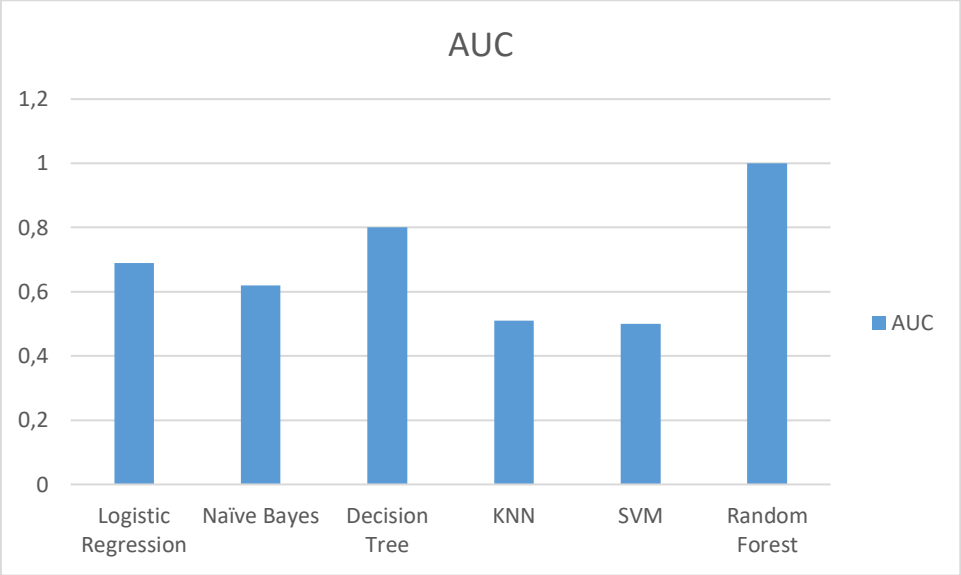
Γράφημα 68: Αποτελέσματα F1-score

Από το Γράφημα 68 καθώς παρατηρούμε ότι τα μεγαλύτερα F1-score έχουν οι αλγόριθμοι Decision Tree και η Λογιστική Παλινδρόμηση. Ο αλγόριθμος SVM έχει F1-score=0.Ο αλγόριθμος με το χαμηλότερο F1-score είναι ο Random Forest.



Γράφημα 69: Αποτελέσματα Accuracy

Παρατηρούμε ότι τη μεγαλύτερη ακρίβεια στην απόδοση την παρουσιάζει ο αλγόριθμος Random Forest ακολουθούμενος από τον Decision tree (βλ. Γράφημα 69). Τη χαμηλότερη ακρίβεια την παρουσιάζει ο Naive Bayes και στη συνέχεια η Λογιστική Παλινδρόμηση.



Γράφημα 70: Αποτελέσματα AUC

Σχετικά με την τιμή της AUC που αντικατοπτρίζει την προγνωστική ικανότητα του εκάστοτε αλγορίθμου, παρατηρούμε ότι ο αλγόριθμος Random Forest έχει την καλύτερη προγνωστική ικανότητα ακολουθούμενος από τον Decision tree και το k-Nearest Neighbors (βλ. Γράφημα 70). Τέλος τη χαμηλότερη ικανότητα πρόγνωσης παρουσιάζουν οι αλγόριθμοι SVM και Naive Bayes αντίστοιχα.

Μετά την αναζήτηση της βιβλιογραφίας βρέθηκαν δέκα δημοσιεύσεις που σχετίζονται με την πρόβλεψη καρδιαγγειακών νοσημάτων χρησιμοποιώντας τεχνικές κατηγοριοποίησης [117-127]. Σε όλες τις δημοσιεύσεις ακολουθείται η ίδια μεθοδολογία ως προς τη διαχείριση του συνόλου δεδομένων δηλαδή η αναζήτηση για ελλιπή δεδομένα, η απαλοιφή ελλιπών τιμών, η αναγνώριση κατηγορικών και ποσοτικών δεδομένων και η κωδικοποίηση των κατηγορικών δεδομένων ώστε να παίρνουν ακέραιες τιμές. Επίσης σε όλες τις μελέτες γίνεται διαχωρισμός του συνόλου δεδομένων σε training set και test set.

Από την άλλη πλευρά έξι δημοσιεύσεις δεν κάνουν κάποιον έλεγχο για την ανισορροπία των κλάσεων [117-123] ενώ οι υπόλοιπες επιλύουν το συγκεκριμένο πρόβλημα χρησιμοποιώντας τεχνικές υπερδειγματοληψίας ή υποδειγματοληψίας [124-127]. Το κοινό στοιχείο είναι ότι σε κάθε μελέτη ένας κατηγοριοποιητής δίνει μεγαλύτερες ή μικρότερες τιμές από ότι θα δώσει ο αντίστοιχος κατηγοριοποιητής σε κάποια άλλη. Τέλος σημαντικό στοιχείο είναι

ότι η μελέτες που χρησιμοποιούν τεχνικές υπερδειγματοληψίας ή υποδειγματοληψίας βελτιώνεται η απόδοση των κατηγοριοποιητών.

Τέλος ένα σημαντικό στοιχείο είναι ότι τα δεδομένα των μελετών προέρχονται από κλινικές και νοσοκομεία της εκάστοτε χώρας στην οποία γίνεται η μελέτη. Αυτό είναι στην ουσία και ο περιορισμός της παρούσας μελέτης καθώς το σύνολο δεδομένων έχει ληφθεί από το διαδίκτυο και δεν προέρχεται από κάποιο νοσηλευτικό ίδρυμα του ελληνικού συστήματος υγείας. Ο παραπάνω περιορισμός αποτελεί και την μελλοντική επέκταση της έρευνας του παρόντος κεφαλαίου. Δηλαδή να δημιουργηθούν σύνολα δεδομένων τα οποία ανταποκρίνονται στην ελληνική πραγματικότητα.

9.4 Συμπεράσματα

Στο παρόν κεφάλαιο της διατριβής πραγματοποιήθηκε μία έρευνα στο τομέα της εξόρυξης γνώσης από ιατρικά δεδομένα που είχε σκοπό την ανάπτυξη διαφόρων μεθόδων πρόβλεψης για την δυνατότητα διάγνωσης των ασθενών με καρδιαγγειακά νοσήματα. Οι μέθοδοι πρόβλεψης που χρησιμοποιήθηκαν είναι: Λογιστική Παλινδρόμηση, Naive Bayes Classifier, Δένδρα αποφάσεων, Αλγόριθμος Κ κοντινότερων γειτόνων, Αλγόριθμος SVM (Support Vector Machine) και Random Forest.

Για να χρησιμοποιηθούν τα δεδομένα από τις μεθόδους εξόρυξης γνώσης έπρεπε πρώτα να υποστούν μια προεπεξεργασία, η οποία και απαιτούσε το περισσότερο χρόνο για την υλοποίηση της συγκεκριμένης έρευνας. Σύμφωνα με τη προεπεξεργασία που πραγματοποιήθηκε, δημιουργήθηκε ένα κατώφλι, με μία δυαδική μεταβλητή-κλάση η οποία έδειχνε την ύπαρξη ή τη μη ύπαρξη ασθένειας.

Σύμφωνα με τη βιβλιογραφία αλλά και με τα αποτελέσματα των μοντέλων που εφαρμόστηκαν για το σύνολο δεδομένων, η απάντηση στην ερώτηση ποιος αλγόριθμος κατηγοριοποίησης είναι καλύτερος είναι κανένας [128]. Η καταλληλότητα του εκάστοτε αλγορίθμου εξαρτάται από μια σειρά από παράγοντες όπως [129] [130]:

- ✓ Το μέγεθος του συνόλου δεδομένων.
- ✓ Το πρόβλημα της ανισορροπίας των κλάσεων.

- ✓ Η ύπαρξη και η απαλοιφή coefficients μεταξύ των χαρακτηριστικών.
- ✓ Η δυνατότητα βελτίωσης του εκάστοτε μοντέλου.

Τα αποτελέσματα του παρόντος κεφαλαίου δείχνουν ότι μπορούν να αναπτυχθούν μοντέλα πρόγνωσης μέσα από τις τεχνικές της εξόρυξης δεδομένων τα οποία μπορούν να προβλέψουν με ακρίβεια περιπτώσεων ασθενών με καρδιαγγειακά νοσήματα. Παρόλα αυτά, είναι αναγκαίο ύστερα από την εφαρμογή της εξόρυξης γνώσης και την εξαγωγή των αποτελεσμάτων, τα αποτελέσματα να αξιολογηθούν από ειδικούς στο τομέα της ιατρικής για να υπάρξουν ασφαλέστερα συμπεράσματα.

Κεφάλαιο 10 – Γενικά Συμπεράσματα

10.1 Συμπεράσματα

Ένα ευρύ φάσμα πηγών όπως μητρώα ασθενειών, διοικητικά δεδομένα για την υγεία, ηλεκτρονικοί φάκελοι υγείας, δεδομένα συνταγών φαρμακείων, αιτούμενα ποσά ιατρικών πράξεων, δεδομένα από εφαρμογές κινητών τηλεφώνων κ.ά. συνιστούν πηγές των μεγάλων δεδομένων περιέχοντας επιδημιολογικά δεδομένα, κλινικά δεδομένα, μοτίβα συμπεριφορών υγείας, έκθεση σε φάρμακα και κλινικές εκβάσεις που είναι πιθανόν χρήσιμες στην ιατρική, στην επιδημιολογία και στη βελτίωση της υγείας γενικότερα. Την ίδια στιγμή όμως, τα δεδομένα αυτά είναι συχνά πολλά, ημιτελή, ετερογενή, από πηγές αμφίβολης αξιοπιστίας και αλλάζουν με μεγάλη ταχύτητα ώστε να μην είναι εύκολο να αποθηκευτούν, να υποστούν επεξεργασία και να μετουσιωθούν σε κάτι που να δίνει αξία. Ηθικά και νομικά διλήμματα ως προς την ιδιωτικότητα και την πρόσβαση έρχονται σε σύγκρουση με την άμεση ανάγκη να σχεδιάσουμε, να μελετήσουμε και να αποκτήσουμε πρόσβαση σε νέες καινοτόμες θεραπείες που θα δώσουν λύση σε ανίατες, απειλητικές για τη ζωή ασθένειες.

Για αυτό το λόγο διενεργήθηκε μία εμπειρική μελέτη σχετικά με τη καταγραφή της γνώσης των επιστημόνων της Πληροφορικής Υγείας σχετικά με την τεχνολογία των Big Data. Τα αποτελέσματα της παραπάνω έρευνας δείχνουν ότι ένα μεγάλο ποσοστό του δείγματος ανέφερε ότι γνωρίζει την τεχνολογία των Big Data. Αξιοσημείωτο είναι το γεγονός ότι οι περισσότεροι επαγγελματίες αναφέρουν ότι έχουν ακούσει για τη συγκεκριμένη τεχνολογία και στο χώρο εργασίας. Επιπλέον τα αποτελέσματα της έρευνας έδειξαν ότι επιστήμονες της Πληροφορικής Υγείας συμφωνούν με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας είναι χρήσιμη. Ακόμη συμφωνούν με την άποψη ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα μπορούσε να αυξήσει την αποτελεσματικότητα των παρεχόμενων Υπηρεσιών Υγείας.

Επίσης διενεργήθηκε σε δεύτερη φάση μία έρευνα μελέτη σχετικά με τη καταγραφή της γνώσης των επαγγελματιών υγείας σχετικά με την τεχνολογία των Big Data.Τα αποτελέσματα της έρευνας έδειξαν οι η πλειοψηφία των

ιατρών και των νοσηλευτών δεν έχουν επαρκείς γνώσεις σχετικά με τη συγκεκριμένη τεχνολογία. Επιπλέον, είναι αξιοσημείωτο ότι στα αρκετοί ερωτηθέντες στα σχόλια του ερωτηματολογίου αναφέρουν ότι μία τέτοια τεχνολογία είναι δύσκολο να εφαρμοστεί στα Νοσοκομεία της χώρας. Επίσης, κάποιοι ερωτηθέντες σχολίασαν ότι δεν γνωρίζουν για την τεχνολογία των Big Data ωστόσο τους ακούγεται ως κάτι θετικό. Από τα αποτελέσματα της έρευνας δεν μπορούμε να καταλήξουμε σε ασφαλή συμπεράσματα σχετικά με την άποψη των ερωτηθέντων για τη χρησιμότητα των Big Data και τη συμβολή τους και στη βελτίωση των παρεχόμενων υπηρεσιών υγείας.

Θέλοντας να τονίσουμε τη συμβολή της τεχνολογίας των Big Data στον τομέα της Υγείας προχωρήσαμε στην ανάπτυξη μεθόδων πρόβλεψης για την δυνατότητα διάγνωσης των ασθενών με καρδιαγγειακά νοσήματα. Οι μέθοδοι πρόβλεψης που χρησιμοποιήθηκαν είναι: Λογιστική Παλινδρόμηση, Naive Bayes Classifier, Δένδρα αποφάσεων, Αλγόριθμος Κ κοντινότερων γειτόνων, Αλγόριθμος SVM (Support Vector Machine) και Random Forest. η ανάπτυξη περιλάμβανε όλα τα στάδια προεπεξεργασίας των δεδομένων. Επίσης χρησιμοποιήσαμε συγκεκριμένες μετρικές για τη μέτρηση της απόδοσης των κατηγοριοποιητών. Τέλος προχωρήσαμε και σε βελτιώσεις της απόδοσης των κατηγοριοποιητών χρησιμοποιώντας διασταυρωτική επαλήθευση με την μέθοδο cross-validation ενώ επιλύσαμε και το πρόβλημα της ανισορροπίας των κλάσεων χρησιμοποιώντας τη μέθοδο SMOTE.

10.3 Συμβολή της Διατριβής

Η παρούσα Διδακτορική Διατριβή επιχειρεί να διερευνήσει τις γνώσεις των επιστημόνων της Πληροφορικής υγείας και των επαγγελματιών υγείας σχετικά με τη τεχνολογία των Big Data. Ακόμη προσπαθεί να εκτιμήσει το τεχνολογικό υπόβαθρο των επιστημόνων της Πληροφορικής Υγείας και των επαγγελματιών υγείας σχετικά με τις γνώσεις τους για τη συγκεκριμένη τεχνολογία. Η έρευνα που πραγματοποιήθηκε, θα μπορούσε να χαρακτηριστεί ως καινοτόμα καθώς δεν έχουν γίνει παρόμοιες μελέτες στο χώρος της υγείας στην Ελλάδα. Επίσης, μέσα από τη χρήση της μηχανικής μάθησης προσπαθεί να δείξει πως η τεχνολογία των Big Data μπορεί να βοηθήσει στην έγκαιρη πρόγνωση ασθενειών με τη μέγιστη δυνατή ακρίβεια. Ακόμη δείχνει τη λειτουργία και τη συνεισφορά των κατηγοριοποιητών που

πρόβλεψης. Επιπλέον επιδιώκει, μέσα από την επεξεργασία των δεδομένων, αλλά και τη χρήση τεχνικών που επιλύουν προβλήματα της ανισορροπίας των κλάσεων και βοηθούν στη βελτίωση της απόδοσης των κατηγοριοποιητών, να τονίσει την συμβολή των επιστημόνων της «Πληροφορικής της Υγείας» στη προαγωγή της υγείας του πληθυσμού, την υποστήριξη των επαγγελματιών υγείας και τη βελτίωση των παρεχόμενων υπηρεσιών υγείας με τη χρήση καινοτόμων τεχνολογιών.

10.4 Περιορισμοί της Διατριβής

Ένας από τους περιορισμούς της έρευνας είναι το δείγμα καθώς επιδίωξη ήταν να απαντήσουν περισσότερα άτομα. Ωστόσο διαπιστώθηκε μία σχετική απροθυμία, έλλειψη χρόνου ή μη εξοικείωση της πλειοψηφίας των επαγγελματιών υγείας στη χρήση Η/Υ ώστε να συμπληρώσουν το ηλεκτρονικό ερωτηματολόγιο. Επίσης, για λόγους προσωπικών δεδομένων δεν έγινε καταγραφή των νοσοκομειακών ιδρυμάτων που εργάζονται οι επαγγελματίες υγείας. Παράλληλα η έρευνα έδειξε ότι ενώ η πλειονότητα των ιατρών και νοσηλευτών δεν γνωρίζει για την τεχνολογία των Big Data και θεωρεί ότι δεν μπορεί να εφαρμοστεί στο ΕΣΥ, πιστεύει ότι είναι χρήσιμη και θα συμβάλλει στη βελτίωση των παρεχόμενων υπηρεσιών υγείας. Έτσι δεν μπορούν να βγουν ασφαλή συμπεράσματα για τις απόψεις των επαγγελματιών υγείας σχετικά με τα Big Data. Τέλος ένας ακόμη περιορισμός είναι ότι τα datasets που χρησιμοποιήθηκαν στους κατηγοριοποιητές για την εξόρυξη γνώσης, δεν προέρχονται από δεδομένα ασθενών του Ελληνικού Συστήματος Υγείας άλλα από ελεύθερα αποθετήρια στο διαδίκτυο.

10.5 Μελλοντικές επεκτάσεις της έρευνας

Η παρούσα έρευνα δύναται να επεκταθεί χρησιμοποιώντας μεγαλύτερο δείγμα. Αυτό θα μπορούσε να γίνει εξασφαλίζοντας άδειες από τις διοικήσεις των νοσοκομείων ώστε να συμπληρωθούν τα ερωτηματολόγια στο χώρο εργασίας των επαγγελματιών υγείας. Επίσης η έρευνα θα μπορούσε να διερευνήσει κατά πόσο τα νοσοκομεία έχουν την κατάλληλη υποδομή σε υλικό και λογισμικό ώστε να υποστηρίξουν την τεχνολογία των Big Data. Επιπλέον θα μπορούσε να διερευνηθούν οι ενέργειες που γίνονται από τους φορείς του

ΕΣΥ για την εισαγωγή και την υποστήριξη της τεχνολογίας των Big Data στο χώρο της υγείας.

Όσον αφορά το κομμάτι της εξόρυξης γνώσης, θα μπορούσε να μετρηθεί η απόδοση και η προβλεπτική ικανότητα των κατηγοριοποιητών που εφαρμόστηκαν κάνοντας χρήση δεδομένων ασθενών από κάποιο νοσοκομειακό ίδρυμα της χώρας. Επίσης μία επέκταση της παρούσας διατριβής θα μπορούσε να είναι η χρήση τεχνικών συσχέτισης και συσταδιοποίησης για την ανακάλυψη προτύπων.

Περίληψη

Η μεγαλύτερη πρόκληση των σύγχρονων υπολογιστικών συστημάτων είναι αναμφισβήτητα η αποδοτική αποθήκευση και ανάκτηση πολύ μεγάλου όγκου δεδομένων. Η ανάγκη αυτή έκανε την εμφάνισή της τα τελευταία χρόνια λόγω της έκρηξης δεδομένων που παρατηρείται στο διαδίκτυο και αποκτά ολοένα και μεγαλύτερη σημασία λόγω του πολύ μεγάλου εύρους πληροφοριών που μπορούμε να αντλήσουμε. Ο τομέας της υγειονομικής περίθαλψης και των ιατρικών δεδομένων είναι συνεχώς και ταχέως εξελισσόμενος. Η αξιοποίηση των Big Data στο χώρο της υγείας προσφέρει πολύτιμη πληροφόρηση καθώς παρουσιάζουν απεριόριστες δυνατότητες για αποτελεσματική αποθήκευση, επεξεργασία, sql queries και ανάλυση ιατρικών δεδομένων.

Σκοπός της παρούσας διατριβής είναι η μελέτη τεχνικών εξόρυξης γνώσης για δεδομένα μεγάλου όγκου, που αφορούν το πεδίο της Υγείας. Παράλληλα σκοπός της έρευνας είναι η μελέτη στατιστικών και υπολογιστικών αλγορίθμων ανάλυσης μεγάλου όγκου δεδομένων υγείας που έχουν ως αποτέλεσμα την παραγωγή νέας γνώσης καθώς και την εξαγωγή στατιστικά σημαντικής πληροφορίας για τους επαγγελματίες υγείας. Τέλος, η παρούσα διατριβή διερευνά τις γνώσεις των επιστημόνων της Πληροφορικής Υγείας και των επαγγελματιών υγείας σχετικά με τα Big Data.

Στην παρούσα διδακτορική διατριβή έγινε βιβλιογραφική ανασκόπηση της έννοιας των Big Data. Η ανασκόπηση αυτή περιλαμβάνει τον ορισμό των Big Data ,τα χαρακτηριστικά τους, τα πλεονεκτήματα και τα μειονεκτήματά τους στο χώρο της υγείας. Στη συνέχεια γίνεται αναφορά στην υλοποίηση και στους μηχανισμούς αποθήκευσης των Big Data. Επιπλέον γίνεται αναφορά στα συστήματα ανάλυσης και επεξεργασίας μεγάλου όγκου δεδομένων, στις γλώσσες προγραμματισμού για Big Data, στην εξόρυξη γνώσης δεδομένων στο χώρο της υγείας. Ακόμη γίνεται αναφορά στη χρήση των Big Data στην Ευρώπη και στον κόσμο. Τέλος παρουσιάζονται οι βασικές αρχές του GDPR καθώς και το πώς σχετίζεται με τα Big Data στο χώρο της υγείας. Επίσης διεξήχθησαν δύο εμπειρικές μελέτες.

Η πρώτη μελέτη είχε σαν στόχο την καταγραφή της άποψης των επιστημόνων της Πληροφορικής Υγείας σχετικά με την τεχνολογία των Big Data. Η συλλογή των δεδομένων έγινε με χρήση ερωτηματολογίου. Η

στατιστική ανάλυση έδειξε τη θετική ανταπόκριση του δείγματος σχετικά με την τεχνολογία των Big Data.

Η δεύτερη μελέτη είχε σαν στόχο την καταγραφή της άποψης των Επαγγελματιών Υγείας σχετικά με την τεχνολογία των Big Data. Η συλλογή των δεδομένων έγινε με χρήση ερωτηματολογίου. Η στατιστική ανάλυση δεν έδωσε επαρκείς απαντήσεις καθώς οι ερωτηθέντες έδειξαν θετική στάση απέναντι στα Big Data ενώ απάντησαν ότι δεν γνωρίζουν πολλά για τη συγκεκριμένη τεχνολογία.

Το τελευταίο κομμάτι της διατριβής περιλαμβάνει την ανάπτυξη μεθόδων πρόβλεψης για την δυνατότητα διάγνωσης των ασθενών με καρδιαγγειακά νοσήματα. Οι μέθοδοι πρόβλεψης που χρησιμοποιήθηκαν είναι: Λογιστική Παλινδρόμηση, Naive Bayes Classifier, Δένδρα αποφάσεων, Αλγόριθμος Κ κοντινότερων γειτόνων, Αλγόριθμος SVM (Support Vector Machine) και Random Forest. Η ανάπτυξη περιλάμβανε όλα τα στάδια προεπεξεργασίας των δεδομένων ενώ χρησιμοποιήθηκαν συγκεκριμένες μετρικές για τη μέτρηση της απόδοσης των κατηγοριοποιητών. Τέλος έγιναν βελτιώσεις της απόδοσης των κατηγοριοποιητών χρησιμοποιώντας διασταυρωτική επαλήθευση με την μέθοδο cross-validation ενώ επιλύθηκε και το πρόβλημα της ανισορροπίας των κλάσεων χρησιμοποιώντας τη μέθοδο SMOTE.

Λέξεις κλειδιά: Εμπειρική Μελέτη, Big Data, Επαγγελματίες Υγείας, Εξόρυξη γνώσης, Τεχνικές Κατηγοριοποίησης

ABSTRACT

The biggest challenge of modern computing is undoubtedly the efficient storage and retrieval of very large amounts of data. This need made its appearance in recent years due to the explosion of data that is observed on the Internet and is becoming more important because of the very wide range of information we can learn. The fields of healthcare and medical data are constantly evolving. The use of BigData in healthcare offers valuable information as they are capable for effective storage, data processing, sql queries and health data analysis.

The aim of the current PhD thesis is the study and the implementation of data mining techniques using Big Data, in the field of healthcare. Also, the current PhD thesis is to build and compare classification techniques for cardiovascular diseases. Finally, the research aim of this study is to investigate the perceptions of the Health Informatics Scientists and Health Professionals about the Big Data Technology in Healthcare.

On this Doctoral Thesis, a literature review was conducted in order to record the current status of Big Data in Healthcare. This review records the definition of Big Data, the attributes, the advantages and disadvantages of Big Data in Healthcare. Furthermore the review presents the implementation and storage mechanisms of Big Data. Also, the review records software for analysis and processing Big Data, Big Data programming languages. The use of Big Data in Europe and in the world is also presented. Finally, the basic principles of the GDPR are presented and it's correlation to Big Data in the field of Healthcare. Two empirical studies were also conducted.

The research aim of the first study is to investigate the perceptions of Health Information scientists about the Big Data Technology in Healthcare. A questionnaire was developed in order to measure the aforementioned dimensions. The current study reveals a rather positive attitude toward the usage of Big Data in the Healthcare domain.

The research aim of the second study is to investigate the perceptions of Health Professionals about the Big Data Technology in Healthcare. A questionnaire was developed in order to measure the aforementioned dimensions. The current study did not give sufficient results as the

respondents showed a positive attitude towards Big Data although they do not know much about this technology.

The last part of the thesis refers to the development of data mining techniques for the prediction of cardiovascular diseases. The methods used are: Logistic Regression, Naive Bayes Classifier, Decision Trees, K Nearest Neighbors Algorithm, SVM (Support Vector Machine) Algorithm and Random Forest. The development included all stages of data preprocessing while specific metrics were used to measure the performance of the classifiers. Finally, the performance of the classifiers was improved using cross-validation, while the problem of class imbalance was solved using the SMOTE method.

Keywords: Empirical Study, Big Data, Health Professionals, Data Mining, Classification Techniques.

Βιβλιογραφία

1. Raghupathi W. Data Mining in Health Care. In: Kudyba S, editor. Healthcare Informatics: Improving Efficiency and Productivity. Springer Science & Business Media, Berlin, 2010: pp. 211–223.
2. Burghard C. Big Data and Analytics Key to Accountable Care Success; IDC Health Insights: Framingham, MA, USA, 2012; pp. 1–9.
3. Dembosky, A. Data prescription for better healthcare. Financial Times 2012; 34: pp. 2-3.
4. Feldman B, Martin EM, Skotnes T. Big Data in Healthcare Hype and Hope [Internet]. Dr. Bonnie 360 degree (Business Development for Digital Health), 2012 [updated 2012 Oct 19; cited 2013 Jan 8]; Available from: http://www.riss.kr/link?id=A99883549.
5. Fernandes L, O'Connor M, Weaver V. Big data, bigger outcomes. Journal of AHIMA 2012; 83(10): pp. 38-42.
6. IHTT. Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry [Internet]. 2013 [update 2013 Nov 10; cited 2013 Feb 10]; Available from: http://ihealthtran.com/wordpress/2013/03/iht%C2%B2-releases-bigdata research - reportdownload-today/
7. Canada Inforoute. Big Data Analytics in health, White Paper, Full Report, April 2013.
8. Bellazzi R. Big Data and Biomedical Informatics: A Challenging Opportunity. IMIA Yearbook of Medical Informatics, Thieme, New York, 2014: pp. 8-13.
9. Gantz J, Reinsel D. Extracting value from chaos, IDC iView 2011; 24(3): pp. 1–12.
10. Blaise C. A High-Throughput–Compatible 3D Microtissue Co-Culture System for Phenotypic RNA Screening Applications. Journal of the American Society for Inform. Science & Technology 2013; 63(3): pp. 435–436.
11. LaValle S, Lesser E, Shockley R, Hopkins M, Kruschwitz N. Big data, analytics and the path from insights to value. MITS loan Management Review 2011; 52(2): pp. 20-32.
12. Groves P, Kayyali B, Knon D, Van Kuiken S. The big data revolution in healthcare. McKinsey Quarterly 2012; 2(3): pp. 6-9.
13. Chen H, Chiang R, Storey V. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly 2012; 36(4): pp. 1165-1188.
14. Agrawal D, Das S, El Abbad A. Big data and cloud computing: current state and future opportunities. PVLDB 2011; 3(2): pp. 530-533.
15. Cai L, Zhu Y. The Challenges of data quality and data quality assessment in the Big Data Era,. Data Science Journal 2015; 14(1), pp. 2-4.

16. Mikalef P, Pappas IO, Krogstie J, Giannakos M. Big data analytics capabilities: a systematic literature review and research agenda. Journal of Information Systems and e-Business Management 2018; 16(3): pp. 547-578.
17. DeWitt D, Gray J. Parallel database systems: the future of high performance database systems. Commun ACM 1992; 35(6): pp. 85–98.
18. Tonebraker S. SQL databases vs. NoSQL databases. Communications of the ACM 2010; 53 (4): pp. 10 -11.
19. Lee J, Kao HA, Yang S. Service innovation and smart analytics for industry 4.0 and big data environment. Procedia Cirp Journal 2014; 6(4): pp. 3-8.
20. Usha D, Aslin J. A Survey of Big Data Processing in Perspective of Hadoop and Mapreduce. International Journal of Current Engineering and Technology 2014; 4(2): pp.2-4.
21. Jeffrey D, Sanjay DG. MapReduce: Simplified Data Processing on Large Clusters. COMMUN-ACM 2008; 51(4): pp.107-113.
22. White T. Hadoop: The Definitive Guide. 4th ed. Sebastopol: O'Reilly Media; 2009: pp.24-26.
23. Harshawardhan S, Devendra P. A Review Paper on Big Data and Hadoop in International Journal of Scientific and Research Publications 2014; 4(10): pp1-3.
24. Mridul M, Khajuria A, Dutta S, Kumar N. Analysis of Big Data using Apache Hadoop and Map Reduce. International Journal of Advance Research in Computer Science and Software Engineering 2014; 4(5): pp. 4-6.
25. Yao Y, Wang J, Sheng B, Lin J, Mi N. HaSTE: Hadoop YARN Scheduling Based on Task Dependency and Resource-Demand. Proceedings of IEEE 7th International Conference on Cloud Computing; 2014 Nov 30 – Dec 3; Vancouver, Canada.
26. Isard M, Budiu M, Yu Y, Birrell A, Fetterly D. Dryad: Distributed data-parallel programs from sequential building blocks. Proceedings of EuroSys; 2007 Aug 18-24; Brussels, Belgium.
27. Watson H. Update Tutorial: Big Data Analytics: Concepts, Technology, and Applications. Communications of the Association for Information Systems 2015; 44(2): pp. 21.
28. Watson H. Big data analytics: technologies, and applications. Communications of the Association for Information Systems CAIS 2012; 34(1): pp. 65
29. Suthaharan S. Big data classification: Problems and challenges in network intrusion prediction with machine learning. Performance Evaluation Review 2014; 41(4): pp. 70-73.
30. De Mauro A, Greco M, Grimaldi M. What is big data? A consensual definition and a review of key research topics, AIP 2015; 1644(1): pp.7-35

31. Waller A, Fawcett S E. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. Journal of Business Logistics 2013; 34(2): pp. 77-84.
32. Vatika S, Meenu D. SQL and NoSQL Databases, International Journal of Advanced Research in Computer Science and Software Engineering 2012; 2(8): pp. 20-27.
33. Shih M, Chang M. Design and Analysis of High Performance Crypt-NoSQL. Proceedings of IEEE Conference on Dependable and Secure Computing; 2017 Aug 7-10; Taipei, Taiwan: pp. 52–59.
34. Kalid S, Syed A, Mohammad A, Halgamuge M. Big-Data NoSQL Databases: A Comparison and Analysis of Big-Table. International Journal of Recent Technology and Engineering 2017; 7(6): pp. 89–93.
35. Child H, Geveci B, Schroeder W, Meredith J, Moreland K, Sewell C, Kuhlen T, Bethel EW. Research challenges for visualization software. Procedia Computer Science 2013; 46(5): pp. 34-42.
36. Zhong R, Newman S, Huang G, Lan S. Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. Journal of Computers & Industrial Engineering 2016; 101(4): pp. 572-591.
37. Bach B, Riche N, Carpendale S, Pfister H. The Emerging Genre of Data Comics. IEEE Computer Graphics and Applications 2017; 38(3): pp. 6-13.
38. Mackinlay J. Automating the design of graphical presentations of relational information. ACM Transactions on Graphics 1986; 5(2): pp. 110-141.
39. Friendly M, Denis D. The early origins and development of the scatter plot. Journal of the History of the Behavioral Sciences 2005; 41(2): pp. 103–130.
40. Klein, J. Statistical Visions in Time: A History of Time Series Analysis, Cambridge University Press 1997; 2(6): pp. 1662-1938.
41. Wattenberg M, Fisher D. Analyzing perceptual organization in information graphics. Information Visualization 2004; 3(2): pp. 123–133.
42. Havlak P. Nesting of reducible and irreducible loops. ACM Transactions on Programming Languages and Systems 1997; 19(4): pp. 557–567.
43. Tseng S, Wang K, Lee C. A pre-processing method to deal with missing values by integrating clustering and regression techniques. Applied Artificial Intelligence 2003; 17(5-6): pp. 535–544.
44. Shafique U, Qaiser H. Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research 2014; 12(1): pp. 217–222.
45. Li Y, Zhang C, Zhang S. Cooperative strategy for Web data mining and cleaning. Applied Artificial Intelligence 2003; 17(6): pp. 443–460.
46. Atish P, Huimin Z. Incorporating domain knowledge into data mining classifiers: An application in indirect lending. Decision Support Systems 2008; 46(1): pp. 287–299.

47. Foster D, Stine R. Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy. Journal of the American Statistical Association 2004; 99(466): pp. 303-313.
48. Han J, Kamber M. Data mining: concepts and techniques. 3rd ed. Waltham: Morgan Kaufman Publishers; 2012.
49. Bailey-Kellog C, Ramakrishnan N, Marathe M. Spatial Data Mining to Support Pandemic Preparedness. ACM SIGKDD Explorations 2018; 1(8): pp. 80-82.
50. Wong W, Moore A, Cooper G, Wagner M. What's Strange About Recent Events (WSARE): An Algorithm for the Early Detection of Disease Outbreaks. Journal of Machine Learning Research 2005; 6(4): pp. 1961-199.
51. Tomar D, Agarwal S. A survey on data mining approaches for healthcare. International Journal of Bio-Science and Bio-Technology 2013; 5(5): pp. 241-266.
52. Kob H, Tan G. Data mining applications in healthcare. Journal of Healthcare Information Management 2005; 19(2): pp. 64-72.
53. Farabet C, Couprie C, Najman L, LeCun Y. Learning hierarchical features for scene labeling. IEEE Transactions on Pattern Analysis and Machine Intelligence 2013; 35(8):pp.1915–1929.
54. Berkhin P. A survey of clustering data mining techniques. Grouping multidimensional data 2006; 10(3): pp. 25-71.
55. Sohn S, Kim J. Decision tree-based technology credit scoring for start-up firms: Korean case. Expert Systems with Applications 2012; 39(4): pp. 4007-4012.
56. Eaton D, Murphy K. Exact Bayesian structure learning from uncertain interventions. Artificial Intelligence and Statistics 2007; 2(1):pp. 107-114.
57. Hung J, Zhang L. Revealing online learning behaviors and activity patterns and making predictions with data mining techniques in online teaching. MERLOT Journal of Online Learning and Teaching 2008; 4(4): pp. 426–437.
58. Zhu K. Research based on data mining of an early warning technology for predicting engineering students' performance. World Transactions on Engineering and Technology Education 2014; 12(3):pp. 572–575.
59. Zhu M, Liu W, Yang, Y. Construction algorithm of MPD-JT for Bayesian networks based on full conditional independence. Systems Engineering and Electronics 2010; 32(6): pp. 8-11.
60. Wang C, Li R, Fan M. Mining Positively Correlated Frequent Itemsets. Computer Applications 2007; 27(2): pp. 108-109.
61. Prasad P, Malik L. Using Association Rule Mining for Extracting Product Sales Patterns in Retail Store Transactions. International Journal on Computer Science and Engineering 2011; 3(5): pp. 2177-2182.
62. Christodoulakis A, Karanikas H, Billiris A, Thireos E, Pelekis N. Big data in health care Assessment of the performance of Greek NHS hospitals

using key performance and clinical workload indicators. Archives of Hellenic Medicine 2016; 33(4): pp. 489-497.
63. Turner M. European perspectives on the use of Big Data in health care research. Data and Information Management 2017; 2(3): pp. 1-24.
64. Moseley E, Hsu D, Stone D, Celi L. Beyond Open Big Data: Addressing Unreliable Research. Journal of Medical Internet Research 2014; 16(11): pp. 259.
65. Wasserman L. All of Statistics: A Concise Course in Statistical Inference. 2nd ed. New York: Springer; 2004: pp. 209-226.
66. Boshra I, Bahrami MH. Prediction and Diagnosis of Heart Disease by Data Mining Techniques. Journal of Multidisciplinary Engineering Science and Technology 2015; 2(2): pp. 4-10.
67. Komarek P, Moore A. Making logistic regression a core data mining tool. Proceedings of Fifth IEEE International Conference on Data Mining; 2005 Nov 27-30; Houston, USA.
68. Ng AY, Jordan, MI. On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. Neural Information Processing Systems 2001; 13(2): pp. 3-24.
69. Faisal KM, Mofizur RC, Alamgir H, Kesavm D. Enhanced classification accuracy on naive Bayes data mining models. International Journal of Computer Applications 2011; 28(3): pp. 9–16.
70. Hailu TG. Comparing Data Mining Techniques in HIV Testing Prediction. Addis Ababa, Ethiopia. Intelligent Information Management 2015; 7(1): pp. 153-180.
71. Mahendra T. An Empirical Study of Applications of Data Mining Techniques for Predicting Student Performance in Higher Education. International Journal of Computer Science and Mobile Computing 2013; 2 (2): pp.53 –57.
72. Domeniconi C, Peng J. Gunopulos D. Locally adaptive metric nearest-neighbor classification. Transactions on Pattern Analysis and Machine Intelligence 2002; 24 (9): pp.1281–1285.
73. Hall P, Park BU, Samworth RJ. Choice of neighbor order in nearest-neighbor classification. The Annals of Statistics 2008; 36(5): pp. 2135-2152.
74. Gil-Garcia R, Pons-Porrata A. A New Nearest Neighbor Rule for Text Categorization. Lecture Notes in Computer Science 4225, New York: Springer; 2006: pp. 814–823.
75. Enas G, Choi SC. Choice the smoothing parameter and efficiency of K-Nearest Neighbor classification. Comp and Maths with Apps 1986; 12(2): pp. 235-244.
76. Audibert J, Tsybakov A. Fast learning rates for plug-in classifiers under the margin condition. Annals of Statistics 2007; 35(2); pp. 608–633.

77. Bauer ME, Burk TE, Ek AR, et al. Satellite Inventory of Minnesota's Forest Resources. Photogrammetric Engineering and Remote Sensing 1994; 60(3): pp. 287–298.
78. Mythili T. A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL). International Journal of Computer Applications 2014; 68(16): pp.11-15.
79. Chaudhuri A, De K, Chatterjee D. A Comparative Study of Kernels for the Multi-class Support Vector Machine. IEEE Transactions of Neural Networks 2008; 19(12): pp. 3-7.
80. Kubica J, Goldenberg A, Komarek P, Moore A, Schneider J. A comparison of statistical and machine learning algorithms on the task of link completion. PNAS 2003; 99(18): pp. 8-14.
81. Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines. Cambridge: University Press; 2000.
82. Platt J. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. Cambridge: MIT Press; 2009: pp. 185- 208.
83. Tripoliti E, Fotiadis D, Manis G. Dynamic Construction of Random Forests: Evaluation using Biomedical Engineering Problems. Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine; 2010 Nov 3-5; Corfu, Greece.
84. Zhang H, Wang M. Search for the smallest Random Forest. Statistics and Its Interface 2009: 2(3): pp. 381-388.
85. Wu X, Chen Z. Toward dynamic ensemble: The BAGA Approach. Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications; 2005 Jan 6; Cairo, Egypt.
86. Almarabeh H, Amer E. A study of Data Mining Techniques Accuracy for Healthcare. International Journal of Computer Applications 2017; 168(3): pp 12-17.
87. Japkowicz N. Assessment metrics for imbalanced learning, in Imbalanced Learning: Foundations. Algorithms and Applications. 1st ed. New Jersey: Wiley IEEE Press; 2013.
88. Huang J, Ling C. Using AUC and accuracy in evaluating learning algorithms, IEEE Transactions on Knowledge Data Engineering 2005: 17(3): pp. 299-310.
89. Hellenic Ministry of Health. ESY Annual Report [Internet]. 2013 [updated 2013 Mar 10; cited 2013 May 15]; Available from: http://www.moh.gov.gr/articles/esynet/stoixeia-noshley-tikhskinhshs/ leesynet-2013/2253-ethsia-leitoyrgika-2013.
90. Christodoulakis A, Karanikas H, Billiris A, Thireos E, Pelekis N. Big data in health care Assessment of the performance of Greek NHS hospitals using key performance and clinical workload indicators. Archives of Hellenic Medicine 2016; 33(4):pp. 489-497.

91. Pastorino R, De Vito C, Migliara G. Benefits and challenges of Big Data in healthcare: an overview of the European initiatives. The European Journal of Public Health 2019; 29(3): pp. 23-27.
92. Auffray C, Balling R, Barroso I, et al. Making sense of big data in health research: Towards an EU action plan. Genome Medicine 2016; 71(8): pp.1-13.
93. Bates DW, Saria S, Ohno Machado L, et al. Big Data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff 2014; 33: pp.1123–1131.
94. Williams A, Harland L, Groth S, Chichester C, Willighagen E, Evelo C, Blomberg N, Ecker G, Goble C, Mons B. Open PHACTS: semantic interoperability for drug discovery. Drug Discovery Today 2012; 17 (21): pp.1188-1198.
95. McGregor C. Next Generation Neonatal Health Informatics with Artemis. IOS Press 2011; 10(2): pp. 115-119.
96. Ferrucci DA. Introduction to This is Watson. IBM Journal of Research and Development 2012; 56(3): pp. 2-8.
97. Sweeney L. K-anonymity: a model for protecting privacy. International Journal of Uncertain Fuzziness 2002; 10(5): pp.557–570.
98. Priyank J, Manasi G, Nilay K. Big data privacy: a technological perspective and review. Journal of Big Data 2016; 3(25):pp. 25-27.
99. Moldovan D, Pasca M, Harabagiu S, Surdeanu M. Performance issues and error analysis in an open-domain question answering system. ACM Transactions on Information Systems 2003; 21(2): pp.133–154.
100. Bhandari R, Hans V, Jyothi Ahuja N. Big Data Security: Challenges and Recommendations. International Journal of Computer Sciences and Engineering 2016; 4(1): pp 93-98.
101. Coventry L, Branley D. Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. Maturitas 2018; 2(3): pp. 48-52.
102. Kuner C, Cate F, Millard C, Svantesson J. The challenge of big data for data protection. International Data Privacy Law 2012; 2(2): pp.48-49.
103. Liu L, Lin J. Some special issues of network security monitoring on big data environments. Proceedings of the IEEE 11th International Conference on Dependable, Autonomic and Secure Computing; 2013 Dec 21-22; Chengdu, China.
104. Kaltheuner F, Bietti E. Data Is Power: Towards Additional Guidance on Profiling and Automated Decision-Making in the GDPR. Journal of Information Rights Policy and Practice 2018; 2(2): pp. 1-17.
105. Brodin M. A Framework for GDPR Compliance for Small- and Medium-Sized Enterprises. European Journal for Security Research 2019; 4(2): pp. 243-264.
106. Orel A, Bernik I. GDPR and Health Personal Data; Tricks and Traps of Compliance. Studies in health technology and informatics 2018; 1(8): pp: 155- 159.

107. Mathisen BM, Wienhofen L, Roman D. Empirical Big Data Research: A Systematic Literature Mapping. Computers and Society 2015; 4(2): pp.1–19.
108. Witten H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Biomedical Engineering Online 2006; 5(51): pp. 1-2.
109. Cleveland Heart Lab; 2021 [updated 201 November 6; cited 2021 July 6]. Available from: https://www.clevelandheartlab.com/chl-and-quest-forming-base-for-quests-national-cardiometabolic-center-of-excellence/
110. Faouzi J, Janati H. A Python Package for Time Series Classification. Journal of Machine Learning Research 2020; 21(46): pp. 1-6.
111. Hsieh FY, Bloch D, Larsen M. A simple method of sample size calculation for linear and logistic regression. Statistics in Medicine 1998; 17(14): pp.1623-1634.
112. Buchtala O, Klimek M, Sick B. Evolutionary optimization of radial basis function classifiers for data mining applications. IEEE Transactions on Cybernetics 2005; 35(5): pp. 928–947.
113. Elrahman SM, Abraham A. A Review of Class Imbalance Problem. Journal of Network and Innovative Computing 2013; 1(2): pp.332-340.
114. Garcia V, Sanchez JS, Mollineda, RA. On the effective of preprocessing methods when dealing with different levels of class imbalance. Knowledge-Based Systems 2012; 25(1). pp. 13-21.
115. Kerdprasop N, Kerdprasop K. On the Generation of Accurate Predictive Model from Highly Imbalanced Data with Heuristics and replication Technologies. International Journal of Bio-Science and Bio-Technology 2014; 4(1): pp. 49-64.
116. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 2002; 16(1): pp. 321–357.
117. Akella A, Kaushik V. Machine Learning Algorithms for Predicting Coronary Artery Disease: Efforts Toward an Open Source Solution. Heart Views 2017; 18(3): pp. 109–114.
118. Padmanabhan M, Yuan P, Chada G, Van Nguyen H. Physician-Friendly Machine Learning: A Case Study with Cardiovascular Disease Risk Prediction. Journal of Clinical Medicine 2019; 8(1050): pp.1-13.
119. Joloudari JH, Joloudari EH, Saadatfar H, Ghasemigol M. Coronary Artery Disease Diagnosis; Ranking the Significant Features Using a Random Trees Model. International Journal of Environmental Research and Public Health 2020; 17(731): pp.2-24.
120. Marimuthu M, Abinaya M. A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach. International Journal of Computer Applications 2018; 181(18): pp. 20-25.
121. Hippisley-Cox J. Predicting cardiovascular risk in England and Wales: Prospective derivation and validation of QRISK2. BMJ 2008; (2) 336: pp. 1475–1482.

122. El-Bialy R, Salamay MA, Karam O, Essam Khalifa M. Feature Analysis of Coronary Artery Heart Disease Data Sets. Procedia Computer Science 2015: 65(2): pp. 459-468.
123. Moturi S, Tirumala Rao S, Vemuru S. Predictive Analysis of Imbalanced Cardiovascular Disease Using SMOTE. International Journal of Advanced Science and Technology 2020;29(5): pp. 6301-6311.
124. Long N, Meesad P, Unger H. 'A highly accurate firefly based algorithm for heart disease prediction. Expert Systems with Applications: An International Journal 2015;42(21): pp. 8221–8231.
125. Jagtap A, Malewadkar P, Baswat O, Rambade H. Heart disease prediction using machine learning. International Journal of Research in Engineering Science and Management 2019; 2 (2): pp. 352-355.
126. Yahaya L, Oye N, Garba E. A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques. American Journal of Artificial Intelligence 2020; 4(1): pp. 20-29.
127. Park G, Kim Y. Model for predicting cardiovascular disease: Insights from a Korean cardiovascular risk model. Pulse 2015; 3(2): pp. 153–157.
128. Barot V, Chauhan SS, Patel B. Feature Selection for Modeling Intrusion Detection. Computer Network and Information Security 2016:1(1): pp 35-42.
129. Kiang MY. A comparative assessment of classification methods. Decision Support Systems 2003; 35(2): pp. 441-454
130. Neocleous C, Schizas C. Artificial Neural Network Learning: A Comparative Review. Proceedings of Second Hellenic Conference on AI; 2002 April 11-12; Thessaloniki, Greece.

ΠΑΡΑΡΤΗΜΑ Α –Πίνακες στατιστικής ανάλυσης του Κεφαλαίου 7

Descriptive Statistics						
	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Ηλικία	65	18	63	28,69	9,857	97,154
Valid N (listwise)	65					

Πίνακας 1: Περιγραφικά στοιχεία για την ηλικία του δείγματος

Item Statistics			
	Mean	Std. Deviation	N
Q1	6,28	1,165	67
Q2	6,15	1,246	67
Q3	5,99	1,237	67
Q4	5,88	1,008	67
Q5	5,81	1,373	67

Πίνακας 2: Αποτελέσματα ερωτήσεων κλίμακας Likert

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,814	,812	5

Πίνακας 3: Αποτελέσματα ελέγχου αξιοπιστίας

Tests of Normality						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Ηλικία	,477	66	,000	,125	66	,000
ScoreQ1	,528	66	,000	,114	66	,000
ScoreQ2	,527	66	,000	,115	66	,000
ScoreQ3	,526	66	,000	,117	66	,000
ScoreQ4	,525	66	,000	,115	66	,000
ScoreQ5	,524	66	,000	,119	66	,000

Πίνακας 4: Αποτελέσματα ελέγχου κανονικότητας για τις ποσοτικές μεταβλητές

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Ηλικία is the same across categories of Γνωρίζετε τι είναι τα Big Data;.	Independent-Samples Mann-Whitney U Test	,066	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 5: Αποτελέσματα συσχέτισης Ηλικίας-Γνώσης Big Data

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,384[a]	1	,536		
Continuity Correction[b]	,072	1	,789		
Likelihood Ratio	,395	1	,530		
Fisher's Exact Test				,729	,402
Linear-by-Linear Association	,378	1	,539		
N of Valid Cases	67				
a. 1 cells (25,0%) have expected count less than 5. The minimum expected count is 3,88.					
b. Computed only for a 2x2 table					

Πίνακας 6: Αποτελέσματα συσχέτισης Φύλου-Γνώσης Big Data

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,271[a]	1	,603		
Continuity Correction[b]	,068	1	,795		
Likelihood Ratio	,272	1	,602		
Fisher's Exact Test				,795	,399
Linear-by-Linear Association	,267	1	,606		
N of Valid Cases	65				
a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 10,00.					

Πίνακας 7: Αποτελέσματα συσχέτισης Φύλου - Big Data για Δομημένα Δεδομένα

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)

Pearson Chi-Square	5,647[a]	1	,017		
Continuity Correction[b]	4,383	1	,036		
Likelihood Ratio	6,137	1	,013		
Fisher's Exact Test				,024	,016
Linear-by-Linear Association	5,560	1	,018		
N of Valid Cases	65				

Πίνακας 8: Αποτελέσματα συσχέτισης Φύλου - Big Data για Αδόμητα Δεδομένα

Directional Measures						
			Value	Asymptotic Standard Error[a]	Approximate T	Approximate Significance
Nominal by Nominal	Lambda	Symmetric	,012	,012	.[b]	.[b]
		Φύλο Dependent	,012	,012	.[b]	.[b]
		Αδόμητα δεδομένα (unstructured data) - ελεύθερο κείμενο. Dependent	,012	,012	.[b]	.[b]
	Goodman and Kruskal tau	Φύλο Dependent	,087	,061		,018[c]
		Αδόμητα δεδομένα (unstructured data) - ελεύθερο κείμενο. Dependent	,087	,061		,018[c]

Πίνακας 9: Μέτρηση συσχέτισης Φύλου - Big Data για Αδόμητα Δεδομένα

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2,677[a]	1	,102		
Continuity Correction[b]	1,906	1	,167		
Likelihood Ratio	2,723	1	,099		
Fisher's Exact Test				,129	,083
Linear-by-Linear Association	2,636	1	,104		
N of Valid Cases	65				
a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 11,20.					
b. Computed only for a 2x2 table					

Πίνακας 10: Αποτελέσματα συσχέτισης Φύλου - Big Data για Ημιδομημένα Δεδομένα

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ScoreQ1 is the same across categories of Φύλο.	Independent-Samples Mann-Whitney U Test	,271	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 11: Αποτελέσματα συσχέτισης Φύλου – Σκορ Χρησιμότητας Big Data

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ScoreQ2 is the same across categories of Φύλο.	Independent-Samples Mann-Whitney U Test	,428	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 12: Αποτελέσματα συσχέτισης Φύλου – Σκορ Αποτελεσματικότητας Big Data

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ScoreQ3 is the same across categories of Φύλο.	Independent-Samples Mann-Whitney U Test	,671	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 13: Αποτελέσματα συσχέτισης Φύλου – Σκορ Αποφάσεων για Big Data

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ScoreQ4 is the same across categories of Φύλο.	Independent-Samples Mann-Whitney U Test	,059	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 14: Αποτελέσματα συσχέτισης Φύλου – Σκορ Παροχής Υπηρεσιών Υγείας

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ScoreQ5 is the same across categories of Φύλο.	Independent-Samples Mann-Whitney U Test	,197	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 15: Αποτελέσματα συσχέτισης Φύλου – Σκορ Παροχής Υπηρεσιών Υγείας

Correlations

		Ηλικία	ScoreQ 1	ScoreQ 2	ScoreQ 3	ScoreQ 4	ScoreQ 5
Ηλικία	Pearson Correlation	1	-,015	-,015	-,012	-,014	-,011
	Sig. (2-tailed)		,904	,905	,924	,909	,927
	N	66	66	66	66	66	66
ScoreQ 1	Pearson Correlation	-,015	1	1,000**	1,000**	1,000**	1,000**
	Sig. (2-tailed)	,904		,000	,000	,000	,000
	N	66	67	67	67	67	67
ScoreQ 2	Pearson Correlation	-,015	1,000**	1	1,000**	1,000**	1,000**
	Sig. (2-tailed)	,905	,000		,000	,000	,000
	N	66	67	67	67	67	67
ScoreQ 3	Pearson Correlation	-,012	1,000**	1,000**	1	1,000**	1,000**
	Sig. (2-tailed)	,924	,000	,000		,000	,000
	N	66	67	67	67	67	67
ScoreQ 4	Pearson Correlation	-,014	1,000**	1,000**	1,000**	1	1,000**
	Sig. (2-tailed)	,909	,000	,000	,000		,000
	N	66	67	67	67	67	67
ScoreQ 5	Pearson Correlation	-,011	1,000**	1,000**	1,000**	1,000**	1
	Sig. (2-tailed)	,927	,000	,000	,000	,000	
	N	66	67	67	67	67	67

**. Correlation is significant at the 0.01 level (2-tailed).

Πίνακας 16: Αποτελέσματα συσχέτισης Ηλικίας και των ScoreQ1, ScoreQ2, ScoreQ3, ScoreQ4,ScoreQ5

Chi-Square Tests

	Value	df	Asymptotic	Exact Sig. (2-	Exact Sig. (1-

			Significance (2-sided)	sided)	sided)
Pearson Chi-Square	,644[a]	1	,422		
Continuity Correction[b]	,000	1	1,000		
Likelihood Ratio	,992	1	,319		
Fisher's Exact Test				1,000	,612
Linear-by-Linear Association	,634	1	,426		
N of Valid Cases	67				
a. 2 cells (50,0%) have expected count less than 5. The minimum expected count is ,39.					
b. Computed only for a 2x2 table					

Πίνακας 17: Αποτελέσματα συσχέτισης Φύλου – Χρήση των Big Data στον τομέα της Υγείας

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Ηλικία is the same across categories of Η Τεχνολογία των Big Data θα μπορούσε να εφαρμοστεί στον τομέα της Υγείας;.	Independent-Samples Mann-Whitney U Test	,242[1]	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

[1]Exact significance is displayed for this test.

Πίνακας 18: Αποτελέσματα συσχέτισης Ηλικίας – Χρήση των Big Data στον τομέα της Υγείας

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4,062[a]	1	,044		
Continuity Correction[b]	3,110	1	,078		
Likelihood Ratio	4,113	1	,043		
Fisher's Exact Test				,077	,038
Linear-by-Linear Association	4,000	1	,046		
N of Valid Cases	66				
a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 13,00.					
b. Computed only for a 2x2 table					

Πίνακας 19: Αποτελέσματα συσχέτισης Φύλου – Χρήση των Big Data στο εξωτερικό

Directional Measures				
	Value	Asymptotic	Approximate T[b]	Approximate

				Standard Error[a]		Significance
Nominal by Nominal	Lambda	Symmetric	,153	,154	,939	,347
		Γνωρίζετε περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας, στο εξωτερικό; Dependent	,242	,134	1,599	,110
		Φύλο Dependent	,038	,217	,174	,862
	Goodman and Kruskal tau	Γνωρίζετε περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας, στο εξωτερικό; Dependent	,062	,059		,046[c]
		Φύλο Dependent	,062	,059		,046[c]
a. Not assuming the null hypothesis.						
b. Using the asymptotic standard error assuming the null hypothesis.						
c. Based on chi-square approximation						

Πίνακας 20: Αποτελέσματα συντελεστή συσχέτισης Φύλου – Χρήση των Big Data στο εξωτερικό

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Ηλικία is the same across categories of Γνωρίζετε περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας, στο εξωτερικό;.	Independent-Samples Mann-Whitney U Test	,058	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 21: Αποτελέσματα συσχέτισης Ηλικίας – Χρήση των Big Data στο εξωτερικό

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,051[a]	1	,822		
Continuity Correction[b]	,000	1	1,000		
Likelihood Ratio	,051	1	,821		
Fisher's Exact Test				1,000	,551

Linear-by-Linear Association	,050	1	,823		
N of Valid Cases	66				

a. 1 cells (25,0%) have expected count less than 5. The minimum expected count is 4,33.

b. Computed only for a 2x2 table

Πίνακας 22: Αποτελέσματα συσχέτισης Φύλου – Χρήση των Big Data στην Ελλάδα

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Ηλικία is the same across categories of Γνωρίζετε περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας, στην Ελλάδα;.	Independent-Samples Mann-Whitney U Test	,853	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 23: Αποτελέσματα συσχέτισης Ηλικίας – Χρήση των Big Data στην Ελλάδα

Τελειώνοντας, παρακαλώ καταγράψτε προαιρετικά παρατηρήσεις, προτάσεις ή γενικότερα σχόλια που έχετε σχετικά με την ιδέα μιας εφαρμογής για την Διαχείριση Big Data στον τομέα της Υγείας.
Θα πρέπει να συνοδευτεί με μελέτη για την προστασία των ευαίσθητων αυτών δεδομένων και να μη ξεκινήσει "αυτόνομα"
poli kali kai kainotoma idea
Αντί για γενικές προτάσεις, είναι πιο πρακτικό να επενδύσετε σε συγκεκριμένες εφαρμογές που έχουν βέβαιες πηγές δεδομένων και αξιόπιστο μηχανισμό πρόσβασης σε αυτά. Ο όγκος των δεδομένων πρέπει να είναι πράγματι τόσο μεγάλος ώστε να δικαιολογεί την ύπαρξη αξίας κατά την επεξεργασία τους.
τι σημαίνει εφαρμογή big data? δεν είναι σαφές για αυτόν που συμπληρώνει το ερωτηματολόγιο... με αποτέλεσμα οι απαντήσεις να δίνονται με σχετική επιφύλαξη... αν υπήρχαν κάποια παραδείγματα θα βοηθούμε για μεγαλύτερη ακρίβεια στις απαντήσεις. επίσης δεν υπάρχει καμιά αναφορά σε βασικά θέματα που σχετίζονται με τη διαχείριση δεδομένων στην υγεία όπως διαλειτουργικότητα, πρότυπα, κωδικοποιήσεις, ασφάλεια δεδομένων, συγκατάθεση ασθενή, δυνατότητα data portability (δες νέο κανονισμό για τη προστασία των προσωπικών δεδομένων), συγκατάθεση ασθενή, ανωνυμοποίηση δεδομένων, κλπ. Με τη τωρινή κατάσταση στην Ελλάδα big data δεν εχουμε, έχουμε πολλά σκουπίδια και αδόμητη πληροφορία που δεν επαναχρησιμοποιείται (συχνά σκοπίμως κιόλας...) Τέλος, ο χώρος τη υγείας είναι πολύ ιδιαίτερος και ΔΕΝ πρέπει να αντιμετωπίζεται ως άλλα δεδομένα ηλεκτρονικής διακυβέρνησης, η συναλλαγών. πολύ καλή προσπάθεια κατά τα άλλα.
Η τάση είναι η χρήση της τεχνολογίας των big data. Τα δεδομένα θα προσφέρουν πολλές πληροφορίες, όμως εξαρτάται από το βαθμό πρόσβασης στις πληροφορίες και από το αποτέλεσμα της παρέμβασης αν η χρήση της τεχνολογίας θα είναι προς όφελος των πολιτών ή 0.
Μείζον θέμα μιας τέτοιας εφαρμογής θεωρώ ότι είναι η διασφάλιση των δεδομένων και η χρήση των πληροφοριών που πηγάζουν από αυτήν μόνο από αρμόδιους φορείς.
Καλη συνεχεια στην ερευνα σας!! Μακαρι να υλοποιηθει η ιδεα!! Το να εχεις ανα πασα στιγμη τα δεδομενα της καταστασης απο το ποσους και τι γιατρους εχεις μεχρι δεδομενα ηλικιακα, προβληματων υγειας, πληθυσμου κτλ. ειναι σημαντικο πλεονεκτημα στη δημοσια υγεια και στη διαχειριση καταστασεων κρισεων υγειας οπως μπορει να ειναι το προσφυγικο μια φυσικη καταστροφη. Ποσο μαλλον οταν θες να μιλας evidence based και με μετρησιμα μεγεθη και στατιστικες. Απλα θελει ευλαβεια στην καταχωρηση και σε βαθος παραμετροποιηση και εξειδικευση. Σημαντικο δε και το θεμα της ασφαλειας των δεδομενων για να μην χρησιμοποιηθουν κερδοσκοπικα. Τερασιες προοπτικες η ψηφιοποιηση και επεξεργασια των δεδομενων αλλα δεν ξερω κατα ποσον ειμαστε ετοιμοι για αυτο και δη στην Ελλαδα. Αν και νομιζω οτι αναποφευκτα καποια στιγμη η κατασταση θα μας οδηγησει να μπει μια ταξη στο χαος των δεδομενων και της πληροφοριας.
Θεωρώ ότι για να δημιουργηθεί το Big Data θα πρέπει να δημιουργηθούν όλα τα στοιχεία απο τις δημόσιες υπηρεσίες απο την αρχή και να αποθηκευτούν σε μια καινούργια βάση ώστε και τα προγράμματα και η βάση να είναι εξοπλισμένα με τις τελευταίες εξελίξεις της τεχνολογίας.
Such an application in order to be adopted has to take into serious consideration issues around privacy and data security.
Η χρήση των μεγάλων δεδομένων είναι αρκετά σημαντική για το μέλλον τόσο στον τομέα της υγείας οσο και σε άλλους τομείς. Ένα σημαντικό σχόλιο το οποίο θα μπορούσα να παραθέσω είναι οτι η ασφάλεια των δεδομένων είναι το αρχικό και το κυριότερο στάδιο που κάνεις θα έπρεπε να σταθεί πριν τη μελέτη των μεγάλων δεδομένων.
Μια εφαρμογή διαχείρισης Big Data θα δημιουργούσε πολλά ηθικά διλήμματα

ΠΑΡΑΡΤΗΜΑ Β –Πίνακες στατιστικής ανάλυσης του Κεφαλαίου 8

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Ηλικία	144	21	67	38,92	13,471
Valid N (listwise)	144				

Πίνακας 1: Περιγραφικά στοιχεία για την ηλικία του δείγματος

Item Statistics			
	Mean	Std. Deviation	N
Q1	6,04	1,074	46
Q2	6,00	,943	46
Q3	5,30	1,474	46
Q4	5,50	1,090	46
Q5	5,59	1,240	46

Πίνακας 2: Αποτελέσματα ερωτήσεων κλίμακας Likert

Reliability Statistics	
Cronbach's Alpha	N of Items
,888	5

Πίνακας 3: Αποτελέσματα Cronbach Alpha

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Ηλικία	,111	144	,000	,926	144	,000

Πίνακας 4: Έλεγχος κανονικότητας για Ηλικία

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Ηλικία is the same across categories of Γνωρίζετε τι είναι τα Big Data;.	Independent-Samples Mann-Whitney U Test	,018	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 5: Αποτελέσματα Συσχέτισης Ηλικίας -Γνώσης Big Data

Correlations

		Ηλικία	Γνωρίζετε τι είναι τα Big Data;
Ηλικία	Pearson Correlation	1	-,229**
	Sig. (2-tailed)		,006
	N	144	144
Γνωρίζετε τι είναι τα Big Data;	Pearson Correlation	-,229**	1
	Sig. (2-tailed)	,006	
	N	144	151

**. Correlation is significant at the 0.01 level (2-tailed).

Πίνακας 6: Αποτελέσματα Συντελεστή Συσχέτισης Ηλικίας -Γνώσης Big Data

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,309[a]	1	,578		
Continuity Correction[b]	,150	1	,698		
Likelihood Ratio	,309	1	,578		
Fisher's Exact Test				,615	,349
Linear-by-Linear Association	,307	1	,579		
N of Valid Cases	151				

Πίνακας 7: Αποτελέσματα Συντελεστή Συσχέτισης Φύλου -Γνώσης Big Data

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)

Pearson Chi-Square	,152[a]	1	,697		
Continuity Correction[b]	,049	1	,825		
Likelihood Ratio	,152	1	,697		
Fisher's Exact Test				,739	,412
Linear-by-Linear Association	,151	1	,698		
N of Valid Cases	150				

Πίνακας 8: Αποτελέσματα Συντελεστή Συσχέτισης Επαγγέλματος -Γνώσης Big Data

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,134[a]	1	,714		
Continuity Correction[b]	,003	1	,955		
Likelihood Ratio	,136	1	,712		
Fisher's Exact Test				1,000	,486
Linear-by-Linear Association	,133	1	,715		
N of Valid Cases	149				

Πίνακας 9: Αποτελέσματα Συντελεστή Συσχέτισης Επαγγελματικής Εμπειρίας -Γνώσης Big Data

Tests of Normality						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Διάρκεια σε έτη	,136	137	,000	,897	137	,000

Πίνακας 10: Αποτελέσματα Ελέγχου κανονικότητας για Διάρκεια Επαγγελματικής Εμπειρίας

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Διάρκεια σε έτη is the same across categories of Γνωρίζετε τι είναι τα Big Data;.	Independent-Samples Mann-Whitney U Test	,002	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 11: Αποτελέσματα Συσχέτισης Διάρκειας Επαγγελματικής Εμπειρίας -Γνώσης Big Data

Correlations			
		Διάρκεια σε έτη	Γνωρίζετε τι είναι τα Big Data;
Διάρκεια σε έτη	Pearson Correlation	1	-,278**
	Sig. (2-tailed)		,001

	N	137	137
Γνωρίζετε τι είναι τα Big Data;	Pearson Correlation	-,278**	1
	Sig. (2-tailed)	,001	
	N	137	151

Πίνακας 12: Αποτελέσματα Συντελεστή Συσχέτισης Διάρκειας Επαγγελματικής Εμπειρίας -Γνώσης Big Data

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,019[a]	1	,890		
Continuity Correction[b]	,000	1	1,000		
Likelihood Ratio	,019	1	,890		
Fisher's Exact Test				1,000	,512
Linear-by-Linear Association	,019	1	,890		
N of Valid Cases	148				

Πίνακας 13: Αποτελέσματα Συσχέτισης Φύλου – Μορφής των Big Data (Δομημένα Δεδομένα)

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2,459[a]	1	,117		
Continuity Correction[b]	1,729	1	,189		
Likelihood Ratio	2,443	1	,118		
Fisher's Exact Test				,134	,095
Linear-by-Linear Association	2,443	1	,118		
N of Valid Cases	148				

Πίνακας 14: Αποτελέσματα Συσχέτισης Φύλου – Μορφής των Big Data (Αδόμητα Δεδομένα)

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,077[a]	1	,782		
Continuity Correction[b]	,000	1	,986		
Likelihood Ratio	,077	1	,782		

Fisher's Exact Test				,800	,490
Linear-by-Linear Association	,076	1	,782		
N of Valid Cases	148				

Πίνακας 15: Αποτελέσματα Συσχέτισης Φύλου – Μορφής των Big Data (Ημιδομημένα Δεδομένα)

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Ηλικία is the same across categories of Δομημένα δεδομένα(structured data)..	Independent-Samples Mann-Whitney U Test	,010	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 16: Αποτελέσματα Συσχέτισης Ηλικίας – Μορφής των Big Data (Δομημένα Δεδομένα)

Correlations

		Ηλικία	Δομημένα δεδομένα(structured data).
Ηλικία	Pearson Correlation	1	-,239**
	Sig. (2-tailed)		,004
	N	144	142
Δομημένα δεδομένα(structured data).	Pearson Correlation	-,239**	1
	Sig. (2-tailed)	,004	
	N	142	148

Πίνακας 17: Αποτελέσματα Συντελεστή Συσχέτισης Ηλικίας- Μορφής των Big Data (Δομημένα Δεδομένα)

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Ηλικία is the same across categories of Αδόμητα δεδομένα (unstructured data) - ελεύθερο κείμενο..	Independent-Samples Mann-Whitney U Test	,780	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 18: Αποτελέσματα Συσχέτισης Ηλικίας – Μορφής των Big Data (Αδόμητα Δεδομένα)

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Ηλικία is the same across categories of Ημιδομημένα δεδομένα (semistructured data)..	Independent-Samples Mann-Whitney U Test	,004	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 19: Αποτελέσματα Συσχέτισης Ηλικίας – Μορφής των Big Data (Ημιδομημένα Δεδομένα)

Correlations

		Ηλικία	Ημιδομημένα δεδομένα(semistructured data).
Ηλικία	Pearson Correlation	1	-,213*
	Sig. (2-tailed)		,011
	N	144	142
Ημιδομημένα δεδομένα(semistructured data).	Pearson Correlation	-,213*	1
	Sig. (2-tailed)	,011	
	N	142	148

Πίνακας 20: Αποτελέσματα Συντελεστή Συσχέτισης Ηλικίας- Μορφής των Big Data (Ημιδομημένα Δεδομένα)

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,919[a]	1	,338		
Continuity Correction[b]	,621	1	,431		
Likelihood Ratio	,919	1	,338		
Fisher's Exact Test				,395	,215
Linear-by-Linear Association	,913	1	,339		
N of Valid Cases	147				

Πίνακας 21: Αποτελέσματα Συσχέτισης Επαγγέλματος – Μορφής των Big Data (Δομημένα Δεδομένα)

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	7,878[a]	1	,005		
Continuity Correction[b]	6,528	1	,011		
Likelihood Ratio	8,593	1	,003		
Fisher's Exact Test				,005	,004
Linear-by-Linear Association	7,825	1	,005		
N of Valid Cases	147				

Πίνακας 22: Αποτελέσματα Συσχέτισης Επαγγέλματος – Μορφής των Big Data (Αδόμητα Δεδομένα)

Correlations		
	Ηλικία	Αδόμητα δεδομένα

			(unstructured data) - ελεύθερο κείμενο.
Ηλικία	Pearson Correlation	1	,005
	Sig. (2-tailed)		,951
	N	144	142
Αδόμητα δεδομένα (unstructured data) - ελεύθερο κείμενο.	Pearson Correlation	,005	1
	Sig. (2-tailed)	,951	
	N	142	148

Πίνακας 23: Αποτελέσματα Συντελεστή Συσχέτισης Επαγγέλματος- Μορφής των Big Data (Αδόμητα Δεδομένα)

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,536[a]	1	,464		
Continuity Correction[b]	,218	1	,640		
Likelihood Ratio	,536	1	,464		
Fisher's Exact Test				,598	,320
Linear-by-Linear Association	,533	1	,466		
N of Valid Cases	147				

Πίνακας 24: Αποτελέσματα Συσχέτισης Επαγγέλματος – Μορφής των Big Data (Ημιδομημένα Δεδομένα)

Tests of Normality						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ScoreQ1	,224	148	,000	,899	148	,000

Πίνακας 25: Αποτελέσματα ελέγχου κανονικότητας για Σκορ Χρησιμότητας

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ScoreQ1 is the same across categories of Φύλο.	Independent-Samples Mann-Whitney U Test	,045	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 26: Αποτελέσματα Συσχέτισης Φύλου- Σκορ Χρησιμότητας

Correlations		
	Φύλο	ScoreQ

			1
Φύλο	Pearson Correlation	1	-,177*
	Sig. (2-tailed)		,031
	N	151	148
ScoreQ 1	Pearson Correlation	-,177*	1
	Sig. (2-tailed)	,031	
	N	148	148
*. Correlation is significant at the 0.05 level (2-tailed).			

Πίνακας 27: Αποτελέσματα Συντελεστή Συσχέτισης Φύλου- Σκορ Χρησιμότητας

Tests of Normality						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ScoreQ 2	,223	148	,000	,907	148	,000

Πίνακας 28: Αποτελέσματα ελέγχου κανονικότητας για Σκορ Αποτελεσματικότητας

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ScoreQ2 is the same across categories of Φύλο.	Independent-Samples Mann-Whitney U Test	,009	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 29: Αποτελέσματα Συσχέτισης Φύλου - Σκορ Αποτελεσματικότητας

Correlations			
		Φύλο	ScoreQ 2
Φύλο	Pearson Correlation	1	-,214**
	Sig. (2-tailed)		,009
	N	151	148
ScoreQ 2	Pearson Correlation	-,214**	1
	Sig. (2-tailed)	,009	
	N	148	148
**. Correlation is significant at the 0.01 level (2-tailed).			

Πίνακας 30: Αποτελέσματα Συντελεστή Συσχέτισης Φύλου- Σκορ Αποτελεσματικότητας

Tests of Normality						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ScoreQ	,524	147	,000	,057	147	,000

3						

Πίνακας 31: Αποτελέσματα ελέγχου κανονικότητας για Σκορ Αποφάσεων

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ScoreQ3 is the same across categories of Φύλο.	Independent-Samples Mann-Whitney U Test	,394	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 32: Αποτελέσματα συσχέτισης Φύλου- Σκορ Αποφάσεων

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ScoreQ4	,524	147	,000	,057	147	,000

Πίνακας 33: Αποτελέσματα ελέγχου κανονικότητας για Σκορ Παροχής Υπηρεσιών Υγείας

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ScoreQ4 is the same across categories of Φύλο.	Independent-Samples Mann-Whitney U Test	,924	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 34: Αποτελέσματα ελέγχου κανονικότητας για Σκορ Αποφάσεων

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ScoreQ5	,202	147	,000	,912	147	,000

Πίνακας 35: Αποτελέσματα ελέγχου κανονικότητας για Σκορ Πρόληψης

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ScoreQ5 is the same across categories of Φύλο.	Independent-Samples Mann-Whitney U Test	,235	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 36: Αποτελέσματα συσχέτισης Φύλου- Σκορ Πρόληψης

Tests of Normality						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ScoreQ 1	,224	148	,000	,899	148	,000

Πίνακας 37: Αποτελέσματα ελέγχου κανονικότητας για Σκορ Χρησιμότητας

Correlations			
		Ηλικία	ScoreQ 1
Ηλικία	Pearson Correlation	1	-,519**
	Sig. (2-tailed)		,000
	N	144	141
ScoreQ 1	Pearson Correlation	-,519**	1
	Sig. (2-tailed)	,000	
	N	141	148

Πίνακας 38: Αποτελέσματα ελέγχου συσχέτισης Ηλικίας –σκορ Χρησιμότητας

Correlations				
			Ηλικία	ScoreQ 2
Spearman's rho	Ηλικία	Correlation Coefficient	1,000	-,516**
		Sig. (2-tailed)	.	,000
		N	144	141
	ScoreQ 2	Correlation Coefficient	-,516**	1,000
		Sig. (2-tailed)	,000	.
		N	141	148

Πίνακας 39: Αποτελέσματα ελέγχου συσχέτισης Ηλικίας Σκορ Αποτελεσματικότητας

Correlations				
			Ηλικία	ScoreQ 3
Spearman's rho	Ηλικία	Correlation Coefficient	1,000	-,155
		Sig. (2-tailed)	.	,080
		N	144	129
	ScoreQ	Correlation Coefficient	-,155	1,000

	3	Sig. (2-tailed)	,080	.
		N	129	147

Πίνακας 40: Αποτελέσματα ελέγχου συσχέτισης Ηλικίας - Σκορ Αποφάσεων

Correlations			Ηλικία	ScoreQ 4
Spearman's rho	Ηλικία	Correlation Coefficient	1,000	-,187*
		Sig. (2-tailed)	.	,034
		N	144	128
	ScoreQ 4	Correlation Coefficient	-,187*	1,000
		Sig. (2-tailed)	,034	.
		N	128	147

Πίνακας 41: Αποτελέσματα ελέγχου συσχέτισης Ηλικίας - Σκορ Παροχής Υπηρεσιών Υγείας

Correlations				
			Ηλικία	ScoreQ 5
Spearman's rho	Ηλικία	Correlation Coefficient	1,000	-,396**
		Sig. (2-tailed)	.	,000
		N	144	140
	ScoreQ 5	Correlation Coefficient	-,396**	1,000
		Sig. (2-tailed)	,000	.
		N	140	147

Πίνακας 42: Αποτελέσματα ελέγχου συσχέτισης Ηλικίας - Σκορ Πρόληψης

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ScoreQ1 is the same across categories of Επάγγελμα.	Independent-Samples Mann-Whitney U Test	,000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 43: Αποτελέσματα ελέγχου συσχέτισης Επαγγέλματος – Σκορ χρησιμότητας

Correlations				
			Επάγγελμα	ScoreQ 1
Επάγγελμα	Pearson Correlation		1	,342**
	Sig. (2-tailed)			,000
	N		150	147
ScoreQ1	Pearson Correlation		,342**	1
	Sig. (2-tailed)		,000	
	N		147	148
**. Correlation is significant at the 0.01 level (2-tailed).				

Πίνακας 44: Συντελεστής συσχέτισης Επαγγέλματος – Σκορ χρησιμότητας

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ScoreQ2 is the same across categories of Επάγγελμα.	Independent-Samples Mann-Whitney U Test	,001	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 45: Αποτελέσματα ελέγχου συσχέτισης Επαγγέλματος – Σκορ Αποτελεσματικότητας Υπηρεσιών Υγείας

Correlations

		Επάγγελμα	ScoreQ2
Επάγγελμα	Pearson Correlation	1	,291**
	Sig. (2-tailed)		,000
	N	150	147
ScoreQ2	Pearson Correlation	,291**	1
	Sig. (2-tailed)	,000	
	N	147	148

Πίνακας 46: Συντελεστής συσχέτισης Επαγγέλματος – Σκορ Αποτελεσματικότητας Υπηρεσιών Υγείας

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ScoreQ3 is the same across categories of Επάγγελμα.	Independent-Samples Mann-Whitney U Test	,630	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 47: Αποτελέσματα ελέγχου συσχέτισης Επαγγέλματος – Σκορ Αποφάσεων

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ScoreQ4 is the same across categories of Επάγγελμα.	Independent-Samples Mann-Whitney U Test	,642	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 48: Αποτελέσματα ελέγχου συσχέτισης Επαγγέλματος – Σκορ Παροχής Υπηρεσιών Υγείας

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ScoreQ5 is the same across categories of Επάγγελμα.	Independent-Samples Mann-Whitney U Test	,029	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

Πίνακας 49: Αποτελέσματα ελέγχου συσχέτισης Επαγγέλματος – Σκορ Πρόληψης

Correlations			
		Επάγγελμα	ScoreQ4
Επάγγελμα	Pearson Correlation	1	,099
	Sig. (2-tailed)		,254
	N	150	134
ScoreQ4	Pearson Correlation	,099	1
	Sig. (2-tailed)	,254	
	N	134	147

Πίνακας 50: Συντελεστής συσχέτισης Επαγγέλματος – Σκορ Αποτελεσματικότητας Υπηρεσιών Υγείας

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,246[a]	1	,620		
Continuity Correction[b]	,012	1	,914		
Likelihood Ratio	,246	1	,620		
Fisher's Exact Test				,711	,455
Linear-by-Linear Association	,244	1	,621		
N of Valid Cases	150				

Πίνακας 51: Αποτελέσματα ελέγχου συσχέτισης Επαγγέλματος -Περιπτώσεις χρήσης Big Data στο εξωτερικό

Chi-Square Tests					
	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,078[a]	1	,780		
Continuity Correction[b]	,000	1	1,000		
Likelihood Ratio	,078	1	,780		
Fisher's Exact Test				1,000	,545
Linear-by-Linear Association	,077	1	,781		
N of Valid Cases	150				

Πίνακας 52: Αποτελέσματα ελέγχου συσχέτισης Επαγγέλματος -Περιπτώσεις χρήσης Big Data στην Ελλάδα

ΠΑΡΑΡΤΗΜΑ Γ- Εμπειρική διερεύνηση της άποψης των επιστημόνων της Πληροφορικής Υγείας σχετικά με την χρήση της Τεχνολογίας των Big Data (Μεγάλα Δεδομένα) στον χώρο της Υγείας

Εμπειρική διερεύνηση της άποψης των επαγγελματιών υγείας σχετικά με την χρήση της Τεχνολογίας των Big Data (Μεγάλα Δεδομένα) στον χώρο της Υγείας.

Φύλο

- Άντρας
- Γυναίκα

Ηλικία σε έτη: _____

Β.Γνώσεις σχετικά με τα Big Data(Μεγάλα Δεδομένα)

Ποια ήταν η κύρια πηγή ενημέρωσής σας για τα Big Data (Μπορείτε να επιλέξετε περισσότερες της μιας απαντήσεις):

- Εφημερίδες-Περιοδικά
- Πανεπιστημιακές Παραδόσεις – Σεμινάρια
- Διαδίκτυο
- Βιβλία
- Μέσα Κοινωνικής Δικτύωσης
- Χώρος Εργασίας
- Άλλο

Γνωρίζετε τι είναι η τεχνολογία των Big Data:

- Ναι
- Όχι

Σε ποιον από τους παρακάτω τομείς πιστεύετε ότι μπορεί να χρησιμοποιηθεί η τεχνολογία των Big Data (Μπορείτε να επιλέξετε περισσότερες της μιας απαντήσεις):

- Ενέργεια
- Δημόσιος τομέας
- Έρευνα και Εκπαίδευση
- Υγειονομική περίθαλψη
- Χρηματοοικονομικές Υπηρεσίες
- Εμπόριο

Ποια από τα παρακάτω θεωρείτε ότι μπορούν να χρησιμοποιηθούν ως πηγές για τη συλλογή δεδομένων και την δημιουργία των Big Data (Μπορείτε να επιλέξετε περισσότερες της μιας απαντήσεις);

- Ηλεκτρονικές Συναλλαγές
- Μηνύματα ηλεκτρονικού ταχυδρομείου
- Μέσα κοινωνικής δικτύωση
- Δεδομένα Αισθητήρων
- Δεδομένα Υγείας
- Διαδίκτυο

Ποια θεωρείτε ότι είναι η μορφή των Big Data(Μπορείτε να επιλέξετε περισσότερες της μιας απαντήσεις);

- Δομημένα δεδομένα(structured data).
- Αδόμητα δεδομένα(unstructured data).
- Ημιδομημένα δεδομένα (semistructured data).
- Δεν γνωρίζω

Ποια εργαλεία διαχείρισης και ανάλυσης για Big Data γνωρίζετε;

- Hadoop
- MongoDB
- Cassandra
- Cisco Common Big Data Platform
- Άλλα

Γ. Big Data στην υγεία

Η Τεχνολογία των Big Data θα μπορούσε να εφαρμοστεί στον τομέα της Υγείας;

- Ναι
- Όχι

Γνωρίζετε περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας, στο εξωτερικό;

- Ναι
- Όχι

Γνωρίζετε περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας, στην Ελλάδα;

- Ναι
- Όχι

Ποια από τα παρακάτω θεωρείτε ότι μπορούν να χρησιμοποιηθούν ως πηγές για τη συλλογή Big Data στον τομέα της Υγείας;

- Ηλεκτρονικός Φάκελος Ασθενή

- Συστήματα Κλινικών Αποφάσεων
- Ηλεκτρονική Συνταγογράφηση
- Επιστημονικά περιοδικά
- Μέσα κοινωνικής δικτύωσης

Εμπειρική μελέτη

Για την καταγραφή της άποψης των επιστημόνων σχετικά με μία εφαρμογή που θα Διαχειρίζεται (Συλλογή - Επεξεργασία - Ανάλυση - Μοντελοποίηση) Big Data στον τομέα της Υγείας, θα μετρηθούν η

"Αντιλαμβανόμενη Χρησιμότητα", η "Αντιλαμβανόμενη Ευκολία Χρήσης", το "Συγκριτικό Πλεονέκτημα", η "Συμβατότητα", "Στάση ως προς την Χρήση" και η "Πρόθεση Χρήσης".

Η μέτρηση των παραπάνω παραγόντων/διαστάσεων γίνεται απαντώντας στο κατά πόσο συμφωνείτε ή διαφωνείτε με τις προτάσεις/δηλώσεις που υπάρχουν σε αυτές. Παρακαλώ χρησιμοποιείστε τη παρακάτω κλίμακα για να δηλώσετε τη προτίμησή σας σε κάθε μια από τις διατυπώσεις:

1 - Διαφωνώ απόλυτα

2 - Διαφωνώ

3 - Διαφωνώ Μερικώς

4 - Ούτε διαφωνώ / Ούτε συμφωνώ

5 - Συμφωνώ Μερικώς

6 - Συμφωνώ

7 - Συμφωνώ απόλυτα

Θεωρώ ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα βοηθήσει τους Επαγγελματίες Υγείας είναι χρήσιμη.

Διαφωνώ Απόλυτα 1 2 3 4 5 6 7 Συμφωνώ Απόλυτα

Θεωρώ ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα μπορούσε να αυξήσει την αποτελεσματικότητα των παρεχόμενων Υπηρεσιών Υγείας.

Διαφωνώ Απόλυτα 1 2 3 4 5 6 7 Συμφωνώ Απόλυτα

Θεωρώ ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα βοηθήσει τους Επαγγελματίες Υγείας στη διαδικασία λήψης αποφάσεων.

Διαφωνώ Απόλυτα 1 2 3 4 5 6 7 Συμφωνώ Απόλυτα

Θεωρώ ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα δώσει τη δυνατότητα παροχής υπηρεσιών υγείας προσαρμοσμένων στις ανάγκες των ασθενών.

Διαφωνώ Απόλυτα 1 2 3 4 5 6 7 Συμφωνώ Απόλυτα

Θεωρώ ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας προσφέρει αποτελεσματικότερη πρόληψη στους πληθυσμούς υψηλού κινδύνου.

Διαφωνώ Απόλυτα 1 2 3 4 5 6 7 Συμφωνώ Απόλυτα

ΠΑΡΑΡΤΗΜΑ Δ- Εμπειρική διερεύνηση της άποψης των Επαγγελματιών Υγείας σχετικά με την χρήση της Τεχνολογίας των Big Data (Μεγάλα Δεδομένα) στον χώρο της Υγείας

Εμπειρική διερεύνηση της άποψης των επαγγελματιών υγείας σχετικά με την χρήση της Τεχνολογίας των Big Data (Μεγάλα Δεδομένα) στον χώρο της Υγείας.

Φύλο

- Άντρας
- Γυναίκα

Ηλικία σε έτη: _____

Επάγγελμα

- Ιατρός
- Νοσηλευτής

Ειδικότητα: _____

Αν ναι, διάρκεια (σε έτη): _____

Β.Γνώσεις σχετικά με τα Big Data(Μεγάλα Δεδομένα)

Γνωρίζετε τι είναι τα Big Data:

- Ναι
- Όχι

Ποια ήταν η κύρια πηγή ενημέρωσής σας για τα Big Data (Μπορείτε να επιλέξετε περισσότερες της μιας απαντήσεις):

- Εφημερίδες-Περιοδικά
- Πανεπιστημιακές Παραδόσεις – Σεμινάρια
- Διαδίκτυο
- Βιβλία
- Μέσα Κοινωνικής Δικτύωσης
- Χώρος Εργασίας
- Άλλο

Σε ποιον από τους παρακάτω τομείς πιστεύετε ότι μπορεί να χρησιμοποιηθεί η τεχνολογία των Big Data (Μπορείτε να επιλέξετε περισσότερες της μιας απαντήσεις):

- Ενέργεια
- Δημόσιος τομέας
- Έρευνα και Εκπαίδευση

- Υγειονομική περίθαλψη
- Χρηματοοικονομικές Υπηρεσίες
- Εμπόριο
- Δεν γνωρίζω

Ποια από τα παρακάτω θεωρείτε ότι μπορούν να χρησιμοποιηθούν ως πηγές για τη συλλογή δεδομένων και την δημιουργία των Big Data (Μπορείτε να επιλέξετε περισσότερες της μιας απαντήσεις);

- Ηλεκτρονικές Συναλλαγές
- Μηνύματα ηλεκτρονικού ταχυδρομείου
- Μέσα κοινωνικής δικτύωση
- Δεδομένα Αισθητήρων
- Δεδομένα Υγείας
- Διαδίκτυο

Ποια θεωρείτε ότι είναι η μορφή των Big Data(Μπορείτε να επιλέξετε περισσότερες της μιας απαντήσεις);

- Δομημένα δεδομένα(structured data).
- Αδόμητα δεδομένα(unstructured data).
- Ημιδομημένα δεδομένα (semistructured data).
- Δεν γνωρίζω

Γ. Big Data στην υγεία

Η Τεχνολογία των Big Data θα μπορούσε να εφαρμοστεί στον τομέα της Υγείας;

- Ναι
- Όχι

Γνωρίζετε περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας, στο εξωτερικό;

- Ναι
- Όχι

Αν ναι, αναφέρετε:_____

Γνωρίζετε περιπτώσεις χρήσης της τεχνολογίας των Big Data στον χώρο της Υγείας, στην Ελλάδα;

- Ναι
- Όχι

Αν ναι, αναφέρετε:_____

Ποια από τα παρακάτω θεωρείτε ότι μπορούν να χρησιμοποιηθούν ως πηγές για τη συλλογή Big Data στον τομέα της Υγείας;

- Ηλεκτρονικός Φάκελος Ασθενή
- Ζωτικά Σημεία
- Ηλεκτρονική Συνταγογράφηση
- Συστήματα PACS
- Επιστημονικά περιοδικά
- Διαδίκτυο
- Μέσα κοινωνικής δικτύωσης
- Δεδομένα Ασφαλιστικών Εταιρειών
- Δεν γνωρίζω

Εμπειρική μελέτη

Για την καταγραφή της άποψης των επιστημόνων σχετικά με μία εφαρμογή που θα Διαχειρίζεται (Συλλογή - Επεξεργασία - Ανάλυση - Μοντελοποίηση) Big Data στον τομέα της Υγείας, θα μετρηθούν η

"Αντιλαμβανόμενη Χρησιμότητα", η "Αντιλαμβανόμενη Ευκολία Χρήσης", το "Συγκριτικό Πλεονέκτημα", η "Συμβατότητα", "Στάση ως προς την Χρήση" και η "Πρόθεση Χρήσης".

Η μέτρηση των παραπάνω παραγόντων/διαστάσεων γίνεται απαντώντας στο κατά πόσο συμφωνείτε ή διαφωνείτε με τις προτάσεις/δηλώσεις που υπάρχουν σε αυτές. Παρακαλώ χρησιμοποιείστε τη παρακάτω κλίμακα για να δηλώσετε τη προτίμησή σας σε κάθε μια από τις διατυπώσεις:

1 - Διαφωνώ απόλυτα

2 - Διαφωνώ

3 - Διαφωνώ Μερικώς

4 - Ούτε διαφωνώ / Ούτε συμφωνώ

5 - Συμφωνώ Μερικώς

6 - Συμφωνώ

7 - Συμφωνώ απόλυτα

Θεωρώ ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα βοηθήσει τους Επαγγελματίες Υγείας είναι χρήσιμη.

Διαφωνώ Απόλυτα 1 2 3 4 5 6 7 Συμφωνώ Απόλυτα

Θεωρώ ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα μπορούσε να αυξήσει την αποτελεσματικότητα των παρεχόμενων Υπηρεσιών Υγείας.

Διαφωνώ Απόλυτα 1 2 3 4 5 6 7 Συμφωνώ Απόλυτα

Θεωρώ ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα βοηθήσει τους Επαγγελματίες Υγείας στη διαδικασία λήψης αποφάσεων.

Διαφωνώ Απόλυτα 1 2 3 4 5 6 7 Συμφωνώ Απόλυτα

Θεωρώ ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας θα δώσει τη δυνατότητα παροχής υπηρεσιών υγείας προσαρμοσμένων στις ανάγκες των ασθενών.

Διαφωνώ Απόλυτα 1 2 3 4 5 6 7 Συμφωνώ Απόλυτα

Θεωρώ ότι η εφαρμογή της τεχνολογίας των Big Data στον τομέα της Υγείας προσφέρει αποτελεσματικότερη πρόληψη στους πληθυσμούς υψηλού κινδύνου.

Διαφωνώ Απόλυτα 1 2 3 4 5 6 7 Συμφωνώ Απόλυτα