ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Πρόγραμμα Μεταπτυχιακών Σπουδών

Επιστήμη και Τεχνολογία της Πληροφορικής και των Υπολογιστών

Ειδίκευση Λογισμικού και Πληροφοριακών Συστημάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Οπτικοποίηση τεχνικών εξόρυξης διεργασιών

Καούνη Αντωνία
Α.Μ. : 19016

Επιβλέποντες:
Γεώργιος Μιαούλης (Καθηγητής)
Γεωργία Θεοδωροπούλου (Υποψήφια Διδάκτωρ)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Οπτικοποίηση τεχνικών εξόρυξης διεργασιών

Καούνη Αντωνία
Α.Μ. : 19016

Εισηγητής: Γεώργιος Μιαούλης, Καθηγητής

Εξεταστική Επιτροπή:

1. **Μιαούλης Γεώργιος (Καθηγητής)**
2. **Βουλόδημος Αθανάσιος (Επίκουρος Καθηγητής)**
3. **Μπαρδής Γεώργιος (Λέκτορας)**

Ημερομηνία εξέτασης : 23/4/2021

Georgios Miaoulis
Digitally signed by Georgios Miaoulis
DN: c=GR, l=Athens, o=University of West Attica, ou=Class B - Private Key created and stored in software CSP, sn=Miaoulis, givenName=Georgios, serialNumber=3179983822, cn=Georgios Miaoulis, email=gmiaoul@uniwa.gr
Date: 2021.04.27 11:13:11 +03'00'

Athanasios Voulodimos
Digitally signed by Athanasios Voulodimos
Date: 2021.04.26 17:50:43 +03'00'

Georgios Bardis
Digitally signed by Georgios Bardis
Date: 2021.04.26 23:41:46 +03'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

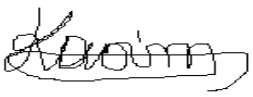
Η κάτωθι υπογεγραμμένη Καούνη Αντωνία του Κωνσταντίνου, με αριθμό μητρώου 19016, φοιτήτρια του Προγράμματος Μεταπτυχιακών Σπουδών «Επιστήμη και Τεχνολογία της Πληροφορικής και των Υπολογιστών», ειδίκευση «Λογισμικού και Πληροφοριακών Συστημάτων» του Τμήματος ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ της Σχολής ΜΗΧΑΝΙΚΩΝ του Πανεπιστημίου Δυτικής Αττικής, δηλώνω ότι:

«Είμαι συγγραφέας αυτής της μεταπτυχιακής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της, είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Επιθυμώ την απαγόρευση πρόσβασης στο πλήρες κείμενο της εργασίας μου μέχρι 31/10/2021 και έπειτα από αίτηση μου στη Βιβλιοθήκη και έγκριση του επιβλέποντα καθηγητή.

Η Δηλούσα



ΕΥΧΑΡΙΣΤΙΕΣ

Με την ολοκλήρωση της μεταπτυχιακής διπλωματικής μου εργασίας, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες σε όλους όσους συνέβαλλαν στην εκπόνησή της.

Ευχαριστώ θερμά τον επιβλέποντα καθηγητή μου, Δρ. Μιαούλη Γεώργιο, για την εμπιστοσύνη που μου έδειξε αναθέτοντάς μου το συγκεκριμένο θέμα, για την επιστημονική του καθοδήγηση και τις υποδείξεις του που με βοήθησαν να ολοκληρώσω την εργασία μου.

Επιπλέον, ιδιαίτερες ευχαριστίες θα ήθελα να απευθύνω στην υποψήφια διδάκτορα κα Θεοδωροπούλου Γεωργία για τη συνεχή υποστήριξη, καθοδήγηση και πολύτιμη βοήθεια που μου παρείχε στην επίλυση των προβλημάτων που αντιμετώπισα καθόλη τη διάρκεια της συγγραφής της παρούσας εργασίας.

Τέλος, θα ήθελα εκφράσω την ευγνωμοσύνη μου στην οικογένειά μου για όλη τη συμπαράσταση και την κατανόησή που έδειξαν κατά τη διάρκεια των μεταπτυχιακών σπουδών μου, που συνέπεσε με τη δύσκολη κατάσταση της πανδημίας που επηρέασε όλους μας.

ΠΕΡΙΛΗΨΗ

Η εργασία αυτή αφορά την ανάπτυξη προγράμματος σε Python στο οποίο έγινε η οπτικοποίηση δεδομένων που έχουν να κάνουν με την εξόρυξη διεργασιών. Χρησιμοποιώντας βιβλιοθήκες οπτικοποίησης όπως η matplotlib ή η plotly δημιουργήθηκαν διαγράμματα διαφόρων μορφών (bar charts, scatter charts κ.ά) για την οπτικοποίηση τόσο των δεδομένων που χρησιμοποιούνται στην εξόρυξη διεργασιών (αρχεία γεγονότων) όσο και των τεχνικών αναπαράστασης και ανάλυσης διεργασιών.

A B S T R A C T

This work covers the development of a Python program that visualizes process mining data. Using visualization libraries like matplotlib or plotpy, we created different kinds of charts (bar charts, scatter charts etc.) for both the visualization of the data in process mining (event files) and the analysis and visualization techniques of processes.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ / ΣΧΗΜΑΤΩΝ

ΕΙΣΑΓΩΓΗ

Αυτή τη στιγμή ζούμε στην εποχή των μεγάλων δεδομένων. Η εποχή των μεγάλων δεδομένων είναι μία εποχή που περιγράφεται από την ταχέως αναπτυσσόμενη ποσότητα δεδομένων, η οποία είναι πολύ μεγαλύτερη από αυτήν που οι περισσότεροι άνθρωποι θα φαντάζονταν ποτέ.

Αυτό που ορίζει το σήμερα ως την εποχή των big data είναι ότι οι εταιρείες, οι κυβερνήσεις και οι μη κερδοσκοπικές οργανώσεις έχουν βιώσει μια αλλαγή στη συμπεριφορά. Μπορούν και χρησιμοποιούν όλα τα δεδομένα που είναι δυνατόν να συλλέξουν, για έναν τρέχοντα ή μελλοντικό άγνωστο σκοπό, ώστε να βελτιώσουν την επιχείρησή τους, τη χώρα τους και τις επιστήμες. Η πρόκληση δεν είναι η συλλογή των δεδομένων, αλλά η εύρεση των σωστών δεδομένων και η χρήση υπολογιστών για την αύξηση της γνώσης στον τομέα ενδιαφέροντος, καθώς και ο προσδιορισμός προτύπων.

Ορισμένες βασικές τεχνολογίες και ανάγκες της αγοράς μας οδήγησαν σε αυτό το σημείο όπου ο όγκος των δεδομένων που συλλέγονται, αποθηκεύονται και εξετάζονται σε αναλυτικές δραστηριότητες, έχει αυξηθεί με τεράστιο ρυθμό. Αυτό οφείλεται σε πολλούς παράγοντες, όπως είναι το πρωτόκολλο Internet Protocol 6 (IPv6), ο βελτιωμένος τηλεπικοινωνιακός εξοπλισμός, οι τεχνολογίες όπως η RFID, το μειωμένο κόστος ανά μονάδα παραγωγής ηλεκτρονικών ειδών, τα κοινωνικά μέσα και το διαδίκτυο. (Dean, 2014, pp. 1-5).

Η οπτικοποίηση δεδομένων για την εξόρυξη γνώσης από τα δεδομένα κατά την τελική επεξεργασία τους, το στάδιο της αξιολόγησης – ερμηνείας είναι πολύ σημαντική, διότι σχετίζεται με το πώς αντιλαμβάνονται οι χρήστες τα αποτελέσματα της όλης διαδικασίας. Έρχεται να συμπληρώσει τα τελικά συμπεράσματα που προκύπτουν με αριθμητικές τιμές σε έναν πίνακα, εκμεταλλευόμενη το γεγονός ότι το ανθρώπινο μάτι

αντιλαμβάνεται πιο γρήγορα και εύκολα την πληροφορία όταν αυτή δίνεται ως σχήμα στο χώρο.

Η τεχνική της οπτικοποίησης των δεδομένων γίνεται εύκολα σε παρατηρήσεις που αφορούν δύο ή τρεις μεταβλητές, αντιστοιχίζοντας κάθε μεταβλητή σε μία διάσταση του χώρου. Σε περιπτώσεις περισσότερων μεταβλητών, ο πίνακας μας πρέπει να μεταβληθεί ώστε να μπορεί να γίνει αντιστοίχηση στις τρεις διαστάσεις του χώρου και στη συνέχεια να αναπαρασταθεί γραφικά.

Η παρούσα εργασία αφορά την ανάπτυξη προγράμματος σε Python, στο οποίο θα γίνεται οπτικοποίηση δεδομένων που έχουν να κάνουν με την εξόρυξη διεργασιών. Χρησιμοποιώντας βιβλιοθήκες της Python, όπως η pm4py, για τη διαχείριση των δεδομένων από τα Event Logs, και την matplotlib έχουν δημιουργηθεί διαγράμματα διαφόρων μορφών (bar charts, scatter charts κ.ά) για την οπτικοποίηση τόσο των δεδομένων που χρησιμοποιούνται στην εξόρυξη διεργασιών (αρχεία γεγονότων) όσο και των τεχνικών αναπαράστασης και ανάλυσης διεργασιών.

ΚΕΦΑΛΑΙΟ 1

1.1 Η γλώσσα Python

Η Python είναι μία υψηλού επιπέδου, γλώσσα προγραμματισμού. Δημιουργήθηκε από τον Ολλανδό Γκίντο Βαν Ρόσσουμ (Guido van Rossum) στο ερευνητικό κέντρο Centrum Wiskunde & Informatica (CWI) το 1989 και κυκλοφόρησε για πρώτη φορά το 1991.[1] Είναι διερμηνευόμενη (interpreted), η υλοποίησή της δηλαδή γίνεται μέσω ενός διερμηνέα, είναι γενικού σκοπού (general-purpose) και ανήκει στις γλώσσες προστακτικού προγραμματισμού (Imperative programming). Υποστηρίζει τόσο το διαδικαστικό (procedural programming) όσο και το αντικειμενοστραφές (object – oriented programming) προγραμματιστικό υπόδειγμα (programming paradigm). [1]

Ο κύριος στόχος της είναι η αναγνωσιμότητα του κώδικά της και η ευκολία χρήσης της. Το συντακτικό της επιτρέπει στους προγραμματιστές να εκφράσουν έννοιες σε λιγότερες γραμμές κώδικα από ότι θα ήταν δυνατόν σε γλώσσες όπως η C++ ή η Java.[2] Διακρίνεται λόγω του ότι έχει πολλές βιβλιοθήκες που διευκολύνουν ιδιαίτερα αρκετές συνηθισμένες εργασίες και για την ταχύτητα εκμάθησης της.

Οι διερμηνευτές της Python είναι διαθέσιμοι για εγκατάσταση σε πολλά λειτουργικά συστήματα, επιτρέποντας στην Python την εκτέλεση κώδικα σε ευρεία γκάμα συστημάτων. Χρησιμοποιώντας εργαλεία τρίτων, όπως το Py2exe, ο κώδικας της Python μπορεί να πακεταριστεί σε αυτόνομα εκτελέσιμα προγράμματα για μερικά από τα πιο δημοφιλή λειτουργικά συστήματα, επιτρέποντας τη διανομή του βασισμένου σε Python λογισμικού για χρήση σε αυτά τα περιβάλλοντα χωρίς να απαιτείται εγκατάσταση του διερμηνευτή της Python.

[1] Αγγελιδάκης Ν.
[2] McConnell S.

Η Python αναπτύσσεται ως ανοιχτό λογισμικό (open source) και η διαχείρισή της γίνεται από τον μη κερδοσκοπικό οργανισμό Python Software Foundation. Ο κώδικας διανέμεται με την άδεια Python Software Foundation License, η οποία είναι συμβατή με την GPL. Το όνομα της γλώσσας προέρχεται από την ομάδα των Άγγλων κωμικών Μόντυ Πάιθον και δεν έχει καμιά σχέση με το φίδι πύθωνα, παρότι το λογότυπό της παραπέμπει σε κάτι τέτοιο.

Η Python ενδείκνυται ως γλώσσα εισαγωγική στον προγραμματισμό και στην επιστήμη των υπολογιστών γενικότερα, με δεδομένη την απλότητα στη σύνταξή της. Διεθνώς πολλά πανεπιστήμια αλλά και η δευτεροβάθμια εκπαίδευση σε πολλές χώρες (και στη δική μας), υιοθετούν μια Python first προσέγγιση εισαγωγής στον προγραμματισμό. Στη δευτεροβάθμια εκπαίδευση της χώρας μας το πρώτο βήμα έγινε με τα Επαγγελματικά Λύκεια στα οποία έχουν εισαχθεί τα σχετικά μαθήματα (Αρχές Προγραμματισμού Υπολογιστών στη Β Λυκείου και Προγραμματισμός Υπολογιστών στη Γ' Λυκείου, στον τομέα Πληροφορικής).[3]

Υπάρχει διαθέσιμο πλούσιο υλικό και στην Ελληνική γλώσσα, πέραν της διεθνούς βιβλιογραφίας και πηγών. Για παράδειγμα στις ιστοσελίδες διαδικτυακών μαθημάτων mathesis.cup.gr και coursity.gr υπάρχουν διαδικτυακά μαθήματα εισαγωγής στον προγραμματισμό με Python, το περιεχόμενο των οποίων διατίθεται ελεύθερα. Στο mathesis.cup.gr υπάρχει και μάθημα προχωρημένου προγραμματισμού με Python.

[3] Αράπογλου, κλπ

1.2 Χρήση της Python

Η Python έχει χρησιμοποιηθεί για διάφορους σκοπούς, όπως :

1. Data Science, συμπεριλαμβανομένων των machine learning, data analysis, and data visualization. Γενικά η Python είναι εξαιρετική για την διαχείριση και οπτικοποίηση δεδομένων.
2. Web Development, όπου κάποιος μπορεί να χτίσει το backend κομμάτι μιας διαδικτυακής εφαρμογής με Python.
3. Scripting, η γλώσσα είναι ιδιαίτερα καλή κατά τη σύνταξη σύντομων "scripts" ή μικρών ad hoc προγραμμάτων που χρησιμοποιούνται για την αυτοματοποίηση εργασιών, με εφαρμογές π.χ. στην διαχείριση συστημάτων.[4]

Η γλώσσα Python, όπως αναφέρθηκε παραπάνω χρησιμοποιείται και για την ανάλυση δεδομένων και την οπτικοποίηση αυτών. Έχει σαφή σύνταξη, μεγάλη κοινότητα και καλό documentation, όλα online και δωρεάν. Λόγω της καλής σύνταξης μπορούμε να δαπανήσουμε λιγότερο χρόνο στην επιδιόρθωση σφαλμάτων και την αντιμετώπιση προβλημάτων που συνήθως προκύπτουν στη συγγραφή κώδικα.

Η μεγάλη δύναμη της Python είναι το πλήθος των ελεύθερων third-party βιβλιοθηκών. Στις αρχές του 2019, το επίσημο Python repository περιείχε γύρω στις 165 χιλιάδες projects. Σχεδόν για κάθε περίπτωση, υπάρχει διαθέσιμη μια υψηλής ποιότητας βιβλιοθήκη. Μεταξύ αυτών υπάρχουν μερικές εξαιρετικές βιβλιοθήκες για την επιστήμη των δεδομένων, καλύπτοντας κάθε στάδιο της ανάλυσης των δεδομένων. [5]

Σε έρευνα το 2019 για τη δημοτικότητα των γλωσσών προγραμματισμού η Python και η R συνεχίζουν να κυριαρχούν. Η νέα καταχώρηση σε σχέση με το 2018 στα αποτελέσματα της

[4] https://medium.com/@GoldenGatePro/Python-libraries-data-science-bbc98c1bb148 (ανακτήθηκε 3/1/2021)
[5] https://docs.Python.org/3/library/index.html (ανακτήθηκε 3/1/2021)

έρευνας ήταν Javascript, η οποία έλαβε αξιοσέβαστο μερίδιο 6,8%. Το μερίδιο της Julia έχει αυξηθεί, ενώ οι περισσότερες άλλες γλώσσες έχουν μειωθεί.

Εδώ παρουσιάζονται οι κύριες γλώσσες προγραμματισμού ταξινομημένες κατά δημοτικότητα. (Πηγή : Python leads the 11 top Data Science, Machine Learning platforms: Trends and Analysis (kdnuggets.com) (ανακτήθηκε 8/1/2021)

Πίνακας 1 : Κύριες γλώσσες προγραμματισμού κατά δημοτικότητα

Platform	2019 % share	2018 % share	% change
Python	65.8%	65.6%	0.2%
R Language	46.6%	48.5%	-4.0%
SQL Language	32.8%	39.6%	-17.2%
Java	12.4%	15.1%	-17.7%
Unix shell/awk	7.9%	9.2%	-13.4%
C/C++	7.1%	6.8%	3.7%
Javascript	6.8%	na	na
Other programming and data languages	5.7%	6.9%	-17.1%
Scala	3.5%	5.9%	-41.0%
Julia	1.7%	0.7%	150.4%
Perl	1.3%	1.0%	25.2%
Lisp	0.4%	0.3%	46.1%

ΚΕΦΑΛΑΙΟ 2

2.1 Python και data analysis

Προκειμένου να γίνει η ανάλυση δεδομένων πρέπει πρώτα να γίνει συλλογή, προ επεξεργασία και μετασχηματισμός αυτών σε κατάλληλη μορφή. Το ερευνώμενο σύνολο δεδομένων πρέπει να είναι κατάλληλο ώστε η ανάλυση τους να μπορεί να αποκαλύψει μόνο τα πρότυπα που πράγματι εμφανίζονται στα δεδομένα.

Το σύνολο δεδομένων που ερευνούμε, πρέπει να είναι αρκετά μεγάλο για να περιέχει αυτά τα πρότυπα παραμένοντας να εξορυχθεί σε ένα αποδεκτό χρονικό διάστημα. Συνήθως λοιπόν, το ερευνώμενο σύνολο καθαρίζεται. Το καθάρισμα δεδομένων διαγράφει τις παρατηρήσεις που περιέχουν θόρυβο και αυτές με ελλιπή ή άσχετα δεδομένα. (Simmi Bagga., Dr. G.N. Singh, 2012).

Η ανάλυση δεδομένων με τη βοήθεια της Python βρίσκει εφαρμογή σε διάφορους τομείς, οι σημαντικότεροι των οποίων είναι:

1. Η Ιατρική

Τα τελευταία χρόνια, η εξόρυξη δεδομένων χρησιμοποιείται ευρέως στους τομείς της ιατρικής, όπως η βιοϊατρική, το DNA ,η γενετική και η φαρμακευτική. Στον τομέα της γενετικής, ο σκοπός είναι να κατανοήσουμε την χαρτογράφηση της σχέσης μεταξύ της μεταβολής των ακολουθιών του ανθρώπινου DNA και την προδιάθεση στην αρρώστια. Η εξόρυξη δεδομένων είναι ένα σημαντικό εργαλείο που μπορεί να βοηθήσει στην βελτίωση της διάγνωσης, της πρόληψης και της θεραπείας των ασθενειών.

Εξαιτίας της αύξησης των βιοϊατρικών ερευνών, η μεγάλη κλίμακα γονιδιακών προτύπων και λειτουργιών πρέπει να εξετασθεί. Τα εργαλεία της εξόρυξης δεδομένων μπορούν να βοηθήσουν σε μεγάλο βαθμό για να μελετήσουμε την σύσταση του DNA και να βρούμε ποικίλα πρότυπα και λειτουργίες αυτού.

Ένας από τους κύριους στόχους που σχετίζεται με την ανάλυση δεδομένων του DNA είναι η σύγκριση ποικίλων

ακολουθιών και η αναζήτηση ομοιοτήτων μεταξύ των δεδομένων του DNA. Η σύγκριση κυρίως περιλαμβάνει την γονιδιακή ακολουθία υγιών και βλαβερών ιστών για να βρει την διαφορά ανάμεσα σε αυτούς τους δύο τύπους. Αυτό μπορεί να επιτευχθεί ανακτώντας τις τάξεις υγιών αλλά και βλαβερών γονιδιακών ακολουθιών και μετά βρίσκοντας τις συχνά εμφανιζόμενες μορφές των δύο τάξεων. Αυτή η ανάλυση βοηθάει στο να βρίσκουμε τις ομοιότητες και τις διαφορές στις γενετικές ακολουθίες.

Στην βιοϊατρική, ερευνάται αν οι περισσότερες ασθένειες προκαλούνται από ένα συνδυασμό των γονιδίων. Η μέθοδος της συσχέτισης χρησιμοποιείται για να καθορίσει την συνύπαρξη ομάδων των γονιδίων και επίσης μπορούμε να εξετάσουμε την αλληλεπίδραση και την σχέση μεταξύ των γονιδίων.

Υπάρχουν διάφοροι συνδυασμοί γονιδίων που συμβάλλουν στις ασθένειες, αλλά αυτά τα γονίδια ενεργοποιούνται σε διαφορετικά επίπεδα. Η ανάλυση μονοπατιού (path analysis) χρησιμοποιείται για να συνδέει διαφορετικά γονίδια με διαφορετικά στάδια κατά την εξέλιξη της ασθένειας. Η ανάλυση μονοπατιού διαδραματίζει ένα σπουδαίο ρόλο στην γενετική.

2. Η Οικονομία

Άλλος τομέας που εφαρμόζεται η εξόρυξη δεδομένων είναι η οικονομία. Τα οικονομικά δεδομένα κυρίως συλλέγονται από τράπεζες και από άλλους οικονομικούς οργανισμούς. Τα δεδομένα αυτά συνήθως είναι αξιόπιστα, ολοκληρωμένα και έχουν υψηλή ποιότητα και απαιτούν συστηματική μέθοδο για την ανάλυση αυτών. Η συνεισφορά της εξόρυξης δεδομένων στην επιστήμη της οικονομίας συναντάται στην συλλογή και κατανόηση των δεδομένων, στην βελτίωση δεδομένων (data refinement), στην δημιουργία και εκτίμηση ενός μοντέλου και στην ανάπτυξη αυτού. Η σωστή ανάλυση των οικονομικών δεδομένων μας διευκολύνει στο να παίρνουμε καλύτερες αποφάσεις ενεργώντας σύμφωνα με την ανάλυση της αγοράς. Τα εργαλεία και οι τεχνικές της εξόρυξης

δεδομένων βοηθούν στο να αναλύσουμε τα οικονομικά δεδομένα με τους παρακάτω τρόπους:

Τα δεδομένα που συλλέγονται από διάφορα οικονομικά ιδρύματα, όπως οι τράπεζες, συγκεντρώνονται αρχικά στην αποθήκη δεδομένων (data warehouse). Οι τεχνικές της πολυδιάστατης ανάλυσης δεδομένων χρησιμοποιούνται για την ανάλυση τέτοιων δεδομένων που συλλέγονται στην αποθήκη δεδομένων για τις γενικές ιδιότητές του.

Μία άλλη εφαρμογή της εξόρυξης δεδομένων σχετίζεται με την πρόβλεψη αποπληρωμής δανείου και πολιτικές πίστωσης του πελάτη. Μέθοδοι της εξόρυξης, όπως η επιλογή χαρακτηριστικών (feature selection) βοηθάει στην ταυτοποίηση ποικίλων χαρακτηριστικών όπως το επίπεδο εισοδήματος του πελάτη, την εξόφληση ανάλογα με τα έσοδα, την πιστωτική του ιστορία κτλ. Με την επεξεργασία αυτών των χαρακτηριστικών, η τράπεζα μπορεί να αποφασίσει για τις πολιτικές δανειοδότησης βάσει των σχετικά χαμηλών κινδύνων.

Οι τεχνικές της συσταδοποίησης και της ταξινόμησης βοηθούν τα χρηματοπιστωτικά ιδρύματα να ομαδοποιούν διάφορους πελάτες που έχουν κοινά χαρακτηριστικά. Η αποτελεσματική συσταδοποίηση και οι μέθοδοι φιλτραρίσματος βοηθούν τις τράπεζες να ταυτοποιούν μία ομάδα πελατών, να συσχετίζουν ένα νέο πελάτη με την παρούσα ομάδα και να τους παρέχουν κοινά οφέλη.

3. Οι Τηλεπικοινωνίες

Η τηλεπικοινωνιακή βιομηχανία αναπτύσσεται πολύ γρήγορα, όπως και η τεχνολογία. Αυτές τις μέρες οι τηλεπικοινωνιακές υπηρεσίες έχουν επεκταθεί από τοπικές και μεγάλης απόστασης τηλεπικοινωνίες, σε συσκευές τηλεειδοποίησης, κινητό τηλέφωνο, και ηλεκτρονικό ταχυδρομείο.

Εξαιτίας των εξελίξεων στις τηλεπικοινωνιακές τεχνολογίες και για να δουλέψουν αποτελεσματικά αυτές οι τεχνολογίες, οι

τεχνικές της ανάλυσης και εξόρυξης δεδομένων ενσωματώνονται σε αυτές τις τεχνολογίες για να παράγουν αποδοτικά αποτελέσματα.

Η ανάλυση δεδομένων βοηθάει στην διάκριση τηλεπικοινωνιακών προτύπων, στην καταπολέμηση παράνομων δραστηριοτήτων και επίσης βοηθάει στην καλύτερη χρήση των πόρων και στη βελτίωση της ποιότητας των υπηρεσιών.

Η εξόρυξη δεδομένων βελτιώνει τις τηλεπικοινωνιακές υπηρεσίες με την εξής διαδικασία :

- Τα τηλεπικοινωνιακά δεδομένα που συλλέγονται, περιλαμβάνουν τον τύπο κλήσης, την τοποθεσία του καλούντος και του κληθέντος, τον χρόνο κλήσης, την διάρκεια κλήσης κλπ.
- Η πολυδιάστατη ανάλυση βοηθά στον προσδιορισμό και στην σύγκριση του φορτιού του συστήματος, κίνηση δεδομένων, και κέρδος κλπ.
- Η ανάλυση μπορεί να δείξει διαγράμματα και γράφους των πόρων του συστήματος, του προορισμού κλπ κάνοντας χρήση των εργαλείων οπτικοποίησης της εξόρυξης δεδομένων.

Τέτοια εργαλεία όπως η συσχετισμένη οπτικοποίηση και η συσταδοποίηση παρέχουν χρήσιμες υπηρεσίες στην ανάλυση των δεδομένων τηλεπικοινωνίας.

Το κυρίως πρόβλημα που αντιμετωπίστηκε από την βιομηχανία τηλεπικοινωνιών είναι οι παράνομες δραστηριότητες. Αυτές οι δραστηριότητες μπορεί να έχουν να κάνουν με σκόπιμες κλήσεις κατά την ώρα αιχμής, περιοδικές κλήσεις κ.α. με αποτέλεσμα να επιδρούν αρνητικά στην επίδοση του δικτύου επικοινωνιών.

Μέθοδοι όπως η συσταδοποίηση και η ανάλυση ακραίων τιμών, συνεισφέρει στην ανίχνευση παράνομων προτύπων βελτιώνοντας την αποτελεσματικότητα των υπηρεσιών τηλεπικοινωνίας. Εκμεταλλευόμενοι τα εργαλεία της ανάλυσης και

εξόρυξης δεδομένων είναι δυνατή η δημιουργία προφίλ των πελατών και ο εντοπισμός βλαβών στο δίκτυο.

(Γούλου Ζ., 2010)

4. Εκπαίδευση

Τα τελευταία χρόνια, υπάρχει ένα αυξανόμενο ενδιαφέρον στη χρήση της εξόρυξης δεδομένων για τη διερεύνηση επιστημονικών ερωτημάτων μέσα στην εκπαιδευτική έρευνα. Αυτό το πεδίο ονομάζεται EDM (Educational Data Mining) και ορίζεται ως ο τομέας της επιστημονικής έρευνας γύρω από την ανάπτυξη μεθόδων ώστε να γίνουν ανακαλύψεις μέσα στα μοναδικά είδη δεδομένων που προέρχονται από εκπαιδευτικές τοποθεσίες.

Χρησιμοποιώντας αυτές τις μεθόδους υπάρχει καλύτερη κατανόηση των μαθητών και των πλατφορμών στις οποίες μαθαίνουν. Για παράδειγμα, στην εξόρυξη δεδομένων σχετικά με το πώς οι μαθητές επιλέγουν να χρησιμοποιούν εκπαιδευτικό λογισμικό, μπορεί να αξίζει να εξεταστούν ταυτόχρονα τα δεδομένα στο επίπεδο της πληκτρολόγησης, στο επίπεδο απάντησης, στο επίπεδο συνεδρίας, στο επίπεδο σπουδαστών, στο επίπεδο τάξης και στο επίπεδο σχολείου. (Baker, 2008)

Οι νέες εφαρμογές EDM θα επικεντρωθούν στο να επιτρέπουν στους μη τεχνικούς χρήστες να χρησιμοποιούν και να συμμετέχουν σε εργαλεία και δραστηριότητες εξόρυξης δεδομένων, καθιστώντας τη συλλογή και επεξεργασία δεδομένων πιο προσιτή σε όλους τους χρήστες EDM. Παραδείγματα περιλαμβάνουν εργαλεία στατιστικής και οπτικοποίησης που αναλύουν τα κοινωνικά δίκτυα και την επιρροή τους στα μαθησιακά αποτελέσματα και την παραγωγικότητα.[6] Για την ανάλυση δεδομένων η εισαγωγή των δεδομένων στην Python μπορεί να γίνει από αρχείο (πχ. xls, csv, txt) και η επεξεργασία τους μπορεί να γίνει με ποικίλους τρόπους που υποστηρίζει η βιβλιοθήκη pandas.

[6] https://en.wikipedia.org/wiki/Educational_data_mining (ανακτήθηκε 5/1/2021)

2.2 Ανάλυση δεδομένων σε σχέση με το process mining

Τα αποτελέσματα της εξόρυξης διεργασιών μπορούν να θεωρηθούν ως ακτινογραφίες που αποκαλύπτουν τι πραγματικά συμβαίνει μέσα στις διαδικασίες και μπορούν να χρησιμοποιηθούν για τη διάγνωση προβλημάτων και την πρόταση κατάλληλων μεθόδων διόρθωσης αυτών.

Σύμφωνα με τον Van Der Aalst (2012), «το σημείο εκκίνησης για την εξόρυξη διεργασιών είναι ένα αρχείο καταγραφής συμβάντων, στο οποίο κάθε συμβάν αναφέρεται σε μια δραστηριότητα, ή ένα καλά καθορισμένο βήμα σε κάποια διαδικασία και σχετίζεται με μια συγκεκριμένη περίπτωση ή διαδικασία. Τα συμβάντα που ανήκουν σε μια υπόθεση ταξινομούνται και μπορούν να θεωρηθούν ως ένα «τρέξιμο» της διαδικασίας. Τα αρχεία καταγραφής συμβάντων ενδέχεται επίσης να αποθηκεύουν πρόσθετες πληροφορίες σχετικά με συμβάντα, όπως τον πόρο, το άτομο ή τη συσκευή που εκτελεί ή ξεκινά τη δραστηριότητα και τη χρονική σήμανση του συμβάντος».

Η εξόρυξη διεργασιών έχει σχέση με τη συνεχιζόμενη εκθετική αύξηση του όγκου δεδομένων-γεγονότων. Για παράδειγμα, σύμφωνα με το McKinsey Global Institute [7] το 2010 οι επιχειρήσεις αποθήκευσαν περισσότερα από 7 exabytes νέων δεδομένων σε δίσκους, ενώ οι καταναλωτές αποθήκευσαν περισσότερα από 6 exabytes νέων δεδομένων σε συσκευές, όπως υπολογιστές και φορητούς υπολογιστές.

Η ανάλυση δεδομένων βελτιώνει την απόδοση κι επειδή τα περισσότερα αρχεία καταγραφής συμβάντων περιέχουν χρονικές σημάνσεις, η επανάληψη μπορεί να χρησιμοποιηθεί για την επέκταση του μοντέλου με πληροφορίες απόδοσης. Εμφανίζει τη μεταβλητότητα μιας διαδικασίας και βοηθά στη βελτίωση της απόδοσης γιατί δεν παρουσιάζει εξειδανικευμένες καταστάσεις,

[7] http://www.mckinsey.com/mgi

αλλά τη μεταβλητότητα στην πάροδο του χρόνου σε πραγματικές συνθήκες.

Επιπρόσθετα, η εξόρυξη διεργασιών μπορεί να συμβάλει στη βελτίωση της αξιοπιστίας των συστημάτων και των διαδικασιών και την πρόβλεψη προβλημάτων πριν αυτά εμφανιστούν. Για παράδειγμα, οι κανονισμοί σε διάφορες χώρες απαιτούν δοκιμαστικά συστήματα υπό ρεαλιστικές συνθήκες. Η Philips Healthcare χρησιμοποίησε την εξόρυξη διεργασιών για διάγνωση σφαλμάτων για να εντοπίσει πιθανές αστοχίες στα συστήματα ακτίνων Χ.

Μαθαίνοντας από προηγούμενη αστοχία του συστήματος, η διάγνωση σφαλμάτων μπόρεσε να βρει τη βασική αιτία για νέα αναδυόμενα προβλήματα και να προβλέψει ότι ένας σωλήνας ακτίνων Χ στο πεδίο πρόκειται να αποτύχει ανακαλύπτοντας μοτίβα βλαβών στα αρχεία καταγραφής συμβάντων του μηχανήματος, με αποτέλεσμα ο σωλήνας να αντικατασταθεί πριν το μηχάνημα αρχίσει να δυσλειτουργεί. (Van Der Aalst, 2012).

Πολλές πηγές δεδομένων σήμερα ενημερώνονται σε σχεδόν πραγματικό χρόνο και υπάρχει επαρκής υπολογιστική ισχύς για την ανάλυση των συμβάντων καθώς αυτά συμβαίνουν. (Van Der Aalst, 2011).

Η διαδικασία εξόρυξης μέσα από την ανάλυση δεδομένων στοχεύει στην αντιμετώπιση προβλημάτων με τη δημιουργία άμεσης σύνδεσης μεταξύ των υποθετικών μοντέλων που προκύπτουν και των πραγματικών συμβάντων. Οι τεχνικές ανάλυσης των δεδομένων επιτρέπουν την προβολή των διαφορετικών αποτελεσμάτων που προκύπτουν από διαφορετικές οπτικές γωνίες. (Van Der Aalst, 2010).

ΚΕΦΑΛΑΙΟ 3

3.1 Process mining – Εξόρυξη διαδικασιών

Με το πέρασμα του χρόνου, η πρόσβαση στο Internet έχει γίνει προσιτή σε ολοένα και περισσότερους ανθρώπους. Αυτό με τη σειρά του οδήγησε στο να αναπτυχθούν περισσότεροι ιστότοποι και να χρησιμοποιηθούν βάσεις δεδομένων για την αποθήκευση των δεδομένων. Με τη δημιουργία εμπορικών και κοινωνικών ιστοσελίδων υπήρξαν τα πρώτα άλματα στις απαιτήσεις και ανάγκες για αποθήκευση και διαχείριση μεγάλου όγκου δεδομένων.

Σήμερα, το πλήθος των διαθέσιμων δεδομένων είναι τεράστιο και αυξάνεται εκθετικά κάθε μέρα. Η μείωση στο κόστος συλλογής και της δυσκολίας στη συλλογή και αποθήκευση των δεδομένων συνετέλεσε σημαντικά στην ανάπτυξη του πεδίου αυτού.

Ο τεράστιος όγκος δεδομένων, που συσσωρεύεται στις βάσεις δεδομένων και στις αποθήκες δεδομένων (data warehouses), δεν μπορεί να αξιοποιηθεί όπως είναι. Πρέπει αρχικά να γίνουν κάποιες ενέργειες για να δομηθούν κατάλληλα τα δεδομένα, ώστε στη συνέχεια να μπορούν να αξιοποιηθούν προς όποιο τομέα ενδιαφέροντος θέλουν οι επιστήμονες να κατευθυνθούν. (Fayyad, U, Piatetsky-Shapiro, G. & Smyth, P., 1996).

Η Επιστήμη των Δεδομένων (Data Science), είναι ένας καινούριος όρος, ο οποίος ήρθε να αντικαταστήσει προγενέστερους όρους, όπως Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (Knowledge Discovery in Database) ή Εξόρυξη Δεδομένων (Data Mining). Και οι τρεις αυτοί όροι χρησιμοποιούνται σχεδόν εναλλακτικά, για να περιγράψουν μία ημι-αυτοματοποιημένη διαδικασία, σκοπός της οποίας είναι να αναλύσει έναν μεγάλο όγκο δεδομένων που αφορούν ένα συγκεκριμένο πρόβλημα, συνήθως εμπορικού ή επιστημονικού ενδιαφέροντος, για την παραγωγή προτύπων (patterns), όπως συνηθίζεται να λέμε σε ορισμένους τομείς, όπως στη

Στατιστική, στη Μηχανική Μάθηση (Machine Learning) και στην Αναγνώριση Προτύπων (Pattern Recognition). [8]

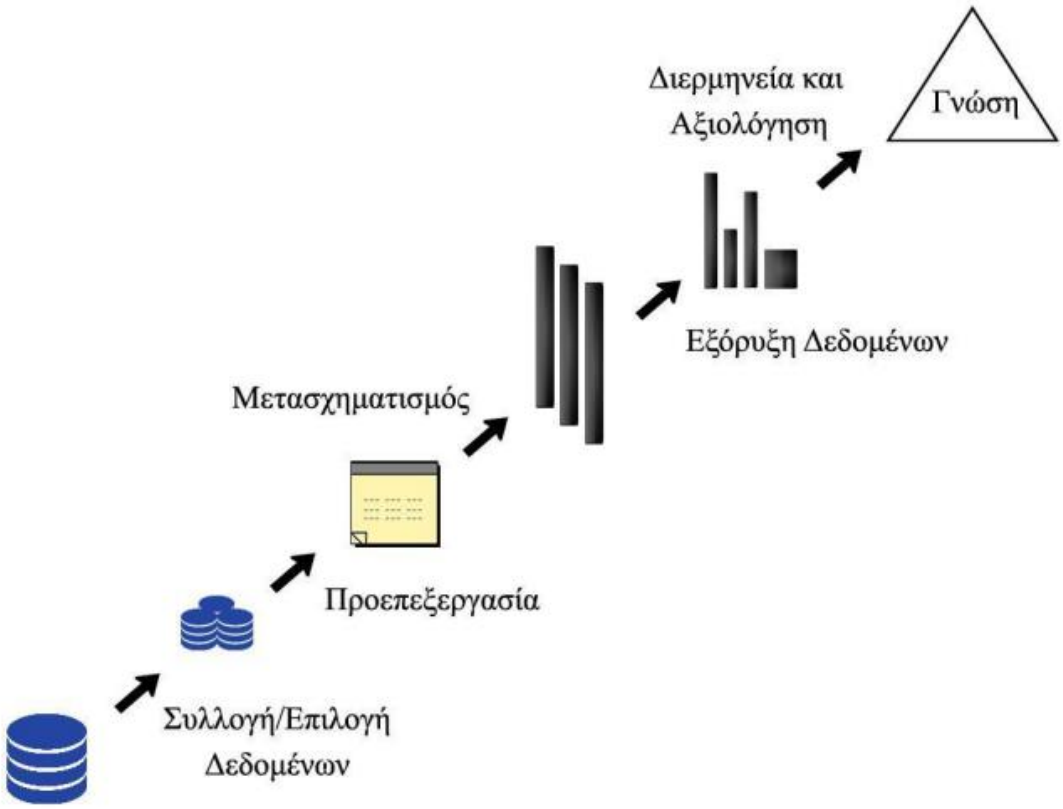
Η άνευ όρων παραγωγή δεδομένων σε εικοσιτετράωρη βάση καλύπτει μια τεράστια γκάμα ανθρώπινων δραστηριοτήτων και όχι μόνο, όπως είναι τα δεδομένα από το καλάθι αγορών, τον ιατρικό φάκελο του ασθενούς, τις συζητήσεις ή και ανακοινώσεις στα κοινωνικά μέσα δικτύωσης, τις τραπεζικές ή και χρηματιστηριακές συναλλαγές, τα ίχνη κινούμενων οχημάτων, τα δεδομένα αισθητήρων από κινητήρες αεροσκαφών, η καταγραφή συνομιλιών σε κέντρα εξυπηρέτησης πελατών κ.λπ.

Τα δεδομένα αυτά διαφέρουν πάρα πολύ μεταξύ τους τόσο σε μορφή (εικόνα, βίντεο, κείμενο, πολυδιάστατα ή πραγματικού χρόνου δεδομένα, ακολουθίες DNA και άλλα πολλά) όσο και στην ταχύτητα συλλογής. Εάν, μάλιστα, δεν υποστούν άμεση ανάλυση, ίσως να είναι ιδιαίτερα δύσκολο να αποθηκευτούν ή να τα επεξεργαστούν οι άνθρωποι, δημιουργώντας έτσι μία καινούρια ερευνητική δράση, γνωστή με τον όρο Μεγάλα Δεδομένα (Big Data).

Η Επιστήμη των Δεδομένων στοχεύει σε αυτή τη φάση να καλύψει τις ανάγκες που δημιουργούνται από αυτόν τον νέο τομέα και να προσφέρει λύσεις για την κλιμακούμενη και αποτελεσματική επεξεργασία out-of core (εκτός μνήμης ή εξωτερικής μνήμης) δεδομένων.

Ο τομέας της εξόρυξης γνώσης από δεδομένα σχετίζεται όμως και με πολλούς άλλους τομείς, όπως η Ανάκτηση Πληροφοριών (ΑΠ), οι Μηχανές Αναζήτησης (Search Engines), τα Συστήματα Υποστήριξης Αποφάσεων - ΣΥΑ (Decision Support Systems - DSS), οι Αποθήκες Δεδομένων (Data Warehouse), τα Συστήματα Άμεσης Ανάλυσης Δεδομένων (Online Analytical Processing - OLAP) και η Αναγνώριση Προτύπου (Pattern Recognition).

[8] https://en.wikipedia.org/wiki/Big_data (ανακτήθηκε 5/1/2021)



Εικόνα 1 : Βασικά στάδια Ανακάλυψης Γνώσης από Βάσεις Δεδομένων

Για την επίλυση προβλημάτων που αφορούν τους παραπάνω τομείς, εκτός από προγράμματα γραμμένα σε Python, θα μπορούσαν να χρησιμοποιηθούν και έτοιμα προγράμματα για process mining, όπως το ProM ή το Disco γιατί και τα δύο προσφέρουν ένα User Interface ιδανικό για process mining. Ο χρήστης μπορεί να κάνει import τα αρχεία δεδομένων του και να τα αναλύσει, επεξεργαστεί και αναπαραστήσει σε γράφους ή animations.

3.2 Big data

Τα δεδομένα (Big Data) έχουν τα ακόλουθα τέσσερα (4) χαρακτηριστικά :

i. Volume (όγκος δεδομένων)
ii. Velocity (ταχύτητα): Η ταχύτητα είναι μια μεγάλη πρόκληση για τους μηχανικούς που επεξεργάζονται τα δεδομένα καθώς δεν είναι μόνο τεράστιος ο όγκος των δεδομένων που αποτελεί πρόκληση, αλλά και το ότι παράγονται με μεγάλο ρυθμό και αλλάζουν πολύ γρήγορα.
iii. Variety (ποικιλία): Δεν υπάρχει μόνο ένα είδος δεδομένων, αλλά πάρα πολλά διαφορετικά είδη, τα οποία ποικίλουν από κείμενο σε εικόνες αλλά και πολλά άλλα είδη. Έτσι, είναι μεγάλη η πρόκληση του να συνδυαστούν όλα αυτά τα δεδομένα από εντελώς διαφορετικές πηγές.
iv. Veracity (εγκυρότητα): Αυτό σημαίνει ότι ποτέ δεν μπορούμε να είμαστε σίγουροι για την ακρίβεια των δεδομένων που έχουν καταγραφεί. Ένα παράδειγμα είναι ότι ποτέ δεν μπορούμε να είμαστε σίγουροι για το ποιος είναι ο χρήστης της συσκευής που παράγει τα δεδομένα, καθώς αυτά μπορεί να διαφοροποιούνται από χρήστη σε χρήστη.

Το αντικείμενο ενδιαφέροντος των Big Data είναι υπερβολικά επικεντρωμένο στα δεδομένα (data), στην αποθήκευση (storage) και στην επεξεργασία (processing) τους και όχι τόσο στις διαδικασίες (processes) που θα ήταν χρήσιμο για μια επιχείρηση να αναλυθούν και να βελτιωθούν.

Για αυτό ακριβώς το λόγο υπάρχει η εξόρυξη διαδικασιών (process mining), η οποία παρέχει μια καινούρια οπτική στα δεδομένα (Big or Small data) και παρέχει τα απαραίτητα εργαλεία για να ξεκινήσει η ανάλυση πραγματικών συμπεριφορών βασιζόμενοι σε πραγματικά γεγονότα (event data) που μπορούν να βρεθούν σε οποιαδήποτε επιχείρηση.

Η εξόρυξη διαδικασιών (process mining) είναι ο συνδετικός κρίκος ανάμεσα στην ανάλυση βασιζόμενη σε μοντέλα (model based analysis) και στην ανάλυση που επικεντρώνεται στα δεδομένα (data oriented analysis), όπως η εξόρυξη δεδομένων (data mining). Με την εξόρυξη διαδικασιών μπορούμε να απαντήσουμε ερωτήματα σχετικά με την απόδοση (performance) των διαδικασιών αλλά και ερωτήματα τα οποία σχετίζονται με την συμμόρφωση (compliance) πάνω σε κάποια μοντέλα. Είναι ο συνδετικός κρίκος ανάμεσα σε πολλά πράγματα και μπορεί να φανεί απίστευτα πολύτιμο.

Ως κοινό στοιχείο ανάμεσα στην εξόρυξη δεδομένων (data mining) και στην εξόρυξη διαδικασιών (process mining) είναι ότι η υλοποίησή τους ξεκινάει από τα δεδομένα (data), αλλά κατά τα άλλα υπάρχουν πολλές διαφορές ανάμεσα τους. Για τις διαδικασίες της εξόρυξης δεδομένων (data mining), οι γραμμές και οι στήλες στους πίνακες με τα δεδομένα μπορούν να σημαίνουν οτιδήποτε. Αντίθετα, στις τεχνικές της εξόρυξης διαδικασιών (process mining), υποθέτουμε ότι τα δεδομένα έχουν συγκεκριμένη μορφή και αναφέρονται σε activities σε συγκεκριμένες χρονικές στιγμές.

Επιπλέον, τα events είναι ταξινομημένα και για αυτό το λόγο αναφερόμαστε σε end-to-end διαδικασίες. Η βασική διαφορά τους είναι ότι η εξόρυξη δεδομένων (data mining) έχει σαν κέντρο τα δεδομένα και όχι τις διαδικασίες. Έτσι, γίνεται κατανοητό ότι η ανακάλυψη διαδικασιών (process discovery), η συμμόρφωση σε ένα μοντέλο (conformance checking) και η ανάλυση για την εύρεση κάποιων σημείων όπου καθυστερεί η ομαλή λειτουργία ενός οργανισμού (bottlenecks discovery) δεν αποτελεί τμήμα της εξόρυξης δεδομένων (data mining), αλλά τμήμα της εξόρυξης διαδικασιών (process mining). Και φυσικά τίποτα από αυτά δεν μπορεί να επιτευχθεί χρησιμοποιώντας την εξόρυξη δεδομένων (data mining).

Η εξόρυξη διαδικασιών (process mining) συνδυάζει μοντέλα διαδικασιών (process models) και δεδομένα γεγονότων (event data) με πολλούς διαφορετικούς καινοτόμους τρόπους. Σαν αποτέλεσμα είναι ότι κάποιος χρησιμοποιώντας την εξόρυξη διαδικασιών

(process mining) μπορεί να ανακαλύψει τι κάνουν στην πραγματικότητα οι άνθρωποι και οι οργανισμοί.

Για παράδειγμα, μοντέλα διαδικασιών μπορούν να ανακαλυφθούν αυτόματα από event data (process model discovery). Η συμμόρφωση δεδομένων (compliance) μπορεί να ελεγχθεί συγκρίνοντας μοντέλα με event data. Διάφορα προβλήματα σε διαδικασίες μπορούν να ανακαλυφθούν ξανά εφαρμόζοντας διάφορα events πάνω σε μοντέλα που έχουν ήδη βρεθεί. Έτσι, η εξόρυξη διαδικασιών (process mining) μπορεί να χρησιμοποιηθεί για την ανακάλυψη και κατανόηση αποκλίσεων, ρίσκων αλλά και προβλημάτων στην καθημερινή λειτουργία μιας εταιρείας ή ενός οργανισμού.

Ένα τυπικό αρχείο γεγονότων καταγράφει τις εκτελούμενες δραστηριότητες μιας ροής εργασίας μαζί με περιπτώσεις (αρχικά στιγμιότυπα διαδικασίας) στις οποίες ανήκουν οι δραστηριότητες. Κάθε περίπτωση σε ένα αρχείο γεγονότων αποτελείται από δραστηριότητες που ταξινομούνται σύμφωνα με τους χρόνους εκτέλεσής τους. [9]

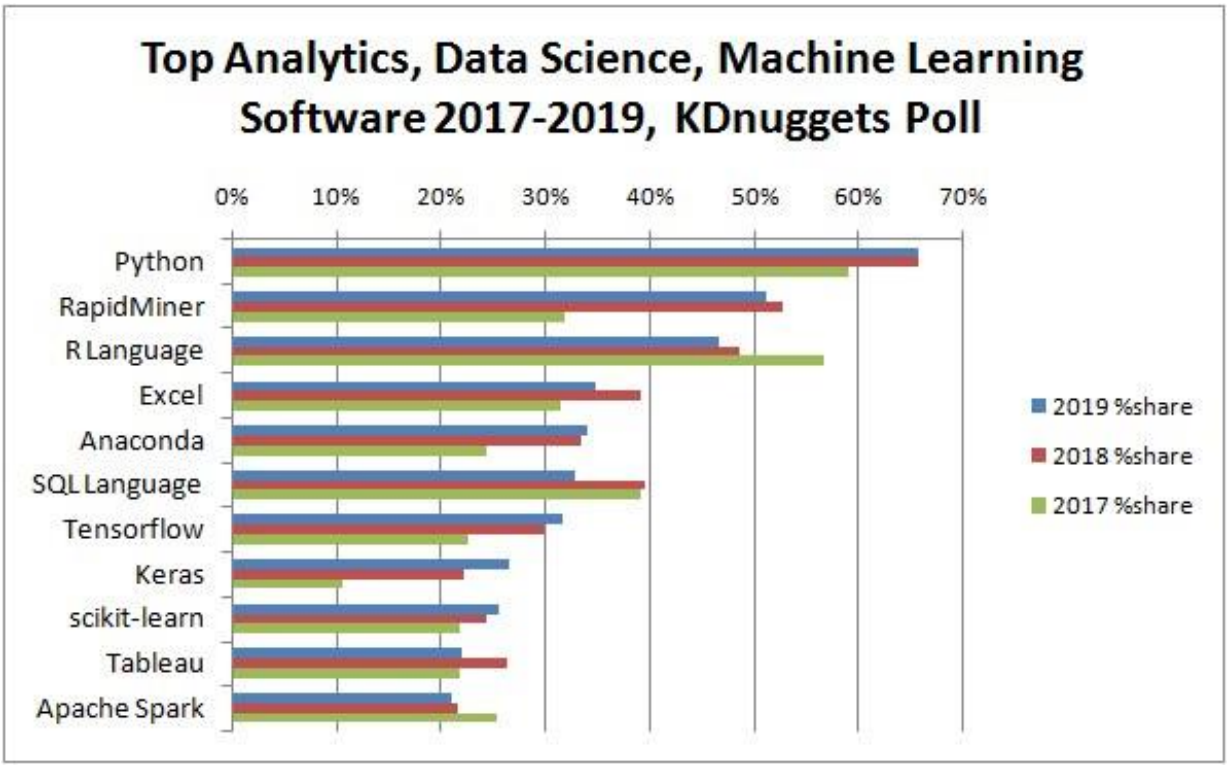
Η όλη επεξεργασία των δεδομένων μέχρι να καταλήξουμε σε συμπεράσματα μπορεί να γίνει με διάφορους τρόπους, χρησιμοποιώντας διάφορες γλώσσες προγραμματισμού ή άλλα λογισμικά.

Σύμφωνα με έρευνα[10], η Python συνεχίζει να ηγείται των κορυφαίων πλατφορμών Επιστήμης Δεδομένων, αλλά οι R και RapidMiner διατηρούν το μερίδιό τους. Σχεδόν το 50% έχουν χρησιμοποιήσει εργαλεία Deep Learning. Η SQL είναι σταθερή. Η ενοποίηση συνεχίζεται.

[9] http://ikee.lib.auth.gr/record/270228/files/GRI-2015-14856.pdf (p.77, ανακτήθηκε 9/4/2021)
[10] Python leads the 11 top Data Science, Machine Learning platforms: Trends and Analysis (kdnuggets.com) (ανακτήθηκε 28/2/2021)

Η 20η ετήσια δημοσκόπηση λογισμικού KDnuggets είχε περισσότερους από 1.800 συμμετέχοντες. Η έρευνα αυτή λοιπόν, κατέληξε για το 2019 στα παρακάτω αποτελέσματα:

Η Python παρέμεινε στην κορυφή, με σχεδόν το ίδιο μερίδιο (65,8% έναντι 65,6%) των ερωτηθέντων με το 2018. Η RapidMiner διατήρησε το μερίδιό της στο 51% περίπου, γεγονός που αντικατοπτρίζει τόσο μια μεγάλη βάση χρηστών όσο και μια επιτυχημένη καμπάνια για να παρακινήσει τους χρήστες της. Το ποσοστό της γλώσσας R παρουσιάζει μείωση τα 2 τελευταία χρόνια, αλλά λιγότερο φέτος σε σχέση με το προηγούμενο έτος. Η SQL είναι σταθερή, με μερίδιο άνω του 30% για πολλά χρόνια.



Εικόνα 2 : Python έναντι άλλων λογισμικών για ανάλυση δεδομένων

3.3 Event logs

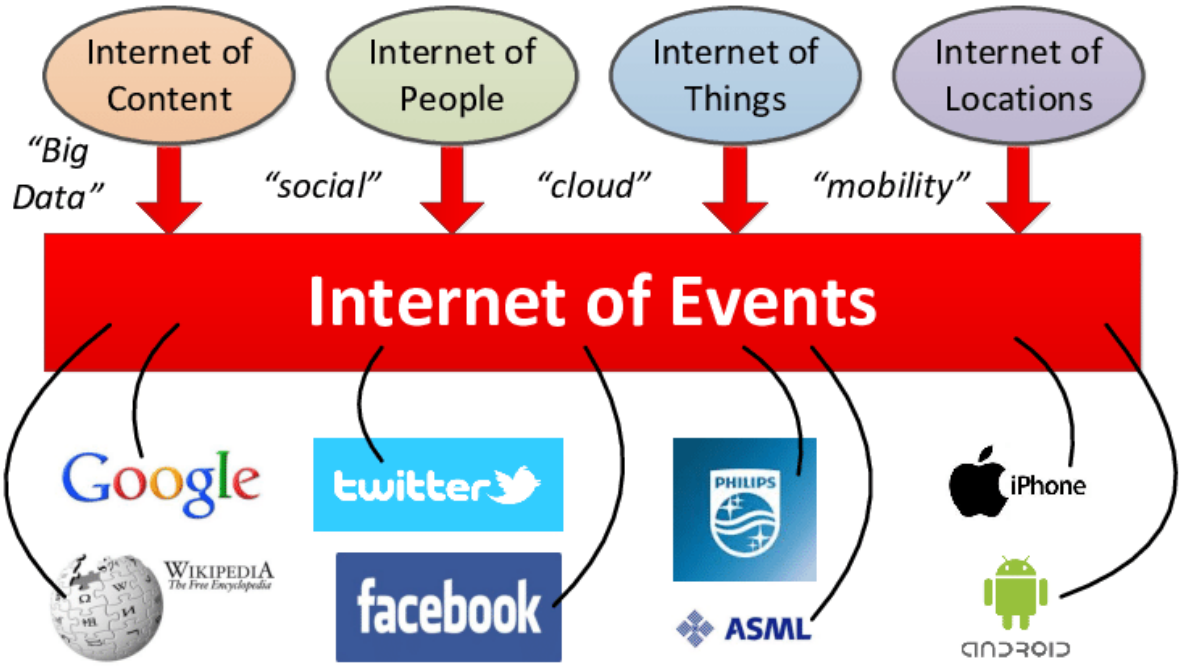
Η εξόρυξη διαδικασιών (process mining) αποτελεί έναν συνδετικό κρίκο ανάμεσα στην εξόρυξη δεδομένων (data mining) και στην διαχείριση διαδικασιών σε επιχειρήσεις (business process management). Συγκεκριμένα αποτελεί μια οικογένεια από τεχνικές που υποστηρίζουν την ανάλυση των διαδικασιών βασισμένες σε διάφορα σύνολα γεγονότων (event logs). Σκοπός της εξόρυξης διαδικασιών (process mining) είναι η κατανόηση αλλά και η βελτίωση της απόδοσης των διαδικασιών μιας επιχείρησης.

Η πιο σημαντική πηγή δεδομένων είναι τα δεδομένα που διακινούνται μέσω του internet. (internet of events).

Υπάρχουν 4 είδη τέτοιων δεδομένων :

1. Internet of content: Στη συγκεκριμένη κατηγορία ανήκουν ιστοσελίδες από το κλασικό internet, όπως είναι γνωστό σε όλους, δηλαδή Google και Wikipedia.
2. Internet of people: Στη συγκεκριμένη κατηγορία ανήκουν τα social media, όπως το Twitter και το Facebook, τα οποία παράγουν τεράστιο όγκο δεδομένων.
3. Internet of things: Στη συγκεκριμένη κατηγορία ανήκουν τα δεδομζνα που παράγονται από διάφορες συσκευές που είναι συνδεδεμένες στο Internet όπως ο εκτυπωτής, η τηλεόραση, οι παιχνιδομηχανές. Στο μέλλον όλο και περισσότερες συσκευές θα συνδέονται στο Internet, όπως το ψυγείο και το πλυντήριο και αυτό θα έχει σαν αποτέλεσμα την παραγωγή ακόμα μεγαλύτερου όγκου δεδομένων.
4. Internet of places: Όταν πχ ο χρήστης χρησιμοποιεί το κινητό του τηλέφωνο, το οποίο έχει πάνω του διάφορους αισθητήρες τοποθεσίας και καταγράφει την θέση του. Αυτό αποτελεί μια πηγή δεδομένων.

(Leon Zhao J.,etc, 2015)



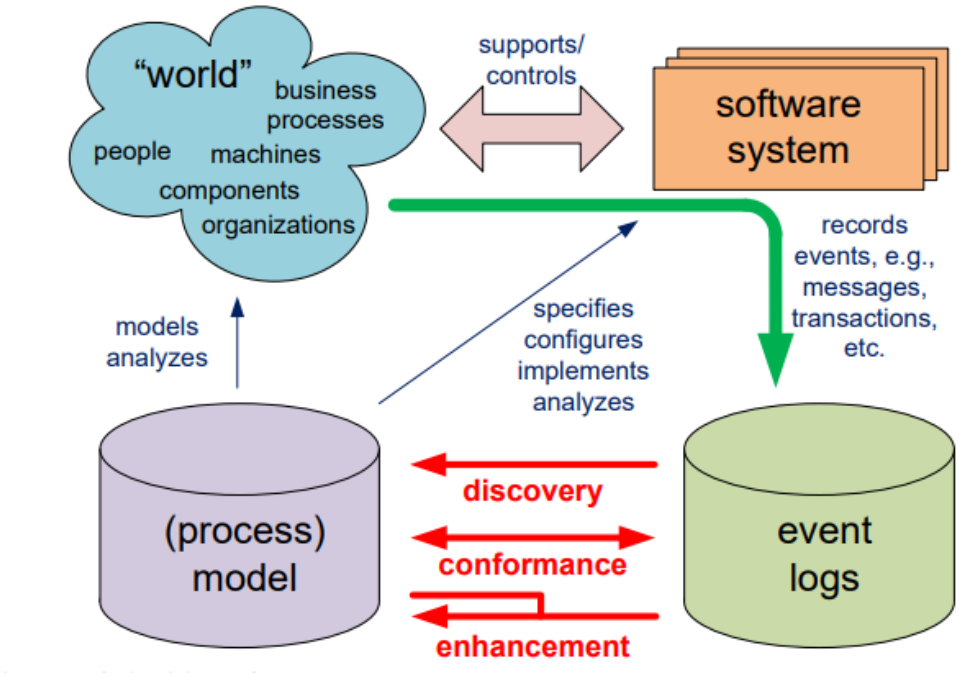
Εικόνα 3 : Πηγές παραγωγής δεδομένων

Ένα αρχείο καταγραφής συμβάντων είναι ένας βασικός πόρος που βοηθά στην παροχή πληροφοριών. Είναι ένα βασικό "βιβλίο καταγραφής" που μπορεί να περιέχει πολλούς διαφορετικούς τύπους πληροφοριών. Ένα event log λοιπόν, αποθηκεύει αυτές τις πληροφορίες για ανάκτηση και περαιτέρω επεξεργασία από διάφορα λογισμικά που παρέχουν εργαλεία για την ανάλυση των περιεχομένων του σε υψηλό επίπεδο και μπορούν τελικά να προσφέρουν βοήθεια στους ενδιαφερόμενους ώστε να προσδιορίσουν τι ακριβώς συμβαίνει στην συγκεκριμένη κατάσταση.

Μπορούμε να έχουμε αρχεία καταγραφής συμβάντων σε κάθε συμβάν που αναφέρεται σε περίπτωση, δραστηριότητα ή και σημείο στο χρόνο. Έτσι :

• Ένα αρχείο καταγραφής συμβάντων μπορεί να θεωρηθεί ως συλλογή από υποθέσεις / περιπτώσεις.

• Μια υπόθεση / περίπτωση μπορεί να θεωρηθεί ως μία ακολουθία από γεγονότα. [11]



Εικόνα 4 : Επιλέγοντας τα σωστά δεδομένα

Αρχεία καταγραφής γεγονότων μπορούμε να ανακτήσουμε από παντού. Για παράδειγμα μπορούμε να πάρουμε event log από :

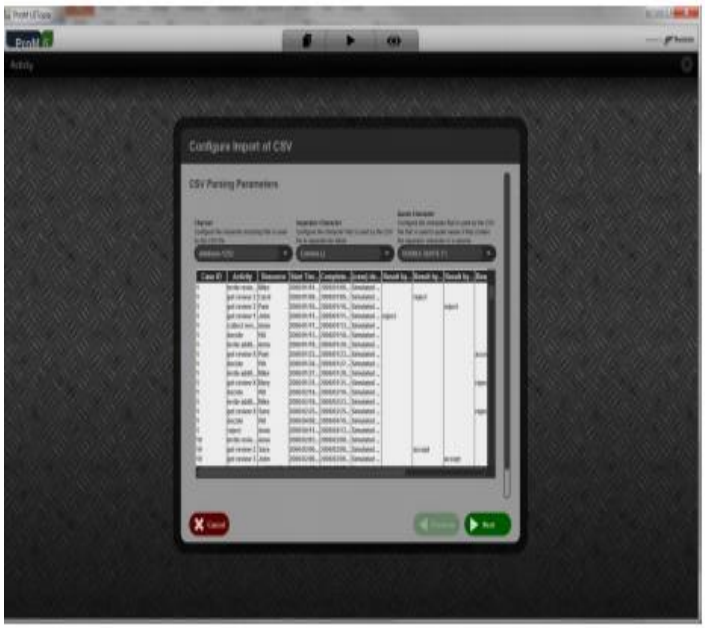
- ένα σύστημα βάσης δεδομένων (π.χ. δεδομένα ασθενών σε νοσοκομείο),
- ένα υπολογιστικό φύλλο με τιμές διαχωρισμένες με κόμμα (CSV),
- ένα αρχείο καταγραφής συναλλαγών (π.χ. ένα σύστημα συναλλαγών),
- ένα επιχειρηματικό σύστημα ERP (SAP, Oracle, κ.α.),
- ένα αρχείο καταγραφής μηνυμάτων (π.χ. από το IBM middleware),

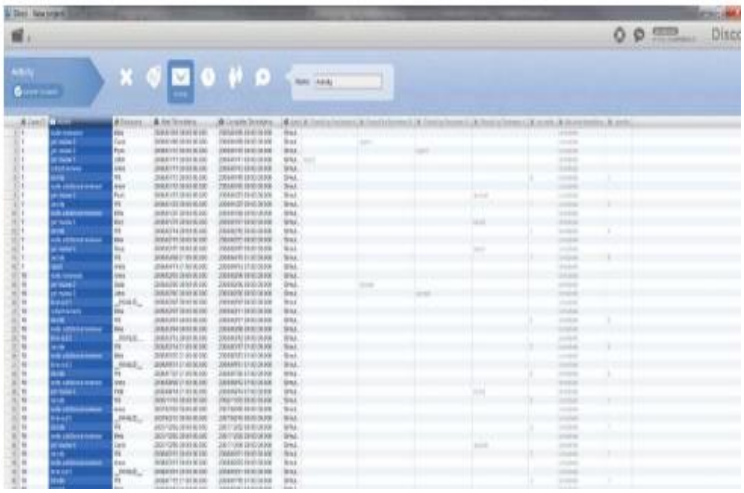
[11] http://www.processmining.org/_media/presentations/event_logs_the_input_for_process_mining.pdf (ανακτήθηκε στις 26/1/2021)

- ένα open API που παρέχει δεδομένα από ιστότοπους, ή κοινωνικά δίκτυα, ή ΜΜΕ.

Η πρόκληση είναι η ανακάλυψη των διαδικασιών σε μια επιχείρηση, η αναπαράσταση τους με μοντέλα και η υποστήριξη σε ολόκληρο το φάσμα του process mining, από τα δεδομένα που λαμβάνονται από τα πληροφοριακά συστήματα των επιχειρήσεων και των οργανισμών και τα αρχεία γεγονότων μέχρι τα μοντέλα που δημιουργούνται. [12]

Οι περισσότερες επιχειρήσεις και οργανισμοί, διαθέτουν δεδομένα γεγονότων "κρυμμένα" στα πληροφοριακά τους συστήματα. Από τη στιγμή που βρεθούν τα δεδομένα η μετατροπή τους και η επεξεργασία τους μπορεί να γίνει εύκολα με λογισμικά όπως το ProM και το Disco που υποστηρίζουν τη μετατροπή αρχείων από διάφορες μορφές σε μορφή *.mmxl και *.xes.





[12] http://ikee.lib.auth.gr/record/270228/files/GRI-2015-14856.pdf (p.84, ανακτήθηκε 9/4/2021)

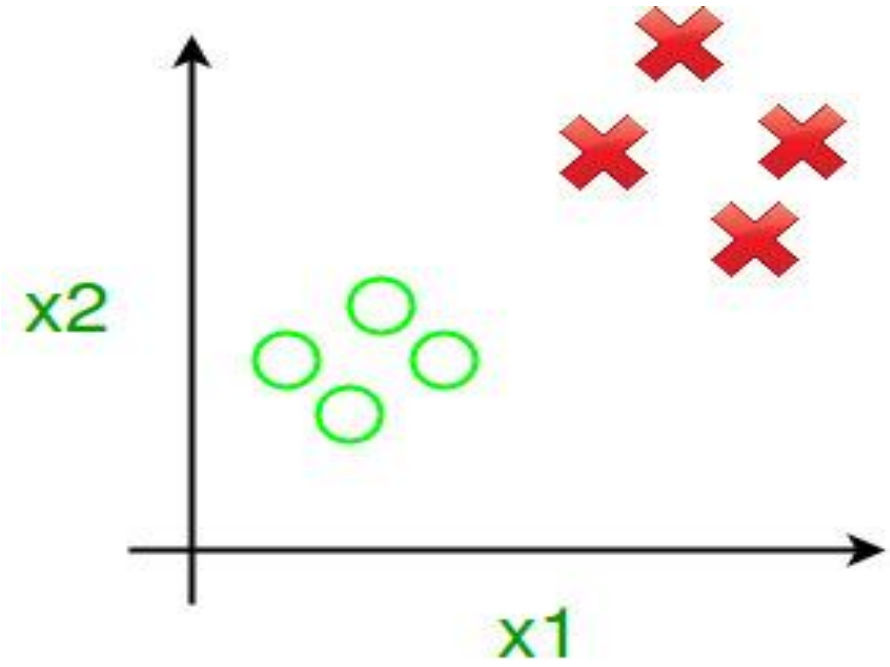
3.4 Μέθοδοι Εξόρυξης Δεδομένων

Υπάρχει μια μεγάλη ποικιλία μεθόδων εξόρυξης δεδομένων. Ανάλογα με το είδος των δεδομένων και το είδος της γνώσης που εξάγεται, αυτές κατατάσσονται σε διαφορετική κατηγορία.

Μερικές βασικές μέθοδοι της Εξόρυξης Δεδομένων παρουσιάζονται παρακάτω.

1. **<u>Κατηγοριοποίηση</u>**

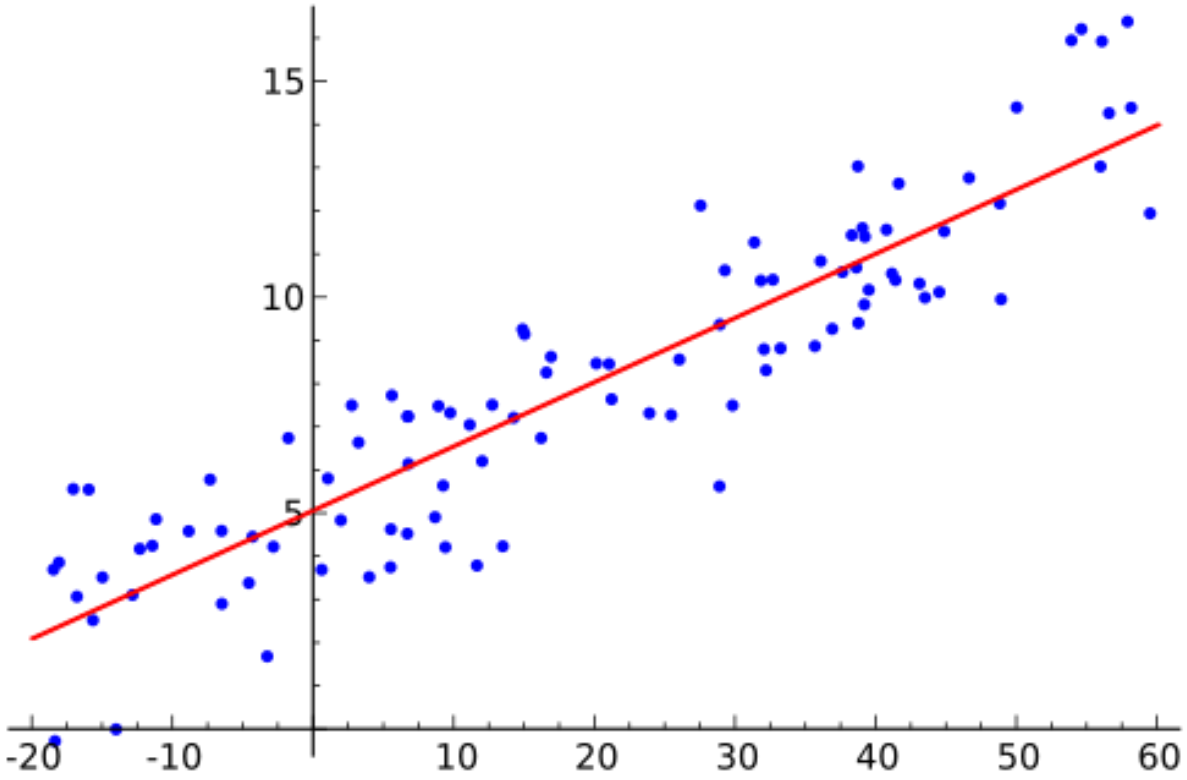
Πρόκειται για μια προγνωστική μέθοδο. Στόχος είναι η δημιουργία ενός μοντέλου – κατηγοριοποιητή (classifier) με βάση τα υπάρχοντα δεδομένα. Στην κατηγοριοποίηση, το αποτέλεσμα που θέλουμε να προβλέψουμε είναι η κλάση των δειγμάτων. Η κλάση μπορεί να πάρει διακριτές τιμές από ένα πεπερασμένο σύνολο.



Εικόνα 5 : πχ κατηγοριοποίησης

2. **Παλινδρόμηση**

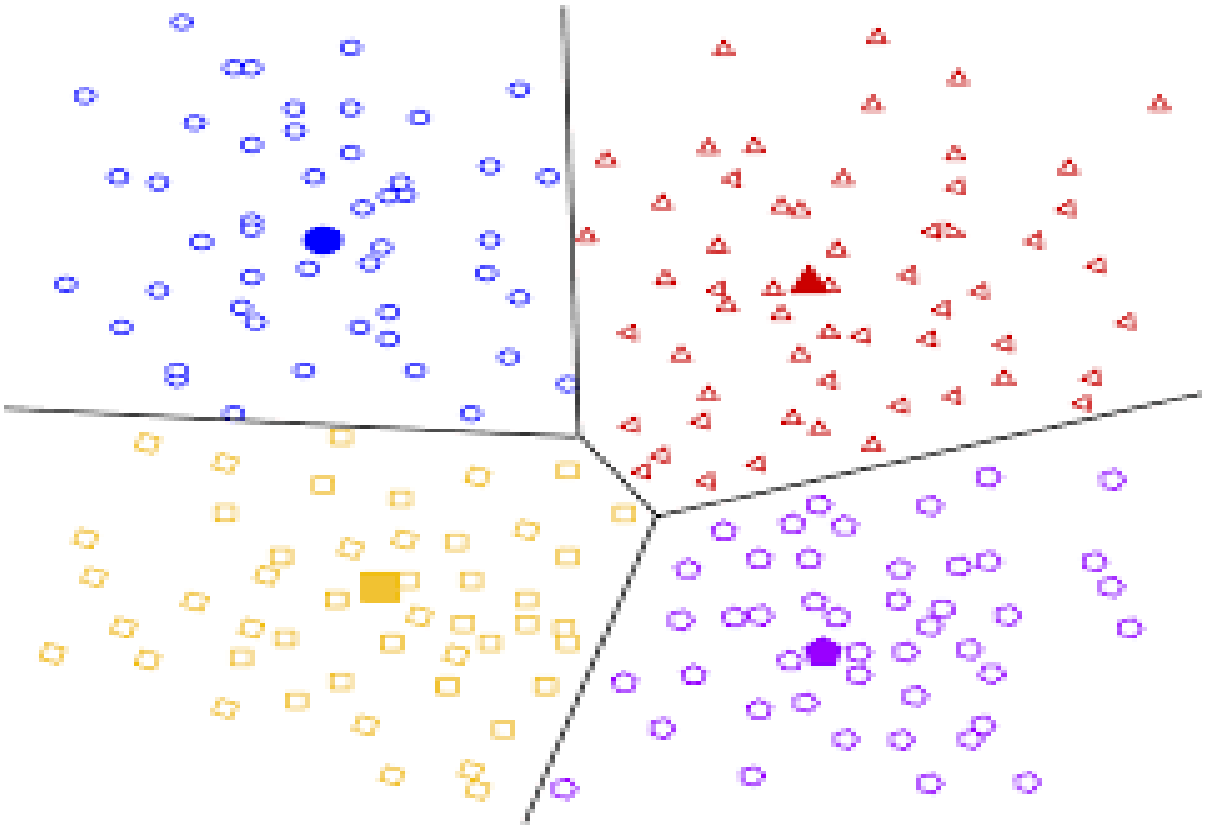
Μια σχετική διαδικασία με την κατηγοριοποίηση είναι η παλινδρόμηση (regression), στόχος της οποίας είναι η εκπαίδευση (training) μιας συνάρτησης, η οποία απεικονίζει ένα αντικείμενο σε μία πραγματική μεταβλητή. Στόχος είναι με βάση κάποιες ανεξάρτητες μεταβλητές (independent variables) να προβλεφθούν οι τιμές μιας εξαρτημένης μεταβλητής (dependent variable).



Εικόνα 6 : Γραμμική Παλινδρόμηση

3. **Συσταδοποίηση**

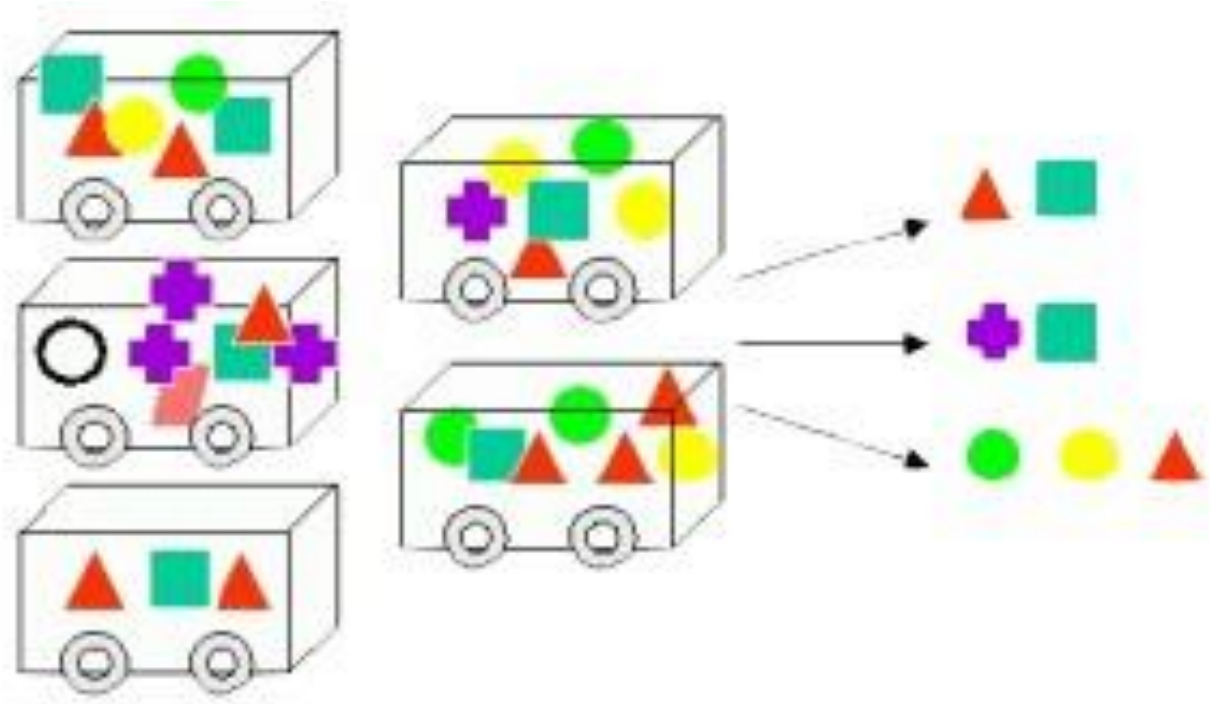
Η συσταδοποίηση (clustering) είναι μια περιγραφική μέθοδος. Έχοντας ένα σύνολο δεδομένων, στόχος της συσταδοποίησης είναι η δημιουργία συστάδων (clusters), δηλαδή ομάδων, οι οποίες θα περιέχουν όμοια ή παρεμφερή δείγματα.



Εικόνα 7 : Συσταδοποίηση (clustering)

4. Εξαγωγή και Ανάλυση Συσχετίσεων

Η εξαγωγή κανόνων συσχέτισης (Mining Association Rules) θεωρείται μια από τις σημαντικότερες διεργασίες εξόρυξης δεδομένων. Οι κανόνες συσχέτισης ανακαλύπτουν κρυμμένες «συσχετίσεις» μεταξύ των γνωρισμάτων ενός συνόλου των δεδομένων. Αυτοί οι συσχετισμοί παρουσιάζονται στη μορφή A ◊ B, όπου τα A και B αποτελούν σύνολα που αναφέρονται στα χαρακτηριστικά του συνόλου δεδομένων που αναλύουμε. Δεδομένου ενός συνόλου από δεδομένα, ένας κανόνας συσχέτισης A ◊B προβλέπει την εμφάνιση των χαρακτηριστικών του συνόλου B δεδομένης της εμφάνισης των χαρακτηριστικών του συνόλου A.

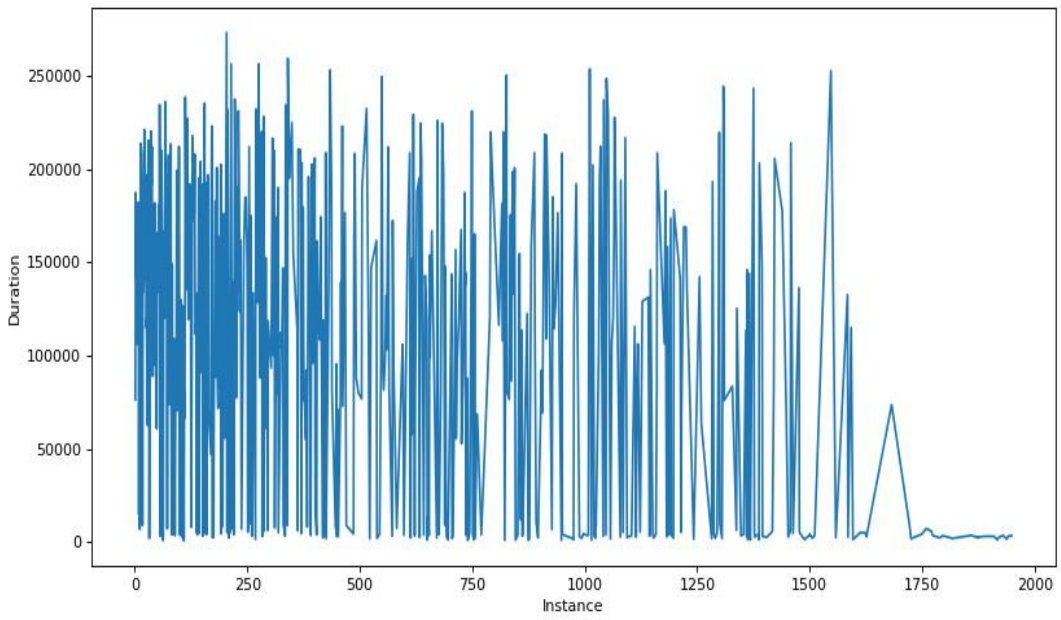


Εικόνα 8 : Data mining rules

5. **Οπτικοποίηση**

Η οπτικοποίηση των δεδομένων συχνά βοηθάει στην καλύτερη κατανόηση όχι μόνο των ίδιων των δεδομένων, αλλά και των συσχετίσεων που μπορεί να υπάρχουν μεταξύ τους. Ωστόσο, οπτικοποίηση μπορεί να γίνει μόνο για συγκεκριμένο αριθμό διαστάσεων. Αυτό σημαίνει ότι για σύνολα δεδομένων με πολλά χαρακτηριστικά, η οπτικοποίηση τους είναι ανέφικτη ή εναλλακτικά αρκούμαστε στην οπτικοποίηση ενός μέρους αυτών.

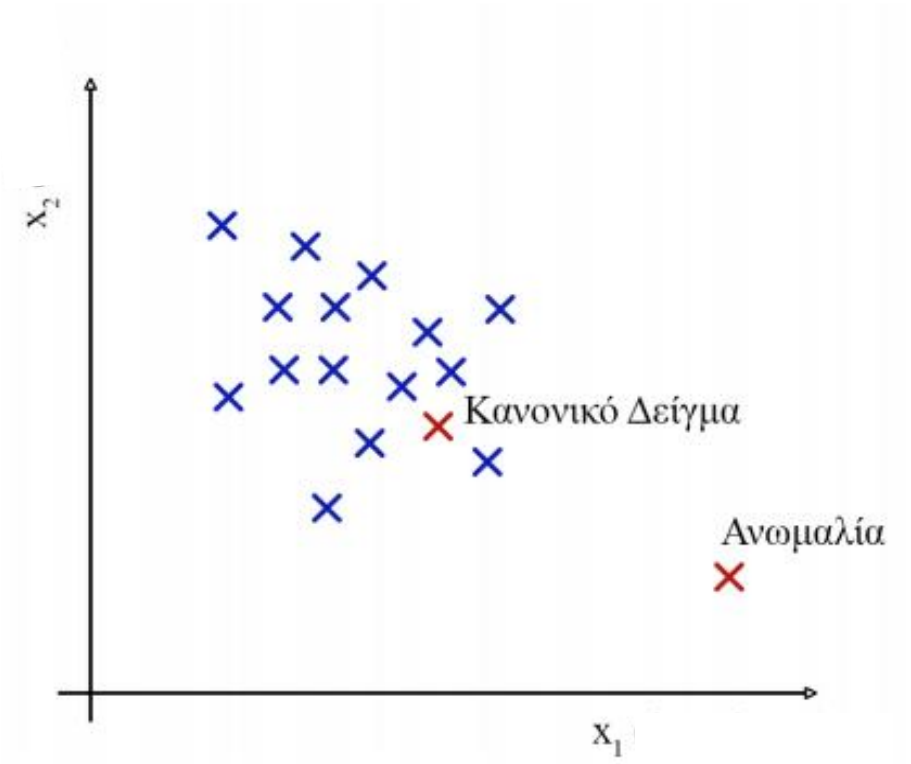
Σε κάθε περίπτωση, οι οπτικοποιήσεις θα πρέπει να συνοδεύονται και από τους αντίστοιχους στατιστικούς ελέγχους, προκειμένου να βεβαιωθούμε για την εγκυρότητα των συσχετίσεων που απεικονίζονται.



Εικόνα 9 : Από την οπτικοποίηση δεδομένων της εργασίας

6. Ανίχνευση Ανωμαλιών

Η ανίχνευση ανωμαλιών εστιάζει στην ανακάλυψη αποκλίσεων στα δεδομένα σε σχέση με αντίστοιχα δεδομένα, τα οποία έχουν συλλεχθεί στο παρελθόν ή με τυπικές τιμές των δεδομένων αυτών.



Εικόνα 10: Ανίχνευση ανωμαλιών

Οι περιορισμοί που εμφανίζουν αυτές οι μέθοδοι μας έχουν οδηγήσει στην εκμετάλλευση των εργαλείων της εξόρυξης για καλύτερα και περισσότερο αξιόπιστα αποτελέσματα.

ΚΕΦΑΛΑΙΟ 4

4.1 Οπτικοποίηση δεδομένων

Ως Οπτικοποίηση δεδομένων (Data Visualization) ορίζεται ουσιαστικά η οπτική αναπαράσταση δεδομένων, τα οποία έχουν εξαχθεί είτε από κείμενο είτε από άλλου τύπου πηγές, με τη μορφή σχηματικών δομών και γραφημάτων. «Κύριος σκοπός τη μεθόδου αυτής είναι, να διαδώσει πληροφορία ξεκάθαρα και αποτελεσματικά με τη χρήση γραφικών μέσων» (Friedman, 2008).

Ιστορικά η οπτικοποίηση δεδομένων είχε αρχίσει να απασχολεί τους επιστήμονες εδώ και αρκετά χρόνια χωρίς όμως να μπορούν οι προτεινόμενες μέθοδοι να εφαρμοστούν στην πράξη λόγω έλλειψης τεχνολογικών μέσων. Μετά τη δεκαετία του 1980 με την ανάπτυξη των προσωπικών ηλεκτρονικών υπολογιστών δημιουργήθηκαν οι κατάλληλες συνθήκες για την εφαρμογή τεχνικών οπτικοποίησης και γραφικής αναπαράστασης δεδομένων. (Kehrer J., Hauser H, 2013).

Οι πιο τυπικές τεχνικές ανάλυσης πληροφορίας μέσω της οπτικοποίησης είναι τα ιστογράμματα, γραφικές παραστάσεις, δεντρικά διαγράμματα κ.ά., μέσω των οποίων μπορεί να γίνει εξαγωγή περεταίρω πληροφορίας και σημαντικών συμπερασμάτων, χωρίς να είναι απαραίτητη η εξειδίκευση του χρήστη σε κάποιο επιστημονικό τομέα. Η φαντασία και η δημιουργικότητα και μόνο μπορούν να παράγουν σημαντικά συμπεράσματα, τα οποία θα ήταν πολύ δύσκολο να εξαχθούν χωρίς τη χρήση τεχνικών οπτικοποίησης. Συμπεράσματα τα οποία σαφώς μπορούν να επεξεργαστούν περεταίρω μέσω στατιστικών αναλύσεων και άλλων συναφών μεθόδων.

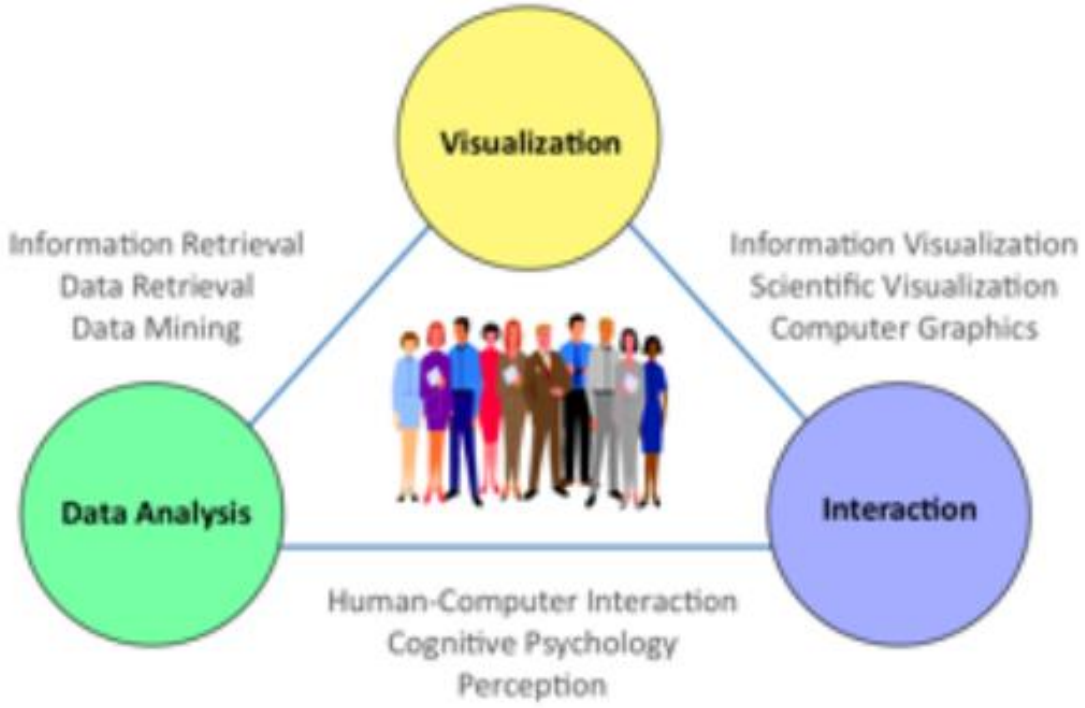
Βασικός σκοπός χρησιμοποίησης τεχνικών οπτικοποίησης, είναι η παρουσίαση πολύπλοκων και σύνθετων δομών δεδομένων, με απλό και κατανοητό τρόπο. Πολλές φορές επιδιώκεται η επεξήγηση των δεδομένων ή η επίλυση ενός συγκεκριμένου προβλήματος. Επίσης, είναι πιθανό να επιχειρείται η εξερεύνηση

των δεδομένων με σκοπό την καλύτερη κατανόησή τους ή ίσως η ανεύρεση στοιχείων που θα ήταν διαφορετικά δύσκολο να επισημανθούν, όπως ο εντοπισμός ακραίων τιμών (outliers), ενώ τέλος είναι δυνατό να προβλεφθούν μελλοντικές τιμές και να απεικονισθούν πιθανές τάσεις. Με την εξέλιξη των υπολογιστικών συστημάτων και τα δεδομένα να γίνονται διαρκώς μεγαλύτερα, μπορεί να γίνει με ασφάλεια η πρόβλεψη ότι η χρήση τεχνικών οπτικοποίησης θα συνεχίσει να μεγαλώνει και να αποκτά διαρκώς αυξανόμενη αξία (Miller, 2017).

Μια αναδυόμενη επιστήμη, η οπτική ανάλυση συνδυάζει αναλυτικό συλλογισμό με την ουσιαστική ικανότητα του ανθρώπινου εγκεφάλου να εσωτερικεύει και να κατανοεί γρήγορα δεδομένα που παρουσιάζονται οπτικά. Μέσω της χρήσης διαδραστικών διεπαφών, η οπτική ανάλυση παρέχει έναν μηχανισμό μέσω του οποίου ο χειριστής, ο μηχανικός και ο υπεύθυνος λήψης αποφάσεων μπορούν να συνεργαστούν σε πραγματικό χρόνο με προσομοίωση και να αναλύσουν πειραματικά και επιχειρησιακά δεδομένα, παρέχοντας άμεσα εικόνα των συσχετίσεων και των σχέσεων που οδηγούν στη βελτιστοποίηση της διαδικασίας. (Soban D., Thornhill D., Salunkhe S., Long A.).

Το Visual Analytics μπορεί να θεωρηθεί ως μια ολοκληρωμένη προσέγγιση που συνδυάζει την οπτικοποίηση, τους ανθρώπινους παράγοντες και την ανάλυση δεδομένων.

Το σχήμα απεικονίζει τους ερευνητικούς τομείς που σχετίζονται με το Visual Analytics. Εκτός από την οπτικοποίηση και την ανάλυση δεδομένων, ειδικά οι ανθρώπινοι παράγοντες, συμπεριλαμβανομένων των τομέων της γνώσης και της αντίληψης, διαδραματίζουν σημαντικό ρόλο στην επικοινωνία μεταξύ του ανθρώπου και του υπολογιστή, καθώς και στη διαδικασία λήψης αποφάσεων.



Εικόνα 11 : Τομείς που σχετίζονται με το Visual Analytics

Η Life Cycling Engineering, η μηχανική του κύκλου ζωής (L.C.E) με βάση την οπτική ανάλυση, δίνει μια συνδυασμένη προοπτική από κάτω προς τα πάνω και από πάνω προς τα κάτω. Έχει σκοπό να καθοδηγήσει όλες τις δραστηριότητες που αφορούν στην ανάπτυξη, κατασκευή, λειτουργία και επεξεργασία ως το τέλος του κύκλου ζωής των προϊόντων. Η μεθοδολογία αξιολόγησης (LCA) χρησιμεύει ως βασική προσέγγιση στο να ποσοτικοποιήσει τις περιβαλλοντικές επιπτώσεις προϊόντων και διαδικασιών. Απαιτεί ειδικές γνώσεις για να ερμηνεύσει τα αποτελέσματά της, που προέρονται από πολύπλοκες αλληλεξαρτήσεις εντός του κύκλου ζωής των προϊόντων, οι οποίες οδηγούν σε πολλαπλές επιπτώσεις.

Οι έρευνες και εργασίες των τελευταίων δεκαετιών σε αυτό τον τομέα οδηγούν σε ένα πλήθος μεθόδων και εργαλείων που στοχεύουν στην καθοδήγηση αυτών των δραστηριοτήτων που βοηθούν στη λήψη αποφάσεων σε διάφορους τομείς.

Ωστόσο, στην πράξη η εφαρμογή του LCE παρουσιάζει ελλείψεις. Αυτό οφείλεται στην ενσωμάτωση της τεχνικής στις επιχειρηματικές διαδικασίες και τη συμμετοχή διαφόρων ιεραρχικών επιπέδων, ειδικά όσον αφορά το επίπεδο διαχείρισης. Ένα άλλο πρόβλημα αναφέρεται στη συνεργασία των ενδιαφερόμενων. Όλα αυτά εμποδίζουν την αποτελεσματικότητα του LCE σε σχέση με τη βιωσιμότητα των προϊόντων ή των διαδικασιών που μελετώνται.

Όταν συνδυάζουμε τις προκλήσεις του LCE με τους στόχους της οπτικής ανάλυσης (VA), προκύπτουν πιθανές συνέργειες. Η οπτική ανάλυση όπως έχουμε αναφέρει στοχεύει σε εφαρμογές όπου οι σύνθετες εξαρτήσεις ενός συστήματος και ενός μεγάλου αριθμού ενδιαφερόμενων μερών και υπευθύνων λήψης αποφάσεων εμπλέκονται. Ως εκ τούτου, η οπτική ανάλυση νοείται ως ολοκληρωμένη προσέγγιση οπτικοποίησης, ανθρώπινης αλληλεπίδρασης και ανάλυσης δεδομένων. [13]

[13] "Life cycle engineering based on visual analytics", Kaluzaa A., Gellricha S., Cerdasa F., Thiedea S., Herrmanna C., ScienceDirect, 25th CIRP Life Cycle Engineering (LCE) Conference, 30 April – 2 May 2018, Copenhagen, Denmark, p.37-39

4.2 Οπτική ανάλυση στην εξόρυξη διεργασιών

Έχοντας στη διάθεσή μας λοιπόν, συνεχώς όλο και μεγαλύτερο όγκο δεδομένων, υπάρχει επείγουσα ανάγκη να βρεθούν νέες και αποτελεσματικές τεχνικές για την κατάλληλη χρήση τους, καθώς η ικανότητα συλλογής και αποθήκευσης των δεδομένων αυξάνεται με ταχύτερο ρυθμό από την ικανότητα ανάλυσής τους.

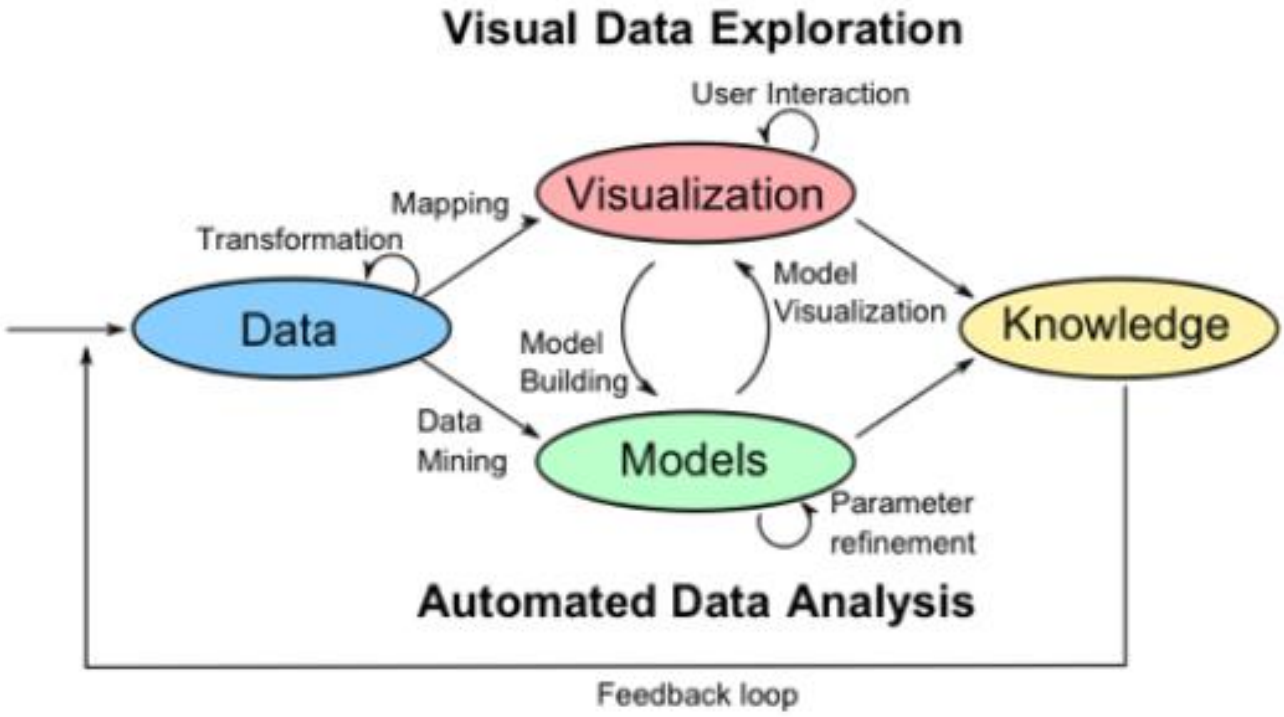
Τις τελευταίες δεκαετίες, αναπτύχθηκε ένας μεγάλος αριθμός μεθόδων αυτόματης ανάλυσης δεδομένων. Ωστόσο, η περίπλοκη φύση πολλών προβλημάτων καθιστά απαραίτητο να συμπεριληφθεί η ανθρώπινη παρέμβαση στο αρχικό στάδιο της ανάλυσης των δεδομένων. Οι μέθοδοι οπτικής ανάλυσης επιτρέπουν στους υπεύθυνους λήψης αποφάσεων να συνδυάσουν την ανθρώπινη ευελιξία, τη δημιουργικότητα και τις γνώσεις τους με τις τεράστιες δυνατότητες αποθήκευσης και επεξεργασίας των σημερινών υπολογιστών για να αποκτήσουν γνώση πιο περίπλοκων προβλημάτων[14].

Ένα κρίσιμο στοιχείο όμως, για να ληφθεί η βέλτιστη απόφαση είναι η κατάλληλη επιλογή των παραμέτρων της διαδικασίας. Πολλές φορές λοιπόν, η επιλογή παραμέτρων που γίνεται συχνά χειροκίνητα, με βάση την ικανότητα, την εμπειρία και τη διαίσθηση του υπευθύνου έχει ως αποτέλεσμα, η βελτιστοποίηση της διαδικασίας να είναι σε αρκετές περιπτώσεις επαναληπτική, και να στερείται ιχνηλασιμότητας. (Soban D., Thornhill D., Salunkhe S., Long A.).

Χρησιμοποιώντας οπτικές τεχνικές, οι άνθρωποι μπορούν να αλληλεπιδρούν άμεσα με τις δυνατότητες ανάλυσης δεδομένων του σημερινού υπολογιστή, επιτρέποντάς τους να λαμβάνουν σε περίπλοκες καταστάσεις, τη βέλτιστη απόφαση ανάμεσα σε άλλες που έχουν αναλυθεί και έχουν παρουσιαστεί τα αναμενόμενα αποτελέσματα τους.

[14] https://www.visual-analytics.eu/faq/ (ανακτήθηκε 22/3/2021)

Το παρακάτω σχήμα μας δείχνει μια αφηρημένη επισκόπηση των διαφόρων σταδίων της Οπτικής Ανάλυσης[15].



Εικόνα 12 : Στάδια οπτικής ανάλυσης

Το πρώτο βήμα είναι συχνά η προεπεξεργασία και ο μετασχηματισμός των δεδομένων, όπως η κανονικοποίηση τους, η ομαδοποίηση ή η ενοποίηση ετερογενών πηγών, ώστε να προκύψουν διαφορετικές αναπαραστάσεις για περαιτέρω εξερεύνηση.

Μετά τον μετασχηματισμό, ο αναλυτής μπορεί να επιλέξει μεταξύ της εφαρμογής οπτικών ή άλλων μεθόδων ανάλυσης. Εάν χρησιμοποιείται μια αυτοματοποιημένη ανάλυση πρώτα, εφαρμόζονται μέθοδοι εξόρυξης δεδομένων για τη δημιουργία μοντέλων των αρχικών δεδομένων. Μόλις δημιουργηθεί ένα μοντέλο, ο αναλυτής πρέπει να αξιολογήσει και να βελτιώσει τα

[15] https://www.visual-analytics.eu/ (ανακτήθηκε 23/3/2021)

μοντέλα, κάτι που μπορεί να γίνει καλύτερα αλληλεπιδρώντας με τα δεδομένα.

Οι οπτικοποιήσεις επιτρέπουν στους αναλυτές να τροποποιούν παραμέτρους ή να επιλέγουν άλλους αλγόριθμους ανάλυσης. Η οπτικοποίηση μοντέλου μπορεί στη συνέχεια να χρησιμοποιηθεί για την αξιολόγηση των ευρημάτων.

Τα παραπλανητικά αποτελέσματα σε ένα ενδιάμεσο βήμα μπορούν έτσι να ανακαλυφθούν σε πρώιμο στάδιο, οδηγώντας σε καλύτερα αποτελέσματα που έχουν μεγαλύτερο βαθμό αξιοπιστίας. Απαιτείται αλληλεπίδραση χρήστη με την οπτικοποίηση για την αποκάλυψη πληροφοριών, για παράδειγμα με μεγέθυνση σε διαφορετικές περιοχές.

Με αυτή την προσέγγιση οπτικοποίησης δίνεται η δυνατότητα για διαδραστική οπτική ανάλυση μεγάλων συνόλων δεδομένων σε παράλληλες συντεταγμένες.[16] Μπορούμε να δημιουργήσουμε "εικονικούς κόμβους" που με το κλικ του ποντικιού σε κάποια στοιχεία δεδομένων, να φωτίζονται και να δίνουν άμεσα μία εικόνα στο διάγραμμα για τη θέση τους και τη σημασία τους. Η νέα αυτή προσέγγιση μπορεί να χειριστεί την οπτικοποίηση και την αλληλεπίδραση με εξαιρετικά μεγάλο σύνολο δεδομένων. (Mao LinHuang, Tze-HawHuang, XuyunZhang).

Οι οπτικοποιήσεις διαδραματίζουν ζωτικό ρόλο στην ανάλυση βιομηχανικών δεδομένων, π.χ. διευκολύνοντας την κατανόηση δεδομένων, την αξιολόγηση μοντέλου και την εξαγωγή γνώσεων από μοντέλα.

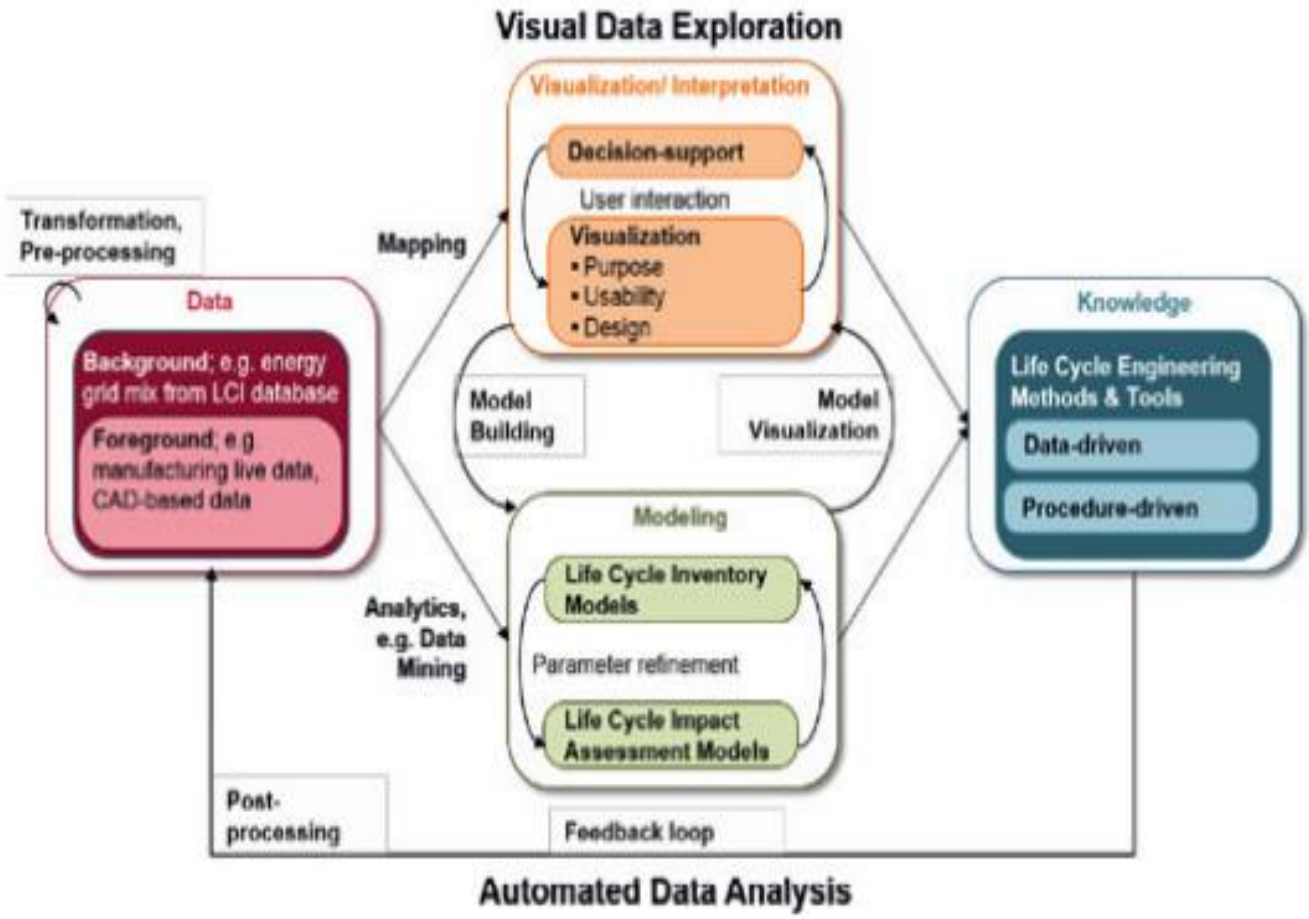
Η διαδικασία της οπτικής ανάλυσης χαρακτηρίζεται από μια ανθρωποκεντρική προσέγγιση, διευκολύνοντας την υπόθεση και μια ενημερωμένη λήψη αποφάσεων μέσω της μείωσης της πολυπλοκότητας των τεράστιων συνόλων δεδομένων. Μέσω της VA ο χρήστης έχει τη δυνατότητα να χειρίζεται τεράστια, δυναμικά

[16] https://www.sciencedirect.com/science/article/abs/pii/S0167739X15000382?via%3Dihub (ανακτήθηκε 1/4/2021)

μεταβαλλόμενα δεδομένα και να εντοπίζει ανωμαλίες, αλλαγές, μοτίβα και σχέσεις, προκειμένου να αποκτήσει νέα γνώση.

Οι Keim et al. (2008) πρότειναν ένα πλαίσιο που δομικά στοιχεία και διαδικασίες της οπτικής ανάλυσης περιγράφουν την αλληλεπίδραση της απόκτησης δεδομένων, μοντέλων, οπτικοποιήσεων και εξαγωγής γνώσεων από αυτά.

Το πλαίσιο αυτό φαίνεται στην παρακάτω εικόνα και παρέχει την κατάλληλη βάση για μια δομημένη χαρτογράφηση της προτεινόμενης διαδικασίας μοντελοποίησης.



Εικόνα 13: LCE και VA

Η μεθοδολογία μπορεί να χωριστεί σε τρία διαδοχικά βήματα [17] :

1. Κατανόηση και προετοιμασία δεδομένων,
2. Δημιουργία μοντέλων και τέλος
3. Οπτικοποίηση και ανάπτυξη γνώσεων.

Το πρώτο βήμα είναι εξοικείωση με το σύνολο των δεδομένων. Κατά τη διάρκεια αυτής της φάσης, γνωστή ως κατανόηση δεδομένων, χρησιμοποιούνται ευρέως μέθοδοι διερευνητικής ανάλυσης δεδομένων για παράδειγμα, το σύνολο των δεδομένων εξετάζεται σε σχέση με τις κατανομές (π.χ. διάμεση, τυπική απόκλιση), συσχετίσεις, ποιότητα δεδομένων (π.χ. τιμές που λείπουν, λάθος τιμές) και ακραίες τιμές για τη δημιουργία περαιτέρω ουσιαστικών χαρακτηριστικών που προέρχονται από τα υπάρχοντα (feature engineering).

Στο δεύτερο βήμα έχουμε τη δημιουργία μοντέλων μέσω μηχανικής μάθησης χρησιμοποιώντας αλγορίθμους μηχανικής μάθησης χωρίς επίβλεψη. Η βασική υπόθεση εδώ για τον προσδιορισμό διαφορετικών μοντέλων είναι ότι παρόμοιες καταστάσεις χαρακτηρίζονται από μικρότερη απόσταση μεταξύ τους από ότι διαφορετικές. Με τη βοήθεια των συλλεγόμενων δεδομένων ανά παρατήρηση, η ομοιότητα των καταστάσεων μπορεί να ποσοτικοποιηθεί.

Η οπτικοποίηση του μοντέλου (π.χ. scatter-plot) υποστηρίζει την αξιολόγηση του σε σχέση με τη σημασία του για συγκεκριμένη περίπτωση χρήσης. Ο επαναληπτικός κύκλος δημιουργίας μοντέλων και οπτικοποίησης τους μπορεί να οδηγήσει σε αποφάσεις μέσω της αξιολόγησης, οι οποίες όμως εξαρτώνται σε μεγάλο βαθμό από την κατανόηση των δεδομένων και τους στόχους της συγκεκριμένης μελέτης – περίπτωσης χρήσης.

[17]https://www.sciencedirect.com/science/article/pii/S2212827120306740 (ανακτήθηκε στις 3/4/2021)

Για να ολοκληρωθεί ο κύκλος της οπτικής ανάλυσης στο τρίτο βήμα, η γνώση της διαδικασίας πρέπει να αποκτηθεί από τα μοντέλα παρατήρησης που αναπτύχθηκαν προηγουμένως (ανάπτυξη γνώσης) και τελικά να οδηγηθούμε σε λήψη αποφάσεων.

Η διαδικασία λήψης αποφάσεων είναι πολύπλοκη και δεν υπάρχει συγκεκριμένο πρότυπο λειτουργικών ενεργειών, που θα οδηγούν σε ενιαία αντιμετώπιση του προβλήματος, όμως η οπτικοποίηση λειτουργεί βοηθητικά στον σκοπό της απόφασης που δεν είναι άλλη παρά η επιλογή της εναλλακτικής λύσης που μεγιστοποιεί μία αντικειμενική συνάρτηση ή ικανοποιεί συγκεκριμένα κριτήρια [18].

[18] https://eclass.uoa.gr/modules/document/file.php/D343/%CE%A0%CE%B1%CF%81%CE%BF%CF%85%CF%83%CE%B9%CE%AC%CF%83%CE%B5%CE%B9%CF%82/%CE%94%CE%B9%CE%AC%CE%BB%CE%B5%CE%BE%CE%B7%201%CE%B7%20%2821-03-07%29.pdf (ανακτήθηκε στις 6/4/2021)

4.3 Τεχνικές οπτικοποίησης

Οι σύγχρονες τεχνικές οπτικοποίησης της πληροφορίας είναι πολύτιμα εργαλεία για την ανάλυση των δεδομένων και την εξαγωγή συμπερασμάτων, καθώς προσφέρουν μια σειρά από πλεονεκτήματα.

Ειδικότερα οι τεχνικές οπτικοποίησης: [19]

- Απεικονίζουν ιδιότητες των δεδομένων με μια άμεσα κατανοητή εικόνα.
- Παρέχουν συμπυκνωμένη πληροφόρηση με μια ματιά.
- Αποκαλύπτουν τάσεις δεδομένων, εξαιρέσεις και ακραίες τιμές, συστάδες δεδομένων και κενά.
- Είναι ικανές να χειρίζονται μεγάλους όγκους δεδομένων.
- Αναπαριστούν την πληροφορία με αντικειμενικό τρόπο. Αντιθέτως, η λεκτική περιγραφή μπορεί να αντανακλά ή να υποκρύπτει υποκειμενικές αντιλήψεις.
- Είναι διαδραστικές και επιτρέπουν στον χρήστη τη διεξαγωγή διαφορετικών αναλύσεων.
- Αποκαλύπτουν κρυμμένη πληροφορία, που θα χρησιμοποιηθεί για την εξαγωγή συμπερασμάτων.
- Τα αποτελέσματα τους, τα οποία αποκαλύπτουν ιδιότητες των δεδομένων, μπορούν να χρησιμοποιηθούν για τον προσανατολισμό της περαιτέρω ανάλυσης με άλλα μέσα.

Κατά καιρούς προτάθηκαν διάφοροι τρόποι ταξινόμησης των τεχνικών οπτικοποίησης. Ο δημοφιλέστερος ίσως τρόπος ταξινόμησης είναι αυτός που προτάθηκε από τους Keim and Kriegel (1996) και επαναδιατυπώθηκε από τον Keim(2002).

Σύμφωνα με τον Keim (2002), οι τεχνικές οπτικής αναπαράστασης μπορούν να ταξινομηθούν με βάση τρία κριτήρια:

i. τον τύπο των δεδομένων
ii. την τεχνική οπτικοποίησης

[19] https://repository.kallipos.gr/bitstream/11419/1232/2/Kef._5.pdf (ανακτήθηκε 5/12021)

iii. την τεχνική αλληλεπίδρασης και στρέβλωσης.

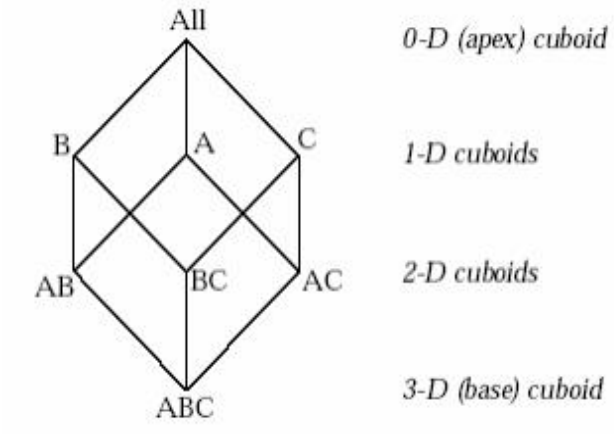
Οι κλασικές μέθοδοι οπτικοποίησης λειτουργούν ικανοποιητικά για δεδομένα χαμηλών διαστάσεων, αλλά αποτυγχάνουν για τα δεδομένα υψηλής διάστασης. Είναι πολύ απλουστευτικές και δεν βοηθούν να αποκτήσουμε βαθύτερη εικόνα των δεδομένων ή διαπιστώνεται ότι είναι υπερβολικά δυσκίνητες και υπολογιστικά ανεπαρκείς.

Όμως, κάθε τεχνική οπτικοποίησης μπορεί να συνδυαστεί με κάθε τεχνική αλληλεπίδρασης για κάθε τύπο δεδομένων. Ένα σύστημα μπορεί να χρησιμοποιεί συνδυασμούς τεχνικών οπτικοποίησης και αλληλεπίδρασης για διάφορους τύπους δεδομένων. οι οποίες είναι αποτελεσματικές για την οπτικοποίηση δεδομένων μεγάλης διάστασης.

4.3.1 Ταξινόμηση τεχνικών οπτικοποίησης ως προς τον τύπο τους

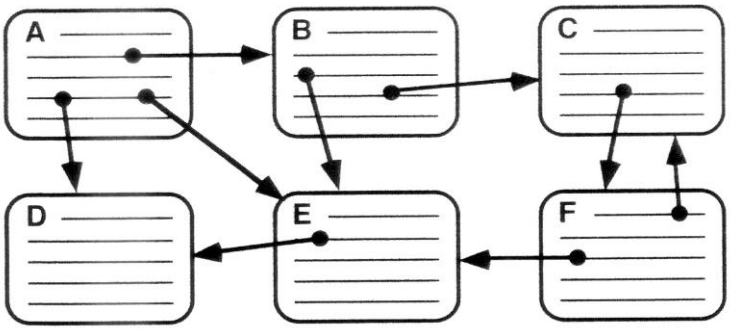
Ως προς τον τύπο τους, τα δεδομένα που θα οπτικοποιηθούν μπορεί να είναι : [8]

- Μονοδιάστατα. Τα δεδομένα έχουν μια διάσταση, πχ χρονικά δεδομένα.
- Δυσδιάστατα. Τα δεδομένα έχουν δύο διαστάσεις, πχ γεωγραφικά σημεία με γεωγραφικό μήκος και πλάτος.
- Πολυδιάστατα. Τα δεδομένα έχουν πολλές διαστάσεις, πχ πίνακες σχεσιακών βάσεων δεδομένων με πολλές στήλες.



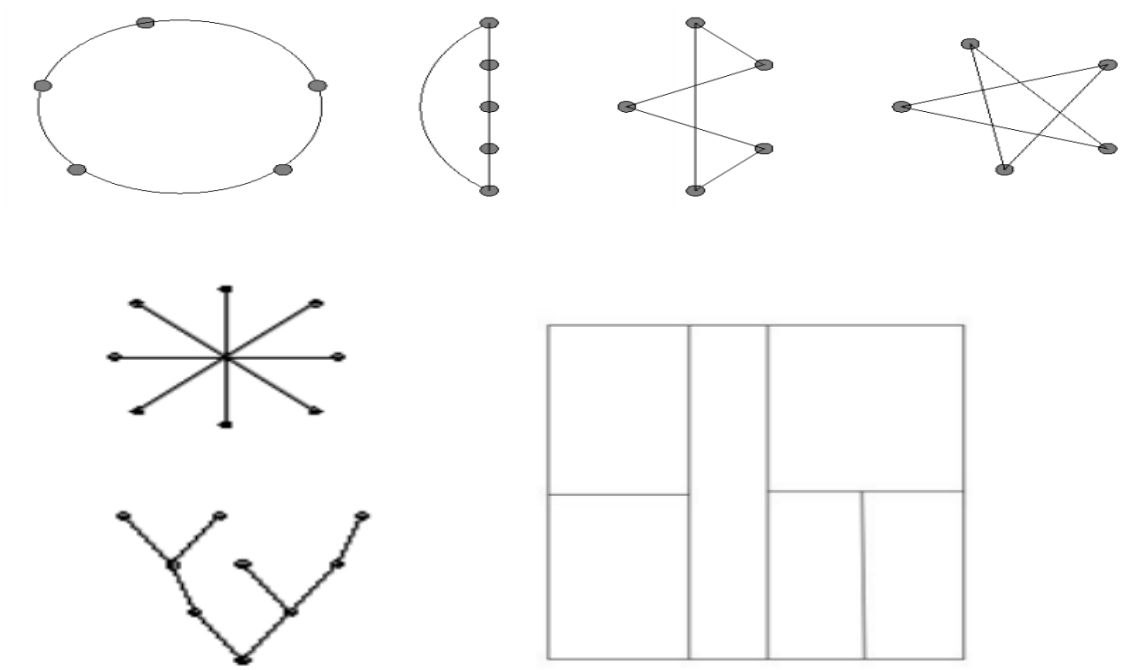
Εικόνα 14 : πχ πολυδιάστατων δεδομένων

- Κείμενο και Υπερκείμενο. Τα δεδομένα είναι αδόμητα και δεν μπορούν να εκφραστούν σε σχέση με διαστάσεις.



Εικόνα 15 : πχ δεδομένων από κείμενο και υπερκείμενο

- Ιεραρχίες και Γράφοι. Τα αντικείμενα συνδέονται μεταξύ τους με σχέσεις. Μπορούν να αναπαρασταθούν με ένα γράφο, όπου τα αντικείμενα είναι οι κόμβοι και οι σχέσεις είναι οι ακμές που τους συνδέουν. Ένα κύριο χαρακτηριστικό των τεχνικών αυτών είναι ότι προσπαθούν να αξιοποιήσουν όσο το δυνατόν πιο αποτελεσματικά τη διαθέσιμη επιφάνεια προβολής ώστε να είναι δυνατή η αναπαράσταση μεγάλου όγκου δεδομένων. Είναι ιδανικές τεχνικές για οπτικοποίηση δεδομένων αρχείου και ιδιαίτερα στην εύρεση ομαδοποιήσεων.

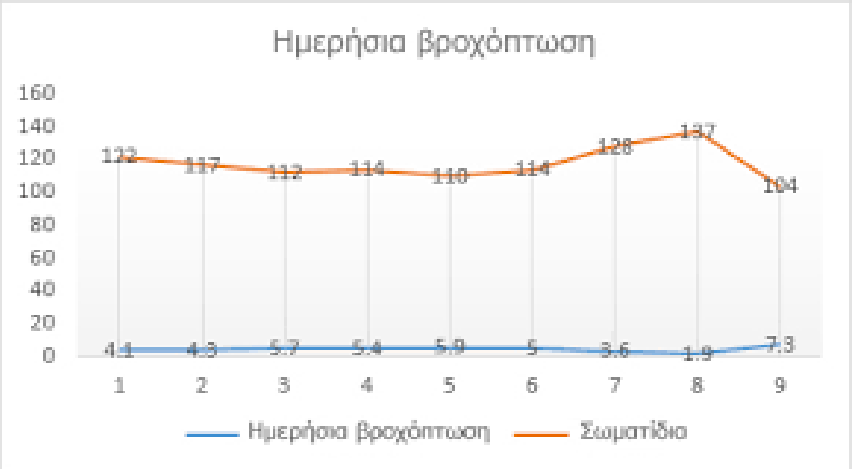


Εικόνα 16: Παραδείγματα γράφων

- Αλγόριθμοι και λογισμικό. Αφορά δεδομένα ροής πληροφοριών σε ένα πρόγραμμα.

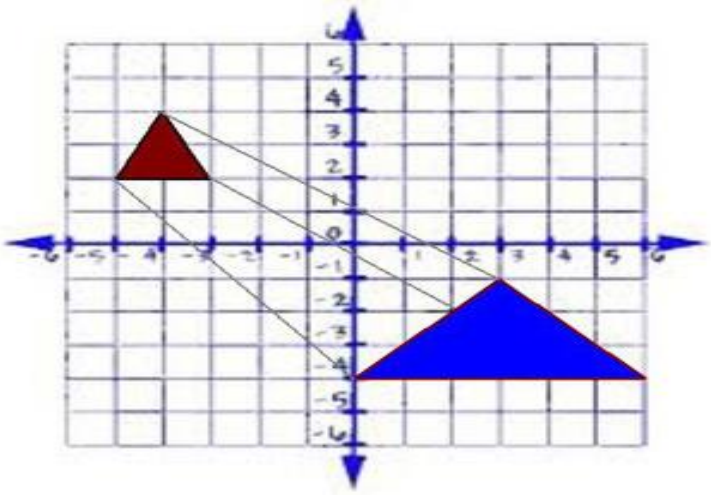
Οι τεχνικές οπτικοποίησης υπάγονται στις παρακάτω κατηγορίες:

1. Τυπικές δύο ή τριών διαστάσεων (Standard 2D/3D displays). Σχετικά απλές τεχνικές, που συνήθως εφαρμόζονται στα πρώτα στάδια της ανάλυσης. Δεν είναι κατάλληλες για την οπτικοποίηση σύνθετων δομών.



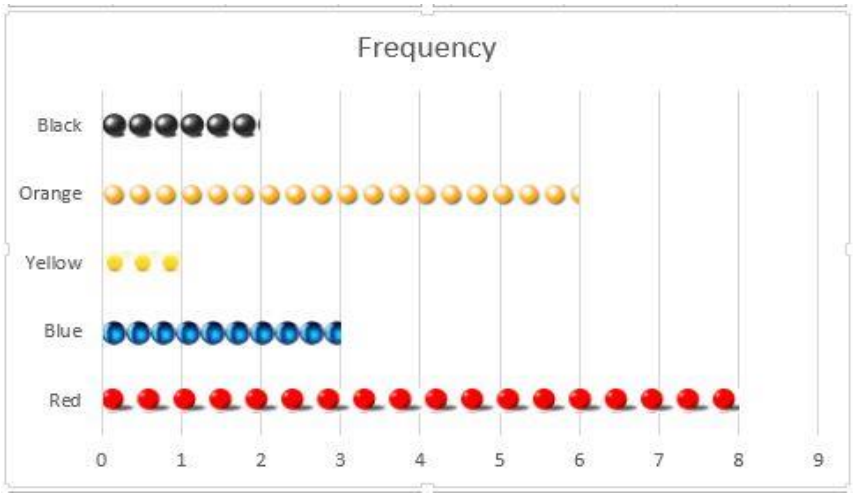
Εικόνα 17: Πχ γραφήματος γραμμής

2. Γεωμετρικού Μετασχηματισμού (Geometrically Tranformed). Πολυδιάστατα δεδομένα μετασχηματίζονται και προβάλλονται με γεωμετρικό τρόπο, ώστε να αποκαλυφθούν πιθανές σχέσεις τους. Περίπτωση τεχνικής γεωμετρικού μετασχηματισμού είναι τα διαγράμματα παράλληλων συντεταγμένων.



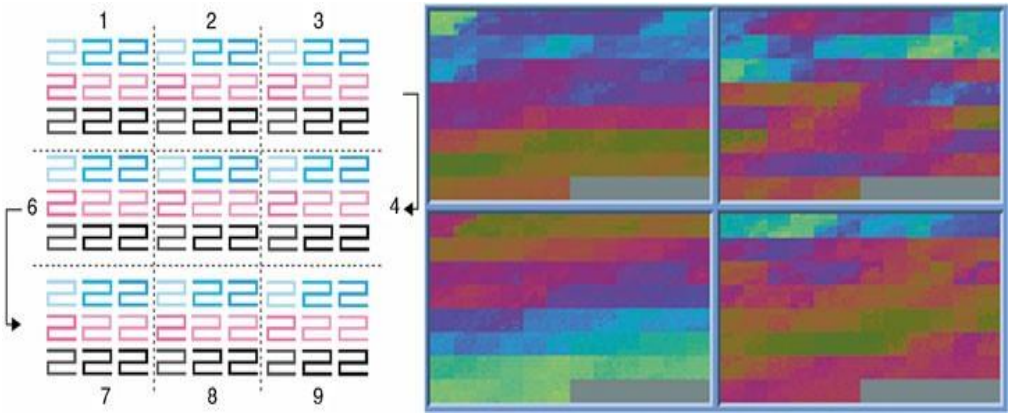
Εικόνα 18 : πχ γραφήματος γεωμετρικού μετασχηματισμού

3. Εικονικογραφικές (Iconic Displays). Κάθε παρατήρηση (αντικείμενο) αντιστοιχίζεται σε μια εικόνα και κάθε τιμή της παρατήρησης αντιστοιχίζεται με ένα χαρακτηριστικό της εικόνας, πχ σχήμα, μέγεθος, χρώμα κλπ. Τεχνικές κατάλληλες για μέτριο πλήθος δεδομένων και σχετικά μικρό αριθμό μεταβλητών.



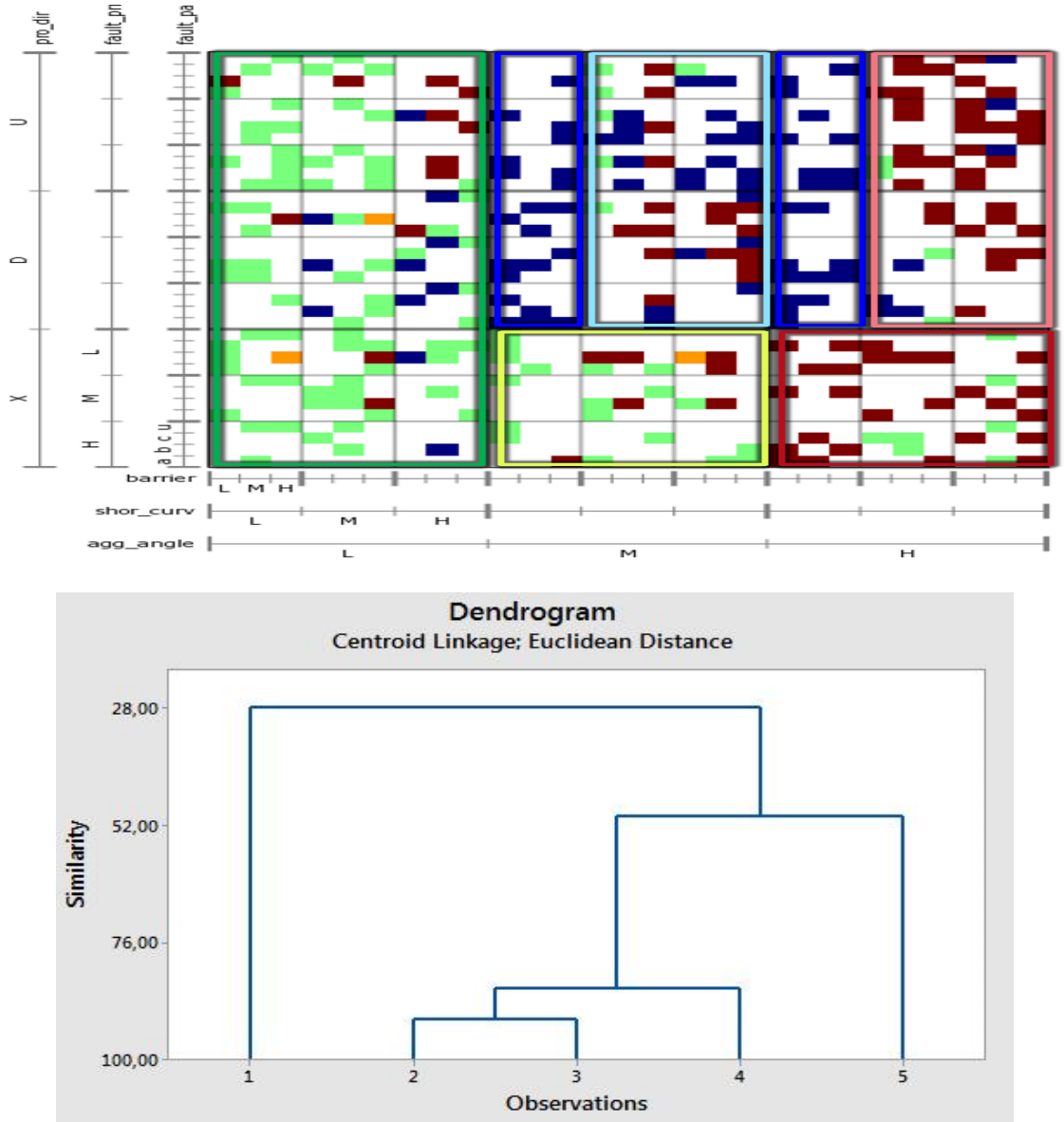
Εικόνα 19 : Πχ εικονογραφήματος

4. Εικονοστοιχείων (Dense Pixel Displays). Κάθε τιμή των δεδομένων αντιστοιχίζεται σε ένα pixel, το οποίο χρωματίζεται ανάλογα με την τιμή. Τα εικονοστοιχεία μιας διάστασης τοποθετούνται σε γειτονικές περιοχές. Τεχνικές κατάλληλες για απεικόνιση περίπου 1.000.000 τιμών, με αδυναμίες όμως στον εντοπισμό σύνθετων δομών δεδομένων.



Εικόνα 20 : πχ dense pixel display

5. Στοίβας (ή ιεραρχικές) (Stacked Displays). Η παρουσίαση των δεδομένων γίνεται στη βάση μιας ιεράρχησης. Οι τύποι των ιεραρχήσεων ποικίλουν. Παράδειγμα τέτοιας τεχνικής είναι η Dimensional Stacking, όπου ο χώρος απεικόνισης χωρίζεται σε τμήματα ανάλογα με δύο διαστάσεις και εντός αυτών των τμημάτων γίνεται απεικόνιση των δεδομένων ανάλογα με δύο άλλες διαστάσεις. Άλλο παράδειγμα, με τελείως διαφορετικό τρόπο ιεράρχησης, είναι τα Δενδρογράμματα, που αποτυπώνουν τη διαδικασία διαδοχικής συγχώνευσης συστάδων.



Εικόνα 21 : πχ Dimensional Stacking - δενδρογράμματος

4.3.2 Ταξινόμηση τεχνικών οπτικοποίησης ως προς τον τρόπο αλληλεπίδρασης

Αναφορικά με τον τρόπο αλληλεπίδρασης και στρέβλωσης οι τεχνικές κατηγοριοποιούνται ως:

1. Δυναμικής προβολής (Dynamic Projections). Συνίσταται στη μεταβολή του τρόπου προβολής των δεδομένων.
2. Διαδραστικής επιλογής (Interactive Filtering). Επιτρέπει την τμηματοποίηση των δεδομένων και την επικέντρωση σε ένα υποσύνολο. Το υποσύνολο των δεδομένων μπορεί να προκύψει είτε με την εκτέλεση κάποιου ερωτήματος είτε με την άμεση επιλογή από τον χρήστη.
3. Διαδραστικής Διαβάθμισης Λεπτομέρειας (Interactive Zooming). Είναι η δυνατότητα προβολής σε διαφορετικό βαθμό λεπτομέρειας. Τα αντικείμενα μπορεί να μεγεθυνθούν ή μπορεί να προβληθεί διαφορετικού τύπου πληροφορία, όπως πχ κείμενο.
4. Διαδραστικής στρέβλωσης (Interactive Distortion). Συνίσταται στην προβολή του συνόλου των δεδομένων με χαμηλό βαθμό λεπτομέρειας, με ταυτόχρονη προβολή τμήματος των δεδομένων με υψηλό βαθμό λεπτομέρειας.
5. Διαδραστικής Σύνδεσης και Χρωματισμού (Interactive Linking and Brushing). Είναι ο συνδυασμός διαφορετικών τεχνικών οπτικοποίησης. Για παράδειγμα, σε ένα σύνολο διαγραμμάτων διασποράς μπορεί να χρωματιστούν και να συνδεθούν ορισμένα σημεία σε όλα τα διαγράμματα. [20]

Πέρα όμως από αυτές έχουν προταθεί και οι παρακάτω κατηγορίες τεχνικών :

- Τυπικές τεχνικές
- Τεχνικές παραμόρφωσης
- Τεχνικές προβολής
- Υβριδικές τεχνικές

[20] https://repository.kallipos.gr/bitstream/11419/1232/2/Kef._5.pdf (ανακτήθηκε 6/1/2021)

ΚΕΦΑΛΑΙΟ 5

Λογισμικά ProM – Disco - Anaconda

Το ProM (το όνομα του οποίου είναι συντομογραφία για το Process Mining framework), είναι ένα open - source πρόγραμμα για αλγόριθμους εξόρυξης διεργασιών. Το ProM παρέχει μια πλατφόρμα λειτουργιών που είναι εύχρηστες και εύκολες στην επέκταση, σε χρήστες και προγραμματιστές αλγορίθμων εξόρυξης διεργασιών.

Αντίστοιχα, το Disco μπορεί να κάνει οπτικοποίηση δεδομένων σε λιγότερο χρόνο από παλιότερα, καθώς μας βοηθά να δημιουργήσουμε «όμορφα» γραφήματα από τα δεδομένα μας με εύχρηστο και σχετικά εύκολο τρόπο.

Το Disco δημιουργήθηκε από πρώην κορυφαίους ακαδημαϊκούς με πολλά χρόνια εμπειρίας στην εξόρυξη διεργασιών. Είναι ένα εργαλείο που ταιριάζει απόλυτα στην επίλυση των προβλημάτων process mining.

Στην παρούσα εργασία, εκτός από πρόγραμμα γραμμένο σε Python θα μπορούσε να χρησιμοποιηθούν έτοιμα προγράμματα για process mining, όπως τα παραπάνω, δηλαδη το ProM ή το Disco.

Και τα δύο αυτά προγράμματα προσφέρουν ένα User Interface ιδανικό για process mining. Ο χρήστης μπορεί να κάνει import τα αρχεία δεδομένων του και να τα αναλύσει, επεξεργαστεί και να τα αναπαραστήσει σε γράφους ή animations. Αυτά τα προγράμματα μπορούν να πραγματοποιήσουν αυτοματοποιημένο process discovery και να δημιουργήσουν όμορφα και ουσιώδη process maps απευθείας από τα δεδομένα.

Ο χρήστης μπορεί να επιλέξει το επιθυμητό επίπεδο αφαίρεσης και να δημιουργήσει φίλτρα από τα activities ή τα paths. Επιπλέον, προσφέρουν user interface στατιστικής ανάλυσης, στην οποία ο χρήστης μπορεί να εξερευνήσει τις σχέσεις μεταξύ των δεδομένων.

Η βιβλιοθήκη pm4py είναι μία βιβλιοθήκη της Python η οποία παρέχει υλοποίηση σε αλγορίθμους εξόρυξης και επεξεργασίας δεδομένων. Είναι open – source και είναι στην ουσία η state-of-the-art βιβλιοθήκη για χειρισμό event data. Χρησιμοποιείται για την ευχρηστία που προσφέρει και για τις πολλές της δυνατότητες.

Μπορούμε να πούμε πως αυτό που καταφέρνουμε με την Python και με όλες τις γλώσσες προγραμματισμού είναι πως γράφοντας μια φορά τον κώδικα του τι θέλουμε να κάνει ένα πρόγραμμα, αυτό θα κάνει κάθε φορά τους ίδιους υπολογισμούς σε δευτερόλεπτα, ανάλογα με τα στοιχεία (datasets) που θα του εισάγουμε κάθε φορά.

Τα εργαλεία της οπτικοποίησης παίζουν σημαντικό ρόλο στην ανάλυση δεδομένων σε σημαντικούς τομείς της επιστήμης όπως πχ. στην βιοϊατρική. Τα εργαλεία αυτά μπορούν να παρουσιάσουν πολύπλοκες δομές γονιδίων σε γραφήματα, δέντρα και αλυσίδες. Η οπτική παρουσίαση βοηθάει στην καλύτερη κατανόηση αυτών των δομών για ανακάλυψη γνώσης και εξερεύνηση των δεδομένων.

Επιπρόσθετα, όπως έχει αναφερθεί αυτά τα εργαλεία βοηθούν στον τομέα της οικονομίας, ώστε τα οικονομικά ινστιτούτα να αναγνωρίζουν τις απάτες και τα εγκλήματα από παραποιημένα δεδομένα από τις διάφορες βάσεις δεδομένων και από το ιστορικό συναλλαγών που έγιναν από τους πελάτες. Οι τεχνικές οπτικοποίησης βοηθούν στην παρουσίαση δεδομένων με διαφορετικές μορφές, όπως γράφοι που βασίζονται σε συγκεκριμένα γνωρίσματα. Προβάλλοντας τα δεδομένα από διάφορες οπτικές γωνίες, η τράπεζα δύναται να διακρίνει τους πελάτες που έχουν επιχειρήσει παράνομες πράξεις και μετά μια λεπτομερής έρευνα αυτών των ύποπτων περιπτώσεων βοηθάει στην εξιχνίαση των απατών και των εγκλημάτων. (Γούλου Ζ., 2010).

Η βιβλιοθήκη pm4py μπορούμε να πούμε πως είναι η state-of-the-art βιβλιοθήκη για χειρισμό event data και χρησιμοποιείται τόσο για την ευκολία χρήσης που προσφέρει, όσο και για τις πολλές της δυνατότητες.

ΚΕΦΑΛΑΙΟ 6

6.1 Δημιουργία εφαρμογής για οπτικοποίηση δεδομένων σε Python

Στην παρούσα εργασία αναπτύχθηκε πρόγραμμα σε Python, στο οποίο έγινε η οπτικοποίηση δεδομένων που έχουν να κάνουν με την εξόρυξη διεργασιών. Χρησιμοποιήθηκαν βιβλιοθήκες οπτικοποίησης όπως η matplotlib ή η plotly δημιουργήθηκαν διαγράμματα διαφόρων μορφών (bar charts, scatter charts κ.ά) για την οπτικοποίηση τόσο των δεδομένων που χρησιμοποιούνται στην εξόρυξη διεργασιών (αρχεία γεγονότων), όσο και των τεχνικών αναπαράστασης και ανάλυσης διεργασιών.

Αναλυτικά, έχουν γίνει τα εξής :

1. Δοθέντος ενός αρχείου γεγονότων τύπου xes ή csv έχει δημιουργηθεί πρόγραμμα το οποίο θα εμφανίζει τα events που περιέχει καθώς και τον αριθμό εμφανίσεων του καθενός. Στη συνέχεια θα εμφανίζει ένα bar chart. Στον άξονα χ τα events, στον άξονα y τη συχνότητα.
2. Δοθέντος ενός αρχείου γεγονότων τύπου xes ή csv έχει δημιουργηθεί πρόγραμμα το οποίο εμφανίζει :
   i. Τα start activities με το πόσες φορές εμφανίζονται μέσα στο log
   ii. Τα end activities με το πόσες φορές εμφανίζονται μέσα στο log
   iii. Τη μέση διάρκεια κάθε δραστηριότητας συνολικά στο log
   iv. Τη διάρκεια των 10 μεγαλύτερων και 10 μικρότερων traces

Για την υλοποίηση αυτών προγραμμάτων χρησιμοποιήθηκε Python 3, καθώς και οι βιβλιοθήκες pm4py, pandas, matplotlib.

Για το (1) χρησιμοποιήθηκε το αρχείο δεδομένων Artificial - Loan Process.xes, ενώ για το (2) χρησιμοποιήθηκε το αρχείο Purchasing-Example.csv.

Γενικότερα, χρησιμοποιήθηκαν και xes και csv αρχεία για να πειραματιστούμε και με τους δύο τύπους. Επιπλέον, επειδή το (2) λυνόταν πολύ εύκολα με τη χρήση dataframe, το αρχείο csv ήταν η καλύτερη επιλογή.

Στην εργασία αυτή χρησιμοποιήθηκαν τα εξής από την pm4py :

- csv_import_adapter, για τα csv αρχεία
- xes_importer, για τα xes αρχεία
- log_converter, για την μετατροπή των δεδομένων σε logs
- start_activities_filter / end_activities_filter, για τον εντοπισμο των start και end activities.

Ο κώδικας αναπτύχθηκε σε Python3 με τη χρήση του Anaconda και των βιβλιοθηκών που προσφέρει.

Το Anaconda είναι μια διανομή πακέτων για την επιστήμη των δεδομένων. Μαζί του, προσφέρεται το πακέτο conda, το οποίο είναι και ένας διαχειριστής περιβάλλοντος. Το conda χρησιμοποιείται για την δημιουργία περιβαλλόντων για την απομόνωση projects που χρησιμοποιούν διαφορετικές εκδόσεις της Python ή διαφορετικές εκδόσεις πακέτων.

Το πλεονέκτημα του Anaconda είναι ότι μπορούμε να σώσουμε τα περιβάλλοντα ανάπτυξης και να τα διανέμουμε μαζί με τα εξαρτημένα πακέτα τους.

Οι βιβλιοθήκες pandas και matplotlib χρησιμοποιήθηκαν για τον χειρισμό των dataframes και το plotting των γραφημάτων αντίστοιχα.

6.2 Υλοποίηση

Ο κώδικας του προγράμματος, το οποίο εμφανίζει τα events που περιέχει το dataset Artificial - Loan Process.xes, τον αριθμό εμφανίσεων του καθενός, όπως επίσης και το bar chart, βρίσκεται στον Αλγόριθμο 1 (Παράρτημα 1).

Αναλυτικά έχουν γίνει τα εξής :

- Αρχικά (γραμμές 1-4) κάνουμε import τις βιβλιοθήκες που θα χρησιμοποιήσουμε.
- Έπειτα (γραμμή 6), φορτώνουμε στη μεταβλητή log το αρχείο και δημιουργούμε ένα Python dictionary, freq_dic (γραμμή 7) το οποίο θα έχει ως κλειδιά το όνομα του activity και σαν τιμές τον αριθμό εμφάνισης του καθενός.
- Στη συνέχεια, για κάθε γραμμή στο αρχείο (γραμμές 8,9) παίρνουμε το event και στη γραμμή 10 παίρνουμε το concept:name από το event το οποίο δηλώνει το όνομα του κάθε activity.
- Στις επόμενες γραμμές (11-14) δημιουργούμε το dictionary, freq_dic προσθέτοντας 1 στην αντίστοιχη τιμή του activity.
- Τέλος, στις γραμμές 16-19 δημιουργούμε το γράφημα και το αποθηκεύουμε. type=listing

Ο κώδικας του προγράμματος που αφορά το dataset Purchasing-Example.csv, το οποίο εμφανίζει τα start activities με το πόσες φορές εμφανίζονται μέσα στο log, τα end activities με το πόσες φορές εμφανίζονται μέσα στο log, τη μέση διάρκεια κάθε δραστηριότητας συνολικά στο log, τη διάρκεια των 10 μεγαλύτερων και 10 μικρότερων traces, βρίσκεται στον Αλγόριθμο 2 (Παράρτημα 2).

Αναλυτικά έχουν γίνει τα εξής :

- Αρχικά, και εδώ, κάνουμε import τις απαραίτητες βιβλιοθήκες (γραμμές 1-13). Αυτή τη φορά όμως, εισάγουμε τα δεδομένα σε ένα dataframe, πράγμα που κάνει την επεξεργασία τους αρκετά ευκολότερη.

- Για την εύρεση των start και end activities (γραμμές 17-21) χρησιμοποιήθηκε package από την pm4py.
- Αφού συλλέξουμε τις δύο κατηγορίες activities από το dataframe, τις κάνουμε plot στις γραμμές (γραμμές 31-37).
- Για τη μέση διάρκεια κάθε δραστηριότητας, έπρεπε πρώτα να βρεθεί ο χρόνος διάρκειας του κάθε activity. Συνεπώς, δημιουργήθηκε η στήλη duration στο dataframe (γραμμές 44-47).
- Έχοντας τη στήλη αυτή μπορούμε εύκολα να βρούμε το μέσο όρο (γραμμή 49) και να τον αναπαραστήσουμε (γραμμές 52-58).
- Τέλος, για να βρεθεί η διάρκεια των 10 μεγαλύτερων και 10 μικρότερων traces, ταξινομούμε τις τιμές των χρόνων διάρκειας των traces (γραμμή 62) και πάλι με πράξεις στο dataframe (γραμμές 70-73) παίρνουμε τα ζητούμενα και τα αναπαριστούμε (γραμμές 76-90).

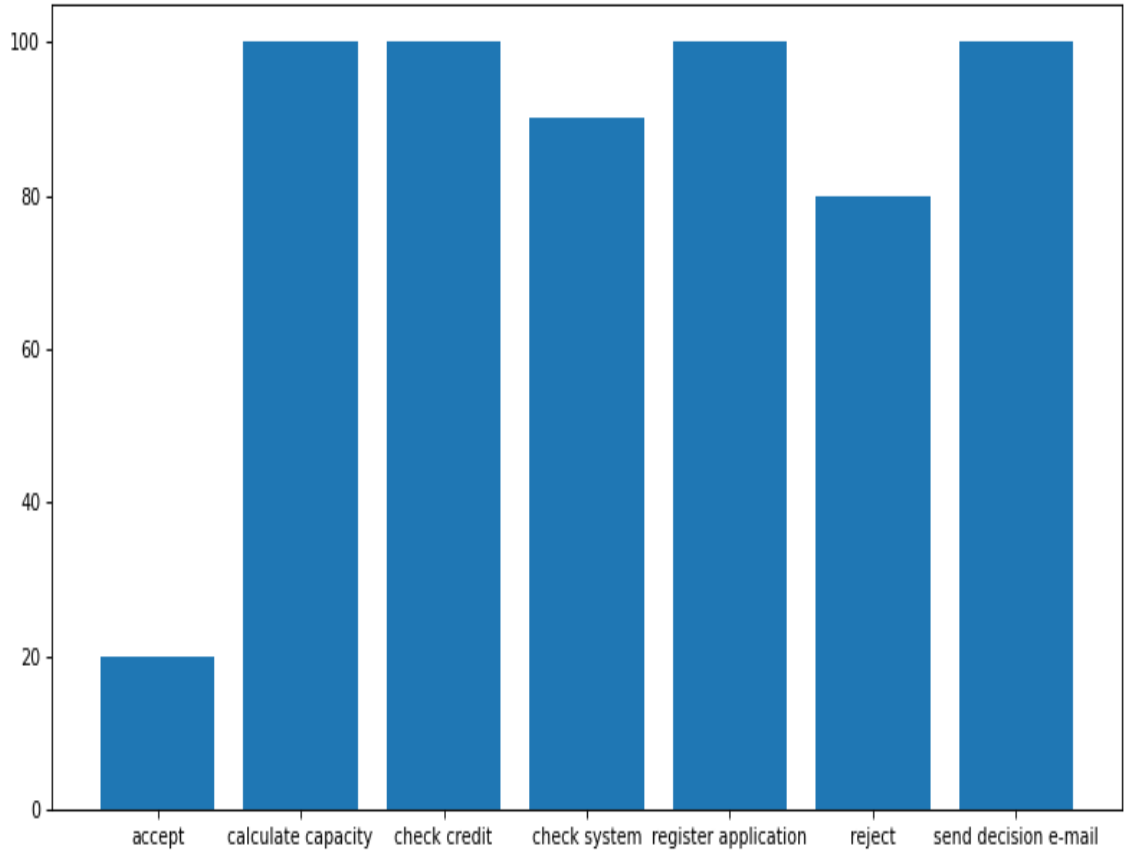
Επιπρόσθετα των παραπάνω διαγραμμάτων, έχουν γίνει και κάποια διαγράμματα με βάση το dataset Purchasing-Example.csv. Είναι διαγράμματα κουκκίδων (επίσης γνωστά και ως Cleveland dot plots) και δείχνουν αλλαγές μεταξύ δύο (ή περισσότερων) σημείων στο χρόνο ή μεταξύ δύο (ή περισσότερων) συνθηκών.

Σε σύγκριση με ένα γράφημα ράβδων, οι κουκκίδες είναι λιγότερο «γεμάτες» και επιτρέπουν μια ευκολότερη σύγκριση μεταξύ των συνθηκών. Η Plotly Express είναι η εύχρηστη διασύνδεση υψηλού επιπέδου με την Plotly, η οποία λειτουργεί σε ποικίλους τύπους δεδομένων και παράγει εύκολα γραφήματα τέτοιου είδους.

6.3 Αποτελέσματα οπτικοποίησης

Παρακάτω δίνονται τα αποτελέσματα της οπτικοποίησης για κάθε περίπτωση όπως έχει αναφερθεί παραπάνω.

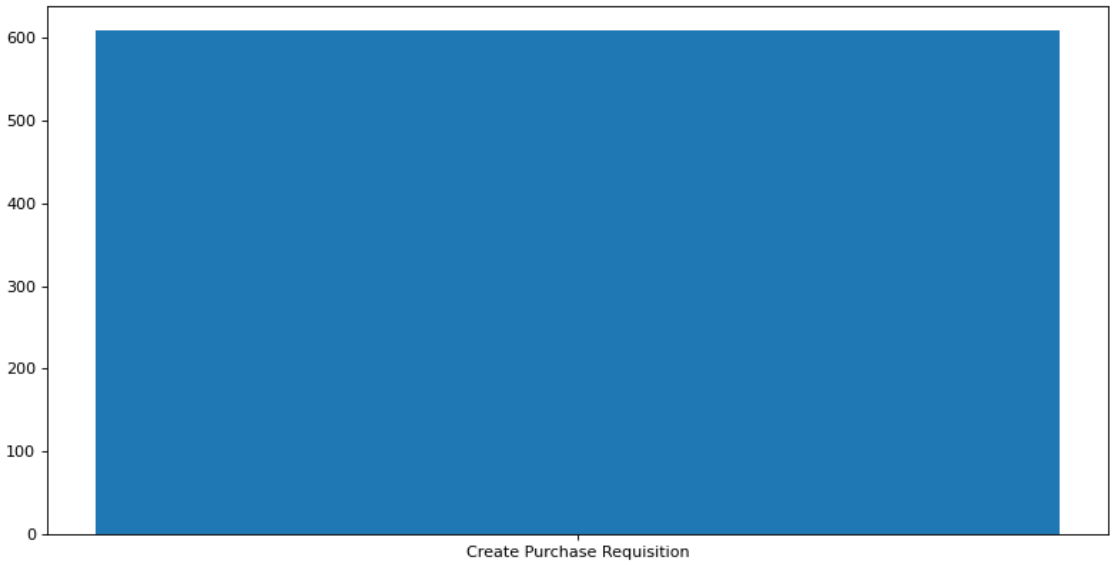
Σε αυτό το γράφημα απεικονίζονται τα 7 activities που υπάρχουν στο dataset Artificial - Loan Process.xes και το πλήθος τους εμφάνισής τους.



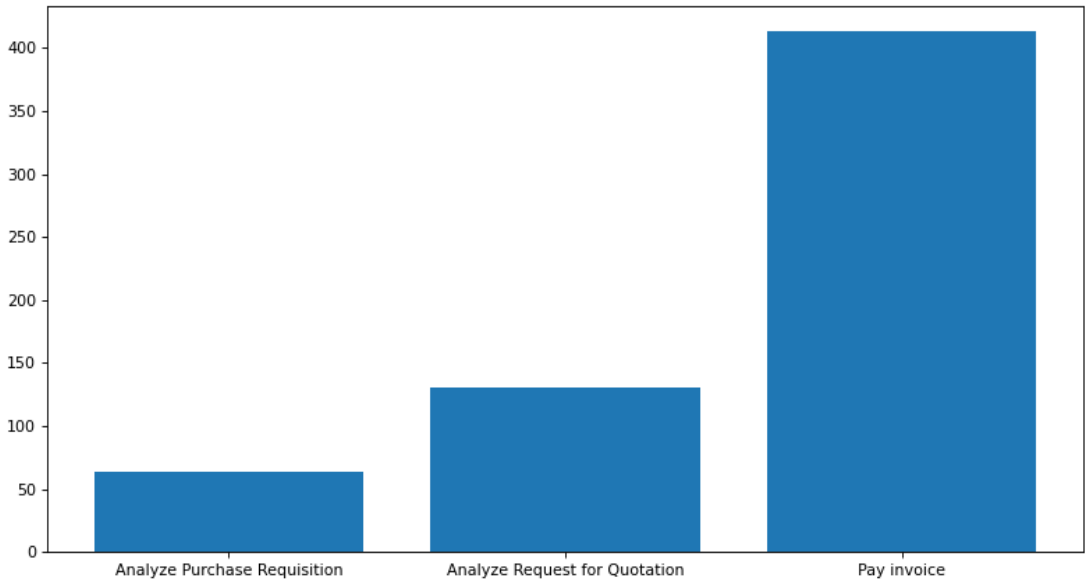
Εικόνα 22 : Πλήθος activities

Τα γραφήματα που ακολουθούν αφορούν το dataset Purchasing-Example.csv

Στα παρακάτων 2 γραφήματα απεικονίζονται τα start και end activities. Είναι πολύ χρήσιμα, διότι μπορούμε να δούμε πως όλες οι διεργασίες ξεκινούν με τον ίδιο τρόπο όμως δεν τελειώνουν με τον ίδιο τρόπο. Έτσι μπορούμε να καταλαβουμε με μία ματιά τι ίσως δεν έχει τελειώσει και τι καθυστερεί.

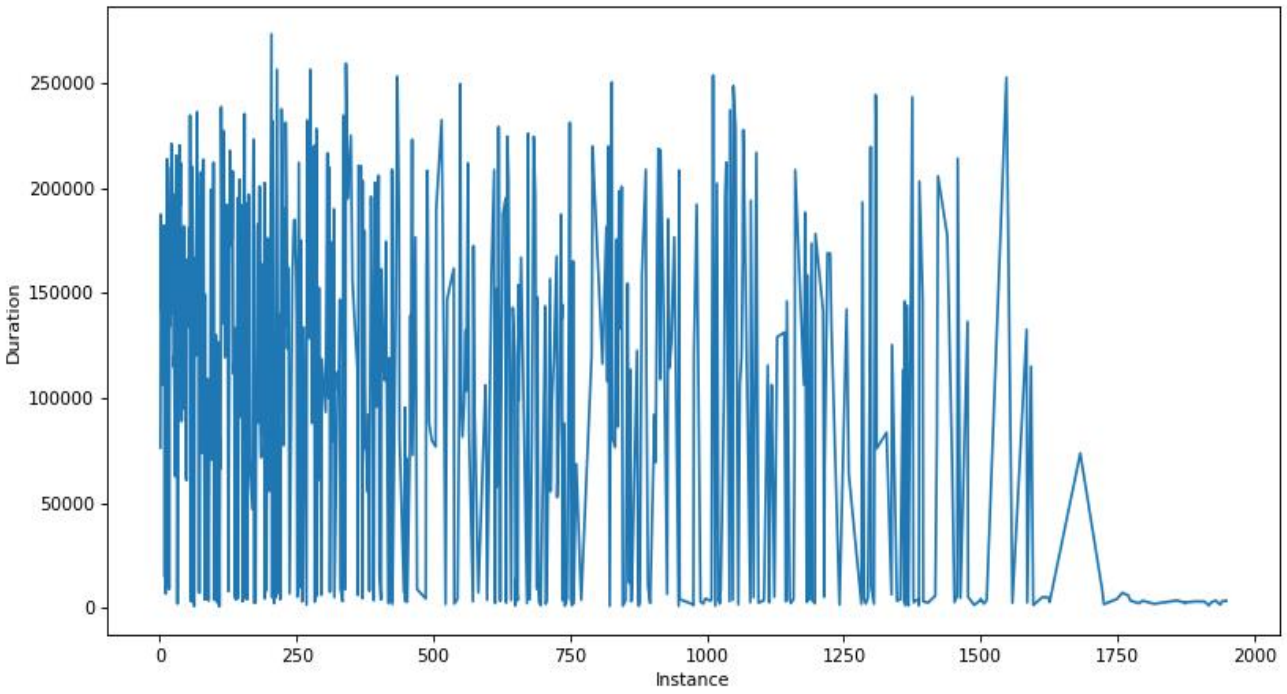


Εικόνα 23 : Start activities



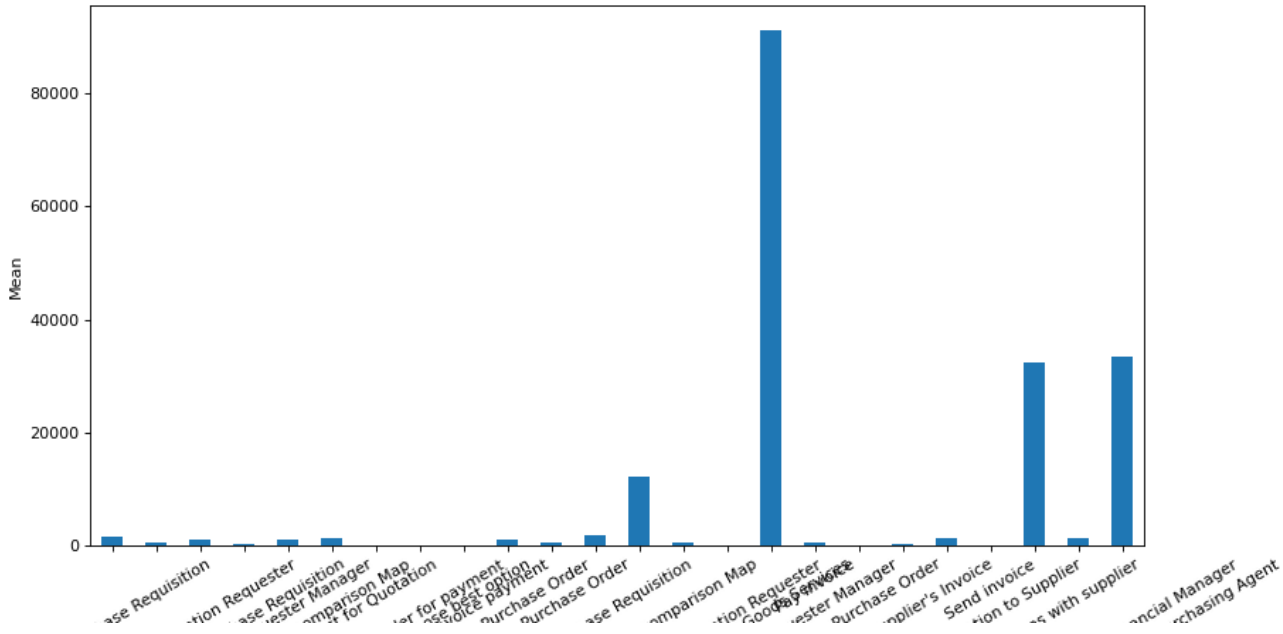
Εικόνα 24: End activities

Στο παρακάτω γράφημα απεικονίζεται η διάρκεια κάθε activity.



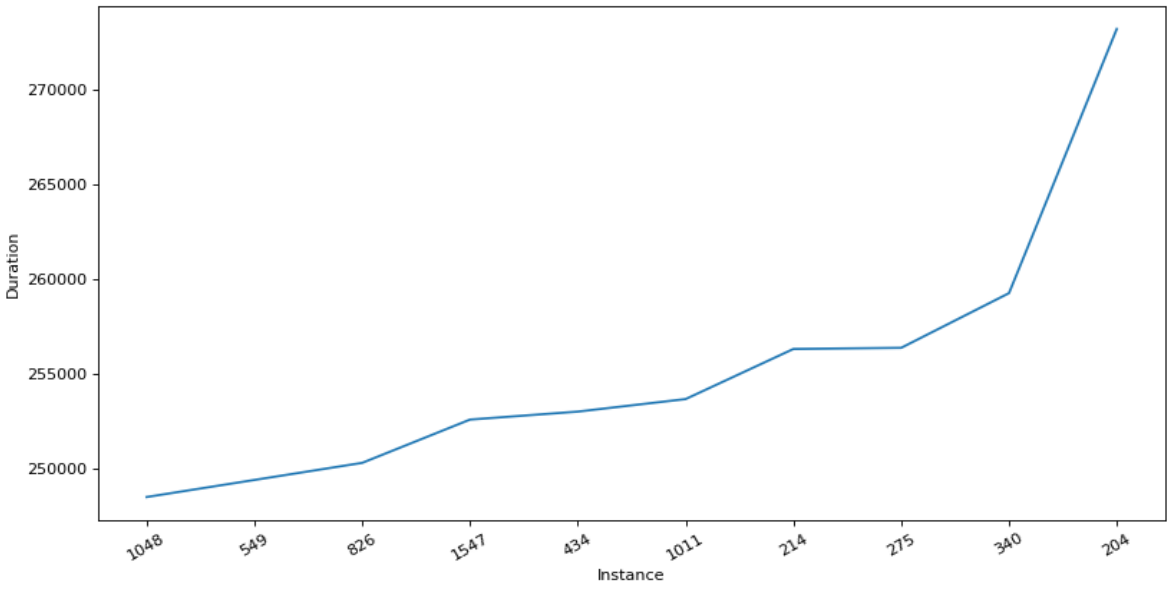
Εικόνα 25: Διάρκεια activities

Σε αυτό το γράφημα απεικονίζεται η μέση διάρκεια κάθε activity και είναι πολύ χρήσιμο διότι μπορούμε να δούμε ποια διεργασία καθυστερεί περισσότερο να ολοκληρωθεί.

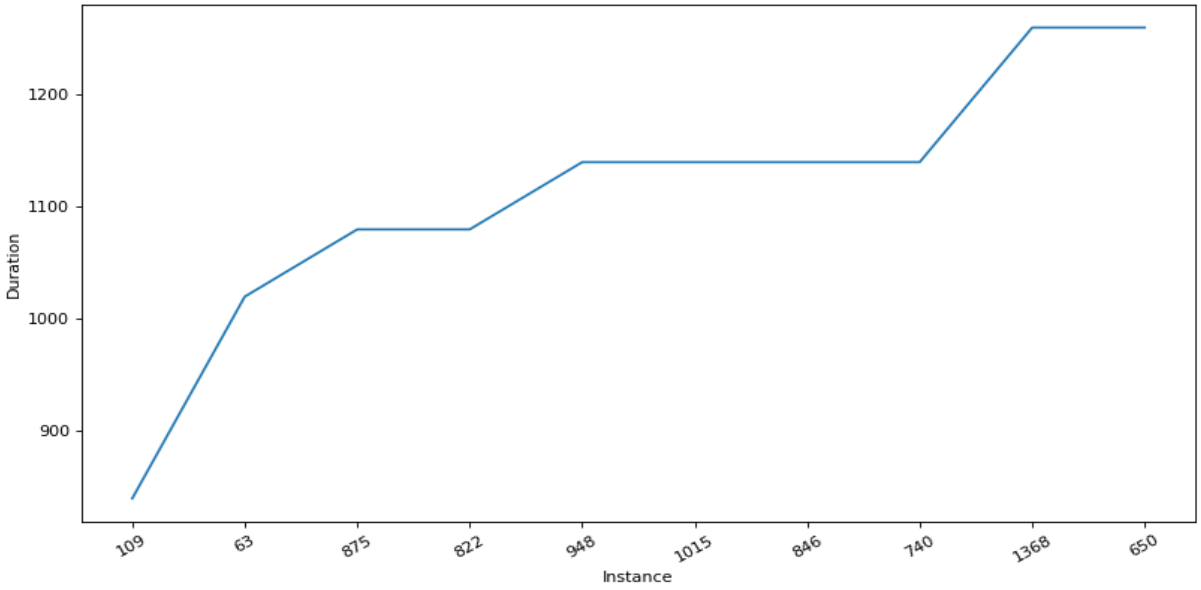


Εικόνα 26 : Μεση διάρκεια activity

Στα γραφήματα που ακολουθούν βλέπουμε τα traces μεγαλύτερης και μικρότερης διάρκειας αντίστοιχα. Αυτό μας βοηθά στο να επικεντρωθούμε στη βελτίωση των συγκεκριμένων ενδεχομένως «προβληματικών» traces.



Εικόνα 27: Μεγαλύτερες 10 διάρκειες trace



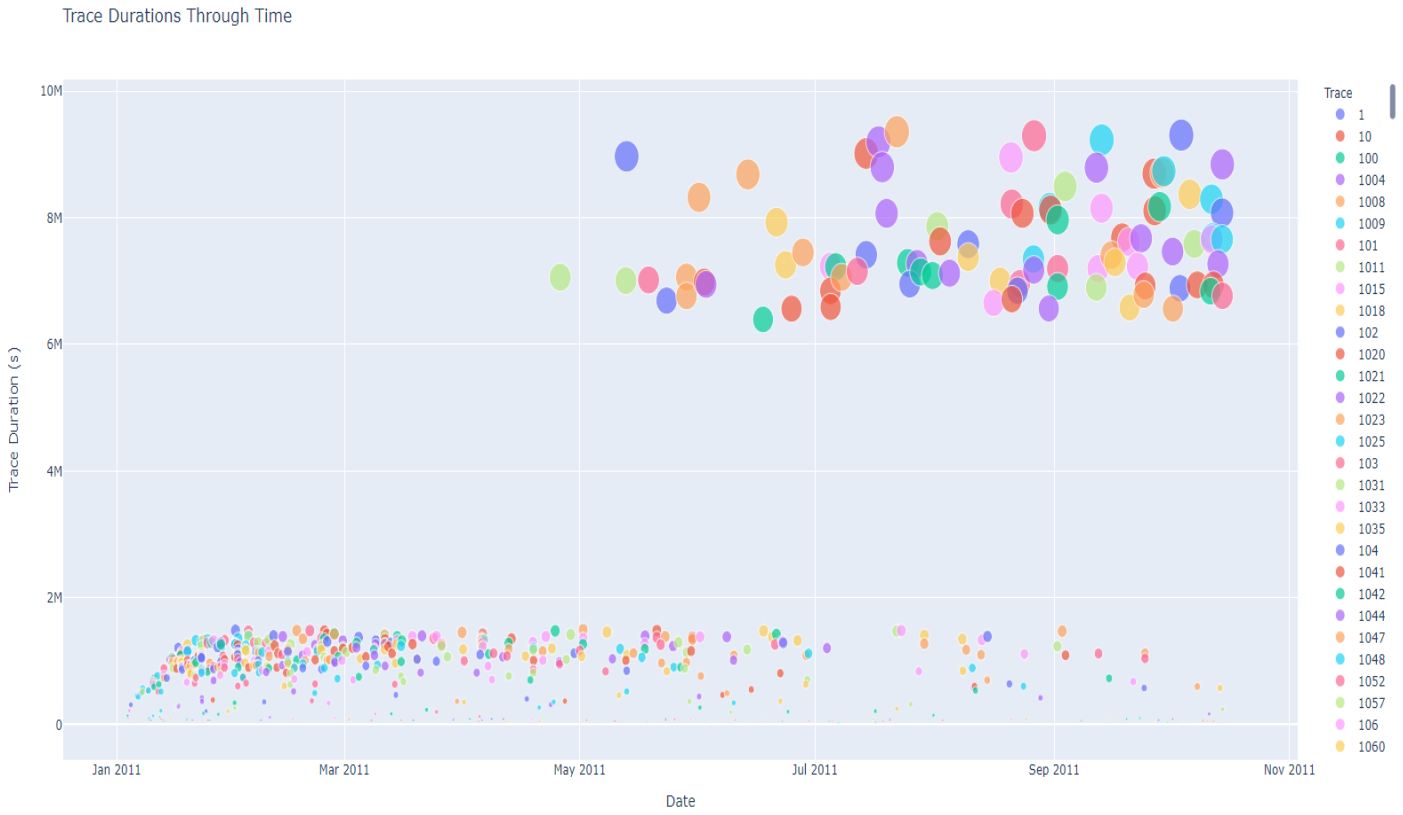
Εικόνα 28 : Μικρότερες 10 διάρκειες trace

Στο διάγραμμα που ακολουθεί έχουμε σχεδιάσει τη διάρκεια των διαφορετικών δραστηριοτήτων στο σύνολο δεδομένων σε σχέση με την ώρα έναρξης. Κάθε σημείο δεδομένων έχει χρωματική κωδικοποίηση βάσει του activity που αντιπροσωπεύει.
Από αυτό το διάγραμμα μπορούμε να δούμε ποιες δραστηριότητες διήρκεσαν περισσότερο και πώς άλλαξε αυτή η διάρκεια τους μήνες του 2011. Στην περίπτωσή μας, οι διάρκειες φαίνεται να ακολουθούν μια ομοιόμορφη κατανομή με το χρόνο, ενώ η δραστηριότητα «Παράδοση αγαθών και υπηρεσιών» φαίνεται να είναι αυτή που διαρκεί περισσότερο.

Εικόνα 29 : Activity durations through time

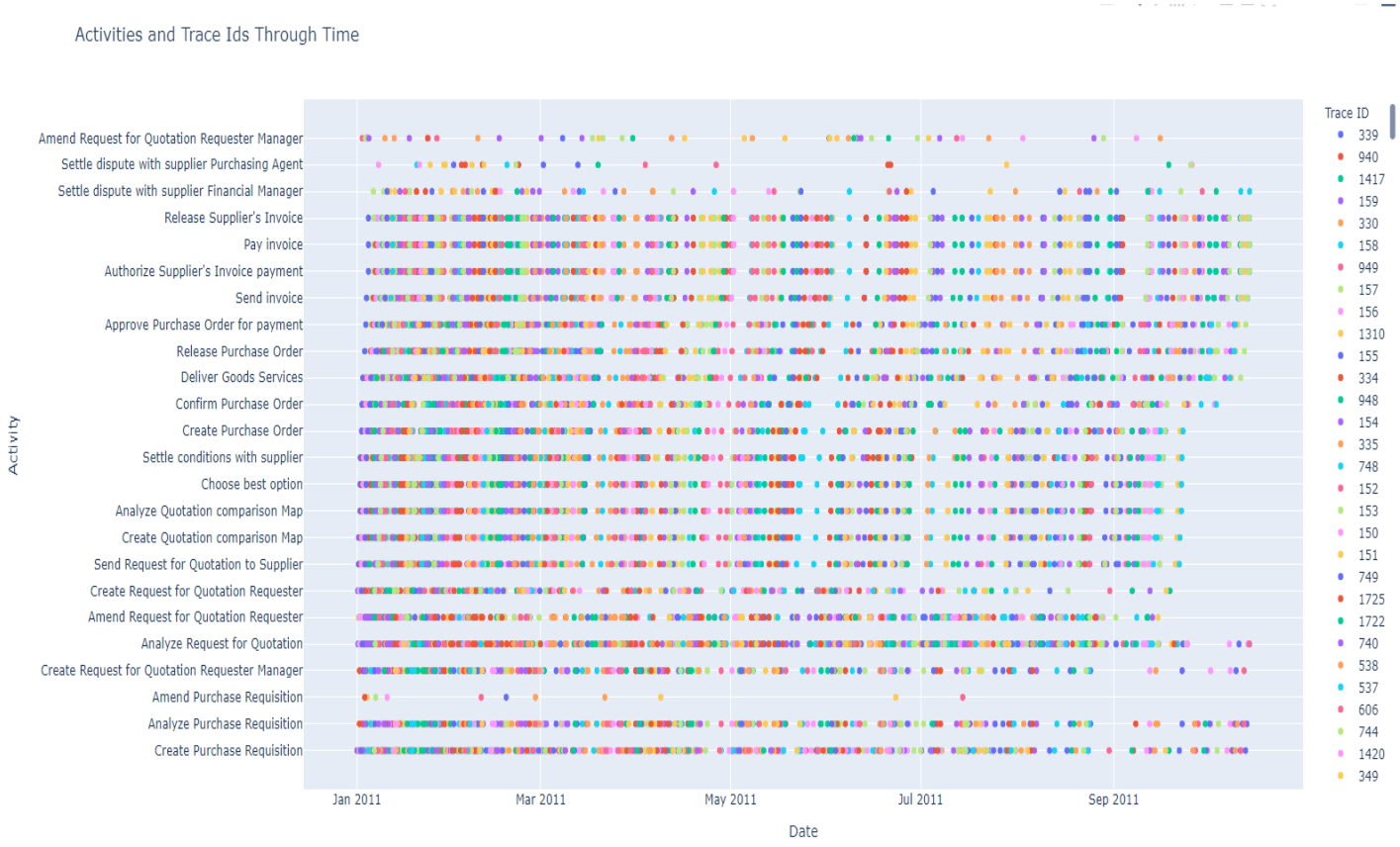
Το παρακάτω διάγραμμα είναι παρόμοιο με το παραπάνω μόνο που τώρα ομαδοποιήσαμε με τον συνολικό χρόνο που χρειάστηκε για να ολοκληρωθούν τα traces. Με αυτό μπορούμε να δούμε ποια traces χρειάστηκαν περισσότερο χρόνο για να τελειώσουν και ποιες ημέρες.

Μια ενδιαφέρουσα παρατήρηση εδώ είναι ότι τα ίχνη μετά τον Μάιο, χρειάστηκαν πολύ περισσότερο χρόνο για να τελειώσουν.



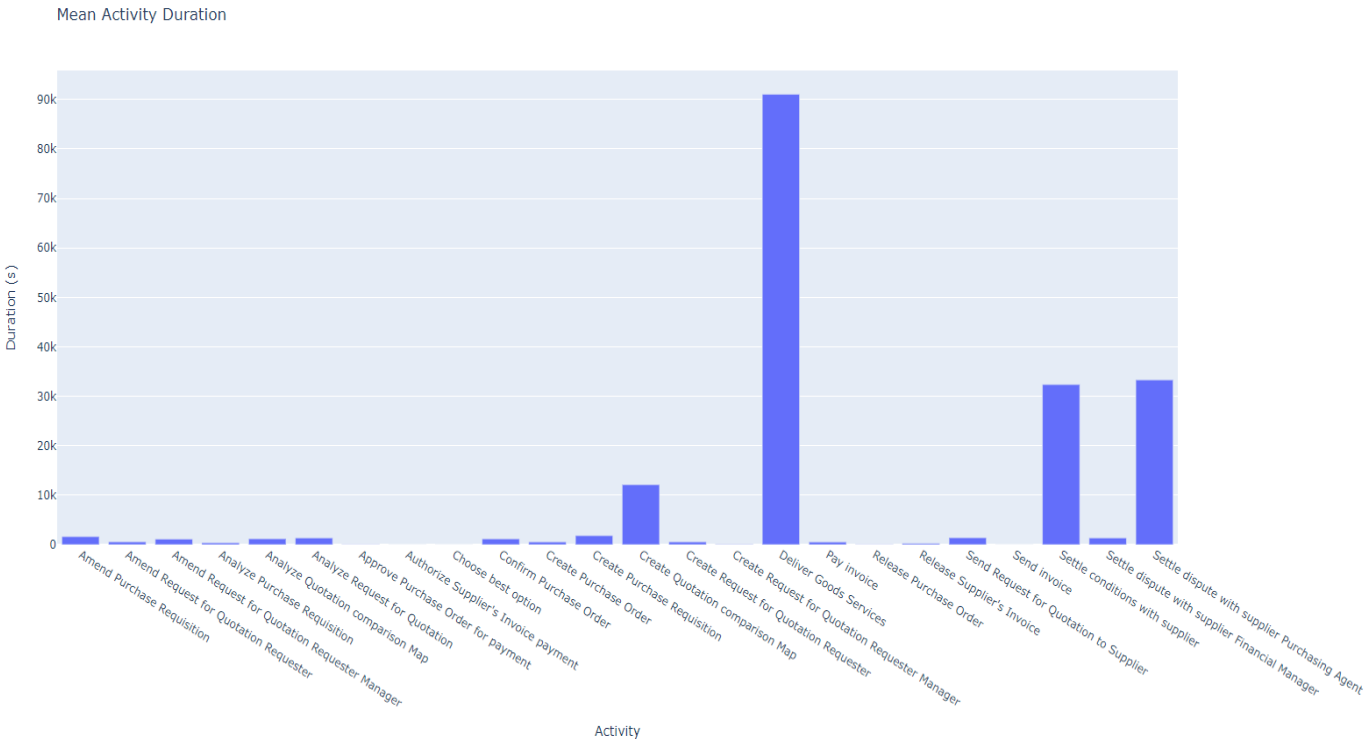
Εικόνα 30 : Trace durations through time

Αυτό είναι ένα διάγραμμα συνδυασμού των ημερομηνιών έναρξης κάθε ίχνους και των δραστηριοτήτων του.



Εικόνα 31 : Activities and Trace Ids through Time

Τέλος, αυτό είναι το ίδιο διάγραμμα με αυτό που υπάρχει παραπάνω (βλ. εικόνα 26), μόνο που τώρα έχει δημιουργηθεί με τη plotly express.



Εικόνα 32 : mean activity duration

Ο κώδικας των τελευταίων 4 γραφημάτων παρατίθεται στο Παράρτημα 3 και έχει γίνει αποθήκευσή τους σε html αρχείο.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Συνοψίζοντας, με τη βοήθεια της οπτικοποίησης μπορούν να απεικονιστούν σε γραφήματα ιδιότητες των δεδομένων, σχέσεις συνάφειας, συγκρίσεις τιμών, γεωγραφική διασπορά συμβάντων, ανοδικές και καθοδικές τάσεις, επιμερισμός συνόλων σε υποσύνολα και πολλές άλλες πληροφορίες. Η πληροφόρηση αυτή παρέχεται με τρόπο κατανοητό μιας και ο ανθρώπινος εγκέφαλος κατανοεί καλύτερα και γρηγορότερα μια πληροφορία, όταν αυτή αποτυπώνεται σε μια εικόνα, παρά όταν περιγράφεται με μορφή αναλυτικού κειμένου.

Επιπλέον, η γραφική απεικόνιση της πληροφορίας είναι καλαίσθητη και σαφώς πιο ευχάριστη από την ανάγνωση κειμένου. Αυτές οι ιδιότητες της οπτικοποίησης την έχουν καταστήσει χρήσιμο εργαλείο για την ανάλυση των δεδομένων και την εξαγωγή συμπερασμάτων.

Χρυσή εποχή των στατιστικών γραφικών αποτέλεσε το δεύτερο μισό του 19ου αιώνα, όταν η επιστήμη της στατιστικής γνώρισε αλματώδη εξέλιξη με τις εργασίες των Gauss και Laplace. Επίσης, το δεύτερο μισό του 20ου αιώνα η οπτικοποίηση των δεδομένων γνωρίζει νέα άνθηση με τη συμβολή του Bertin (1983), που συνδέει στοιχεία των γραφικών με τα χαρακτηριστικά και τις σχέσεις των δεδομένων, καθώς και με τις εργασίες του Tukey (1977), ο οποίος εισήγαγε τη Διερευνητική Ανάλυση των Δεδομένων (Exploratory Data Analysis). Φυσικά, η έλευση της πληροφορική την ίδια εποχή έδωσε νέα τεράστια ώθηση στην οπτικοποίηση των δεδομένων.[21]

Η απεικόνιση δεδομένων με γραφικό τρόπο δεν είναι πάντα μια εύκολη εργασία και δεν υπάρχει μια μαγική συνταγή που να εξασφαλίζει ένα ποιοτικό αποτέλεσμα. Σε μεγάλο βαθμό το αποτέλεσμα εξαρτάται από τη δημιουργικότητα και τη φαντασία του σχεδιαστή. Για ένα σύνολο δεδομένων και για μια εργασία

[21] https://repository.kallipos.gr/bitstream/11419/1232/2/Kef._5.pdf (σελ.2) (ανακτήθηκε 3/3/2021)

ανάλυσης, η επιλογή της κατάλληλης τεχνικής οπτικοποίησης πρέπει όπως έχει ήδη αναφερθεί να λαμβάνει υπόψη της μια σειρά από παράγοντες.

Βασικότερο μειονέκτημα των τεχνικών οπτικοποίησης είναι ότι οι νέοι σύνθετοι τρόποι μπορεί να μην είναι κατανοητοί από εξειδικευμένους χρήστες. Η δυσχέρεια στην κατανόηση τους μπορεί να επιφέρει σύγχυση. Επίσης, υπάρχει ο κίνδυνος της εσφαλμένης ερμηνείας της οπτικής πληροφορίας.[22]

Στην παρούσα εργασία λαμβάνοντας υπόψη τα datasets και το τι πρέπει να βρεθεί ή να παρουσιαστεί, τη φύση των δεδομένων, το πλήθος των διαστάσεων και τη δομή των δεδομένων, έγινε οπτικοποίηση των events, καθώς και του αριθμού εμφανίσεων του καθενός, τα start και end activities και πόσες φορές αυτά εμφανίζονται μέσα στο event log, τη μέση διάρκειά τους με δύο διαφορετικούς τρόπους και τη διάρκεια των 10 μεγαλύτερων και 10 μικρότερων traces, με αντικειμενικό τρόπο, καθώς μόνο στη λεκτική περιγραφή μπορεί να αντανακλάται ή να υποκρύπτεται υποκειμενική αντίληψη.

Επίσης, έχουν δημιουργηθεί διάγραμμα για τη διάρκεια των διαφορετικών δραστηριοτήτων στο σύνολο δεδομένων σε σχέση με την ώρα έναρξης με χρωματική κωδικοποίηση για εύκολη εξαγωγή συμπερασμάτων, διάγραμμα συνδυασμού των ημερομηνιών έναρξης κάθε ίχνους και των δραστηριοτήτων του και διάγραμμα που απεικονίζει ποια traces χρειάστηκαν περισσότερο χρόνο για να τελειώσουν και ποιες ημέρες.

Συμπερασματικά, μπορεί να ειπωθεί πως τα δεδομένα στη σύγχρονη εποχή είναι ασταμάτητα και η έξυπνη ανάλυση τους αποτελεί έναν πολύτιμο πόρο. Η οπτικοποίηση και η οπτική ανάλυση τους ως τεχνική βοηθά στην πληρέστερη κατανόηση τους και σε αρκετές περιπτώσεις στην πρόβλεψη του μέλλοντος.

[22] https://repository.kallipos.gr/bitstream/11419/1232/2/Kef._5.pdf (σελ.3-4) (ανακτήθηκε 2/4/2021)

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Αβούρης Ν., Μ. Κουκιάς, Β. Παλιουράς, Κ. Σγάρμπας, "PYTHON Εισαγωγή στους υπολογιστές"[1]. 3η αναθεωρημένη έκδοση, Πανεπιστημιακές Εκδόσεις Κρήτης, 2016.
2. Αράπογλου Α.,Βραχνός Ε.,Λέκκα Δ.,Κανίδης Ε.,Μακρυγιάννης Π., Μπελεσιώτης Β., Τζήμας Δ., Παπαδάκης Σπ., «Προγραμματισμός Υπολογιστών Γ' Τάξη ΕΠΑ.Λ.» με χρήση *Python 2*, Διδακτικό Υλικό, Εκδόσεις Διόφαντος. (ISBN 978-960-06-5309-0)
3. Γούλου Ζωή.,(2010). Εφαρμογή μεθόδων εξόρυξης δεδομένων στη διαχείριση πελατειακών σχέσεων. Ανακτήθηκε στις 4/1/2021 από http://dspace.lib.uom.gr/bitstream/2159/14808/6/GoulouZoiMsc2012.pdf
4. Καρολίδης Δ., "Μαθαίνετε εύκολα Python", 2η έκδοση, Εκδόσεις Άβακας, 2018.
5. Aalst, W. van der, "Process Mining", Communications of the ACM, August 2012, Vol. 55 No. 8, Pages 76-83
6. Aalst, W. van der. "*Process Mining: Discovery, Conformance and Enhancement of Business Processes*", Springer-Verlag, Berlin, 2011.
7. Aalst, W. van der, Hee, K. van, Werf, J.M. van der, and Verdonk, M. Auditing 2.0: Using process mining to support tomorrow's auditor. *IEEE Computer 43*, 3 (Mar. 2010).
8. Baker Ryan S.J.d, Data Mining for Education, (2008), Carnegie Mellon University, Pittsburgh, Pennsylvania, USA (http://www.columbia.edu/~rsb2162/Encyclopedia%20Chapter%20Draft%20v10%20-fw.pdf)
9. Dean, J. (2014). Big Data, Data Mining and Machine Learning. Hoboken, New Jersey: Wiley.
10. Fayyad, U, Piatetsky-Shapiro, G. & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17 (3), pp. 37-54

11. Filz Marc-André, Gellrich S., Herrmann C., Thiede S., «Data-driven Analysis of Product State Propagation in Manufacturing Systems Using Visual Analytics and Machine Learning» Procedia CIRP Volume 93, 2020, Pages 449-454 (ανακτήθηκε από το https://www.sciencedirect.com/science/article/pii/S2212827120306740 στις 3/4/2021)
12. Friedman, V. (2008) "Data Visualization and Infographics", Smashing Magazine (http://www.smashingmagazine.com)
13. Kaluzaa A., Gellricha S., Cerdasa F., Thiedea S., Herrmanna C., ScienceDirect, "Life cycle engineering based on visual analytics", 25th CIRP Life Cycle Engineering (LCE) Conference, 30 April – 2 May 2018, Copenhagen, Denmark, p.37-39, από το sciencedirect.com (ανακτήθηκε 27/3/2021)
14. Kehrer J., Hauser H., Visualization and Visual Analysis of Multi-faceted Scientific Data: A Survey, PUBLISHED IN IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 19, NO. 3, MARCH 2013
15. Keim, D. A. (2002). Information Visualization and Visual Data Mining. IEEE Transaction on Visualization and Computer Graphics, 7(1), 100-107. doi: 10.1109/2945.981847
16. Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H., . Visual Analytics: Scope and Challenges, in: Simoff, S.J., Böhlen, M.H., Mazeika, A. (Eds.), Visual data mining. Theory, techniques and tools for visual analytics, vol. 4404. Springer, Berlin, 2008, pp. 76–90.
17. Mao LinHuang, Tze-HawHuang, XuyunZhang, "A novel virtual node approach for interactive visual analytics of big datasets in parallel coordinates", Future Generation Computer Systems, Volume 55, February 2016, Pages 510-523
18. McConnell, Steve (30 Νοεμβρίου 2009). *Code Complete, p. 100*. ISBN 9780735636972.
19. Miller, J. D. (2017). Big Data Visualization. Packt Publishing Ltd.Wong P.C and Bergeron R.D., 1994, "30 Years of

Multidimensional Multivariate Visualization, IEEE Computer Society, Washington DC, USA, p.3-33

20. Simmi Bagga., Dr. G.N. Singh., (2012).Applications of Data Mining. Ανακτήθηκε στις 4/1/2021 από http://www.ijsett.com/images/P5.pdf. Αρχειοθετήθηκε 2016-11-23 στο Wayback Machine.
21. Soban D., Thornhill D., Salunkhe S., Long A., "Visual Analytics as an Enabler for Manufacturing Process Decision-making", Procedia CIRP. Volume 56, 2016, Pages 209-214
22. Wong P.C and Bergeron R.D., 1994, "30 Years of Multidimensional Multivariate Visualization, IEEE Computer Society, Washington DC, USA, p.3-33

ΠΗΓΕΣ

1. «History of Python» από Python-course.eu. Ανακτήθηκε 28/12/2020.
2. Learning IPython for Interactive Computing and Data Visualization (2013) Packt Publishing (https://www.packtpub.com/product/learning-iPython-for-interactive-computing-and-data-visualization/9781782169932).
3. https://docs.Python.org/3/library/index.html (https://medium.com/@GoldenGatePro/Python-libraries-data-science-bbc98c1bb148 (ανακτήθηκε 3/1/2021)
4. https://plotly.com/Python/plotly-express/
5. https://medium.com/plotly/introducing-plotly-express-808df010143d
6. «The Making of Python: A Conversation with Guido van Rossum, Part I by Bill Venners» στο artima.com. Δημοσιεύθηκε 13/01/2003. Ανακτήθηκε 29/12/2020. ανακτήθηκε 3/1/2021)
7. https://www.visual-analytics.eu/faq/ (ανακτήθηκε 22/3/2021)
8. Αγγελιδάκης, Νικόλαος Α., Εκπαιδευτικός Πληροφορικής, Μ.Δ.Ε. (M.Sc.) στην Επιστήμη Υπολογιστών (Ηράκλειο, Αύγουστος 2015). "Εισαγωγή στον προγραμματισμό με την Python".
9. Ράτσης, Κωνστανίνος Ρ., Διπλωματική εργασία «Γραφικές μέθοδοι παρουσίασης πολυδιάστατων δεδομένων», Πανεπιστήμιο Πειραιά, 2009
10. Εικόνα 1 https://repository.kallipos.gr/bitstream/11419/2966/1/02_chapter_01.pdf (σελ.15) (ανακτήθηκε 1/1/2021)
11. Εικόνα 2 : Python leads the 11 top Data Science, Machine Learning platforms: Trends and Analysis (kdnuggets.com) (ανακτήθηκε 2/1/2021)
12. Εικόνα 3 :

LEON ZHAO J., City University of Hong Kong, HARRY JIANNAN WANG, University of Delaware, WIL VAN DER AALST, Technische Universiteit Eindhoven, Editorial: "Business Process Intelligence: Connecting Data and Processes", April 2015, ανακτήθηκε στις 25/1/2021 από : https://www.researchgate.net/publication/283024393_EditorialBusiness_Process_Intelligence_Connecting_Data_and_Processes

13. Εικόνα 4 :
Wil van der Aalst, Eindhoven University of Technology, "Event Logs What kind of data does process mining require?", ανακτήθηκε από : http://www.processmining.org/_media/presentations/event_logs_the_input_for_process_mining.pdf στις 26/1/2021
14. Εικόνα 5 : https://www.geeksforgeeks.org/calculate-efficiency-binary-classifier/ (ανακτήθηκε 10/4/2021
15. Εικόνα 6
https://el.wikipedia.org/wiki/%CE%91%CF%80%CE%BB%CE%AE_%CE%B3%CF%81%CE%B1%CE%BC%CE%BC%CE%B9%CE%BA%CE%AE_%CF%80%CE%B1%CE%BB%CE%B9%CE%BD%CE%B4%CF%81%CF%8C%CE%BC%CE%B7%CF%83%CE%B7
(ανακτήθηκε 4/1/2021).
16. Εικόνα 7
https://www.datasciencecentral.com/profiles/blogs/fine-grained-analysis-of-k-mean-clustering-and-where-we-are-using
(ανακτήθηκε 1/1/2021)
17. Εικόνα 8
https://datacadamia.com/data_mining/association
(ανακτήθηκε 1/1/2021)
18. Εικόνα 10 :
https://repository.kallipos.gr/bitstream/11419/1232/2/Kef._5.pdf
(σελ.121) (ανακτήθηκε 23/2/2021)
19. Εικόνα 11 : https://www.vismaster.eu/faq/the-visual-analytics-process/ (ανακτήθηκε 23/2/2021)

20. Εικόνα 12 : https://repository.kallipos.gr/bitstream/11419/2966/1/02_chapter_01.pdf (ανακτήθηκε 15/2/2021)
21. Εικόνα 13,16,17 : Encyclopedia of Database Systems https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-39940-9_1131 (ανακτήθηκε 6/1/2021)
22. Εικόνα 14 : https://www.isko.org/cyclo/hypertext (ανακτήθηκε 14/4/2021)
23. Εικόνα 15 : https://www.google.com/url?sa=i&url=https%3A%2F%2Fnemertes.lis.upatras.gr%2Fjspui%2Fbitstream%2F10889%2F862%2F1%2FNimertis_Gkiza.pdf&psig=AOvVaw3ByrJS6MLPm12owDkEd6Ky&ust=1618469895258000&source=images&cd=vfe&ved=0CAMQjB1qFwoTCIjvxJ6Y_e8CFQAAAAAdAAAAABAJ (ανακτήθηκε 14/4/2021)
24. Εικόνα 18 : https://study.com/academy/lesson/transformations-in-math-definition-graph-quiz.html (ανακτήθηκε 6/1/2021)
25. Εικόνα 19 : https://www.researchgate.net/figure/Dimensional-stacking-plot-displaying-300-simulation-cases-colored-by-outcome-scenario_fig17_311628858 (ανακτήθηκε 6/1/2021)
26. Εικόνα 20 : https://el.wikipedia.org/wiki/%CE%A3%CF%85%CF%83%CF%84%CE%B1%CE%B4%CE%BF%CF%80%CE%BF%CE%AF%CE%B7%CF%83%CE%B7 (ανακτήθηκε 15/2/2021)
27. Εικόνα 21 : https://www.sciencedirect.com/science/article/abs/pii/S0167739X15000382?via%3Dihub
28. https://datacadamia.com/data_mining/association (ανακτήθηκε 15/2/2021)
29. https://www.visual-analytics.eu/ (ανακτήθηκε 22/3/2021)
30. https://www.sciencedirect.com/science/article/pii/S2212827120306740 (ανακτήθηκε 1/4/2021)

Παράρτημα 1

Αλγόριθμος 1: Κώδικας υλοποίησης (1)

```
1 from pm4py.objects.log.importer.xes import factory as xes_importer
2
3 import matplotlib.pylab as plt
4 from matplotlib.pyplot import figure
5
6 log = xes_importer.apply('datasets/Artificial - Loan Process.xes')
7 freq_dic = {}
8 for case_index, case in enumerate(log):
9 for event_index, event in enumerate(case):
10 activity = event["concept:name"]
11 if(activity in freq_dic.keys()):
12 freq_dic[activity] = freq_dic[activity] + 1
13 else:
14 freq_dic[activity] = 1
15
16 figure(num=None, figsize=(12, 6), dpi=80, facecolor='w', edgecolor='k')
17 plt.bar(*zip(*sorted(freq_dic.items())))
18
19 plt.savefig("graph.png")
```

Παράρτημα 2

Αλγόριθμος 2: Κώδικας υλοποίησης (2)

```
1 import os
2 from pm4py.objects.log.adapters.pandas import csv_import_adapter
3 from pm4py.objects.conversion.log import converter as log_converter
4 from pm4py.algo.filtering.log.end_activities import end_activities_filter
5 from pm4py.algo.filtering.log.start_activities import start_activities_filter
6 import pandas as pd
7 from pm4py.util import constants
8
9 import matplotlib.pylab as plt
10 from matplotlib.pyplot import figure
11 from pm4py.algo.filtering.log.start_activities import start_activities_filter
12 from pm4py.algo.filtering.log.end_activities import end_activities_filter
13 import datetime as dt
14
15 csv_file_name = "Purchasing-Example"
16 # PART 1-2
17 dataframe =
```

```
  →
csv_import_adapter.import_dataframe_from_path(os.path.join("datase
ts",
  → csv_file_name + ".csv"), sep=",")
18 dataframe.rename(columns={'Case ID': 'case:concept:name',
'Activity' :
  → 'concept:name', 'Start Timestamp' : 'concept:timestamp'},
inplace=True)
19 event_log = log_converter.apply(dataframe,
  → variant=log_converter.Variants.TO_EVENT_LOG)
20 start_activities = start_activities_filter.get_start_activities(event_log)
21 end_activities = end_activities_filter.get_end_activities(event_log)
22
23 print("Start Activities: ")
24 for start_activity in start_activities.keys():
25     print(start_activity + ": " + str(start_activities[start_activity]))
26
27 print("End Activities: ")
28 for end_activity in end_activities.keys():
29     print(end_activity + ": " + str(end_activities[end_activity]))
30
31 figure(num=None, figsize=(12, 6), dpi=80, facecolor='w',
edgecolor='k')
32 plt.bar(*zip(*sorted(start_activities.items())))
33 plt.savefig("start_activities.png")
34 plt.clf()
```

```
35 plt.bar(*zip(*sorted(end_activities.items())))
36 plt.savefig("end_activities.png")
37 plt.clf()
38 # PART 3 time in seconds
39 #2019-01-09 19:28:00+00:00
40 pattern = "%Y-%m-%d %H:%M:%S %z"
41 # dataframe["concept:timestamp_epoch"] = (dataframe['concept:timestamp'] -
 → dt.datetime(1970,1,1)).dt.total_seconds() * 1000
42 # dataframe["End_epoch"] = (dataframe['End'] -
 → dt.datetime(1970,1,1)).dt.total_seconds() * 1000
43
44 dataframe["concept:timestamp_epoch"] =
 → dataframe['concept:timestamp'].astype('int64') // 1e9
45 dataframe["complete_timestamp_epoch"] = dataframe['Complete
 → Timestamp'].astype('int64') // 1e9
46
47 dataframe["duration"] = dataframe['complete_timestamp_epoch'] -
 → dataframe["concept:timestamp_epoch"]
48 print(dataframe.groupby("concept:name", as_index=False)["duration"].mean())
49 df_reduced = dataframe.groupby("concept:name", as_index=False)["duration"].mean()
50 print(df_reduced)
51 ax = df_reduced['duration'].plot(kind='bar')
52 ax.set_xlabel('Activity') # replace with the labels you want
```

```
53 ax.set_ylabel('Mean')
54 ax.set_xticklabels(df_reduced["concept:name"])
55 plt.xticks(rotation=30)
56 plt.show()
57 plt.savefig("mean_duration.png")
58 plt.clf()
59
60 # PART 4 time in seconds
61 durations = dataframe.groupby("case:concept:name",
  → as_index=True)["duration"].sum()
62 durations_sorted = durations.sort_values()
63 ax = durations.plot(kind='line')
64 ax.set_xlabel('Instance') # replace with the labels you want
65 ax.set_ylabel('Duration')
66 plt.show()
67 plt.savefig("durations.png")
68 plt.clf()
69
70 smallest_10_durations = durations_sorted[:10].values
71 smallest_10_keys = durations_sorted[:10].keys().astype('U').values
72 top_10_durations = durations_sorted.tail(10).values
73 top_10_keys = durations_sorted.tail(10).keys().astype('U').values
74 tickvalues = range(0,10)
75
76 plt.plot(top_10_durations)
```

```
77 plt.xlabel('Instance') # replace with the labels you want
78 plt.ylabel('Duration')
79 plt.xticks(tickvalues, labels = top_10_keys, rotation=30)
80 plt.show()
81 plt.savefig("top_10_durations.png")
82 plt.clf()
83
84 plt.plot(smallest_10_durations)
85 plt.xlabel('Instance') # replace with the labels you want
86 plt.ylabel('Duration')
87 plt.xticks(tickvalues, labels = smallest_10_keys, rotation=30)
88 plt.show()
89 plt.savefig("smallest_10_durations.png")
90 plt.clf()
```

Παράρτημα 3

Στο παράρτημα αυτό παρατίθεται ο κώδικας ανάπτυξης των τελευταίων 4 διαγραμμάτων.

```
import os
import plotly.express as px
import pandas as pd
from pm4py.objects.conversion.log import converter as csv_converter
from pm4py.objects.log.importer.csv import importer as csv_importer
from plotly.subplots import make_subplots
import plotly.graph_objects as go

# Grouping function to get the trace duration
def first_last(df):
    if(len(df) > 0):
        df_grouped = df.iloc[-1]
        df_grouped["trace_duration"] = (df.iloc[-1]["complete_timestamp_epoch"] - df.iloc[0]["time:timestamp_epoch"])
# Endtime of last activity - start time of first
    return df_grouped

log = csv_importer.apply('datasets/Purchasing-Example.csv')
dataframe = csv_converter.apply(log, variant=csv_converter.Variants.TO_DATA_FRAME)
dataframe.rename(columns={'Case ID': 'case:concept:name', 'Activity' : 'concept:name', 'Start Timestamp' : 'time:timestamp'}, inplace=True)
```

Get activity Durations like the last exercise

```
pattern = "%Y-%m-%d %H:%M:%S %z"

dataframe["time:timestamp"] =
pd.to_datetime(dataframe["time:timestamp"])

dataframe['Complete Timestamp'] =
pd.to_datetime(dataframe['Complete Timestamp'])

dataframe["time:timestamp_epoch"] =
dataframe['time:timestamp'].astype('int64') // 1e9

dataframe["complete_timestamp_epoch"] = dataframe['Complete
Timestamp'].astype('int64') // 1e9

dataframe["duration"] = dataframe['complete_timestamp_epoch'] -
dataframe["time:timestamp_epoch"]
```

Base figure

```
fig = px.scatter(dataframe, x="time:timestamp", y="concept:name",
color="case:concept:name", title="Activities and Trace Ids Through
Time", labels={
                "concept:name": "Activity",
                "case:concept:name" : "Trace ID",
                "time:timestamp": "Date"})
fig.write_html('conceptname.html')
```

Durations of each Activity

```
fig = px.scatter(dataframe, y="duration", x = "time:timestamp",
color="concept:name", size="duration", title="Activity Durations
Through Time", labels={
                "duration": "Activity Duration (s)",
                "time:timestamp": "Date"
            })
fig.write_html('activity_durations.html')
```

Trace Durations

```
df_grouped_trace = dataframe.groupby("case:concept:name",
as_index=False).apply(first_last)
fig = px.scatter(df_grouped_trace, y="trace_duration", x =
"time:timestamp", color = "case:concept:name", size="trace_duration",
                                title="Trace Durations Through Time",
labels={
                "trace_duration": "Trace Duration (s)",
                "case:concept:name": "Trace",
                "time:timestamp": "Date"
            })
fig.write_html('trace_durations.html')
```

```
# Mean Durations

df_reduced = dataframe.groupby("concept:name",
as_index=False)["duration"].mean()

fig = px.bar(df_reduced, y="duration", x ="concept:name", title="Mean
Activity Duration", labels={

                "duration": "Duration (s)",

                "concept:name": "Activity"

            })

fig.write_html('mean_durations.html')
```