ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ ΕΣΩΤΕΡΙΚΩΝ

εκδδα
ΕΘΝΙΚΟ ΚΕΝΤΡΟ ΔΗΜΟΣΙΑΣ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΑΥΤΟΔΙΟΙΚΗΣΗΣ

ΕΘΝΙΚΗ ΣΧΟΛΗ ΔΗΜΟΣΙΑΣ ΔΙΟΙΚΗΣΗΣ

ΚΑΙ ΑΥΤΟΔΙΟΙΚΗΣΗΣ

ΚΖ΄ ΕΚΠΑΙΔΕΥΤΙΚΗ ΣΕΙΡΑ
ΤΕΛΙΚΗ ΕΡΓΑΣΙΑ

ΤΙΤΛΟΣ

Η εξόρυξη δεδομένων (Data Mining) και η εφαρμογή της στη λήψη αποφάσεων στη Δημόσια Διοίκηση

ΤΜ. ΕΞΕΙΔΙΚΕΥΣΗΣ: ΨΗΦΙΑΚΗΣ ΠΟΛΙΤΙΚΗΣ

Επιβλέπων:

Απόστολος Ζήβελδης

Σπουδαστής:

Μιχάλης Γεωργακάς

ΑΘΗΝΑ - 2022

Η εξόρυξη δεδομένων (Data Mining) και η εφαρμογή της στη λήψη αποφάσεων στη Δημόσια Διοίκηση

ΔΗΛΩΣΗ

«Δηλώνω ρητά ότι, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής εργασίας, δεν παραβιάζει καθ' οιονδήποτε τρόπο πνευματικά δικαιώματα τρίτων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής».

Αθήνα, 19 / 5 / 2022

Υπογραφή

ΠΕΡΙΛΗΨΗ

Η τεχνολογική πρόοδος, σε συνδυασμό με την εξέλιξη των πληροφοριακών συστημάτων, την αναβάθμιση των δυνατοτήτων του λογισμικού καθώς και την αύξηση του όγκου των αποθηκευμένων δεδομένων, συντέλεσε στη διείσδυση της πληροφορικής σε κάθε δραστηριότητα της σύγχρονης κοινωνίας με αποτέλεσμα τη δημιουργία μιας κοινωνίας βασισμένης στις πληροφορίες. Οι πληροφορίες στην ακατέργαστη τους μορφή αποτελούν δεδομένα, δηλαδή καταγεγραμμένα γεγονότα τα οποία αποθηκεύονται σε ψηφιακά μέσα τις βάσεις δεδομένων. Την τελευταία δεκαετία παρατηρείται η αλματώδης αύξηση στην παραγωγή δεδομένων τόσο από επιχειρήσεις και οργανισμούς όσο και από τους ίδιους τους πολίτες δημιουργώντας ένα απόθεμα κρυμμένης γνώσης που πρέπει να ανακαλυφθεί. Σκοπός της ανακάλυψης γνώσης που προκύπτει μέσα από την εξόρυξη και επεξεργασία των δεδομένων είναι η άντληση κρίσιμων πληροφοριών που θα μπορούσαν να συμβάλλουν στην υποστήριξη λήψης αποφάσεων. Η ύπαρξη τεράστιας ποσότητας δεδομένων, πληροφοριών που με τις κατάλληλες τεχνικές μπορούν να οδηγήσουν σε γνώση, δηλαδή σε αποκάλυψη πληροφοριών που ακόμα δεν έχουν ανακαλυφθεί, αποτελεί ζήτημα που η Δημόσια Διοίκηση οφείλει να αντιμετωπίσει.

Η παρούσα εργασία αποσκοπεί στο να δώσει μια σύντομη και ταυτόχρονα περιεκτική αποτύπωση των τεχνικών και των αλγορίθμων που χρησιμοποιούνται για την εξαγωγή γνώσης από τις βάσεις δεδομένων, με σκοπό την υποστήριξη λήψης αποφάσεων για την παραγωγή δράσεων και δημοσίων πολιτικών από τη Δημόσια Διοίκηση. Για το σκοπό αυτό η εργασία διαρθρώνεται σε δύο μέρη.

Στο πρώτο μέρος, επιχειρείται η θεωρητική αποτύπωση της εξόρυξης δεδομένων, μέσω βασικών διαδικασιών και τεχνικών εξόρυξης όπως αυτές παρουσιάζονται στην διεθνή βιβλιογραφία. Επίσης παρουσιάστηκαν εφαρμογές εξόρυξης και ανακάλυψης γνώσης από βάσεις δεδομένων σε Δημόσιες Διοικήσεις άλλων χωρών με σκοπό την υποστήριξη στη λήψη αποφάσεων.

Στο δεύτερο μέρος της εργασίας παρουσιάζεται η ερευνητική προσέγγιση εξόρυξης δεδομένων μέσω της κατασκευής βάσης δεδομένων από ερευνητικό δείγμα 298 ατόμων που συλλέχθηκε πρωτογενώς για τις ανάγκες της παρούσας εργασίας. Σκοπός ήταν η διερεύνηση των παραγόντων που οδηγούν στην ύπαρξη μηνιαίων ληξιπρόθεσμων οφειλών. Για την ερευνητική προσέγγιση, εφαρμόστηκαν στατιστικές αναλύσεις και

τεχνικές εξόρυξης δεδομένων με σκοπό την ανακάλυψη «κρυμμένης» πληροφορίας η οποία θα ήταν δυνατό να υποστηρίξει τη λήψη αποφάσεων στη Δημόσια Διοίκηση.

Λέξεις κλειδιά: Εξόρυξη Δεδομένων, Ανακάλυψη Γνώσης, Βάσεις Δεδομένων, Κατηγοριοποίηση, Συσταδοποίηση.

ABSTRACT

The technological progress, coupled with the evolution of IT systems, the upgrading of the capabilities of software, as well as the increase of the volume of the stored data, has contributed to the infiltration of informatics in every aspect of the modern society, resulting to the formation of a society based on information. Information, in its unprocessed form constitutes data, that is recorded facts stored in digital means, known as data bases. The last decade there is an exponential increase in the creation of data by enterprises and organizations as well as citizens themselves, accumulating a reserve of "hidden" knowledge, which must be discovered. The purpose of the knowledge discovery which arises through data mining and processing is to obtain crucial information which could contribute to decision making. The existence of huge quantities of information, which could lead to knowledge via suitable techniques, that is the revelation of information which has not been discovered yet, forms an issue that Public Administration has to face.

The current assignment aims at offering a brief and at the same time comprehensive portrayal of the techniques and the algorithms used to extract knowledge by data bases, aiming at the support of decision making in order to produce actions and public policies by the Public Administration. To this purpose this assignment is divided in two parts.

In the first part, a theoretical portrayal of data mining is attempted through basic procedures and techniques of mining as they are presented in international bibliography. Furthermore, methods of mining and knowledge discovery by data bases of other countries Public Administration systems, will be presented, aiming at the support of decision making.

The second part of the assignment presents the exploratory research of data mining through the creation of a data base by a control group of 298 individuals that was collected especially for the purposes of the current assignment. The purpose was the exploration of the factors that lead to past due debts. For the research approach, statistical analysis and data mining techniques were used aiming at the discovery of covert information, which could support decision making in Public Administration.

Keywords: Data Mining, Knowledge Discovery, Data Base, Classification, Clustering

Πίνακας περιεχομένων

Πίνακας Εικονογράφησης

Πίνακας Συντμήσεων και Συντομογραφιών

άρ.	άρθρο
Βλ.	Βλέπε
ΒΨΜ	Βίβλος Ψηφιακού Μετασχηματισμού
ΓΚΠΔ	Γενικός Κανονισμός Προστασίας Δεδομένων
ΔΔ	Δημόσια Διοίκηση
ΕΔ	Εξόρυξη Δεδομένων
ΜΔ	Μεγάλα Δεδομένα
ΜΚΟ	Μη Κυβερνητική Οργάνωση
ΣΔ	Συλλογή Δεδομένων
ΤΝΔ	Τεχνητό Νευρωνικό Δίκτυο
ΤΠΕ	Τεχνολογίες Πληροφορικής και Επικοινωνιών
KDD	Knowledge Discovery in Databases
DM	Data Mining
NN	Neural Network

ΜΕΡΟΣ Α Θεωρητική Προσέγγιση

Κεφάλαιο 1

1.1Εισαγωγή

«Η επιστήμη είναι δεδομένα. Ακριβώς όπως τα σπίτια χτίζονται με τούβλα, έτσι και η επιστήμη χτίζεται με δεδομένα. Αλλά όπως ένας σωρός τούβλα δεν κάνει ένα σπίτι, έτσι και μια συλλογή δεδομένων δεν είναι απαραίτητα επιστήμη».[1] Αυτή η φράση η οποία συνοψίζει τη σχέση μεταξύ δεδομένων και επιστήμης, αποτελεί τη γενεσιουργό αιτία και νοηματοδοτεί το αντικείμενο της Επιστήμης των Δεδομένων ενός ανερχόμενου κλάδου που διαδραματίζει καταλυτικό ρόλο στο παγκόσμιο γίγνεσθαι.

Ως αφετηριακό σημείο της επιστήμης των δεδομένων, η συλλογή, παραγωγή και αποθήκευση δεδομένων δεν αποτελεί όρο καινοφανή. Η πρώτη απόπειρα του ανθρώπου να συλλέξει δεδομένα ανατρέχει χιλιάδες χρόνια πριν, όταν οι άνθρωποι των σπηλαίων χρησιμοποιούσαν ψηλά μπαστούνια για την καταγραφή των αποθεμάτων τους. Αργότερα, η εφεύρεση του αριθμητηρίου, το Κόσκινο του Ερατοσθένη, η δημιουργία της βιβλιοθήκης της Αλεξάνδρειας, η ανακάλυψη του μηχανισμού των Αντικυθήρων, η διεξαγωγή του πρώτου καταγεγραμμένου πειράματος στατιστικής ανάλυσης για τον περιορισμό της εξάπλωσης της πανούκλας στην Ευρώπη από τον J. Graunt, η πρόβλεψη του N. Tesla περί της δυνατότητας του ανθρώπου να έχει πρόσβαση και να αναλύει τεράστιες ποσότητες δεδομένων χρησιμοποιώντας μια μικρή συσκευή που θα χωράει στην τσέπη του, η εφεύρεση από τον F. Pfleumer ενός τρόπου μαγνητικής αποθήκευσης δεδομένων (γεγονός το οποίο αποτέλεσε τη βάση της σύγχρονης τεχνολογίας της ψηφιακής αποθήκευσης δεδομένων), η μηχανή Τούρινγκ, η εμφάνιση της έννοιας της επιχειρησιακής ευφυίας το 1958, η εισαγωγή από την IBM του σχεσιακού μοντέλου βάσης δεδομένων, δηλωτικό του ότι ο καθένας πλέον μπορούσε να χειρίζεται μια βάση δεδομένων, η εμφάνιση του όρου «Big Data» το 1989, η εδραίωση του διαδικτύου και η δημιουργία κβαντικού επεξεργαστή το 2019 αποδεικνύουν ότι η αλληλεπίδραση ανθρώπου-δεδομένων υπήρξε αδιάλειπτη και συνεχώς εξελισσόμενη.

Η κυριαρχία του homo digitalis υποδηλώνει τη σταδιακή μετάβαση από την εποχή της πληροφορίας στην εποχή των δεδομένων. Ήδη από το 1996, οι Fayyad, Piatetsky, Shapiro and Smyth παρατηρούν ότι υπάρχει ένας δραματικός ρυθμός στην παραγωγή

[1] Ανρί Πουανκαρέ, 1854-1912, Γάλλος μαθηματικός

δεδομένων. Η θέση τους επιβεβαιώνεται στο παρόν από τη διαπίστωση ότι ο παραγόμενος ανά τον κόσμο όγκος των δεδομένων αυξάνεται εκθετικά και αναμένεται να ανέλθει σε 175 zettabytes το 2025[2]. Ο άνθρωπος πλέον δεν περιορίζεται στο ρόλο ενός παθητικού καταναλωτή πληροφοριών αλλά μετατρέπεται σε πρωταγωνιστικό υποκείμενο παραγωγής πληροφορίας. Η φρενήρης παραγωγή δεδομένων καλύπτει ένα ευρύτατο φάσμα ανθρώπινων και όχι μόνο δραστηριοτήτων ενώ αυτά διαφέρουν πολλαπλώς μεταξύ τους τόσο σε μορφή όσο και στην ταχύτητα συλλογής. Τα δεδομένα όμως από μόνα τους μοιάζουν με τα τούβλα του Πουανκαρέ.

Η πληροφορία που μπορεί να αντληθεί και να αξιοποιηθεί από την επεξεργασία των δεδομένων αποτελεί την πρόκληση και το μέγιστο διακύβευμα ώστε όντως τα δεδομένα να αποτελέσουν το «νέο πετρέλαιο» (Data is not the oil, but the soil) (McCandless, 2010) ως μείζων ευκαιρία για τις κυβερνήσεις, την οικονομία και τη βιωσιμότητα. Σε αυτό ακριβώς το σημείο αναδεικνύεται η συμβολή της Επιστήμης των Δεδομένων. Αντικείμενό της αποτελεί η διερεύνηση και η ανάλυση ψηφιακών δεδομένων με σκοπό την εξαγωγή κανονικοτήτων και θεωριών, άλλως τη δημιουργία μοντέλων, τα οποία να μπορούν να χρησιμοποιηθούν τόσο για την πρόβλεψη, όσο και για την ερμηνεία-περιγραφή των δεδομένων.

Η πρόβλεψη αναφέρεται στη διαδικασία χρήσης μεταβλητών ή πεδίων μίας βάσης δεδομένων με σκοπό την εκτίμηση άγνωστης ή μελλοντικής τιμής ενός άλλου γνωρίσματος. Η περιγραφή (σε μορφή σύνοψης ή περιληπτικής παρουσίασης) των δεδομένων εστιάζει στην εύρεση κατανοητών από τον άνθρωπο προτύπων, τα οποία περιγράφουν τα δεδομένα (Βερύκιος, Καγκλής, Σταυρόπουλος, 2015). Ορίζεται επομένως ως μία ημι-αυτοματοποιημένη διαδικασία, σκοπός της οποίας είναι να αναλύσει ένα μεγάλο όγκο δεδομένων που αφορούν ένα συγκεκριμένο πρόβλημα, για την παραγωγή προτύπων και να καταλήξει στην «ανακάλυψη της γνώσης».

1.2 Εξόρυξη Δεδομένων

Η τεχνολογική πρόοδος και η εξέλιξη των ηλεκτρονικών υπολογιστών έχει καταστήσει τη χρήση τους απαραίτητη με αποτέλεσμα τη δημιουργία μιας κοινωνίας που βασίζεται στις πληροφορίες. Οι πληροφορίες στην ακατέργαστή τους μορφή ονομάζονται δεδομένα, δηλαδή καταγεγραμμένα γεγονότα τα οποία αποθηκεύονται σε ψηφιακά μέσα τις βάσεις δεδομένων. Την τελευταία δεκαετία παρατηρείται η αλματώδης

[2] Ευρωπαϊκή Στρατηγική για τα Δεδομένα

αύξηση στην παραγωγή δεδομένων τόσο από επιχειρήσεις και οργανισμούς όσο και από τους ίδιους τους πολίτες. Η ύπαρξη τεράστιας ποσότητας πληροφοριών στις βάσεις δεδομένων, πληροφορίες που με τις κατάλληλες τεχνικές μπορούν να οδηγήσουν σε γνώση, δηλαδή σε αποκάλυψη πληροφοριών που ακόμα δεν έχουν ανακαλυφθεί.

Εύλογα διερωτάται κανείς τι σημαίνει ανακάλυψη της γνώσης. Η γνώση συνδέεται με ό,τι εξάγεται από την ανάλυση των δεδομένων. Πρόκειται για την αποκάλυψη ή παραγωγή λειτουργικής γνώσης μέσα από την τελευταία, η οποία γίνεται αντιληπτή από τον άνθρωπο. Αναφέρεται σε ολόκληρη τη διαδικασία, από τη συλλογή δεδομένων μέχρι την αξιοποίηση των αποτελεσμάτων σε πιο πρακτικό επίπεδο. Τα βασικά στάδια της διαδικασίας ανακάλυψης της γνώσης (Han, Kamber, Pei, 2012) είναι η Συλλογή Δεδομένων (Data Collection), η Προεπεξεργασία αυτών (Preprocessing), ο Μετασχηματισμός τους (Transformation), η Εξόρυξη Δεδομένων (Data Mining) και η Διερμηνεία/Αξιολόγησή τους (Interpretation/Evaluation).

Πεδία εφαρμογής της επιστήμης των δεδομένων αποτελούν οι περισσότεροι τομείς στην καθημερινότητα του σύγχρονου ανθρώπου με αποτέλεσμα οι τομείς που αλληλοεπιδρούν με την επιστήμη των δεδομένων να είναι σχεδόν σε κάθε πεδίο της σύγχρονης κοινωνίας άμεσα ή έμμεσα. Ενδεικτικά, πεδία εφαρμογής της επιστήμης των δεδομένων αποτελούν τα χρηματοοικονομικά μιας και αφορούν δεδομένα συναλλαγών που συλλέγονται και διακινούνται σε συνεχώς αυξανομένους όγκους, οι αγορές χρήματος - αξίων καθώς και οι αποφάσεις σε πραγματικό χρόνο (αγορά, πώληση) εν μέσω συνεχομένης ροής δεδομένων, η δημιουργία πιστωτικού προφίλ των τραπεζών καθώς και τα μακροοικονομικά μεγέθη που αφορούν τις εθνικές οικονομίες. Η εκπαίδευση, η υγεία αλλά και η δημόσια πολιτική αποτελούν βασικούς τομείς εφαρμογής της επιστήμης των δεδομένων μιας και αφορά πολιτικές με βάση δεδομένα πολιτών, κανονισμούς και ανοικτά δημόσια δεδομένα.

Στο πεδίο της δημόσιας διοίκησης, η πρόσφατη εμπειρία διαχείρισης των αναγκών που ανέκυψαν από την πανδημία του COVID-19, αφενός επιτάχυνε τη διαδικασία του ψηφιακού μετασχηματισμού, αφετέρου κατέστησε την ανάλυση των δεδομένων καταλύτη για την πραγμάτωση της ψηφιακής μετάβασης. Δεν είναι τυχαίος ο χαρακτηρισμός ότι τα «big data» καθόρισαν τη διακυβέρνηση της πανδημίας. Στην προσπάθεια απλούστευσης και επιτάχυνσης διαδικασιών, δημιουργίας ολοκληρωμένων πληροφοριακών συστημάτων, ανάπτυξης ψηφιακών υπηρεσιών και βελτιστοποίησης της λήψης αποφάσεων προς δημιουργία ολοκληρωμένων δημόσιων πολιτικών, η

συγκέντρωση τεράστιου όγκου δεδομένων μεγάλης κλίμακας απαιτεί τεχνικές εξόρυξης δεδομένων, ώστε η προαναφερόμενη ανακάλυψη της γνώσης να οδηγήσει στην ανασύνταξη του δημοσίου τομέα στις βάσεις της διοίκησης ολικής ποιότητας και στην επιτυχή ανταπόκρισή του στις πρωτόγνωρες προκλήσεις.

1.3 Ορισμός Εξόρυξης δεδομένων

Η Εξόρυξη δεδομένων (data mining) αποτελεί διαδικασία εύρεσης χρήσιμων προτύπων στα δεδομένα με σκοπό τη χρησιμοποίηση των προτύπων που ανακαλύπτονται από τα δεδομένα για την ερμηνεία τρεχόντων και μελλοντικών ενδεχομένων (Dunham, 2002). Η διαδικασία της εξόρυξης δεδομένων περιλαμβάνει την συλλογή και αποθήκευση, την επιλογή και προετοιμασία των δεδομένων, τη δημιουργία και έλεγχο, την ερμηνεία της εγκυρότητας των αποτελεσμάτων καθώς και την εφαρμογή των μοντέλων. Στη διεθνή βιβλιογραφία έχουν προταθεί διάφοροι ορισμοί για την εξόρυξη αλλά και ανακάλυψη γνώσης από δεδομένα.

Ως εξόρυξη δεδομένων (data mining) ορίζεται η διαδικασία χρήσης τεχνικών εκμάθησης υπολογιστών για την αυτόματη ανάλυση και εξαγωγή γνώσης από δεδομένα που περιέχονται σε μια βάση δεδομένων. (Roitzer & Geatz, 2008)

Η εξόρυξη δεδομένων είναι ένα σύνολο τεχνικών και στρατηγικών για την ανακάλυψη πληροφορίας, μοτίβων και προτύπων από μια βάση δεδομένων με τη χρήση αλγορίθμων και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων, όπου η αδόμητη πληροφορία μετασχηματίζεται σε πληροφορία που μπορεί να οδηγήσει σε λήψη αποφάσεων (Han 2012).

Με τον όρο Εξόρυξη Δεδομένων καλείται η διαδικασία μέσω της χρήσης ηλεκτρονικών υπολογιστών για την άντληση χρήσιμων πληροφοριών από μεγάλα σύνολα δεδομένων (Hand et al. 2001), η διαδικασία αναγνώρισης έγκυρων και χρήσιμων προτύπων στα δεδομένα (Fayyad, 1996), και η εξεύρεση σημαντικών και άγνωστων πληροφοριών. Η εξόρυξη γνώσης από μια βάση δεδομένων αναφέρεται στη διαδικασία ανακάλυψης χρήσιμης πληροφορίας από μεγάλα σύνολα δεδομένων. Όμως ο ορισμός που παρουσιάζει με περισσότερη σαφήνεια την έννοα της ανακάλυψης γνώσης από βάσεις δεδομένων δόθηκε από τους Frawley, Piatesky-Shaphiro & Matheus (1991) υποστηρίζοντας πως η *«Ανακάλυψη γνώσης από δεδομένα είναι η ντετερμινιστική*

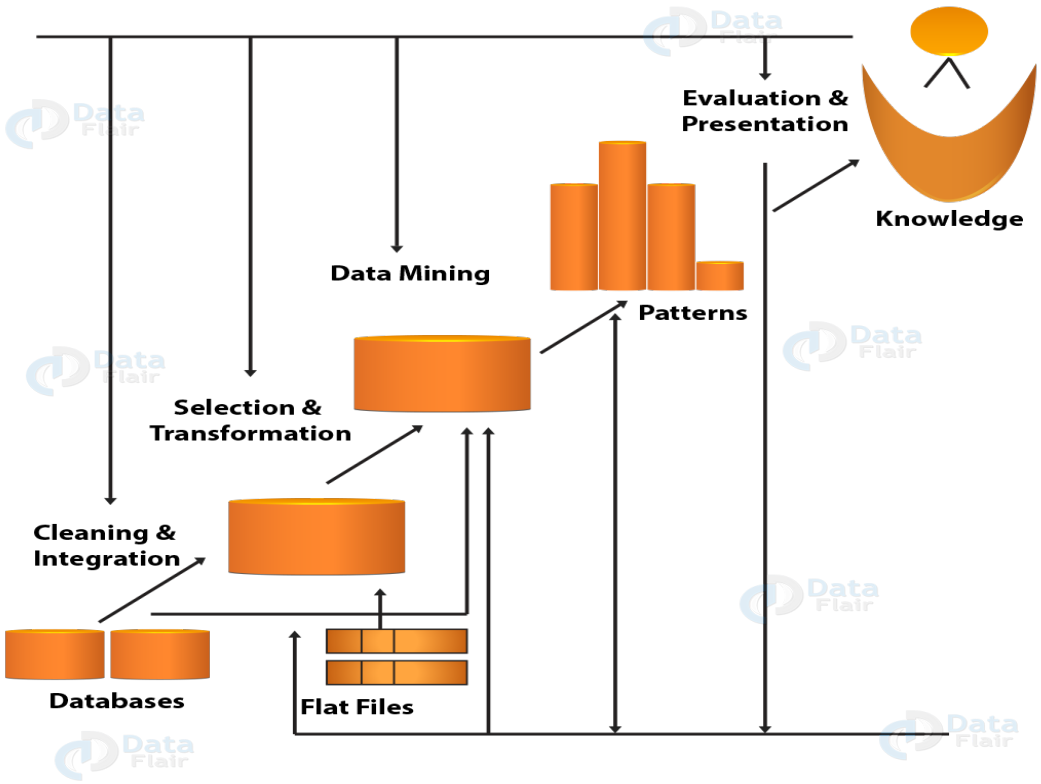
διαδικασία αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων στα δεδομένα».

1.4 Ανακάλυψη γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases, KDD)

Η ανακάλυψη γνώσης σε βάσεις δεδομένων (Knowledge Discovery in Databases – KDD) όρος συναφής με την εξόρυξη δεδομένων, μιας και η εξόρυξη δεδομένων αποτελεί βήμα της KDD διαδικασίας, αναφέρεται στην εφαρμογή της επιστημονικής μεθόδου για την εξόρυξη της πληροφορίας δεδομένου του ότι η εξόρυξη της πληροφορίας σε ένα μοντέλο διαδικασίας KDD ενσωματώνεται με συγκεκριμένη μεθοδολογία (Frawley, Piatesky-Shaphiro & Matheus 1991). Σκοπός αυτής της μεθοδολογίας είναι η εξαγωγή, η προετοιμασία και ο έλεγχός των δεδομένων με σκοπό τη λήψη των αποφάσεων που σχετίζονται με τις απαιτούμενες ενέργειες που είναι απαραίτητο να γίνουν με την ολοκλήρωση της εξόρυξης. Η εξαγωγή και προετοιμασία των δεδομένων αποτελεί χρονοβόρα διαδικασία για την εξόρυξη γνώσης από δεδομένα, ειδικά στην ανάλυση μεγάλου όγκου δεδομένων τα οποία και έχουν αποθηκευτεί σε διάφορες θέσεις.

Όπως διαπιστώνεται από τους ορισμούς η εξόρυξη δεδομένων αναφέρεται εκτός των άλλων στη μάθηση, η οποία αποτελεί μια σύνθετη διαδικασία για το τι μπορούν να μάθουν οι υπολογιστές. Σύμφωνα με τους Merril και Tennyson (1977) μπορούν να διακριθούν τέσσερα επίπεδα μάθησης, τα γεγονότα, οι έννοιες, οι διαδικασίες και οι αρχές. Τα γεγονότα αντικατοπτρίζουν μια απλή δήλωση αληθείας, ενώ οι έννοιες αφορούν μια ομάδα αντικειμένων, συμβόλων ή συμβάντων τα οποία και ομαδοποιούνται εξαιτίας των συγκεκριμένων κοινών χαρακτηριστικών που διαθέτουν. Διαδικασίες αποτελούν οι απαραίτητες ενέργειες που χρειάζονται για την επίτευξη ενός σκοπού. Τέλος οι αρχές αναφέρονται στο υψηλότερο επίπεδο εκμάθησης και αντιπροσωπεύουν τους νόμους και τις γενικές αλήθειες που είναι βασικοί για την εξαγωγή συμπερασμάτων. Δεδομένης της δυνατότητας των υπολογιστών να διαχειρίζονται με μεγάλη ευκολία τις έννοιες δηλαδή την ομάδα των αντικειμένων, συμβόλων ή συμβάντων τα οποία και ομαδοποιούνται εξαιτίας των συγκεκριμένων κοινών χαρακτηριστικών που διαθέτουν το εργαλείο εξόρυξης δεδομένων είναι αυτό που υπαγορεύει την μορφή των εννοιών. Στις δομές των εννοιών συγκαταλέγονται μαθηματικές εξισώσεις, δίκτυα, δένδρα και δίκτυα που ο επιστήμονας της ανάλυσης δεδομένων οφείλει να χρησιμοποιήσει και να

ερμηνεύσει για την εξαγωγή συμπερασμάτων και θεωριών με στόχο την ερμηνεία και την πρόβλεψη.



Εικόνα 1: Διαδικασία Ανακάλυψης Γνώσης από τα δεδομένα. Πηγή: Data Flair

Υπάρχουν τρεις διαφορετικές θεωρήσεις που αφορούν τις έννοιες. Η κλασσική (classical view), η πιθανοτική (probabilistic view) και η παραδειγματική (exemplar view). Στην κλασσική θεώρηση υπάρχει ο ισχυρισμός ότι όλες οι έννοιες έχουν σαφή χαρακτηριστικά καθορισμού, στην πιθανοτική οι έννοιες αντιπροσωπεύονται από κάποια χαρακτηριστικά που ενδεχομένως και να διαθέτουν τα μέλη των εννοιών και στην παραδειγματική θεώρηση όπου ένα δεδομένο στιγμιότυπο προσδιορίζεται ως παράδειγμα μιας συγκεκριμένης έννοιας (Roiger & Geatz, 2008). Οι έννοιες οδηγούν στην καθοδηγούμενη μάθηση (supervised learning) δηλαδή στη δημιουργία μοντέλων κατηγοριοποίησης από σύνολα δεδομένων που περιέχουν παραδείγματα των συνόλων υπό εκμάθηση και αφού δημιουργηθεί το μοντέλο της κατηγοριοποίησης χρησιμοποιείται για να καθορίσει την κατηγοριοποίηση των στιγμιότυπων άγνωστης προέλευσης.

Σε αντίθεση με την καθοδηγούμενη εκμάθηση η μη καθοδηγούμενη εκμάθηση (unsupervised leanring) δημιουργεί μοντέλα από δεδομένα χωρίς να έχουν καθοριστεί οι

αντίστοιχες κατηγορίες. Αυτά τα στιγμιότυπα δεδομένων ομαδοποιούνται βάσει ενός σχήματος ομοιότητας το οποίο καθορίζεται από το σύστημα συσταδοποίησης. Με τη χρήση των κατάλληλων τεχνικών αξιολόγησης εναπόκειται στον ερευνητή για τη σημασία ή μη των δημιουργούμενων κλάσεων.

1.5 Εξόρυξη δεδομένων και καταλληλότατα για την επίλυση ενός προβλήματος.

Η διεξαγωγή ενός ερευνητικού έργου απαιτεί ποιότητα δεδομένων, πηγές με δεδομένα αξιόπιστα, αξιοποιήσιμα, κατάλληλα για το σκοπό και το αντικείμενο του έργου και προς επίρρωση των συμπερασμάτων που θα εξαχθούν αυτά θα πρέπει να έχουν συλλεχθεί βάσει συμφωνημένων διαδικασιών και πρακτικών. Στο τετράπτυχο δεδομένα-πληροφορία-γνώση-σοφία που παρουσιάστηκε πρώτη φορά ως «Πυραμίδα ή Ιεραρχία της Γνώσης» (Ackoff, 2015), τα δεδομένα αποτελούν το πρωταρχικό επίπεδο που οδηγεί στη γνώση για τον κόσμο, άλλως μαζί με την πληροφορία αποτελούν το φορέα διακίνησης της γνώσης· γι' αυτό το λόγο η σωστή επιλογή των δεδομένων είναι κομβική συνιστώσα για την εξαγωγή επιστημονικών ευρημάτων και συμπερασμάτων. Στο ίδιο πλαίσιο, η διαλειτουργικότητα των δεδομένων, η γνησιότητα, η δομή και η ακεραιότητά τους βαίνει καθοριστική για την αξιοποίηση της αξίας των δεδομένων. Τούτων δοθέντων και δεδομένου ότι η ελεύθερη πρόσβαση στη γνώση υλοποιείται κυρίως μέσα από πολιτικές ανοικτών δεδομένων, δηλαδή δεδομένων στα οποία η πρόσβαση γίνεται με τους ελάχιστους δυνατούς τεχνικούς, νομικούς και οργανωτικούς περιορισμούς, η έρευνα/αναζήτηση των δεδομένων για την εκπόνηση της παρούσας εστίασε και σε βάσεις ανοιχτών δεδομένων.

Η ίδια η σημασία της ανοικτής διάθεσης των δεδομένων εξάλλου αναδεικνύεται σε μεγάλο βαθμό με τη διάδοση νεότερων τεχνολογιών όπως: οι τεχνικές εξόρυξης πληροφορίας μέσα από τη μαζική επεξεργασία μεγάλων συνόλων δεδομένων, οι σύγχρονες τεχνικές οπτικοποίησης της πληροφορίας μέσω χαρτών, τα Ανοικτά Διασυνδεδεμένα Δεδομένα και η χρήση της γνώσης του κοινού «πληθοπορισμός» (Howe 2006). Για τη δημιουργία ή και τη βελτίωση των δεδομένων, οι προγραμματιστικές διεπαφές (APIs) που επιτρέπουν τη δημιουργία πληροφορίας σε πραγματικό χρόνο. Τα χαρακτηριστικά δε, των ανοιχτών δεδομένων, δηλαδή η διαθεσιμότητα σε μορφή πρακτικά αναγνώσιμη, η προσβασιμότητα, η δυνατότητα επαναχρησιμοποίησης, αναδιανομής και ανάμειξής τους με άλλα σύνολα δεδομένων και η καθολική συμμετοχή, αποτελούν βασικό παράγοντα προτίμησής.

Η λήψη μιας απόφασης αναφορικά με το αν θα χρησιμοποιηθεί ή όχι η εξόρυξη δεδομένων ως στρατηγική για την επίλυση ενός προβλήματος δεν είναι μια εύκολη διαδικασία. Χρειάζεται να ληφθούν υπόψη κάποιες παράμετροι αναφορικά με το αν θα αποτελέσει η εξόρυξη δεδομένων στρατηγική επιλογή επίλυσης ενός προβλήματος. Σχετίζεται με το αν το κόστος επεξεργασίας δεδομένων είναι μικρότερο από το πιθανό όφελος που θα προκύψει από την εφαρμογή από ενός προγράμματος εξόρυξης γνώσης, αν υπάρχουν ή όχι δεδομένα με πλούτο πληροφορίας ικανά για ανάλυση, αν υπάρχουν δεδομένα τα οποία να περιέχουν κρυφή γνώση και αν μπορεί να καθοριστεί με σαφήνεια το πρόβλημα όπου με τις τεχνικές εξόρυξης δεδομένων μπορεί ο ερευνητής να οδηγηθεί σε συμπεράσματα που θα τον διευκολύνουν για τη λήψη απόφασης.

Για την απάντηση των ερωτημάτων θα πρέπει να ληφθούν υπόψη οι γενικοί τύποι γνώσης οι οποίοι και θα υποστηρίξουν την επιλογή για εφαρμογή ή όχι τεχνικών εξόρυξης δεδομένων. Η επιφανειακή γνώση (shallow knowledge) η οποία μπορεί εύκολα να αποθηκευτεί και να υποστεί επεξεργασία σε μια βάση δεδομένων. Στην περίπτωση αυτή οι γλώσσες ερωτημάτων όπως η SQL αποτελούν ενδεδειγμένο εργαλείο για την εξαγωγή επιφανειακής γνώσης από δεδομένα. Η πολυδιάστατη γνώση (multidimensional knowledge) όπου τα δεδομένα είναι αποθηκευμένα σε πολυδιάστατη μορφή και είναι απαραίτητη η χρήση εργαλείων On Line Analytical Processing (OLAP). Η κρυφή γνώση (hidden knowledge) η οποία αντιπροσωπεύει μοτίβα ή κανονικότητες δεδομένων και που είναι αδύνατον να βρεθούν με τη χρήση μιας γλώσσας ερωτημάτων και χρειάζεται η χρήση αλγορίθμων εξόρυξης για να βρεθούν αυτά τα μοτίβα και η γνώση σε βάθος (deep knowledge) η οποία είναι αποθηκευμένη σε μια βάση δεδομένων και για τον εντοπισμό της χρειάζονται συγκεκριμένες κατευθύνσεις σχετικά με το τι πρέπει να ψάξει ο ερευνητής.

1.6 Διαδικασία εξόρυξης δεδομένων

Η διαδικασία εξόρυξης δεδομένων περιλαμβάνει βήματα που πρέπει να εφαρμοσθούν προκειμένου να αντληθεί πληροφορία από τα δεδομένα. Εφαρμόζονται τέσσερα στάδια για την εξόρυξη δεδομένων. Στο πρώτο στάδιο εντάσσεται η συγκέντρωση του συνόλου των απαιτούμενων δεδομένων τα οποία θα αναλυθούν. Η ποσότητα των προς ανάλυση δεδομένων ποικίλλει (Agrawal, Imielinski & Swami, 1993). Τα δεδομένα είναι δυνατό να αποτελούν μια μεγάλη ποσότητα εγγραφών και από διαφορετικά αρχεία και βάσεις δεδομένων ή μια ποσότητα εγγραφών μικρότερη σε έκταση. Η προσπέλαση των δεδομένων με σκοπό την εξόρυξή τους μπορεί να

πραγματοποιηθεί είτε από αποθήκες δεδομένων, είτε από σχεσιακές βάσεις δεδομένων, είτε από ένα και μόνο επίπεδο, δηλαδή αρχείο ή φύλλο εργασίας. Το δεύτερο στάδιο περιλαμβάνει την ίδια τη διαδικασία της εξόρυξης των δεδομένων. Πριν την εισαγωγή των δεδομένων ο ερευνητής καλείται να αποφασίσει για το είδος της εκμάθησης των δεδομένων αν θα είναι καθοδηγούμενη ή μη καθώς και ποια θα χρησιμοποιηθούν για τον έλεγχο του μοντέλου και ποια για την κατασκευή του (Fayyad, Haussler & Stolroz). Το είδος των χαρακτηριστικών που θα χρησιμοποιηθούν από τον κατάλογο των διαθέσιμων γνωρισμάτων αλλά και τη ρύθμιση εκείνων των παραμέτρων που θα χρησιμοποιηθούν με σκοπό την κατασκευή του μοντέλου. Το επόμενο στάδιο περιλαμβάνει την ερμηνεία των αποτελεσμάτων και τον καθορισμό αν το αποτέλεσμα που έχει εξαχθεί είναι σημαντικό και χρήσιμο. Τελικό στάδιο είναι η ίδια η εφαρμογή των αποτελεσμάτων στη διαδικασία λήψης απόφασης.

Κεφάλαιο 2

2.1 Μέθοδοι εξόρυξης δεδομένων

Οι μέθοδοι εξόρυξης δεδομένων μπορούν να ταξινομηθούν σε καθοδηγούμενες και μη καθοδηγούμενες, δηλαδή σε καθοδηγούμενη εκμάθηση και μη καθοδηγούμενη εκμάθηση. Εφαρμόζοντας τη στρατηγική της καθοδηγούμενης εκμάθησης δημιουργούνται μοντέλα από τα χαρακτηριστικά εισόδου για την πρόβλεψη των τιμών των χαρακτηριστικών εξόδου της επεξεργασίας (Jiawei, 2001) . Υπάρχουν αλγόριθμοι καθοδηγούμενης εξόρυξης δεδομένων που επιτρέπουν την ύπαρξη ενός χαρακτηριστικού εξόδου και εργαλεία καθοδηγούμενης εκμάθησης που δίνουν τη δυνατότητα στον ερευνητή να καθορίσει περισσότερα του ενός χαρακτηριστικά εξόδου γνωστά ως εξαρτημένες μεταβλητές (dependent variables) δεδομένου του ότι η τιμή τους εξαρτάται από τις τιμές ενός ή και περισσότερων χαρακτηριστικών εισόδου όπου ονομάζονται ανεξάρτητες μεταβλητές (independent variables). Η μη καθοδηγούμενη εκμάθηση δεν περιλαμβάνει χαρακτηριστικά εξόδου με αποτέλεσμα με αποτέλεσμα τα χαρακτηριστικά που χρησιμοποιούνται για την κατασκευή μοντέλων να αφορούν αποκλειστικά ανεξάρτητες μεταβλητές. Οι καθοδηγούμενες στρατηγικές εκμάθησης περιλαμβάνουν την κατηγοριοποίηση, την πρόβλεψη, την εκτίμηση και τους κανόνες συσχέτισης (association rules).

2.1.1 Κατηγοριοποίηση (classification)

Στην κατηγοριοποίηση (classification) δύναται η κατασκευή μοντέλων που μπορούν να αντιστοιχίζουν καινούρια στιγμιότυπα κλάσης η οποία ανήκει σε ένα σύνολο ορισμένων κλάσεων και η εξαρτημένη μεταβλητή είναι πάντα μεταβλητή κατηγορίας. Στην στρατηγική της κατηγοριοποίησης δεν περιλαμβάνεται μοντέλα που ερμηνεύουν μελλοντική συμπεριφορά παρά μόνο του παρόντος, δεδομένου ότι τη δυνατότητα αυτή διατίθεται μόνο στα μοντέλα πρόβλεψης (Phyu,2009). Στην κατηγοριοποίηση ζητούμενο αποτελεί η κατασκευή μοντέλου με σκοπό την ταξινόμηση νέων και άγνωστων στιγμιότυπων του προβλήματος δοθέντων ο αριθμός των κλάσεων ενός προβλήματος και των ιδιοτήτων- γνωρισμάτων όπου κάθε στιγμιότυπο του προβλήματος διαθέτει και ενός συνόλου εκπαιδευτικών στιγμιότυπων (training set) για τα οποία είμαστε σε θέση εξαρχής να γνωρίζουμε στην κλάση στην οποία ανήκουν. Για την επιτυχή έκβαση της διαδικασίας απαιτείται οι κλάσεις να είναι προκαθορισμένες, στη διάρκεια της

κατηγοριοποίησης να μη μεταβάλλονται και στην ποιότητα των στιγμιότυπων εκπαίδευσης, στο κατά πόσο τα στιγμιότυπα αυτά είναι αντιπροσωπευτικά.

Τέλος χρειάζεται ένα ακόμα σύνολο στιγμιότυπων για τον έλεγχο της απόδοσης της διαδικασίας κατηγοριοποίησης, δηλαδή τη μέτρηση της ακρίβειας με την οποία η διαδικασία της κατηγοριοποίησης ταξινομεί νέα, άγνωστα στιγμιότυπα του προβλήματος. Ο δείκτης απόδοσης ισούται με τον αριθμό των στιγμιότυπων συνόλου ελέγχου στα οποία προβλέφθηκε με ακρίβεια η κλάση ως προς τον αριθμό ων στιγμιότυπων του συνόλου ελέγχου. Η απόδοση ισούται με τον αριθμό των στιγμιότυπων του συνόλου ελέγχου για τα οποία ο ταξινομητής προέβλεψε με ακρίβεια την κλάση προς το συνολικό αριθμό των στιγμιότυπων του συνόλου ελέγχου (Roiger & Geatz, 2007).

2.1.2 Εκτίμηση (estimation)

Στην εκτίμηση (estimation) τα χαρακτηριστικά εξόδου είναι αριθμητικά και όχι χαρακτηριστικά κατηγοριών και σκοπός είναι ο καθορισμός της τιμής ενός άγνωστου χαρακτηριστικού εξόδου.

2.1.3 Πρόβλεψη (prediction)

Στην πρόβλεψη (prediction) κύριος σκοπός είναι ο καθορισμός μελλοντικών αποτελεσμάτων (Shapiro & Smyth, 1996). Οι τιμές που μπορούν να λάβουν τα χαρακτηριστικά εξόδου της πρόβλεψης μπορούν να είναι είτε αριθμητικά, είτε να αναφέρονται σε κατηγορίες. Το μοντέλο πρόβλεψης δομείται μέσω του συνόλου εκπαίδευσης. Εκτός του συνόλου εκπαίδευσης το σύνολο ελέγχου αποτελείται από παραδείγματα όπου η τιμή πρόβλεψης του γνωρίσματος είναι εξαρχής γνωστή και ισούται με το ένα τρίτο των παραδειγμάτων του συνόλου εκπαίδευσης που χρησιμοποιείται για την αξιολόγηση του μοντέλου πρόβλεψης. Το σύνολο ελέγχου αποσκοπεί στον έλεγχο της απόδοσης του μοντέλου πρόβλεψης δηλαδή στην ακρίβεια με την οποία το μοντέλο προβλέπει την τιμή ενός άγνωστου έως εκείνη τη στιγμή γνωρίσματος στα νέα στιγμιότυπα του προβλήματος.

2.1.4 Κανόνες συσχέτισης (association rules)

Οι κανόνες συσχέτισης ή συσχετισμού είναι δηλώσεις που βοηθούν στην εμφάνιση της πιθανότητας σχέσεων μεταξύ στοιχείων δεδομένων, μέσα σε μεγάλα σύνολα δεδομένων σε διάφορους τύπους βάσεων δεδομένων. Για την ανακάλυψη ενδιαφερόντων συσχετισμών μεταξύ των χαρακτηριστικών που υπάρχουν στις βάσεις δεδομένων χρησιμοποιούνται τεχνικές εξόρυξης δεδομένων με κανόνες συσχετισμού. Οι

κανόνες συσχετισμού είναι απαραίτητοι για την εύρεση συσχετίσεων μεταξύ των διαφορετικών αντικειμένων. Ο κανόνας συσχέτισης μεταξύ δύο αντικειμένων Α και Β δηλώνει πως συνεπάγεται η παρουσία του Α και Β στο ίδιο στιγμιότυπο του προβλήματος (Piatetsky-Shapiro 1991).

Σε αντίθεση με τους παραδοσιακούς κανόνες παραγωγής μπορούν να έχουν περισσότερα του ενός χαρακτηριστικά εξόδου και το χαρακτηριστικό εξόδου ενός κανόνα μπορεί να είναι και χαρακτηριστικό εισόδου κάποιου άλλου κανόνα. Η εξαγωγή κανόνων συσχέτισης πραγματοποιείται με τη χρήση αλγορίθμων, οι οποίοι αποδεικνύονται αρκετά αποδοτικοί (Tan, Steinbach, Kumar, 2005). Μετά την εύρεση και ανάλυση των κανόνων πρέπει να διερευνηθεί αν και πόσο αυτοί οι κανόνες είναι έγκυροι και σημαντικοί. Υπάρχουν δύο συντελεστές για την εκπλήρωση αυτού του σκοπού η υποστήριξη (support) και η εμπιστοσύνη (confidence). Η υποστήριξη (support) είναι ίση με το ποσοστό του συνόλου των στιγμιότυπων, Ν που ικανοποιεί το συνδυασμό Α και Β.

support = [AB]/N

Η εμπιστοσύνη (confidence) ισούται με το ποσοστό του συνόλου των στιγμιότυπων στα οποία όταν ισχύει το Α ισχύει και το Β.

confidence = [AB]/[A]

Οι κανόνες συσχετισμού αποτελούν δημοφιλή τεχνική γιατί η εφαρμογή τους είναι εφικτή η διερεύνηση των πιθανών συνδυασμών των σημαντικών ομαδοποιήσεων αντικειμένων.

2.1.5 Συσταδοποίηση (clustering)

Στη στρατηγική της μη καθοδηγούμενης συσταδοποίησης δεδομένου του ότι δεν υπάρχει εξαρτημένη μεταβλητή οποία καθοδηγεί τη διαδικασία της εκμάθησης, το ίδιο το πρόγραμμα εκμάθησης κατασκευάζει μια δομή γνώσης με τη χρήση μέτρου ποιότητας των συστάδων με σκοπό τη συσταδοποίηση ή τμηματοποίηση στιγμιότυπων σε δύο ή και περισσότερες κλάσεις (Βαζιργιάννης & Χαλκίδη, 2003). Οι χρήσεις της μη καθοδηγούμενης συσταδοποίησης περιλαμβάνουν το κατά πόσο είναι εφικτό να μπορούν να εντοπιστούν στα δεδομένα με τη μορφή εννοιών σχέσεις, τον καθορισμό του βέλτιστου συνόλου χαρακτηριστικών εισόδου της καθοδηγούμενης εκμάθησης, η αξιολόγηση της πιθανής απόδοσης του μοντέλου εκμάθησης και ο εντοπισμός ασυνήθιστων στιγμιότυπων (outliers). Υπάρχει πιθανότητα τα ασυνήθιστα στιγμιότυπα

να είναι σημαντικά για το αποτέλεσμα μιας έρευνας με αποτέλεσμα όταν αυτό είναι δυνατό να πρέπει να εντοπιστούν (Roiger & Geatz, 2008).

Ζητούμενο στην μέθοδο της συσταδοποίησης είναι ο διαχωρισμός των στιγμιότυπων σε συστάδες (clusters), όπου τα στιγμιότυπα με συναφή χαρακτηριστικά να ανήκουν στο ίδιο τμήμα με στόχο την εύρεση των ιδιοτήτων του κάθε τμήματος (Achtert, Böhm, Kriegel, Kröger, Müller-Gorman, Zimek, 2007). Για την εξαγωγή κανόνων σχετικά με τη συμπεριφορά των αντικειμένων κάποιου τμήματος δεν είναι απαραίτητο να εξεταστούν οι ανεξάρτητες εγγραφές του συνόλου των δεδομένων παρά μόνο η εξέταση των χαρακτηριστικών του συγκεκριμένου τμήματος.

Με αυτόν τον τρόπο τα στοιχεία που ανήκουν στο ίδιο τμήμα θα συμπεριφέρονται κατά τρόπο ενιαίο καθώς θα έχουν παρόμοια χαρακτηριστικά. Συνεπώς, κανόνας που είναι έγκυρος για κάποιο από τα στοιχεία ενός τμήματος δύναται να είναι έγκυρος για όλα τα στοιχεία του τμήματος αυτού (Kailing, Kriegel, Kröger, 2004). Διαφορά της ταξινόμησης από τη συσταδοποίηση αποτελεί το γεγονός πως στην ταξινόμηση οι κλάσεις έχουν προκαθοριστεί, ενώ στην τμηματοποίηση οι κλάσεις δεν είναι προκαθορισμένες και τα στιγμιότυπα διαχωρίζονται σε τμήματα βάσει των ομοιοτήτων που παρουσιάζουν μεταξύ τους ως προς τα κύρια γνωρίσματα της τμηματοποίησης. Συνεπώς, κατά τη διαδικασία εφαρμογής τμηματοποίησης σε ένα σύνολο δεδομένων, δεν υπάρχει συγκεκριμένο σύνολο παραδειγμάτων το οποίο θα μπορούσε να υποδείξει ποιες είναι οι επιθυμητές σχέσεις που πρέπει να ισχύουν στα δεδομένα (Roiger & Geatz, 2007).

2.2 Τεχνικές εξόρυξης δεδομένων

Για την εφαρμογή των μεθόδων εξόρυξης δεδομένων απαιτούνται συγκεκριμένες τεχνικές εξόρυξης. Κάθε τεχνική εξόρυξης δεδομένων ορίζεται από αλγόριθμο και συγκεκριμένη δομής γνώσεων. Στις τεχνικές καθοδηγούμενης εξόρυξης πληροφορίας περιλαμβάνονται μέθοδοι όπως τα δέντρα αποφάσεων, οι μέθοδοι συσταδοποίησης, τα νευρωνικά δίκτυα, οι κανόνες συσχετισμού και οι στατιστικές μέθοδοι.

2.2.1 Τα δένδρα αποφάσεων

Τα δέντρα αποφάσεων κατασκευάζονται χρησιμοποιώντας εκείνα τα χαρακτηριστικά τα οποία είναι σε θέση να διαφοροποιούν τις έννοιες που πρόκειται να διδαχτούν (Collins 2020). Για την κατασκευή ενός δέντρου αποφάσεων αρχικά επιλέγεται ένα υποσύνολο στιγμιότυπων από ένα σύνολο δεδομένων εκπαίδευσης. Εν συνεχεία αυτό το υποσύνολο χρησιμοποιείται από τον αλγόριθμο για την κατασκευή ενός

δέντρου αποφάσεων (Rossi and Tsoukias, 2009). Τα υπόλοιπα στιγμιότυπα του συνόλου δεδομένων εκπαίδευσης χρησιμοποιούνται για τον έλεγχο της ακρίβειας του δέντρου που κατασκευάστηκε. Σε περίπτωση που το δέντρο αποφάσεων δεν κατηγοριοποιεί τα στιγμιότυπα εσφαλμένα η διαδικασία τερματίζεται. Αν κάποιο στιγμιότυπο κατηγοριοποιείται εσφαλμένα τότε προστίθεται στο επιλεγμένο σύνολο του στιγμιότυπων εκπαίδευσης με σκοπό την κατασκευή καινούριου δένδρου. Η διαδικασία συνεχίζεται είτε μέχρι να κατασκευαστεί ένα δέντρο το οποίο κατηγοριοποιεί σωστά όλα τα επιλεγμένα στιγμιότυπα, είτε όταν έχει χρησιμοποιηθεί το σύνολο των δεδομένων εκπαίδευσης για την κατασκευή του δέντρου. Τα δέντρα απόφασης αποτελούν τεχνική που χρησιμοποιείται ευρέως στην ταξινόμηση και πρόβλεψη (Roiger & Geatz 2007). Ένα δέντρο απόφασης αντιπροσωπεύει μια σειρά από Αν/Τότε (If/then) κανόνων, εκκινώντας από τη ρίζα του δέντρου και καταλήγοντας στα φύλλα του. Οι εσωτερικοί κόμβοι ενός δέντρου απόφασης ενσωματώνουν τα γνωρίσματα του προβλήματος, οι ακμές περιέχουν τις δυνατές τιμές των γνωρισμάτων και τα φύλλα περιέχουν τις πιθανές κλάσεις του προβλήματος. Απαραίτητο για την κατασκευή ενός ΔΑ είναι ένα σύνολο από στιγμιότυπα εκπαίδευσης, όπου κάθε στιγμιότυπο περιγράφεται από γνωρίσματα και την κλάση του προβλήματος στην οποία ανήκει (Rossi and Tsoukias, 2009).

Οι αλγόριθμοι κατασκευής ενός δένδρου απόφασης ακολουθούν μια συγκεκριμένη διαδικασία. Παρατίθεται μία απλοποιημένη έκδοση αλγορίθμου, η οποία χρησιμοποιεί ολόκληρο το σύνολο του στιγμιότυπων εκπαίδευσης για την κατασκευή ενός δέντρου αποφάσεων τα βήματα του αλγορίθμου έχουν ως εξής.

Η διαδικασία που ακολουθείται από τους αλγόριθμους κατασκευής ενός δέντρου απόφασης είναι η εξής. Εκκινώντας από τη ρίζα του δέντρου ο αλγόριθμος διαχωρίζει το σύνολο των στιγμιότυπων εκπαίδευσης σε υποσύνολά βάσει της βέλτιστης ιδιότητας (best attribute) του κόμβου (Rokach & Maimon, 2007). Με αυτόν τον τρόπο προκύπτει πλήθος υποσυνόλων όπου το καθένα ενσωματώνει λιγότερα παραδείγματα βάσει του αρχικού συνόλου. Για τα επιμέρους υποσύνολα εφαρμόζεται επαναληπτικά η διαδικασία, με χρήση των εναπομεινάντων γνωρισμάτων, οπότε η διάσπαση των στιγμιότυπων συνεχίζεται και σταματά όταν τα στιγμιότυπα του υποσυνόλου είτε ανήκουν στην ιδία κλάση, είτε έχουν εξαντληθεί όλα τα γνωρίσματα.

Εκτός του συνόλου στιγμιότυπων εκπαίδευσης, το σύνολο των στιγμιότυπων ελέγχου, εξετάζει την απόδοση του δένδρου. Συγκεκριμένα, ελέγχεται η ακρίβεια

κατασκευής του δένδρου σχετικά με την ταξινόμηση (Roiger & Geatz, 2003). Ο καθορισμός της ακρίβειας του δένδρου αφορά το πλήθος τον λανθασμένων απαντήσεων.

Για την ταξινόμηση ενός νέου στιγμιότυπου του προβλήματος ο ερευνητής πρέπει να διατρέξει το δένδρο από τη ρίζα ακολουθώντας τα μονοπάτια, ενώ κάθε φορά η επιλογή του κατάλληλου μονοπατιού καθορίζεται με εφαρμογή της συνθήκης ελέγχου του εκάστοτε κόμβου στις τιμές των γνωρισμάτων του στιγμιότυπου προς ταξινόμηση. Όταν καταλήξει η διαδικασία σε κάποιο φύλλο, η κλάση αυτού αποτελεί τη ζητούμενη κλάση του στιγμιότυπου. Τα Δένδρα αποφάσεων αποτελούν τα δημοφιλέστερα μοντέλα κατηγοριοποίησης και αναπαριστούν ένα μοντέλο πρόβλεψης το οποίο δομείται από μια σειρά αποφάσεων του τύπου ναι/όχι και μεγαλύτερο/μικρότερο. Αποτελείται από ενδιάμεσους κόμβους και φύλλα. Τα φύλλα είναι οι κόμβοι στο τελευταίο επίπεδο και αυτή είναι η καταληκτική διαδικασία.

Η κάθε εγγραφή διαθέτει ένα σύνολο από γνωρίσματα/χαρακτηριστικά (attributes). Ένα από τα γνωρίσματα είναι η κλάση/κατηγορία (class). Για την υλοποίηση ενός αλγορίθμου δένδρου απόφασης πρώτα επιλέγεται το κατάλληλο χαρακτηριστικό με σκοπό να αποτελέσει την κορυφή του δένδρου, και στη συνέχεια επαναλαμβάνεται η διαδικασία για κάθε σημείο του κόμβου που προκύπτει. Όταν όλες οι εγγραφές έχουν καταλήξει στη ίδια παράμετρο, έχει γίνει χρήση όλων των χαρακτηριστικών και δεν υπάρχουν υποσύνολα που να περιέχουν περισσότερες από μια εγγραφές η διαδικασία τερματίζεται.

2.2.2 Αλγόριθμοι δένδρων απόφασης

Για την κατασκευή ενός δένδρου απόφασης με την εφαρμογή του αλγορίθμου ID3[3] πρώτα υπολογίζουμε το πληροφοριακό κέρδος για κάθε μεταβλητή, και ως ρίζα του δένδρου ορίζεται η μεταβλητή με το μεγαλύτερο πληροφοριακό κέρδος (Yeturu, 2020). Θα δημιουργηθούν τόσα κλαδιά όσες είναι και οι διακριτές τιμές του επιλεγόμενου κόμβου και στη συνέχεια θα πραγματοποιηθεί διαχωρισμός του συνόλου των δεδομένων σε τόσα υποσύνολα όσα και οι διακριτές τιμές της επιλεχθείσας μεταβλητής. Επιλέγοντας μια συγκεκριμένη τιμή από το υποσύνολο και εφόσον ο κόμβος που επιλεγεί αντιστοιχεί σε μια τιμή τότε επιλέγεται ο επόμενος κόμβος (Quinlan, 1993). Σε περίπτωση που αντιστοιχεί σε περισσότερες της μια τιμές υπολογίζεται το πληροφοριακό κέρδος των

[3] Ο αλγόριθμος ID3 (Iterative Dichotomiser) είναι ένας αλγόριθμος ταξινόμησης που ακολουθεί προσέγγιση κατασκευής ενός δέντρου αποφάσεων επιλέγοντας ένα καλύτερο χαρακτηριστικό που αποδίδει μέγιστο κέρδος πληροφοριών ή ελάχιστη εντροπία.

άλλων μεταβλητών σχετικά με το συγκεκριμένο υποσύνολο και γίνεται επιλογή του κόμβου που διαθέτει το μεγαλύτερο πληροφοριακό κέρδος. Η διαδικασία επαναλαμβάνεται μέχρις ότου να μην υπάρχει δυνατότητα να δημιουργηθούν νέα φύλλα.

Ο αλγόριθμος C4.5 αποτελεί επέκταση του αλγορίθμου ID3. Ο αλγόριθμος C4.5 είναι ένας αρκετά συχνά εφαρμοζόμενος αλγόριθμος στην Εξόρυξη Δεδομένων και. λειτουργεί ως ταξινομητής δένδρων αποφάσεων (Shavlik, Mooney, Towell, 1990). Ο C4.5 είναι ένας αλγόριθμος που χρησιμοποιείται για τη δημιουργία ενός δέντρου αποφάσεων και είναι πολύ χρήσιμος για τη δημιουργία μιας απόφασης, η οποία βασίζεται σε ένα δείγμα δεδομένων. (Witten, Frank, Hall, 2011). Όταν δημιουργείται ένα δέντρο απόφασης με τη βοήθεια του αλγόριθμου C4.5, τότε μπορεί να χρησιμοποιηθεί για την ταξινόμηση του συνόλου δεδομένων, και αυτός είναι ο κύριος λόγος για τον οποίο το C4.5 είναι επίσης γνωστός ως στατιστικός ταξινομητής. Ο αλγόριθμος C4.5 σε αντίθεση με τον ID3 δεν υστερεί όταν λείπουν πολλά δεδομένα, και δεν έχει εφαρμογή μόνο σε διακριτά δεδομένα (Collins 2020).

2.2.3 Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα αποτελούν ένα μαθηματικό μοντέλο που προσπαθεί να μιμηθεί τον ανθρώπινο εγκέφαλο. Συχνά η γνώση αναπαρίσταται ως ένα σύνολο διασυνδεδεμένων επεξεργαστούν που είναι διαρθρωμένη σε επίπεδα. Αυτοί οι κόμβοι επεξεργαστές αναφέρονται και ως νεύρο-κόμβοι κάτι που παραπέμπει ευθέως τους νευρώνες του εγκεφάλου. Κάθε κόμβος έχει μία σταθμισμένη σύνδεση με διάφορους άλλους κόμβους των παρακείμενων επιπέδων (Chester, 1993). Οι μεμονωμένοι κόμβοι παίρνουν την είσοδο που λαμβάνεται από τους άλλους συνδεδεμένους κόμβους και χρησιμοποιούν τους συντελεστές στάθμισης σε συνδυασμό με μια απλή συνάρτηση για να υπολογίσουν τις τιμές εξόδου. Η εκμάθηση νευρωνικών δικτύων μπορεί να είναι καθοδηγούμενη η μη καθοδηγούμενη (Roiger & Geatz 2003). Συντελείται με την τροποποίηση των συντελεστών στάθμισης των συνδέσεων του δικτύου καθώς ένα σύνολο στιγμιότυπων εισόδου διαβιβάζεται κατ’ επανάληψή μέσω του δικτύου. Μετά την εκπαίδευση τα άγνωστα στιγμιότυπα που διαβιβάζονται μέσω του δικτύου κατηγοριοποιούνται σύμφωνα με τις τιμές που φαίνονται στο επίπεδο εξόδου.

Η είσοδος στους κόμβους των νευρωνικών δικτύων πρέπει να είναι αριθμητική και να ανήκει στο κλειστό διάστημα μηδέν έως ένα. Επίσης απαιτείται μια μέθοδος μετατροπής για τα αριθμητικά δεδομένα που βρίσκονται έξω από το εύρος του διαστήματος. Υπάρχουν πολλές επιλογές για την μετατροπή δεδομένων κατηγοριών. Μια

απλή και άμεση τεχνική είναι η διαίρεση του εύρους του διαστήματος σε ίσα τμήματα (Civco, 1991).

Το νευρωνικό δίκτυο είναι ένα σύστημα παράλληλης επεξεργασίας πολλών διασυνδεδεμένων κόμβων επεξεργαστών. (Collins 2020). Η είσοδος στους κόμβους του δικτύου περιορίζεται σε αριθμητικές τιμές που ανήκουν στο κλειστό διάστημα μηδέν έως ένα και για αυτό τα δεδομένα κατηγοριών θα πρέπει να μετασχηματιστούν πριν την εκπαίδευση του δικτύου. Η ανάπτυξη ενός νευρωνικού δικτύου περιλαμβάνει πρώτα την εκπαίδευση του δικτύου έτσι ώστε να εκτελεί τους επιθυμητούς υπολογισμούς και στη συνέχεια την εφαρμογή του εκπαιδευμένου δικτύου για την επίλυση νέων προβλημάτων (Ripley, Brian 1996). Κατά τη διάρκεια της φάσης εκμάθησης τα δεδομένα εκπαίδευσης χρησιμοποιούνται για την τροποποίηση των συντελεστών στάθμισης των συνδέσεων μεταξύ των κόμβων έτσι ώστε να προκύπτει το καλύτερο δυνατό αποτέλεσμα για τον κόμβο εξόδου (Schmidhuber, 2015). Η αρχιτεκτονική νευρωνικών δικτύων με τροφοδότηση προς τα εμπρός χρησιμοποιείται συνήθως για καθοδηγούμενη εκμάθηση. Τα νευρωνικά δίκτυα με τροφοδοτηση προς τα εμπρός περιέχουν ένα σύνολο κόμβων διατεταγμένων σε επίπεδα και σταθμισμένες συνδέσεις μεταξύ κόμβων παρακείμενων επιπέδων. Τα νευρωνικά δίκτυα με τροφοδότηση προς τα εμπρός συχνά εκπαιδεύονται με ένα σχήμα εκμάθησης με οπισθοδιάδοση (Roiger & Geatz 2003). Η λειτουργία της εκμάθησης με οπισθοδιάδοση, περιλαμβάνει τροποποιήσεις στις τιμές των συντελεστών στάθμισης ξεκινώντας από το επίπεδο έξοδο και κινούμενη προς τα πίσω μέσω των κρυφών επιπέδων του δικτύου. Η γενετική εκμάθηση μπορεί επίσης να εφαρμοστεί στην εκπαίδευση δικτύων με τροφοδότηση προς τα εμπρός. Η αρχιτεκτονική αυτοοργανούμενων νευρωνικών δικτύων είναι ένα δημοφιλές μοντέλο για μη καθοδηγούμενη συσταδοποίηση. Η εκμάθηση ενός αυτοοργανούμενου νευρωνικού δικτύου γίνεται με την ύπαρξη πολλών κόμβων εξόδου που διεκδικούν τα στιγμιότυπα εκπαίδευσης. Επομένως οι συντελεστές στάθμισης εισόδου του κόμβου που επικρατεί τροποποιούνται έτσι ώστε να ταιριάζουν περισσότερο με το τρέχον στιγμιότυπο εκπαίδευσης (Roiger & Geatz2003).

Όταν ολοκληρωθεί η καθοδηγούμενη εκμάθηση αποθηκεύονται οι κόμβοι εξόδου που έχουν επιτύχει τα περισσότερα στιγμιότυπα. Στη συνέχεια ακολουθεί ο έλεγχος στο δίκτυο και οι συστάδες που δημιουργούνται από τα δεδομένα του δείγματος ελέγχου αναλύονται για να βοηθήσουν στον προσδιορισμό της σημασίας των στοιχείων που ανακαλύφθηκαν. Ένα βασικό ζήτημα που αφορά τα νευρωνικά δίκτυα είναι η αδυναμία

να εξηγήσουν τι προέκυψε από την εκμάθηση. Παρόλα αυτά τα νευρωνικά δίκτυα εφαρμόζονται με επιτυχία για την επίλυση προβλημάτων στο χώρο των επιχειρήσεων και των επιστημών.

2.2.4 Bayesian ταξινομητές

Θεμελιώδης στατιστική προσέγγιση για το πρόβλημα της ταξινόμησης προτύπων αποτελεί η θεωρία απόφασης του Bayes. Βασισμένη στον ποσοτικό προσδιορισμό των ανταλλαγών/αποφάσεων ταξινόμησης κάνουν χρήση της πιθανότητας και του κόστους που ακολουθεί τις συγκεκριμένες αποφάσεις. Σύμφωνα με τη θεωρία απόφασης το πρόβλημα απόφασης βασίζεται σε πιθανολογικούς όρους με γνωστές τις σχετικές τιμές πιθανοτήτων (Schervish, 1995). Οι ταξινομητές Bayes αναφέρονται σε στατιστικούς ταξινομητές που έχουν τη δυνατότητα να υπολογίσουν την πιθανότητα κάποιου στιγμιότυπου ενός προβλήματος να ανήκει σε μια από τις προκαθορισμένες κλάσεις του προβλήματος. Σύμφωνα με το θεώρημα του Bayes: Αν P ο διαμοιρασμός της πιθανότητας, D η συλλογή στιγμιότυπων για τα οποία η κλάση τους είναι γνωστή, h η υπόθεση ότι τα δεδομένα D ανήκουν σε συγκεκριμένη κλάση C και είναι γνωστή η P(h) εκ των προτέρων (apriori) πιθανότητα η υπόθεση h να είναι ορθή με P(D) η πιθανότητα να παρατηρηθούν τα δεδομένα D και P(D|h) η εκ των υστέρων (posteriori) πιθανότητα να παρατηρηθούν τα δεδομένα D με την προϋπόθεση ότι η D είναι σωστή, τότε ο υπολογισμός της πιθανότητας P(D|h) δηλαδή η πιθανότητα η υπόθεση h να είναι σωστή δίνεται από τη σχέση: P(h|D)=P(D|h)*P(h)/P(D) (Lee &Peter, 2012) .

Οι ταξινομητές Bayes διαθέτουν χαμηλό ρυθμό λάθους συγκρινόμενοι με τους υπόλοιπους ταξινομητές και παρουσιάζουν μεγάλη ακρίβεια και ταχύτητα ιδίως σε μεγάλες βάσεις δεδομένων (Jaynes, 2003). Ωστόσο, εξαιτίας των σφαλμάτων που είναι πιθανό να γίνουν στις υποθέσεις όπως η υπόθεση ανεξαρτησίας προς την κατανομή των κλάσεων, μπορεί να αυξήσει την πιθανότητα χαμηλού ρυθμού λάθους των Bayesian ταξινομητών (McGrayne, 2011). Η χρησιμότητα των συγκεκριμένων ταξινομητών είναι μεγάλη μιας και εξαιτίας του θεωρήματος του Bayes προσφέρουν θεωρητική αιτιολόγηση για άλλους ταξινομητές.

2.2.4.1Ταξινομητής (Naïve) Bayes

Ο Ταξινομητής (Naïve) Bayes αποτελεί απλούστερη εκδοχή του βασικού Bayes αλγορίθμου, η οποία χρησιμοποιείται για ταξινόμηση των στιγμιότυπων ενός προβλήματος σε προκαθορισμένες κλάσεις του προβλήματος. Σύμφωνα με τον ταξινομητή (Naïve) Bayes κάθε στιγμιότυπο X του προβλήματος αποτελείται από

σύνολο γνωρισμάτων $x_1,x_2,...x_n$, δηλαδή $X=<x_1,x_2,...x_n>$. Αν το πρόβλημα διαθέτει m κλάσεις, c1,c2,...cm και δοθεί ένα άγνωστο στιγμιότυπο X του προβλήματος προς επίλυση που δεν γνωρίζουμε σε ποια κλάση ανήκει, ο ταξινομητής μπορεί να προβλέψει αν το X. ανήκει στην κλάση με τη μεγαλύτερη posteriori πιθανότητα αναθέτοντας ένα άγνωστο στιγμιότυπο X του προβλήματος στην κλάση Ci μόνο εφόσον $P(Ci|X)> P(Cj|X)$ για $1\leq j\leq m$, ώστε να μεγιστοποιείται η πιθανότητα P(Ci|X), η οποία δίνεται από τη σχέση $P(Ci|X)=P(X|Ci)*P(Ci)/P(X)$. Στον τύπο το P(X) είναι σταθερό για όλα τα στιγμιότυπα με αποτέλεσμα αυτό που χρειάζεται να μεγιστοποιηθεί να είναι το μέρος $P(X|C_i)*P(C_i)$ (McGrayne, 2011).

2.2.5 Τμηματοποίηση διαμερισμού (partitional clustering)

Η τμηματοποίηση διαμερισμού αποτελεί μορφή τμηματοποίησης βασισμένη στην άμεση αποσύνθεση όλων των δεδομένων σε ένα σύνολο μη σχετιζόμενων clusters. Το εφαρμοζόμενο κριτήριο για την αποσύνθεση αυτή αποτελεί η ελαχιστοποίηση των μέτρων ανομοιότητας στα δείγματα που βρίσκονται μέσα σε κάθε ένα από τα τμήματα και η μεγιστοποίηση της ανομοιογένειας των διαφορετικών τμημάτων (Emre & Celebi, 2014). Η πιο δημοφιλής μέθοδος τμηματοποίησης διαμερισμού αποτελεί η μέθοδος K μέσου (K-means) που στόχο έχει την ελαχιστοποίηση της μέσης τετραγωνικής απόστασης των δεδομένων από τα κέντρα των τμημάτων. Σύμφωνα με τη συγκεκριμένη μέθοδο: $E_k=\Sigma_k||x_k-m_{c(xk)}||^2$ όπου: $c(x_k)$ είναι ο δείκτης του πλησιέστερου στο x_k κέντρου.

Ο αριθμός των τμημάτων που χρησιμοποιεί ο αλγόριθμος K-Means είναι σταθερός και έχει δοθεί εξαρχής. Η μέθοδος που ακολουθεί αρχικά θεωρεί ένα σύνολο K από σημεία ως κέντρα των K τμημάτων όπου κάθε κέντρο αντιπροσωπεύει μια συστάδα με χρήση της Ευκλείδειας απόστασης για την αντιστοίχιση των υπόλοιπων στιγμιότυπων στα πλησιέστερα κέντρα των συστάδων δηλαδή κάθε ένα εκ των σημείων αντιστοιχεί στο τμήμα όπου το κέντρο βρίσκεται πλησιέστερα και υπολογίζονται τα νέα κέντρα των τμημάτων κάνοντας χρήση των μέσων όρων των σημείων τους. Στη συνέχεια αντιστοιχούνται τα σημεία στο τμήμα του οποίου το κέντρο βρίσκεται εγγύτερα επαναλαμβάνοντας τα βήματα μέχρι τα όρια των τμημάτων να σταματήσουν να μεταβάλλονται, δηλαδή οι καινούριες μέσες τιμές να είναι ίδιες με τις μέσες τιμές των προγενέστερων επαναλήψεων ή εφόσον η συνάρτηση σταματήσει να μεταβάλλεται σημαντικά.

2.2.6 Στατιστική Παλινδρόμηση

Η στατιστική παλινδρόμηση είναι μια τεχνική καθοδηγούμενης εκμάθησης η οποία γενικεύει ένα σύνολο αριθμητικών δεδομένων δημιουργώντας μία μαθηματική εξίσωση που συσχετίζει ένα ή περισσότερα χαρακτηριστικά εισόδου με ένα χαρακτηριστικό εξόδου (Κιόχος, 1993). Με την ανάλυση παλινδρόμησης (regression analysis) μελετάται η σχέση μεταξύ δύο ή και περισσότερων μεταβλητών με στόχο την πρόβλεψη των τιμών της μιας μεταβλητής δια μέσου των τιμών της άλλης. Στα προβλήματα παλινδρόμησης διακρίνονται δύο είδη μεταβλητών αυτά των ανεξάρτητων ή επεξηγηματικών ή ελεγχόμενων μεταβλητών (independent explanatory, predictor, variables) και των εξαρτημένων μεταβλητών ή μεταβλητών απόκρισης. (dependent, response variables). Στις πειραματικές έρευνες ανεξάρτητη μεταβλητή Χ καλείται η μεταβλητή η οποία μπορεί να ελεγχθεί δηλαδή που υπάρχει η δυνατότητα καθορισμού των τιμών της ενώ εξαρτημένη μεταβλητή Υ καλείται η μεταβλητή η οποία υφίσταται το αποτέλεσμα των μεταβολών των ανεξάρτητων μεταβλητών. Στις μη πειραματικές έρευνες (δειγματοληψία) η διάκριση δεν είναι πάντα σαφής εξαιτίας του γεγονότος ότι καμία μεταβλητή δεν είναι ελεγχόμενη αλλά όλες είναι τυχαίες.

2.2.6.1 Απλή Γραμμική Παλινδρόμηση

Ένα μοντέλο γραμμικής παλινδρόμησης χαρακτηρίζεται από ένα χαρακτηριστικό εξόδου του οποίου η τιμή καθορίζεται από το γραμμικό άθροισμα σταθμισμένων τιμών χαρακτηριστικών εισόδου. Αν και η παλινδρόμηση μπορεί να είναι και μη γραμμική χρησιμοποιείται περισσότερο για τη δημιουργία γραμμικών μοντέλων. Η γραμμική παλινδρόμηση είναι κατάλληλη όταν τα δεδομένα μπορούν να μου την υλοποιηθούν με μια γραμμική συνάρτηση. Απλή γραμμική παλινδρόμηση ονομάζεται η παλινδρόμηση κατά την οποία χρησιμοποιούνται οι τιμές μιας μόνο μεταβλητής για την πρόβλεψη της μεταβλητής κριτηρίου. Στόχος είναι η περιγραφή της σχέσεως των Χ και Υ με ένα μοντέλο της μορφής (Montgomery & Peck, 1991).

$$\mathbf{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

Όπου Y_i η τιμή της εξαρτημένης μεταβλητής και X_i η τιμή της ανεξάρτητης μεταβλητής. Η τιμή β_0 δίνει το σημείο στο οποίο η ευθεία παλινδρόμησης τέμνει τον κατακόρυφο άξονα, δηλαδή το σημείο $(0, \beta_0)$ το οποίο προκύπτει αν στην εξίσωση η τιμή της μεταβλητής Χ λάβει την τιμή μηδέν. Ο συντελεστής της μεταβλητής Χ, δηλαδή ο αριθμός β_1, καλείται κλίση της ευθείας παλινδρόμησης και ε_i είναι το σφάλμα (error) ή το κατάλοιπο (residual). Το σφάλμα ε_i παριστάνει τη διαφορά μεταξύ της πραγματικής

τιμής της Y και της τιμής της πρόβλεψης που προκύπτει από την εξίσωση. (Kutner et al., 2005).

2.2.6.2 Πολλαπλή Γραμμική Παλινδρόμηση

Το μοντέλο πολλαπλής παλινδρόμησης, συνδέει την εξαρτημένη μεταβλητή Y με p ανεξάρτητες μεταβλητές X. (Kutner et al., 2005; Cohen, Cohen, West & Aiken, 2003; Montgomery & Peck, 1991):

$$\mathbf{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i.}$$

Σε αυτό το μοντέλο, ο δείκτης i δηλώνει τη μονάδα, από την οποία προέρχονται οι παρατηρήσεις επί της μεταβλητής Y, καθώς και από την οποία ελήφθησαν οι p ανεξάρτητες μεταβλητές. Ο δεύτερος δείκτης δηλώνει την ίδια την ανεξάρτητη μεταβλητή. Το μέγεθος του δείγματος n, και ο δείκτης p δηλώνει τον αριθμό των ανεξάρτητων μεταβλητών. Υπάρχουν p+1 παράμετροι οι οποίες πρέπει να υπολογιστούν. Για την περιγραφή του μοντέλου σε μορφή πινάκων απαιτούνται τέσσερις πίνακες ο n×1 πίνακας-στήλη, Y των παρατηρήσεων επί της εξαρτημένης μεταβλητής Yi, ο n×p πίνακας X από μια στήλη σε μονάδες, με την ένδειξη 1, η οποία ακολουθείται από τις p στήλες των παρατηρήσεων επί των ανεξάρτητων μεταβλητών, ο p×1 πίνακας β των προς εκτίμηση παραμέτρων και ο n×1 πίνακας ε των τυχαίων σφαλμάτων (Kutner et al., 2005). Το γραμμικό μοντέλο μπορεί να αναπαρασταθεί με τη μορφή:

$$\mathbf{Y = X\beta + \varepsilon}$$

2.2.6.3 Συντελεστής προσδιορισμού R^2

Η αξιολόγηση του κατά πόσο αποδοτικά ένα μοντέλο γραμμικής παλινδρόμησης ερμηνεύει τη διακύμανση της τιμής της εξαρτημένης μεταβλητής γίνεται με τον υπολογισμό του συντελεστή προσδιορισμού (coefficient of determination R^2 ,R-squared). Ο συντελεστής προσδιορισμού λαμβάνει τιμές από το 0 έως 1 και εκφράζει το ποσοστό της διακύμανσης που ερμηνεύει γραμμικό μοντέλο παλινδρόμησης. Ο συντελεστής R^2 δε διαθέτει μονάδες μέτρησης επειδή εκφράζει ποσοστό. Όταν η τιμή του συντελεστή τείνει προς τη μονάδα, τόσο πληρέστερα επιτυγχάνει το μοντέλο να ερμηνεύσει τη διακύμανση της ανεξάρτητης μεταβλητής.

2.2.6.4 Μη γραμμικά μοντέλα

Τα μη γραμμικά μοντέλα βασίζονται στην παραδοχή ύπαρξης μη γραμμικής σχέσης μεταξύ της εξαρτημένης μεταβλητής. Το μη γραμμικό μοντέλο έχει τη γενική μορφή Yi=f(Xi,β)+εi, όπου Xi είναι το διάνυσμα των προβλεπουσών μεταβλητών και β

το διάνυσμα των παραμέτρων. Η μορφή αυτή είναι συναφής με τη μορφή των γραμμικών μοντέλων με τη διαφορά ότι η αναμενόμενη συνάρτηση είναι μη γραμμική (Bates & Watts, 1988).

2.2.6.5 Πολυωνυμικά μοντέλα

Σε ένα πολυωνιμικό μοντέλο υπάρχει μια ανεξάρτητη μεταβλητή, το πολυώνυμό δευτέρου βαθμού το οποίο αναπαρίσταται με τη μορφή (Rawlings, Pantula & Dickey, 1998): $\mathbf{E(Y)=\beta +\beta X_1+\beta X_2}$

Σε αντίθεση με το γραμμικό μοντέλο, στο πολυωνυμικό μοντέλο, εκτός του X, υπάρχει και ο όρος X^2. Το συγκεκριμένο μοντέλο μπορεί να θεωρηθεί και ως μια ειδική περίπτωση ενός μοντέλου πολλαπλής παλινδρόμησης, όπου είναι X1=X και $X_2=X_2$. Συμπερασματικά, ένα πολυωνυμικό μοντέλο ανώτερης τάξης έχει τη μορφή:

$$\mathbf{E(Y)=\beta +\beta X+\beta X_2+...+\beta X_p \ (2) \ όπου \ p\geq 3.}$$

2.2.6.6 Εκθετικά μοντέλα

Στα εκθετικά μοντέλα παλινδρόμησης (exponential regression models) υπάρχει μόνο μια εξαρτημένη μεταβλητή. Η γενική μορφή ενός μοντέλου εκθετικής παλινδρόμησης με τους όρους σφάλματος οι οποίοι είναι κανονικά κατανεμημένοι είναι η εξής:

$$\mathbf{Y =\beta_0 exp(\beta X)+\varepsilon_i}$$

όπου β_0 και β_1 είναι οι παράμετροι του μοντέλου, οι τιμές X_i είναι γνωστές και τα σφάλματα ε_i είναι ανεξάρτητα και ακολουθούν την κανονική κατανομή με μέση τιμή μηδέν και διακύμανση σ^2, δηλαδή ε_i : N (0, σ^2). Η συνάρτηση απόκρισης (response function) έχει τη μορφή: (Rawlings, Pantula & Dickey, 1998)

$$\mathbf{f\ (X,\beta)= \beta_0 exp(\beta_1 X\ )\ \ όπου\ \beta = (\beta_0,\beta_1)}$$

Η γενική μορφή έχει τη μορφή Y = β + β exp(β X)+εi, όπου είναι β = (β 0,β1 ,β2), με συνάρτηση απόκρισης :

$$\mathbf{f(X,\beta)=\beta +\beta exp(\beta X)}$$

2.2.6.7 Λογιστικά μοντέλα

Η λογιστική παλινδρόμηση (Logistic regression) αποτελεί μοντέλο ταξινόμησης των τιμών μιας μεταβλητής απόκρισης Y βάσει της θεωρίας των πιθανοτήτων. Στο μοντέλο αυτό η μεταβλητή Y συνήθως έχει λαμβάνει δύο τιμές και στοχεύεται η

πρόβλεψη της έκβασης αυτής από πλήθος προβλεπτικών μεταβλητών είτε ονομαστικών, είτε ποσοτικών είτε τακτικών. Όμως η διαφορά μεταξύ λογιστικής και γραμμικής παλινδρόμησης έγκειται στη φύση της επιλεγμένης μεταβλητής απόκρισης, η οποία στην πρώτη μπορεί να είναι κατηγορική, (τακτική ή ονομαστική), στη δεύτερη είναι αποκλειστικά ποσοτική. Ενώ στη γραμμική παλινδρόμηση η εκτίμηση των παραμέτρων α και b_i πραγματοποιείται με τη μέθοδο των ελαχίστων τετραγώνων, στη λογιστική παλινδρόμηση η εκτίμηση των παραμέτρων πραγματοποιείται με τη μέθοδο του λόγου πιθανοφάνειας δηλαδή επιλέγοντας τις πιο πιθανοφανείς τιμές των παραμέτρων, που θα οδηγήσουν στα παρατηρούμενα αποτελέσματα. Συνεπώς στην πρώτη υπάρχει αποδοχή ύπαρξης ομοιογένειας (ομοσκεδαστικότητας) στα υπολείμματα των αποκρίσεων ενώ στη δεύτερη αναπτύσσεται πάντα ετεροσκεδαστικότητα σε κάθε προβλεπόμενη τιμή εξαιτίας του μεταβαλλόμενου ποσοστού διακύμανσης που αναλογεί σε αυτή (Bates & Watts, 1988).

2.2.6.8 Σιγμοειδή μοντέλα ανάπτυξης (Sigmoidal Growth Models)

Στα Σιγμοειδή μοντέλα ανάπτυξης τα δεδομένα έχουν την τάση να προσαρμόζονται σε καμπύλες S-μορφής. Χαρακτηριστικό των S καμπυλών είναι ότι αυξάνουν με μεταβαλλόμενο ρυθμό, εκκινώντας από αρχική τιμή ανάπτυξης, προσεγγίζοντας ασυμπτωτικά μια τελική τιμή (Rawlings, Pantula & Dickey, 1998. Ο ρυθμός ανάπτυξης αρχικά βαίνει αυξανόμενος έως ότου φτάσει μέγιστη τιμή, ενώ στη συνέχεια βαίνει μειούμενος, το οποίο και προκαλεί σημείο καμπής στο γράφημα. Τα πιο δημοφιλή μοντέλα σιγμοειδούς ανάπτυξης είναι το μοντέλο Gompertz, και το λογιστικό μοντέλο ανάπτυξης.

Κεφάλαιο 3

3.1 Μεγάλα Δεδομένα στη Δημόσια Διοίκηση

Την τελευταία δεκαετία η ανάγκη επεξεργασίας, αρχειοθέτησης εξόρυξης και ανάλυσης μεγάλου όγκου δεδομένων στη Δημόσια Διοίκηση λαμβάνει εκθετική πρόοδο με αποτέλεσμα να είναι απαραίτητες νέες μέθοδοι και τεχνικές για τη διαχείρισή τους. Η εξέλιξη των ΤΠΕ (Τεχνολογιών Πληροφοριών και Επικοινωνιών) και υπολογιστικού νέφους δημιούργησαν καινούριες απαιτήσεις αποθήκευσης και επεξεργασίας των παραγόμενων δεδομένων με αποτέλεσμα τη δημιουργία πολύπλοκων εσωτερικών διαδικασιών που η Δημόσια Διοίκηση οφείλει να διαχειριστεί. Εκτός των εσωτερικών διαδικασιών που καλείται να αντιμετωπίσει η Δημόσια Διοίκηση, οφείλει να αφουγκραστεί τη γνώμη των πολιτών ως προς το παραχθέν έργο της. Η εξέλιξη και ανάπτυξη του διαδικτύου δημιούργησαν ένα νέο σημείο διεπαφής Πολίτη - Κράτους καθιστώντας τον σύγχρονο πολίτη ενεργό, οδηγώντας με αυτόν τον τρόπο τη Δημόσια Διοίκηση στην εξωστρέφεια. Η αναβάθμιση του Διαδικτύου από web 1.0 σε web 2.0, η ανάπτυξη των μέσων κοινωνικής δικτύωσης και η ύπαρξη διαδικτυακών forum οδήγησε τους πολίτες στην ευκολότερη πρόσβασή τους και συμμετοχή σε έναν ευρύ κατάλογο θεμάτων που πολλές φορές πρώτοι εκκινούν απέναντι στη Διοίκηση.

Το όφελος από την ανάλυση μεγάλου όγκου δεδομένων από τη Δημόσια Διοίκηση με στόχο τη λήψη των κατάλληλων αποφάσεων που θα οδηγήσουν σε μεγαλύτερη αποτελεσματικότητα του κράτους έχει άμεσο αντίκτυπο στην καθημερινότητα και ποιότητα ζωής των πολιτών μιας και δημιουργείται μηχανισμός άμεσης αντίληψης των αναγκών της κοινωνίας, ενώ ταυτόχρονα απαιτούνται λιγότεροι πόροι για την ικανοποίηση αυτών των αναγκών. Προκειμένου τα Μεγάλα Δεδομένα να χρησιμοποιούνται στο έπακρο και πάντα στο πλαίσιο των γενικών κανονισμών προστασίας δεδομένων, απαιτείται οι Δημόσιες Διοικήσεις των χωρών της Ε.Ε. να δημιουργήσουν ένα καινοτόμο πλαίσιο που αφορά τα νομικά θέματα που προκύπτουν από τα προσωπικά δεδομένα.

3.2 Εξόρυξη δεδομένων και λήψη αποφάσεων στο Δημόσιο Τομέα

Οι Δημόσιες Διοικήσεις έχουν κεντρική θέση τόσο στη δημιουργία όσο και στη διαχείριση της γνώσης. Σύμφωνα με την Jordan (2015), η κυβέρνηση και κατ' επέκταση η Διοίκηση, έχει δύο ρόλους στο έργο της γνώσης: την παραγωγή γνώσης και τη διαχείριση της γνώσης. Τα «μεγάλα δεδομένα» (Big Data) και η εξόρυξη πληροφορίας

από αυτά, μπορούν να επιτρέψουν στη Δημόσια Διοίκηση να προβλέπει τις απαιτήσεις για παραγωγή δημόσιας πολιτικής, ακόμη και πριν οι πολίτες εκφράσουν τις απόψεις τους (Jordan, 2014). Σύμφωνα με τον Henninger (2013) υπάρχει ιδιαίτερη σημασία ενημέρωσης του δημόσιου τομέα, ειδικά για την εξυπηρέτηση του δημόσιου συμφέροντος. Οι πληροφορίες του δημόσιου τομέα χρησιμοποιούνται για την ενημέρωση των δημοσίων πολιτικών και αποφάσεων, την υποστήριξη της οικονομικής ανάπτυξης, την αύξηση της συμμετοχής των πολιτών, τη δημιουργία εμπιστοσύνης στις σχέσεις κράτους - πολίτη και την πρόσβαση στη γνώση. Ο Δημόσιος Τομέας δημιουργεί πληροφορίες συλλέγοντας δεδομένα από τους πολίτες και τους θεσμούς χρησιμοποιώντας τα για την υποστήριξη στη λήψη αποφάσεων και τον σχεδιασμό και την υποβολή εκθέσεων (Henninger, 2013).

Πολλοί τομείς του Δημόσιου Τομέα επηρεάζονται από τον μεγάλο όγκο δεδομένων. Οι Alves, Martinez και Viegas (2012) συνειδητοποίησαν τη σημασία των δεδομένων σε πραγματικό χρόνο και πρότειναν ένα Ευφυές Σύστημα Μεταφορών, για την παροχή πληροφοριών σχετικά με αναμενόμενους χρόνους ταξιδιού. Υπάρχουν ήδη διαφορετικές προσπάθειες δημιουργίας έξυπνων πόλεων, όπως η χρήση πληροφοριών που αποκτήθηκαν από αναζητήσεις ιστοτόπων, επισκέψεις και σχόλια πολιτών για τη δημιουργία νέας γνώσης για τις πόλεις και την προσφορά νέων προσεγγίσεων στην αστική διαχείριση (Arribas-Bel, 2014). Η προσοχή εστιάζεται επίσης στον πολίτη ως κεντρικό παράγοντα σε διαφορετικούς τομείς του δημόσιου τομέα (Annoni, Ferrari, & Salini, 2006; Corallo et al., 2015). Η ενθάρρυνση της συμμετοχής των πολιτών (Henninger, 2013) διαδραματίζει βασικό ρόλο στη σύγχρονη ηλεκτρονική διακυβέρνηση, υποδεικνύοντας ότι τα δεδομένα και οι δημόσιες πληροφορίες χρησιμοποιούνται για την αύξηση της εμπιστοσύνης των πολιτών προς το κράτος και τη συμμετοχή των πολιτών στη σύγχρονη δημοκρατία.

Η δεύτερη διάσταση, η λειτουργία των της ανακάλυψης γνώσης μέσα από βάσεις δεδομένων στον δημόσιο τομέα περιλαμβάνει την υγειονομική περίθαλψη, τα μέσα κοινωνικής δικτύωσης και τα ανοιχτά δεδομένα. Σύμφωνα με τους Huang et al. (2015), οι τεχνικές εξόρυξης έχουν πολλές επιτυχημένες εφαρμογές στους τομείς της υγειονομικής περίθαλψης. Οι ερευνητές επισημαίνουν πολλές περιπτωσιολογικές μελέτες που χρησιμοποιούν πληροφορίες που αντλήθηκαν από μεγάλες βάσεις δεδομένων στην υγειονομική περίθαλψη, όπως αρχεία υγειονομικής περίθαλψης, συστήματα συστάσεων, ανάλυση επιδημιολογικών δεδομένων μέσω διαδικτύου και

χρήση μεγάλων δεδομένων για τον προσδιορισμό της ποιότητας του αέρα (Huang et al., 2015; Jordan, 2015).

Επιπρόσθετα, οι ερευνητές χρησιμοποιήσουν δευτερογενείς πηγές δεδομένων, εθνικά σύνολα δεδομένων από διαφορετικά στατιστικά κέντρα και έρευνες για να μελετήσουν τα θέματα που αφορούν τον τομέα της φροντίδας υγείας (McDermott & Turk, 2015). Ως εκ τούτου, η εξόρυξη και ανάλυση των δεδομένων δημιουργεί τεράστιες προκλήσεις και δυνατότητες στον κλάδο της υγειονομικής περίθαλψης, ενθαρρύνοντας τη μελλοντική έρευνα σε αυτόν τον τομέα (Wong et al., 2015). Η χρήση των μέσων κοινωνικής δικτύωσης δεν περιορίζεται μόνο σε μεμονωμένους χρήστες, αλλά οι τοπικές αρχές επωφελούνται από τη χρήση των μέσων κοινωνικής δικτύωσης καθώς οι αποφάσεις που λαμβάνονται σε δημοτικό επίπεδο έχουν άμεσο αντίκτυπο στην καθημερινή ζωή των πολιτών.

Ως εκ τούτου, η αλληλεπίδραση μεταξύ πολιτών και τοπικών αρχών θεωρείται σημαντική και η χρήση δημοτικών ιστοσελίδων συσχετίζεται με την εμπιστοσύνη στις τοπικές αρχές (Levon & Steinfeld, 2015). Επιπλέον, οι δημοτικές σελίδες στο Facebook περιέχουν σημαντικές δυνατότητες για τη βελτίωση της επικοινωνίας μεταξύ πολιτών και αρχών. Παραδείγματος χάριν, στα μέσα κοινωνικής δικτύωσης η εξόρυξη δεδομένων θα μπορούσε να προσφέρει μια πιθανή λύση στο πρόβλημα της μείωσης των πόρων και των οικονομικών απαιτήσεων εντός των πόλεων (Kennedy et al., 2015). Όπως υποστηρίζουν οι Kamel, Boulos και Al-Shorbaji (2014), το Διαδίκτυο των Πραγμάτων παρέχει νέες δυνατότητες για τη βελτίωση των έξυπνων πόλεων. Οι Napoli και Karaganis (2010) τονίζουν ότι η δημόσια πολιτική πρέπει να διαμορφώνεται με δημόσια διαθέσιμα δεδομένα (ή ανοιχτά δεδομένα) για να αυξηθεί το επίπεδο διαφάνειας και η προσβασιμότητά της για τη λήψη αποφάσεων υποστηρίζοντας ότι η διαφάνεια των δεδομένων και η πρόσβαση σε αυτά είναι θεμελιώδεις για τη χάραξη δημόσιας πολιτικής.

Αντίστοιχα, υπάρχει η απαίτηση οι Δημόσιες Διοικήσεις να είναι ανοιχτές και διαφανείς και, όπως δήλωσε ο Henninger (2013), γιατί τα οφέλη από τη χρήση ανοιχτών δεδομένων είναι πολλά, όχι μόνο για τους κρατικούς φορείς, αλλά για το κοινωνικό σύνολο, όπως οφέλη από την πρόληψη των περιβαλλοντικών επιπτώσεων.

3.3 Προκλήσεις και κίνδυνοι από τη χρησιμοποίηση μεγάλων δεδομένων στο Δημόσιο Τομέα.

Τα σύνολα μεγάλων δεδομένων προκαλούν ποικίλες απαιτήσεις και υπάρχουν πολλά προβληματικά ζητήματα στη διαχείρισή τους. Η εξόρυξη και ανάλυση μεγάλων σε όγκο δεδομένων συμβάλλει στους τομείς συλλογής, ανάκτησης, συνολικής επεξεργασίας και οπτικοποίησης των πολύπλοκων συνόλων δεδομένων. Οι κυβερνήσεις και οι φορείς που διαχειρίζονται τα σύνολα δεδομένων, πρέπει να αναπτύξουν τεχνολογίες, δομές διαχείρισης και δυνατότητες προγραμματισμού για να είναι σε θέση να χειριστούν την εκθετική αύξηση των της παραγόμενης πληροφορίας. Οι ανησυχίες σχετικά με το απόρρητο και τους κινδύνους ασφάλειας των Big Data είναι πιθανό να αυξηθούν καθώς αυξάνεται ο όγκος των δεδομένων. Επομένως, οι πιθανοί κίνδυνοι σχετικά με το απόρρητο και την ασφάλεια πρέπει να αξιολογούνται σε όλους τους τομείς όπου εφαρμόζονται τεχνικές ανακάλυψης γνώσεις από δεδομένα.

Το απόρρητο αποτελεί μία από τις μεγαλύτερες προκλήσεις των συνόλων δεδομένων στο μέλλον, καθώς ομάδες πολιτών διστάζουν να διατηρούν τις πληροφορίες τους με τη μορφή ηλεκτρονικών αρχείων, κάτι που αποτελεί σημαντικό εμπόδιο εάν απαιτούνται σχετικά δεδομένα για τη συγκέντρωση μιας ψηφιακής υπηρεσίας. Οι αυστηροί κανόνες και κανονισμοί σχετικά με την προστασία δεδομένων, μπορούν επίσης να επηρεάσουν την αντίληψη των χρηστών για το απόρρητο. Επιπλέον, έμπειρο και εξειδικευμένο στελεχιακό δυναμικό πρέπει να είναι σε θέση να αντιμετωπίσει τις υπάρχουσες προκλήσεις. Επομένως, νέες προσεγγίσεις διαχείρισης, δομές και πλαίσια πολιτικής είναι απαραίτητες για να ξεπεραστούν οι προκλήσεις που προσδιορίστηκαν και να μετατραπεί η ανακάλυψη γνώσης από τα δεδομένα τα σε πολύτιμη γνώση.

Η συστηματική βιβλιογραφική ανασκόπηση αποκαλύπτει αρκετές προκλήσεις διαχείρισης που σχετίζονται με τα Big Data. Αυτές οι προκλήσεις διαχείρισης μεγάλων δεδομένων περιλαμβάνουν προκλήσεις επεξεργασίας δεδομένων, διαχείρισης και διασύνδεσης δεδομένων μεταξύ των οργανισμών και απουσία τεχνολογιών και τεχνογνωσίας που απαιτούνται για τη διαχείριση μεγάλων ποσοτήτων δεδομένων. Οι τεράστιοι όγκοι δεδομένων περιέχουν δυνατότητες, αλλά η συλλογή, η αρχειοθέτηση και η ανάκτησή τους εξακολουθεί να αποτελεί το σημαντικότερο προς επίλυση ζήτημα (Cao, 2012; Chen & Zhang, 2014; Dobre & Xhafa, 2014; Einav & Levin, 2014; Özköse, Arı, & Gencer, 2015).

Όπως αναφέρεται από τους Raad, Al Bouna και Chbeir (2015), είναι δύσκολο να χειριστεί κανείς τους συνεχώς αυξανόμενους όγκους δεδομένων. Ομοίως, οι Jiao et al. (2013) σημειώνουν ότι τα χαρακτηριστικά του όγκου, της ποικιλίας και της ταχύτητας δημιουργούν προκλήσεις για τη διαχείριση μεγάλων δεδομένων και την ανακάλυψη γνώσης από αυτά. Οι τεράστιες ποσότητες και το εύρος διαφορετικών τύπων δεδομένων, καθώς και η υψηλή ταχύτητα παραγωγής δεδομένων δημιουργούν νέα προς αντιμετώπιση θέματα που αφορούν την ανάλυση και ανακάλυψη γνώσης από τα.

Κατά συνέπεια, η επεξεργασία των δεδομένων αποτελεί μια καινούρια πρόκληση (Pandey & Dhoundiyal, 2015) και έχουν προταθεί αρκετές μέθοδοι όπως η μείωση δεδομένων και η τεχνητή νοημοσύνη προκειμένου να αντιμετωπιστούν οι προκλήσεις που προκύπτουν από τη συνεχή παραγωγή δεδομένων (Quick & Choo, 2014). Η εξόρυξη δεδομένων από βάσεις δεδομένων μεγάλης κλίμακας, δημιουργεί νέες απαιτήσεις ειδικά στον τομέα της ανάπτυξης και οπτικοποίησης των δεδομένων (El Kadiri et al., 2015).

Όσον αφορά την ανάλυση δεδομένων, η χρήση βάσεων δεδομένων και η ερμηνεία των αποτελεσμάτων είναι επίσης ζητήματα για τη διαχείριση μεγάλων δεδομένων (Arribas-Bel, 2014· El Kadiri et al., 2015). Υπάρχουν ζητήματα που σχετίζονται με τη μετατροπή των Μεγάλων Δεδομένων σε πολύτιμη γνώση και την κατανόηση των αποτελεσμάτων των Μεγάλων Δεδομένων (Anna & Nikolay, 2015· Chen et al., 2014). Σύμφωνα με τον Jordan (2015), το πρόβλημα δεν έγκειται απαραίτητα στην ενσωμάτωση των δεδομένων αλλά στη διαθεσιμότητα, την ανάλυση, την εύρεση και κατανόηση των σχέσεων και τη μετατροπή της σε χρήσιμη γνώση. Η διαχείριση μεγάλων δεδομένων απαιτεί τεχνογνωσία και δεξιότητες με σκοπό να αντληθεί πολύτιμη πληροφορία σύμφωνα με τον Kernaghan (Kernaghan, 2014). Υπάρχουν επίσης απαιτήσεις που σχετίζονται με την αρχειοθέτηση, τη διανομή και την ανάκτηση εγγραφών σε διαφορετικές πλατφόρμες (Aljunid et al., 2012· Chen & Zhang, 2014). Η εργασία πάνω στα δεδομένα, ενδέχεται να δημιουργήσει αλλοιώσεις στην πληροφορία επειδή τα δεδομένα έχουν συγχωνευθεί από διαφορετικές πηγές.

Ανάλογα με το σύνολο δεδομένων, μια διακύμανση στη μπορεί να προκαλέδει προβλήματα στη γενίκευση των ευρημάτων από διαφορετικές αναλύσεις και στη συνολική διαχείριση των δεδομένων (Hoagwood et al., 2015). Για να είναι σε θέση ο διαχειριστής της διαδικασίας ανάλυσης δεδομένων να διαχειρίζεται τις αυξανόμενες ποσότητες δεδομένων, πρέπει οι φορείς του δημόσιου τομέα πρέπει να αναπτύξουν

τεχνολογίες, δομές διαχείρισης και δυνατότητες προγραμματισμού (de Miranda Santo, Coelho, dos Santos, & Filho, 2006; Fuchs, Höpken, & Lexhagen, 2014).

Επιπλέον, σε αντίθεση με τον Δημόσιο Τομέα, τρίτα μέρη, όπως ομάδες δημοσίου συμφέροντος ή μη κερδοσκοπικοί οργανισμοί, δεν οφείλουν να ακολουθήσουν συγκεκριμένους κανονισμούς σχετικά με τη διαχείριση δεδομένων. Έτσι, πιθανότατα θα προκύψουν περαιτέρω θέματα και επιπτώσεις στη χάραξη των δημόσιων πολιτικών.

Στη Δημόσια Υγεία, η ποιότητα των δεδομένων πρέπει να λαμβάνεται υπόψη, καθώς η συγχώνευση των συνόλων δεδομένων επεκτείνεται, με αποτέλεσμα να γίνεται ολοένα και πιο πολύπλοκη (McDermott & Turk, 2015). Επομένως, είναι σημαντικό να διευκρινιστεί ποια σύνολα δεδομένων είναι συγκρίσιμα. Τα δεδομένα κακής ποιότητας μπορούν να οδηγήσουν σε απροσδόκητο και σημαντικό κόστος για τα κράτη (Ohemeng & Ofosu-Adarkwa, 2015).

Οι κίνδυνοι της ιδιωτικής ζωής και της ασφάλειας είναι επίσης πιθανό να συσχετίζονται με το μέγεθος, την ποικιλία και την πολυπλοκότητα των δεδομένων (Kshetri, 2014), επομένως είναι κατανοητό ότι οι ερωτήσεις σχετικά με το απόρρητο και την πληροφορική αποτελούν σημαντικό μέρος του διαλόγου (Kernaghan, 2014) και υπάρχει ανησυχία μεταξύ των ερευνητών σχετικά με τον τρόπο με τον οποίο λαμβάνεται υπόψη το απόρρητο των δεδομένων κατά τη διαχείριση των μεγάλων δεδομένων (Terry, 2015; Truyens & Van Eecke, 2014). Η ενθάρρυνση των ατόμων και άλλων ενδιαφερόμενων μερών να δημιουργήσουν μια αξιόπιστη υποδομή κοινής χρήσης δεδομένων είναι απαραίτητη, καθώς το απόρρητο μπορεί να παραβιαστεί κατά τη διάρκεια αυτής της διαδικασίας (Chen et al., 2014· El Kadiri et al., 2015).

Παρόμοιες προκλήσεις εντοπίζονται επίσης σε άλλες μελέτες (Batty et al., 2012; Fuchs et al., 2014). Υπάρχουν διάφορες απόψεις σχετικά με τον πιθανό κίνδυνο πρόσβασης των πολιτών στα σύνολα δεδομένων, όπως οι αρνητικές επιπτώσεις της απώλειας της εμπιστοσύνης του πολίτη που συχνά δηλώνεται όταν συζητείται η διαφάνεια και η εμπιστοσύνη (Henninger, 2013). Επιπλέον, υπάρχει μια σύγκρουση μεταξύ των απαιτήσεων για προστασία της ιδιωτικής ζωής και εκείνων για περισσότερο άνοιγμα και διαφάνεια (Henninger, 2013). Τελικά, οι σημαντικές ερωτήσεις σχετικά με τα Μεγάλα Δεδομένα είναι πιθανό να είναι ηθικές και όχι τεχνικές (Jordan, 2014). Επιπλέον, ο Jordan (2014) δηλώνει ότι η ηθική γνώση είναι το κλειδί, δεδομένου του ισχυρού ρόλου που θα διαδραματίσει η ανάλυση δεδομένων στη δημόσια διοίκηση. Ως

εκ τούτου, η ταχεία ανάπτυξη της πληροφορικής καταδεικνύει την ανάγκη για περισσότερη έρευνα στον τομέα της ηθικής και της πληροφορικής στον δημόσιο τομέα (Kernaghan, 2014).

Υπάρχει ανάγκη για μελλοντική έρευνα σχετικά με την ανακάλυψη εξόρυξη δεδομένων και την ανακάλυψη γνώσης μέσα από μεγάλα σύνολα δεδομένων και τους κινδύνους ιδιωτικότητας και ασφάλειας, σχετικά με τις διαφορές στους νόμους και τους κανονισμούς μεταξύ των χωρών (Kshetri, 2014). Ιδιέταιρη προσοχή δίνεται επίσης στον τομέα των Μεγάλων Δεδομένων και στη μελλοντική τους σημασία (Chen & Zhang, 2014). Οι Fan et al. (2015) επισημαίνουν αρκετές μελλοντικές κατευθύνσεις μελέτης, όπως η επιλογή κατάλληλων πηγών δεδομένων και μεθόδων ανάλυσης, η αντιμετώπιση της ετερογένειας μεταξύ των πηγών δεδομένων και η βελτίωση των πλαισίων καθώς εξελίσσεται η τεχνολογία. Επιπλέον, οι μελετητές προτείνουν τον εμπλουτισμό μοντέλων μελλοντικής λήψης αποφάσεων Big Data (Daniell, Morton, & Ríos Insua, 2015; Fosso Wamba et al., 2015) ενώ οι Fuchs και Horak (2008) εξέτασαν την ανακάλυψη γνώσης σε πραγματικό χρόνο ως μελλοντική δυνατότητα.

Οι Stough και McBride (2014) αναγνωρίζουν την αυξανόμενη σημασία της χρήσης Big Data σε διαφορετικούς κλάδους. Οι π μελέτες εστιάζουν την προσοχή στον μεταβαλλόμενο ρόλο των κυβερνήσεων στην παροχή ανοιχτών δεδομένων στους πολίτες και σε ηθικές ανησυχίες σχετικά με το απόρρητο και τη διανομή δεδομένων (Sieber & Johnson, 2015; Washington, 2014). Τόσο από πλευράς χρήστη όσο και από πλευράς παρόχου υπηρεσιών, τα Big Data περιέχουν πολλές μελλοντικές ευκαιρίες. Ο Shin (2015), πρότεινε τα Μεγάλα Δεδομένα θα πρέπει να αναπτυχθούν με μια προσέγγιση με επίκεντρο το χρήστη αντί για παραδοσιακές πρακτικές που βασίζονται στη θεωρία. Οι Zhong, Huang, Müller Arisona, Schmitt και Batty (2014) υπογραμμίζουν τη χρήση Big Data για τον κλάδο των μεταφορών, μέσω της χρήσης εργαλείων εξόρυξης δεδομένων και ανάλυσης. Επιπλέον, ο ρόλος των Μεγάλων Δεδομένων στην επιστήμη της υγείας τονίζεται στην έρευνα σχετικά με την επίτευξη αποτελεσματικότητας (Huang, Keser, Leland, & Shachat, 2003; Jordan, 2015).

Ο Jordan (2015) εκφράζει επίσης τη σημασία της εξόρυξης δεδομένων δηλώνοντας ότι οι εφαρμογές που δημιουργήθηκαν ειδικά για έρευνα που περιλαμβάνει πρόσβαση σε αρχεία υγείας θα επέτρεπαν στους ερευνητές να έχουν πρόσβαση σε όλα τα δεδομένα υγειονομικής περίθαλψης εάν ο ασθενής το εγκρίνει. Προβλέπεται ότι τέτοια δεδομένα θα δημιουργήσουν διορατικότητα και ικανότητα εύρεσης τάσεων και

προτύπων, δημιουργώντας γνώσεις ωφέλιμες για τους πολίτες. Επιπλέον, οι Kennedy et al. (2015) υποστηρίζουν ότι πρέπει να βρεθούν λύσεις με σκοπό την υποστήριξη σε φορείς του δημόσιου τομέα που έχουν περιορισμένη πρόσβαση στη χρήση ψηφιακών μεθόδων. Οι σωστά επεξεργασμένες αναλύσεις μπορεί να οδηγήσουν σε πολλά πλεονεκτήματα για τους πολίτες. Τα δεδομένα παράγονται συνεχώς με διάφορες μορφές όπως κινητά τηλέφωνα και καθώς το περιεχόμενο σαρώνεται, καταγράφεται και αρχειοθετείται (Teri, 2014). Η δημόσια πληροφόρηση είναι επομένως αναπόσπαστο κομμάτι της οικονομικής ανάπτυξης καθώς παρέχει στις Δημόσιες Υπηρεσίες πληροφορίες κατάλληλες για την παροχή βέλτιστων υπηρεσιών (Henninger, 2013). Με τον ίδιο τρόπο, οι Koerten και Veenswijk (2013) υποστηρίζουν ότι η ανοιχτή πρόσβαση σε δημόσιες πληροφορίες είναι ζωτικής σημασίας για την οικονομική ευημερία.

3.4 Δημόσιος Τομέας και ανακάλυψη γνώσης από δεδομένα.

Στο πλαίσιο του δημόσιου τομέα, η εξόρυξη δεδομένων με σκοπό την ανακάλυψη γνώσης παραμένει αρκετά ασαφής, αλλά μπορεί να παραχθεί με αναφορά στον ορισμό των πληροφοριών του δημόσιου τομέα, ο οποίος ορίζει την «πληροφορία, συμπεριλαμβανομένων όλων των πληροφοριών σε οποιαδήποτε μορφή, και των υπηρεσιών που παράγονται, δημιουργήθηκαν, συλλέγονται, επεξεργάζονται, διατηρούνται, συντηρούνται, διαδίδονται ή χρηματοδοτούνται από ή για δημόσιες οντότητες» (Henninger, 2013).

Τα «προϊόντα πληροφοριών σε οποιαδήποτε μορφή» θα μπορούσαν να ερμηνευθούν ως «δεδομένα οποιουδήποτε τύπου», που είναι χαρακτηριστικό των μεγάλων Δεδομένων. Ο όγκος τέτοιων «προϊόντων πληροφοριών» είναι σχεδόν βέβαιο πως θα αυξηθεί εκθετικά. Έτσι, τα Big Data αντικατοπτρίζονται στην ουσία στις πληροφορίες του δημόσιου τομέα. Επιπλέον, τα «Ανοιχτά Δεδομένα», δεδομένα στα οποία οι Δημόσιοι Οργανισμοί καθιστούν προσβάσιμα για την προώθηση της διαφάνειας και της οικονομικής ανάπτυξης αποτελούν μέρος των Μεγάλων Δεδομένων του δημόσιου τομέα. Αν και δεν υπάρχει κοινή συμφωνία για την έννοια ή τον ορισμό των ανοιχτών δεδομένων (Ohemeng & Ofosu-Adarkwa, 2015), φαίνεται να υπάρχει συμφωνία σχετικά με το τι συνιστά ανοιχτά δεδομένα. Τα ανοιχτά δεδομένα είναι όλες οι πληροφορίες που σχετίζονται με τις δημόσιες επιχειρήσεις σε ένα δημοκρατικό περιβάλλον. Συνεπώς, ο τύπος, η ποσότητα και η ταχύτητα ανάπτυξης αυτών των πληροφοριών αντικατοπτρίζουν το περιεχόμενο του τι είναι μεγάλα δεδομένα.

Τεχνικές, όπως η εξόρυξη δεδομένων, δημιουργούν δυνατότητες για να ρίξουν φως στα δεδομένα, διευρύνοντας έτσι την επιρροή των Big Data στον δημόσιο τομέα. Σύμφωνα με τη διεθνή αρθρογραφία, η λήψη αποφάσεων αποτελεί το μεγαλύτερο μέρος αυτής της εξέλιξης που προσανατολίζεται στα δεδομένα. Ως εκ τούτου, η λήψη αποφάσεων με βάση τεχνικών εξόρυξης δεδομένων και ανακάλυψης γνώσης από τα δεδομένα μπορεί να μετατραπεί σε απαραίτητη διαδικασία σχεδόν σε κάθε Δημόσιο Οργανισμό, αποκαλύπτοντας νέες ευκαιρίες και οφέλη. Στο μέλλον, είναι κατανοητό ότι τα ανοιχτά δεδομένα και τα μέσα κοινωνικής δικτύωσης θα είναιοι πτυχές που οδηγούν τα Μεγάλα Δεδομένα καθώς και τα δύο παρουσιάζουν εκθετική ανάπτυξη.

3.5 Ανακάλυψη γνώσης από Δεδομένα και Κοινωνία των Πολιτών.

Μελέτες σχετικά με την εφαρμογή ανακάλυψης γνώσης με βάση τεχνικών εξόρυξης σε βάσεις δεομένων στο δημόσιο τομέα, αποκαλύπτουν μια μεγάλη ποικιλία παραδειγμάτων για το πώς οι κρατικοί φορείς έχουν χρησιμοποιήσει τα μεγάλα δεδομένα στην πράξη προς όφελος της Δημόσιας Διοίκησης και των πολιτών. Οι τομείς στους οποίους εφαρμόζονται ποικίλλουν και εκκινούν από την αστική διαχείριση έως την υγειονομική περίθαλψη και τη συμμετοχή των πολιτών. Στην αστική διαχείριση, οι Fuchs et al. (2014) εισάγουν την έννοια της υποδομής γνώσης με τη βοήθεια του Big Data analytics, το οποίο είναι πληροφοριακό σύστημα βασισμένο σε νοημοσύνη για τη διαχείριση φράσεων πριν και μετά το ταξίδι. Δεδομένα όπως η αναζήτηση ιστότοπου, η κράτηση και τα σχόλια χρησιμοποιούνται πλήρως και η διαθεσιμότητα ενός τέτοιου συστήματος αξιολογήθηκε στην πόλη Are της Σουηδίας.

Παρόμοιες πρακτικές μπορούν να βρεθούν και σε άλλους τομείς. Φαίνεται ότι υπάρχει συμφωνία για τη σημασία της ροής και της προσβασιμότητας πληροφοριών καθώς και για τον ρόλο που πρέπει να έχει για τη βελτίωση της διαφάνειας, την ενίσχυση της διαφάνειας και την αύξηση της εμπιστοσύνης και τ δέσμευσης των πολιτών με τη Δημόσια Διοίκηση και τις αποφάσεις της. Οι Ohemeng και Ofosu-Adarkwa (2015) εκφράζουν ανησυχίες σχετικά με την ανισορροπία μεταξύ κράτους και κοινωνίας των πολιτών στη δημιουργία ενός περιβάλλοντος ανοιχτών δεδομένων. Έχει δοθεί μεγάλη έμφαση στον ρόλο της κυβέρνησης, ωστόσο, οι μελετητές επαναλαμβάνουν επανειλημμένα την εμπλοκή των πολιτών ως βασική πτυχή της εποχής της ηλεκτρονικής διακυβέρνησης και της κοινωνίας της πληροφορίας. Οι διάφοροι ρόλοι του δημόσιου τομέα είναι σημαντικοί όταν συζητείται το μέλλον των δεδομένων και της

εκμετάλλευσής τους, ως εκ τούτου, υπάρχει αίτημα για μεγαλύτερη εστίαση στον ρόλο της κοινωνίας των πολιτών και όχι στον ρόλο της κυβέρνησης.

Το μέλλον των δεδομένων και κατά συνέπεια της εξόρυξης γνώσης μέσω αυτών, υπόσχεται νέες εξελίξεις, ειδικά στον τομέα της υγείας, της βιομηχανίας και της εφαρμογής δημοσίων πολιτικών. Προηγούμενες μελέτες (Huang et al., 2015), σχετικά με τις δυνατότητες για την ανακάλυψη γνώσης από Μεγάλα Δεδομένα στον τομέα της υγειονομικής περίθαλψης περιλαμβάνουν τη βοήθεια των γιατρών να βελτιστοποιήσουν το χρόνο με τους ασθενείς και την ποιότητα των υπηρεσιών τους, βελτιώνοντας την ακρίβεια και την ταχύτητα της επεξεργασίας πληροφορίας ασθενούς μέσω της χρήσης σωστών και ταχέως διαθέσιμων πληροφοριών. Επιπλέον, η ανάλυση των δεδομένων υποστηρίζει νέες διαδικασίες στην καθιέρωση ηλεκτρονικών συνταγών μέσω ηλεκτρονικής επικοινωνίας μεταξύ ασθενούς και παρόχου υγειονομικής περίθαλψης κάτι που ήδη εφαρμόζεται και στην Ελλάδα.

Εκτός του τομέα φροντίδας υγείας, η ηλεκτρονική διακυβέρνηση, επιτρέπει στους χρήστες την άμεση συμμετοχή. Στην Ευρωπαϊκή Ένωση υπάρχουν σχετικές πρωτοβουλίες αλλά το πεδίο εφαρμογής είναι περιορισμένο, επειδή δεν μπορούν όλοι οι πολίτες να συμμετέχουν στην ηλεκτρονική διακυβέρνηση. Ένα σαφές πλεονέκτημα της ηλεκτρονικής διακυβέρνησης και της ανακάλυψης γνώσης από τα δεδομένα είναι ότι επιτρέπει τη διαφάνεια στην εφαρμογή των Δημοσίων Πολιτικών. Επιπλέον, απαιτείται ανάπτυξη στις ηλεκτρονικές υπηρεσίες με τέτοιον τρόπο ώστε να επιτρέπεται στους χρήστες να έχουν πρόσβαση σε αυτές τις υπηρεσίες μέσω μιας εύκολα κατανοητής γραφικής διεπαφής χρήστη-συστήματος.

3.6 Τομείς εφαρμογής της εξόρυξης δεδομένων στο Δημόσιο Τομέα.

Στόχος της ενότητας είναι να διερευνήσει τη δυνατότητα χρήσης της εξόρυξης δεδομένων σε δημόσιους οργανισμούς ως εργαλείο για τη βελτίωση της αποτελεσματικότητάς τους. Στην εργασία θα επικεντρωθούμε σε τομείς εφαρμογών της εξόρυξης δεδομένων για δημόσιους οργανισμούς. Σε αυτή την ενότητα εξετάζονται τομείς εφαρμογής στην οικονομία, υγειονομική περίθαλψη, δικαιοσύνη και άμυνα, εργασία και κοινωνική πρόνοια, ηλεκτρονική διακυβέρνηση, εκπαίδευση και μεταφορές. Οι τρέχουσες εφαρμογές αναζητήθηκαν στο Διαδίκτυο (Google) με τη χρήση λέξεων: εξόρυξη δεδομένων, κυβέρνηση και δημόσια διοίκηση. Πραγματοποιήθηκε επίσης αναζήτηση στις βάσεις δεδομένων Emerald και Science Direct.

3.6.1 Οικονομία

Διεθνώς οι φορολογικές υπηρεσίες χρησιμοποιούν την εξόρυξη δεδομένων για να δημιουργήσουν ένα μοντέλο πρόβλεψης που θα μπορούσε να βελτιώσει τη διαχείριση ερωτημάτων και την επιλογή ελέγχου απαντώντας σε ερωτήσεις όπως "Ποιος είναι πιθανό να μην μπορεί να ανταποκριθεί φορολογικά και κατά πόσο;" και "Ποιες φορολογικές δηλώσεις είναι πιθανό να μην συμμορφώνονται. Τέτοια μοντέλα μειώνουν την ευκαιρία για απάτη. Με αυτόν τον τρόπο η φοροδιαφυγή και εισφοροδιαφυγή θα μπορούσε να εντοπιστεί με τη μέθοδο του δέντρου αποφάσεων. Επίσης, η ανάλυση συσχέτισης θα μπορούσε να εντοπίσει ομάδες φόρων που οι φοροφυγάδες συνήθως προσπαθούν να αποφύγουν. Τα νευρωνικά δίκτυα έχουν παράσχει πολύτιμες γνώσεις για τους αναλυτές που προβλέπουν φορολογικά έσοδα, τα οποία είναι εξαιρετικά σημαντικά καθώς οι προϋπολογισμοί των φορέων, η υποστήριξη για την εκπαίδευση και οι βελτιώσεις στην υποδομή εξαρτώνται από την ακρίβειά τους (Hansen et.al, 1997).

Το σύστημα εξόρυξης δεδομένων έχει σχεδιαστεί ειδικά για να βοηθά τους φορείς χάραξης πολιτικής να έχουν πρόσβαση στα δεδομένα με τρόπο που χρησιμοποιείται γενικά από τις τράπεζες για την ανάλυση της αποτελεσματικότητας των προγραμμάτων δανεισμού (Makulowich, 1999). Τεχνικές εξόρυξης δεδομένων και ανακάλυψη γνώσης από σύνολα δεδομένω, μπορούν να χρησιμοποιηθούν για την αύξηση της αποτελεσματικότητας των προγραμμάτων που ενθαρρύνουν τις μικρές επιχειρήσεις και τις καινοτομίες. Το Υπουργείο Οικονομίας προσφέρει συνήθως πιστωτικά όρια για ιδιοκτήτες μικρών επιχειρήσεων και καινοτόμους επιχειρηματίες.

3.6.2 Φροντίδα υγείας

Ο Sund (2002) περιγράφει δύο χρήσεις της εξόρυξης δεδομένων στο σύστημα υγειονομικής περίθαλψης της Φινλανδίας. Η μέθοδος γενικευμένων ακολουθιών συμβάντων χρησιμοποιείται για την ανάπτυξη και την εφαρμογή δεικτών απόδοσης που βασίζονται σε μητρώο για τη μέτρηση της αποτελεσματικότητας της χειρουργικής θεραπείας. Επίσης, η περιγραφή έννοιας/τάξης χρησιμοποιείται για την αξιολόγηση και σύγκριση της αποτελεσματικότητας των παρόχων υγειονομικής περίθαλψης.

Η εξόρυξη δεδομένων χρησιμοποιείται για τον εντοπισμό απάτης στην υγειονομική περίθαλψη. Το σύστημα βαθμολογεί κάθε στοιχείο και επεξεργάζεται δεδομένα για να δημιουργήσει έναν «δείκτη υποψίας» όλων των παρόχων. Για να εντοπίσουν ύποπτους παρόχους, οι αναλυτές επιλέγουν από πολλά πρότυπα

συμπεριφοράς κατάλληλα για μια συγκεκριμένη ομάδα ομοτίμων και, στη συνέχεια, συνδυάζουν μοτίβα για να δημιουργήσουν ένα μοντέλο ανάλυσης. Το σύστημα μπορεί επίσης να αναπτυχθεί για το προφίλ των ασθενών–διευκολύνοντας την «ανάλυση συνδέσμων» μεταξύ ιατρών που εμπλέκονται σε απάτη. Σύστημα εξόρυξης δεδομένων χρησιμοποιείται για την ανίχνευση ψευδών αξιώσεων προς το κράτος με χρήση αριθμών ταυτότητας που έχουν κλαπεί από ασθενείς με και την πρόληψη σφαλμάτων πληρωμής όπως δείκτες παρακολούθησης για ασθενείς που εισήχθησαν άσκοπα, ασθενείς που εξήλθαν και εισήχθησαν ξανά την ίδια ημέρα, διαγνώσεις που πραγματοποιούνται με διαφορετικούς τρόπους και εσφαλμένοι κωδικοί διάγνωσης.

Η εξόρυξη δεδομένων αποτελεί διαδικασία για την ανάπτυξη της υγειονομικής περίθαλψης σε χώρες όπως η Γερμανία, η Αυστρία (Haux et.al., 2002) και η Τσεχική Δημοκρατία (Zvarova et.al., 2002). Είναι επίσης αναπόσπαστο μέρος της συνεχούς βελτίωσης της ποιότητας και της έρευνας του The National Emergency Medical Extranet Project που στοχεύει στη βελτίωση της επείγουσας κλινικής περίθαλψης μέσω υποστήριξης πληροφοριών σε πραγματικό χρόνο και στην παροχή οφέλους μέσω της υποστήριξης πληροφοριών για πρωτοβουλίες δημόσιας υγείας (Barthell et.al., 2003).

Ορισμένα από τα έργα εξακολουθούν να στοχεύουν στην επίδειξη της χρησιμότητας των τεχνικών εξόρυξης δεδομένων. Για παράδειγμα, το Data Mining for Toxic Hazard Analysis αξιολογεί τη χρησιμότητα των τεχνικών εξόρυξης δεδομένων για την αξιολόγηση χημικού κινδύνου τροφίμων. Οι τεχνικές εξόρυξης δεδομένων χρησιμοποιούνται για τη δημιουργία κανόνων πρόβλεψης κινδύνων. Τεχνικές εξόρυξης δεδομένων χρησιμοποιούνται για την εφαρμογή τεχνικών μηχανικής μάθησης με σκοπό την παραγωγή κανόνων πρόβλεψης από βάσεις δεδομένων (DSS Consulting, 2003).

3.6.3 Εργασία και κοινωνική πρόνοια

Τα δεδομένα απογραφής είναι μια από τις πιο ολοκληρωμένες βάσεις δεδομένων. Συνήθως γίνεται συλλογή δεδομένων του πληθυσμού και συλλέγονται στατιστικά στοιχεία απαραίτητα για το σχεδιασμό και την εφαρμογή Δημόσιων Πολιτικών. Το σύστημα εξόρυξης δεδομένων χρησιμοποιείται για τον υπολογισμό των δεικτών στέρησης που χρησιμοποιούν για τη μέτρηση της συσχέτισης μεταξύ του επιπέδου στέρησης και μιας ποικιλίας δεικτών υγείας (Klosgen et.al., 2003).

Το KESO (Απόσπαση Γνώσης για Στατιστικές Υπηρεσίες) είναι ένα πρωτότυπο Σύστημα Εξόρυξης Δεδομένων που χρησιμοποιεί η Eurostat. Στόχος του είναι η

παραγωγή ενός Συστήματος Εξόρυξης Δεδομένων που λύνει τις ανάγκες των αναλυτών στατιστικών δεδομένων (Siebes, 1996).

3.6.4 Ηλεκτρονική Διακυβέρνηση

Το Διαδίκτυο προσφέρει τεράστιες ευκαιρίες στη Δημόσια Διοίκηση για την παροχή ποιοτικότερου περιεχομένου και υπηρεσιών με σκοπό να αλληλεπιδρά με τους πολίτες, τις επιχειρήσεις και Δημόσιους Οργανισμούς. Προτείνεται οι νέες τεχνολογίες βάσης δεδομένων και εξόρυξης δεδομένων να γίνουν ο καταλύτης για την ενθάρρυνση της ανταλλαγής πληροφοριών και την υποστήριξη της συνεργασίας και της έρευνας μεταξύ των αστυνομικών τμημάτων, των σωφρονιστικών γραφείων, των κοινωνικών υπηρεσιών και των δικαστηρίων, που στο παρελθόν ήταν δύσκολο να διεξαχθούν (Chen, 2003).

3.6.5 Παιδεία

Στον τομέα της εκπαίδευσης διερευνήθηκε η σχέση μεταξύ της δομής του προγράμματος σπουδών και της τυποποιημένης απόδοσης των τεστ για να κατανοήσει πώς η σειρά μαθημάτων επηρεάζει τις βαθμολογίες των εξετάσεων. Το σύστημα εξόρυξης δεδομένων χρησιμοποιείται για την αποκάλυψη μοτίβων στις τάξεις για τον προσδιορισμό των επιπτώσεων της δομής του προγράμματος σπουδών στη μάθηση, για τη διερεύνηση της σχέσης μεταξύ των διαδοχικών βαθμολογιών των τεστ διαφημίσεων των τάξεων και για τη μεγιστοποίηση της δομής του προγράμματος σπουδών για την εξασφάλιση πιο αποτελεσματικής μάθησης (SPSS, 2003). Στις Η.Π.Α. το σύστημα εξόρυξης δεδομένων προέβλεψε τη δυνατότητα επιστροφής στο σχολείο για κάθε μαθητή που είναι εγγεγραμμένος σε κοινοτικό κολέγιο στη Silicon Valley. Το έργο εφαρμόζει νευρωνικό δίκτυο, CART και C5.0 για την επιλογή της καλύτερης πρόβλεψης που ακολουθείται από ανάλυση ομαδοποίησης (Luan, 2001). Παρόμοια εφαρμογή που αναπτύχθηκε χρησιμοποιείται και στο Πανεπιστήμιο Baylor (Campanelli, 2002).

3.6.6 Μεταφορές

Στη Γαλλία και συγκεκριμένα στο Παρίσι, χρησιμοποιήθηκαν τεχνικές εξόρυξης με σκοπό την πρόβλεψη του είδους μεταφοράς που θα χρησιμοποιούσαν οι πολίτες για να πραγματοποιήσουν συγκεκριμένα ταξίδια. Τα δεδομένα που χρησιμοποιήθηκαν ήταν μια λεπτομερής έρευνα των μέσων μεταφοράς που χρησιμοποιούσε ο πληθυσμός του Ile-de-France, με 400.000 αρχεία. Καθώς τα δεδομένα δεν προορίζονταν αρχικά να χρησιμοποιηθούν για εξόρυξη δεδομένων, έπρεπε να γίνει πολλή εργασία

προεπεξεργασίας. Οι κανόνες δημιουργήθηκαν χρησιμοποιώντας τον αλγόριθμο C4.5 και αποδείχθηκαν ακριβείς καλύπτοντας μεγάλο πληθυσμό. Η δοκιμή με ένα σύνολο δεδομένων επικύρωσης επιβεβαίωσε την ποιότητά τους.

ΜΕΡΟΣ Β Ερευνητική Αποτύπωση

Κεφάλαιο 4

4.1. Εξόρυξη σε Ανοιχτά Δεδομένα

Ανοιχτά Δεδομένα ονομάζονται τα δεδομένα που μπορούν ελεύθερα να χρησιμοποιηθούν, να επαναχρησιμοποιηθούν και να αναδιανεμηθούν από οποιονδήποτε, με την επιφύλαξη απαίτησης απόδοσης και κοινής χρήσης[4]. Τα δεδομένα πρέπει να είναι διαθέσιμα στο σύνολό τους και με τη διαδικασία λήψης μέσω Διαδικτύου (Jackson 2012). Τα δεδομένα πρέπει να είναι διαθέσιμα σε λειτουργική και τροποποιήσιμη μορφή. Επιπρόσθετα, τα δεδομένα πρέπει να παρέχονται με όρους που επιτρέπουν την επαναχρησιμοποίηση και την αναδιανομή, συμπεριλαμβανομένης της προσθήκης και άλλων συνόλων δεδομένων.

Η πρόσβαση στα ανοιχτά δεδομένα πρέπει να είναι καθολική ώστε όλοι να μπορούν να χρησιμοποιούν, να επαναχρησιμοποιούν και να τα αναδιανέμουν και δεν πρέπει να γίνονται διακρίσεις σε ή σε άτομα ή ομάδες. Παραδείγματος χάριν, «μη εμπορικοί» περιορισμοί που θα απέτρεπαν την «εμπορική» χρήση ή περιορισμοί χρήσης για συγκεκριμένους σκοπούς δεν επιτρέπονται (Gavelin, Burall, Wilson, Karin, Simon, Richard, 2009). Η σημαντικότητα σαφήνειας του τι είναι και τι δεν είναι ανοιχτό δεδομένο έχει άμεση σχέση με τον όρο διαλειτουργικότητα. Η διαλειτουργικότητα υποδηλώνει την ικανότητα διαφορετικών συστημάτων και οργανισμών να συνεργάζονται (διαλειτουργούν). Σε αυτή την περίπτωση, είναι η ικανότητα να διαλειτουργούν ή να αναμιγνύονται διαφορετικά σύνολα δεδομένων.

Η διαλειτουργικότητα είναι σημαντική επειδή επιτρέπει σε διαφορετικά στοιχεία να συνεργάζονται. Αυτή η ικανότητα συνιστώσας και «σύνδεσης» στοιχείων είναι απαραίτητη για την κατασκευή μεγάλων, πολύπλοκων συστημάτων. Χωρίς διαλειτουργικότητα αυτό θα ήταν αδύνατο. Ο πυρήνας ενός «κοινού» δεδομένων (ή κώδικα) είναι ότι ένα κομμάτι «ανοιχτού» υλικού που περιέχεται σε αυτό μπορεί να αναμιχθεί ελεύθερα με άλλο «ανοιχτό» υλικό. Αυτή η διαλειτουργικότητα είναι το κλειδί για την πραγματοποίηση των κύριων πρακτικών πλεονεκτημάτων του «ανοιχτού»: τη δραματικά βελτιωμένη ικανότητα συνδυασμού διαφορετικών συνόλων δεδομένων μαζί και ως εκ τούτου για την ανάπτυξη περισσότερων και καλύτερων προϊόντων και

[4] http://opendatahandbook.org/guide/en/what-is-open-data/

υπηρεσιών (αυτά τα οφέλη συζητούνται λεπτομερέστερα στην ενότητα «γιατί» ανοιχτά δεδομένα). Η παροχή ενός σαφούς ορισμού της διαφάνειας διασφαλίζει ότι όταν λαμβάνονται δύο ανοιχτά σύνολα δεδομένων από δύο διαφορετικές πηγές, αυτά θα μπορούν να συνδυαστούν.

4.2. Έρευνα στην πλατφόρμα data.gov.gr

Για την εξόρυξη των δεδομένων επιχειρήθηκε άντλησή τους από την ηλεκτρονική πλατφόρμα δεδομένων του διαδικτυακού τόπου data.gov.gr με σκοπό την ανακάλυψη γνώσης σε έναν τομέα για την υποστήριξη αποφάσεων που θα οδηγήσει στην παραγωγή μιας νέας δημόσιας Πολιτικής, όμως η μορφή των δεδομένων καθώς και η έλλειψη δεδομένων σε συγκεκριμένους τομείς και η ύπαρξη μόνο δεδομένων χρονοσειρών με μικρό βάθος ετών κατέστησε αυτό το εγχείρημα ανεπιτυχές μιας και τα ερευνητικά αποτελέσματα θα περιοριζόταν στην ύπαρξη και μόνο αποτελεσμάτων περιγραφικής στατιστικής χωρίς τη δυνατότητα βαθύτερης ανάλυσης και ανακάλυψης γνώσης από αυτά τα σετ δεδομένων.

4.3. Δημιουργία Βάσεως Δεδομένων

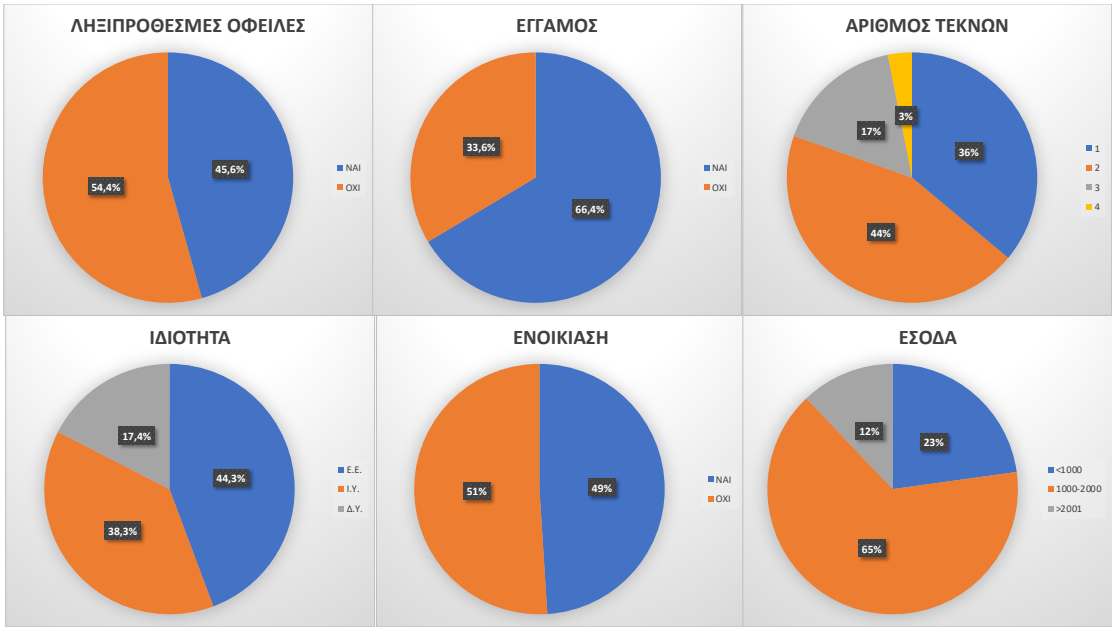
Βάσει αυτών των ερευνητικών προβληματισμών και ερωτημάτων επιχειρήθηκε η δημιουργία βάσεως δεδομένων με σκοπό την εξόρυξη και ανακάλυψης γνώσης μέσω προχωρημένων τεχνικών εξόρυξης. Για το λόγο αυτό δημιουργήθηκε σειρά ερωτημάτων και διανεμήθηκε ερωτηματολόγιο που κατασκευάστηκε στο Google Forms. Κατόπιν αυτού οι φόρμες μεταφέρθηκαν σε υπολογιστικό φύλλο του Microsoft Office 365 και συγκεκριμένα στο Excel με σκοπό των καθαρισμό των δεδομένων για τη μεταφορά τους στο περιβάλλον της πλατφόρμας Rapidminer 9.1 όπου και πραγματοποιήθηκε η εξόρυξη, ανάλυση και οπτικοποίηση των δεδομένων.

Το RapidMiner είναι πλατφόρμα λογισμικού για την επιστήμη των δεδομένων ενσωματώνοντας ένα ολοκληρωμένο περιβάλλον για την προετοιμασία, την εξόρυξη δεδομένων (data mining) τη μηχανική μάθηση (machine learning) , τη βαθιά εκμάθησή (deep learning), την εξόρυξη κειμένου (text mining) και τη δημιουργία προγνωστικών. Συνδυάζει την τεχνολογία και τη δυνατότητα εφαρμογής τεχνικών ώστε να δίνεται η δυνατότητα εξυπηρέτησης και ενσωμάτωσης προς τον τελικό χρήστη των σύγχρονων τεχνικών εξόρυξης δεδομένων. Η δομή των εργασιών στην συγκεκριμένη πλατφόρμα αναπτύσσεται μέσω γραφικού περιβάλλοντος (user interface) που σκοπό έχει τον διαρκές έλεγχο των διεργασιών που βρίσκονται σε εξέλιξη, ενημερώνοντάς τον χρήστη και για

ενδεχόμενα σφάλματα κάνοντας προτάσεις αυτόματα σε περίπτωση προβλημάτων. Η διαδικασία αυτή καθίσταται εφικτή μέσω συγκεκριμένης διαδικασίας (meta-data transformation), με την οποία μετατρέπονται τα βασικά μεταδεδομένα στο στάδιο του σχεδιασμού με τρόπο ώστε η μορφή του αποτελέσματος να μπορεί εξαρχής να προβλεφθεί εντοπίζοντας λύσεις σε περίπτωση ακατάλληλων συνδυασμών δεδομένων. Επιπλέον η πλατφόρμα παρέχει τη δυνατότητα του καθορισμού των σταδίων της διαδικασίας έχοντας τον έλεγχο κάθε ενδιαμέσου αποτελέσματος.

4.4. Στατιστική Ανάλυση δεδομένων

Ακολουθούν τα περιγραφικά στατιστικά για τις μεταβλητές του δείγματος οι οποίες είναι οι ληξιπρόθεσμες οφειλές, η επαγγελματική ιδιότητα, το φύλο, η οικογενειακή κατάσταση, ο αριθμός των τέκνων, η ενοικίαση ή όχι διαμερίσματος και τα μηνιαία έσοδα. Το δείγμα αποτελείται από 298 άτομα εκ των οποίων ληξιπρόθεσμες οφειλές δεν έχουν 162 άτομα (54,4%) ενώ ληξιπρόθεσμες οφειλές έχουν 136 (45,6%). Αναφορικά με την επαγγελματική τους ιδιότητα 132 άτομα (44,3%) είναι ελεύθεροι επαγγελματίες 114 ιδιωτικοί υπάλληλοι (38,3%) και 52 (17,4%) είναι Δημόσιοι Υπάλληλοι. Αναφορικά με το φύλλο 150 άτομα (50,3%) είναι Άνδρες ενώ 148 άτομα (49,7%) είναι γυναίκες. Όσον αφορά την οικογενειακή κατάσταση100 άτομα (33,6%) είναι ανύπαντροι ενώ 198 (66,4% είναι παντρεμένοι). Όσον αφορά την ενοικίαση σπιτιού 152 άτομα 51% δεν ενοικιάζουν σπίτι ενώ 146 άτομα 49% νοικιάζουν σπίτι. Σχετικά με τα έσοδα 68 άτομα (22,8%) έχουν εισόδημα κάτω από 1000 ευρώ, 194 άτομα (65%) έχουν εισόδημα από 1000 έως 2000 ευρώ, ενώ και 36 άτομα (12,1%) έχουν εισόδημα μεγαλύτερο από 2000 ευρώ.



Εικόνα 2: Περιγραφικά στατιστικά του δείγματος

Από τη στατιστική ανάλυση που έγινε και ειδικότερα από τις συσχετίσεις δίνεται ο πίνακας των συσχετίσεων από των οποίο προκύπτουν οι στατιστικώς σημαντικές συσχετίσεις. Υπάρχει στατιστικώς σημαντική συσχέτιση για αρκετές μεταβλητές με τις τιμές του δείκτη που αναφέρονται στον πίνακα. Όσο μεγαλύτερη η τιμή του δείκτη τόσο πιο έντονη η συσχέτιση. Συσχετίσεις με θετικό πρόσημο είναι θετικές συσχετίσεις και αποκαλύπτει συμπεριφορά των τιμών ανάλογη, ενώ συσχετίσεις με αρνητικό πρόσημο είναι αρνητικές συσχετίσεις και αποκαλύπτουν συμπεριφορά των τιμών αντιστρόφως ανάλογη. Στατιστικώς σημαντική αρνητική συσχέτιση έχουν τα έσοδα με τις ληξιπρόθεσμες οφειλές, οι ληξιπρόθεσμες οφειλές με την ενοικίαση καθώς και οι ληξιπρόθεσμες οφειλές με την ηλικία και η κατηγορία παντρεμένος με τον αριθμό τέκνων.

	ΛΗΞΙΠΡΟΘΕΣΜΕΣ ΟΦΕΙΛΕΣ	ΗΛΙΚΙΑ	ΙΔΙΟΤΗΤΑ	ΦΥΛΟ	ΠΑΝΤΡΕΜΕΝΟΣ	ΑΡΙΘΜΟΣ ΤΕΚΝΩΝ	ΕΝΟΙΚΙΑΣΗ	ΕΣΟΔΑ
ΛΗΞΙΠΡΟΘΕΣΜΕΣ ΟΦΕΙΛΕΣ	1							
ΗΛΙΚΙΑ	-0,582	1						
ΙΔΙΟΤΗΤΑ	-0,123	-0,082	1					
ΦΥΛΟ	0,060	-0,034	0,088	1				
ΠΑΝΤΡΕΜΕΝΟΣ	-0,034	0,153	-0,047	0,081	1			
ΑΡΙΘΜΟΣ ΤΕΚΝΩΝ	-0,179	0,302	0,046	0,111	0,761	1		
ΕΝΟΙΚΙΑΣΗ	0,715	-0,198	-0,062	0,127	0,156	0,053	1	
ΕΣΟΔΑ	-0,664	0,250	-0,005	0,045	0,236	0,352	0,089	1

Εικόνα 3: Συσχέτιση των μεταβλητών

Εφαρμόστηκε λογαριθμιστική ανάλυση παλινδρόμησης με εξαρτημένη μεταβλητή την ύπαρξη ληξιπρόθεσμων οφειλών και ανεξάρτητες όλες τις υπόλοιπες προκειμένου να διερευνηθεί αν υπάρχει στατιστικώς σημαντική επίδραση και προβλεπτική ικανότητα των μεταβλητών επί της ύπαρξης ληξιπρόθεσμων οφειλών. Τα αποτελέσματα δίνονται στους πίνακες που ακολουθούν και συγκεκριμένα, από τον πίνακα model summary προκύπτει ότι οι συντελεστές τύπου R^2 όπως ο Cox & Snell R Square και Nagelkerke R Square έχουν τιμή από 0,3 έως 0,42. Οι συντελεστές αυτοί προσπαθούν να μιμηθούν τη συμπεριφορά του δείκτη R^2 της κλασσικής απλής γραμμικής παλινδρόμησης και η ερμηνεία είναι περίπου ανάλογη. Έτσι με βάσει αυτές τις τιμές των δεικτών του πίνακα προκύπτει ότι η προβλεψιμότητα της δημιουργίας ληξιπρόθεσμων οφειλών με βάσει το πακέτο των προβλεπτικών μεταβλητών που επελέγησαν είναι της τάξης του 43% είναι οι υπόλοιπο είναι συστημικοί παράγοντες που δεν έχουν ληφθεί υπόψιν στην παλινδρόμηση ή τυχαία σφάλματα.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	295,520[a]	,321	,429

Εικόνα 4: Τιμές συντελεστών Cox & Snell R Square και Nagelkerke R Square

Ακολουθεί ο πίνακας Variables in the Equation, από τον οποίο δίνονται οι συντελεστές B, τα τυπικά σφάλματα (SE), η τιμή του δείκτη Wald, η στατιστική σημαντικότητα sig και ο δείκτης Odds ratio στη στήλη Exp(B). Αξιολογώντας την έκτη στήλη του πίνακα, τη στήλη sig η οποία είναι η τιμή p value της παλινδρόμησης προκύπτουν ποιες είναι οι μεταβλητές εκείνες που έχουν στατιστικώς σημαντική προβλεψιμότητα των ληξιπρόθεσμων οφειλών. Από τον πίνακα προκύπτει ότι το είδος της εργασίας έχει στατιστικώς σημαντική επίδραση, η μεταβλητή Rental δηλαδή η ενοικίαση έχει στατιστικώς σημαντική επίδραση, και το εισόδημα αλλά μόνο σε κάποιες κατηγορίες ουσιαστικά διαφοροποιείται σημαντικά η κατηγορία εισοδήματος <1000 ευρώ ενώ δε διαφοροποιούνται μεταξύ τους οι υπόλοιπες κατηγορίες δηλαδή είναι σημαντικό αν έχει κάποιος μηνιαίο εισόδημα κάτω από 1000 ευρώ ενώ, αν είναι πάνω από 1000 ευρώ δε δημιουργείται στατιστικώς σημαντική διαφοροποίηση, που σημαίνει πως τα 1000 ευρώ είναι το φράγμα για το αν κάποιος οδηγηθεί σε ληξιπρόθεσμες οφειλές ή όχι.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	,000	,014	,000	1	,982	1,000
	JOB			9,707	2	,008	
	JOB(1)	1,365	,464	8,670	1	,003	3,916
	JOB(2)	1,353	,463	8,550	1	,003	3,870
	GENDER(1)	-,364	,296	1,512	1	,219	,695
	MARITAL(1)	-,382	,493	,599	1	,439	,683
	CHILDREN	-,337	,240	1,959	1	,162	,714
	RENTAL(1)	-1,634	,316	26,811	1	,000	,195
	INCOME			8,623	2	,013	
	INCOME(1)	22,202	6195,658	,000	1	,997	4386468497 ,083
	INCOME(2)	21,190	6195,658	,000	1	,997	1595358591 ,790
	Constant	-20,956	6195,658	,000	1	,997	,000

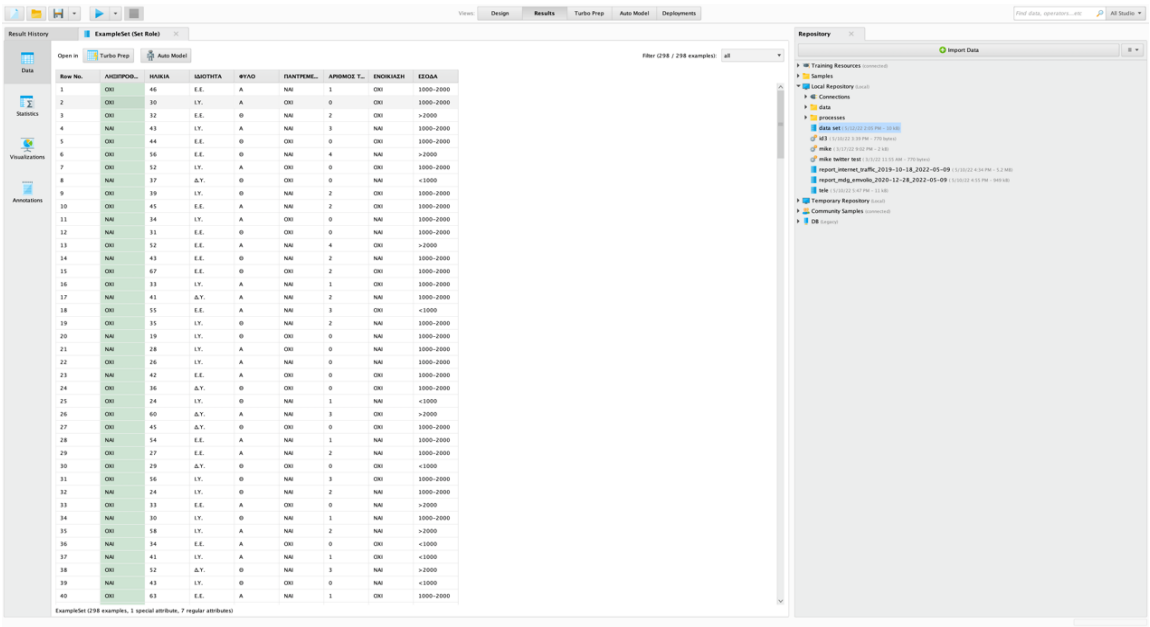
Εικόνα 5: Στατιστική σημαντικότητα μεταβλητών του δείγματος

Το ερευνητικό ερώτημα που οδήγησε στη συλλογή των σχετικών δεδομένων είναι η διερεύνηση ύπαρξης σχέσης μεταξύ των μηνιαίων ληξιπρόθεσμων οφειλών εφόσον υπάρχουν, ενός νοικοκυριού, συναρτήσει του ατομικού εισοδήματος, της επαγγελματικής ιδιότητας, της οικογενειακής κατάστασης και ύπαρξη τέκνων, την ύπαρξη ή όχι ιδιόκτητης οικίας και του ατομικού μηνιαίου εισοδήματος το οποίο θα διερευνηθεί και μέσω αλγορίθμων και τεχνικών εξόρυξης δεδομένων.

4.5. Ορισμός Κλάσης σε ένα πεδίο της βάσης δεδομένων

Όπως έχει αναφερθεί κλάση ορίζεται ένα πεδίο του σετ δεδομένων πάνω στο οποίο απαντούνται οι ερωτήσεις και είναι απαραίτητο να οριστεί πριν εφαρμοστεί οποιοδήποτε μοντέλο. Σε αντίθετη περίπτωση θα εμφανιστεί συντακτικό λάθος. Στη βάση δεδομένων ορίζουμε ως κλάση τις Ληξιπρόθεσμες Οφειλές με την ακόλουθη διαδικασία. Στο πεδίο design εισάγεται η βάση δεδομένων και από το πεδίο operator εισάγεται ο καθορισμός ρόλου (set role) που στη συγκεκριμένη μελέτη τίθενται οι Ληξιπρόθεσμες οφειλές και διασυνδέοντας το παράδειγμα (example) με την έξοδο, εκκινώντας το συγκεκριμένο κύκλωμα.

Το σετ δεδομένων παραμένει ανεπηρέαστο και η μόνη αλλαγή είναι πως στο σετ έχει τεθεί η στήλη Ληξιπρόθεσμες οφειλές ως κλάση.



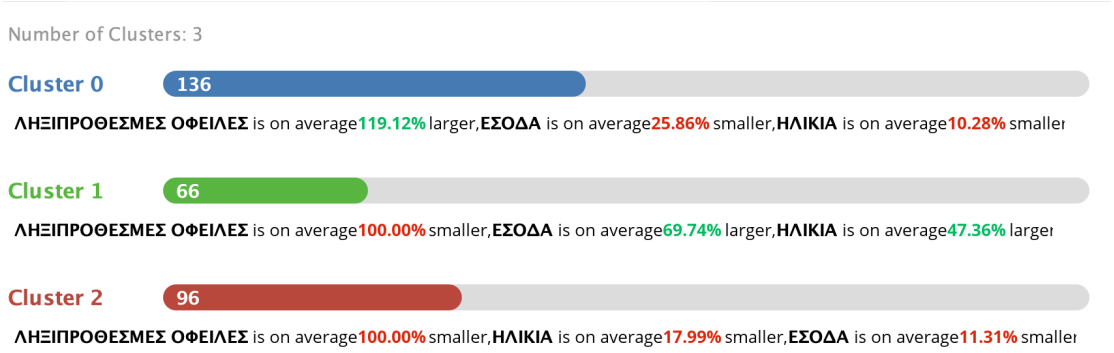
Εικόνα 6: Ορισμός Κλάσης

4.6. Αλγόριθμος k-means.

Στη βάση δεδομένων εφαρμόστηκε ο αλγόριθμος k-means που ανήκει στην κατηγορία της επίπεδης συσταδοποίησης με σκοπό να παραχθεί ένα σύνολο συσταδοποιήσεων. Η ομαδοποίηση k-means είναι μια μέθοδος που στοχεύει να χωρίσει n παρατηρήσεις σε k αριθμό συστάδων στα οποία κάθε παρατήρηση ανήκει στη συστάδα με τον πλησιέστερο μέσο όρο (κέντρα συστάδων ή κέντρο συστάδων), που χρησιμεύει ως πρωτότυπο. Αυτό έχει ως αποτέλεσμα την κατάτμηση του χώρου δεδομένων σε κελιά

Voronoi[5]. Η ομαδοποίηση k-means ελαχιστοποιεί τις διακυμάνσεις εντός της συστάδας (τετράγωνες Ευκλείδειες αποστάσεις). Ο μέσος όρος βελτιστοποιεί τα τετραγωνικά σφάλματα, ενώ μόνο η γεωμετρική διάμεσος ελαχιστοποιεί τις Ευκλείδειες αποστάσεις.

Σε αυτή τη μέθοδο, τα σημεία δεδομένων εκχωρούνται σε συστάδες με τέτοιο τρόπο ώστε το άθροισμα των τετραγωνικών αποστάσεων μεταξύ των σημείων δεδομένων και του κέντρου να είναι όσο το δυνατόν μικρότερο. Είναι σημαντικό να σημειωθεί ότι η μειωμένη ποικιλομορφία εντός των συστάδων οδηγεί σε περισσότερα πανομοιότυπα σημεία δεδομένων εντός της ίδιας συστάδας. Από το σύνολο του δείγματος δημιουργήθηκαν τρεις κλάσεις όπου η κλάση 0 περιλαμβάνει 136 σημεία δεδομένων, η κλάση 1 περιλαμβάνει 66 σημεία δεδομένων και η κλάση 2 , 96 άτομα σε σύνολο δείγματος 298 ατόμων.



Εικόνα 7: Δημιουργία Κλάσεων

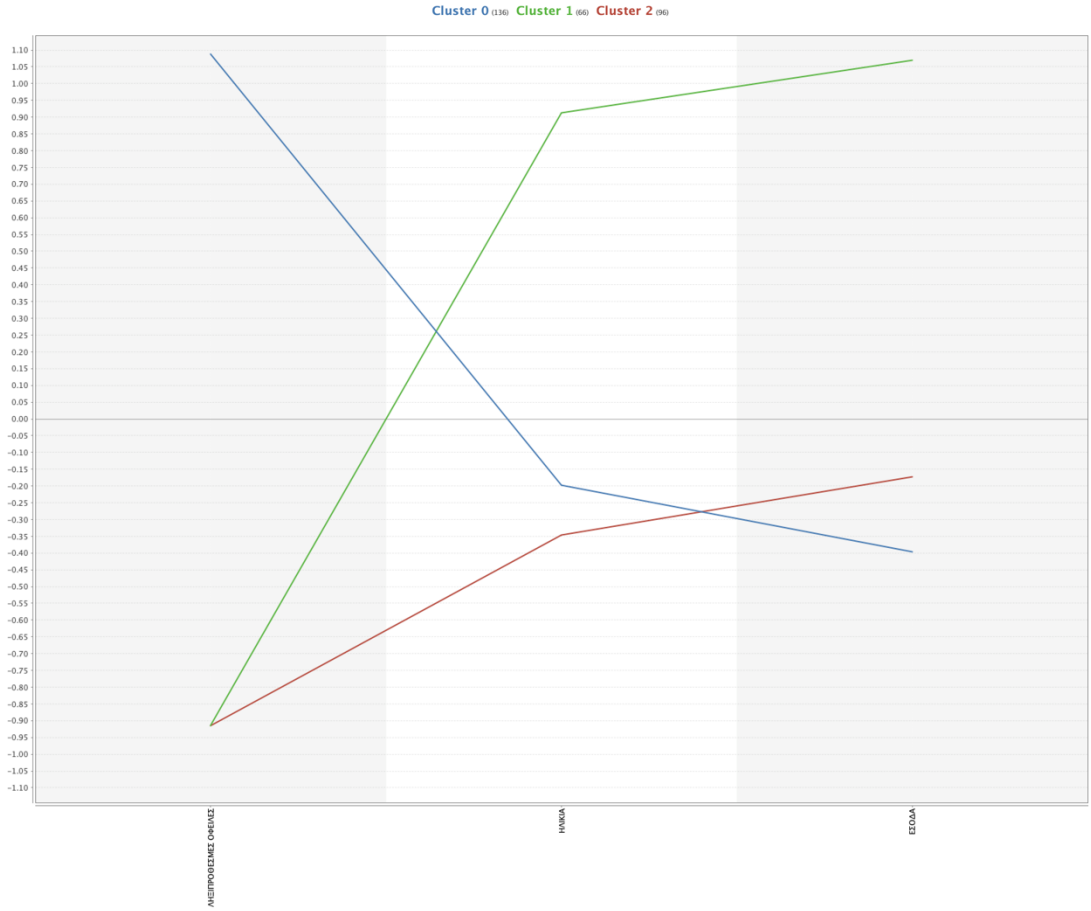
Στον πίνακα κεντροειδών (centroids table) παρατηρείται πως η κλάση 0 έχει σημαντική απόσταση από τις κλάσεις 1 και 2 σε ότι αφορά την κατηγορία των ληξιπρόθεσμων οφειλών. Αντιθέτως στην κατηγορία Ηλικία, η κλάση ένα διατηρεί τη μεγαλύτερη απόσταση από τα κέντρα της κλάσης 0 και 2. Στην κατηγορία Έσοδα η κλάση ένα με 1.070 διατηρεί μεγαλύτερη απόσταση από τις κλάσεις 0 και 2 οι οποίες παρουσιάζουν σχετικά κοντά κέντρα.

Cluster	ΛΗΞΙΠΡΟΘΕΣΜΕΣ ΟΦΕΙΛΕΣ	ΗΛΙΚΙΑ	ΕΣΟΔΑ
Cluster 0	1.090	-0.198	-0.397
Cluster 1	-0.915	0.913	1.070
Cluster 2	-0.915	-0.347	-0.173

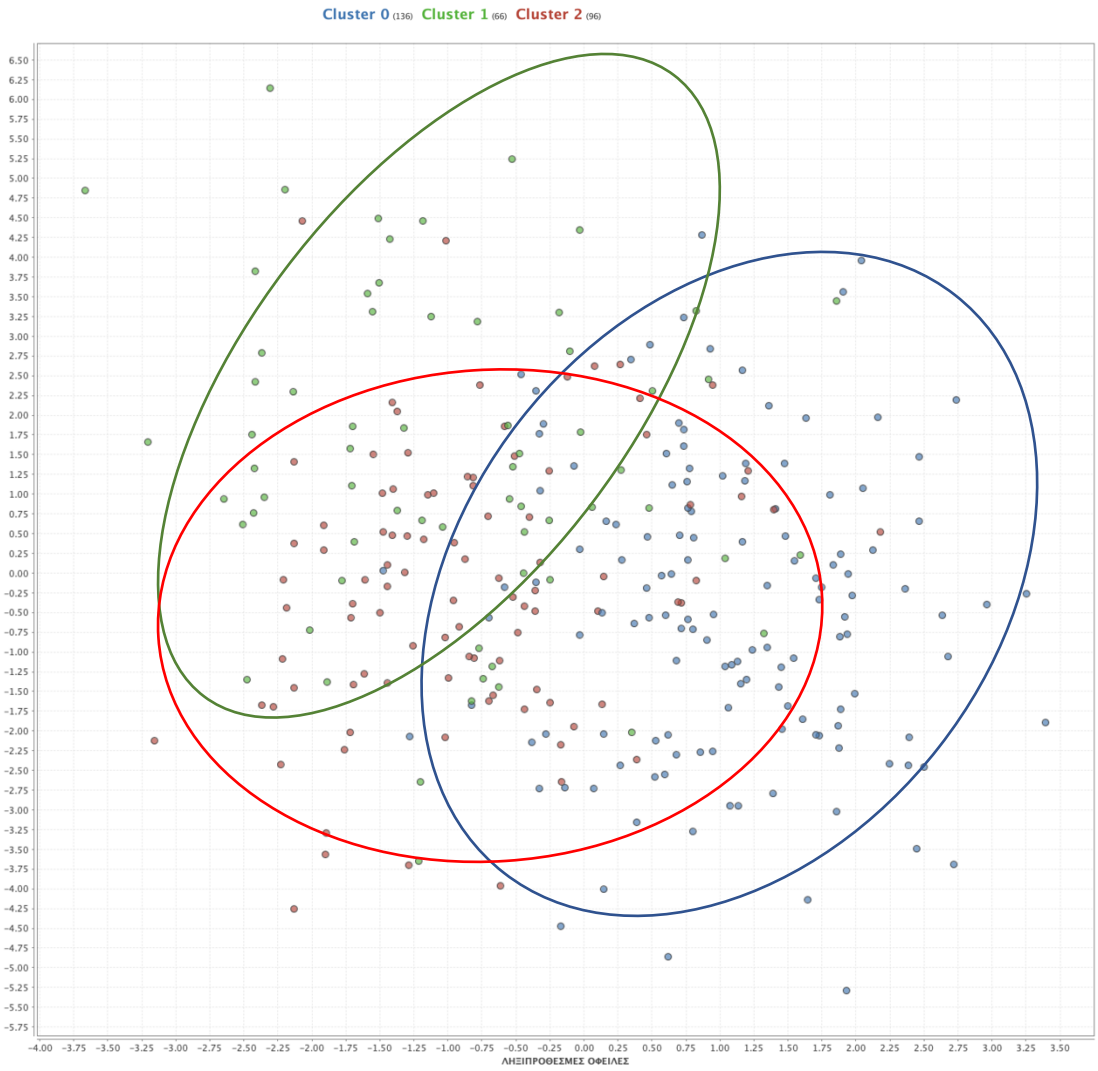
Εικόνα 8: Αποστάσεις μεταξύ των κέντρων των κλάσεων (centroids)

5 Στη μαθηματική επιστήμη, ένα διάγραμμα Voronoi είναι ένα διαχωρισμός ενός επιπέδου σε περιοχές που βασίζονται στην απόσταση από τα σημεία ενός συγκεκριμένου υποσυνόλου του επιπέδου.

Στο γράφημα αυτές οι διαφορές οπτικοποιούνται αναλυτικότερα και παρουσιάζονται οι μεταβολές των κέντρων των κλάσεων σε σχέση με τις Ληξιπρόθεσμες Οφειλές, την Ηλικία και τα Έσοδα .



Εικόνα 9: Γραφική παράσταση των κέντρων των κλάσεων



Εικόνα 10: Διασπορά του δείγματος και αντιστοιχία σε κλάσεις

4.7. Εκπαίδευση δένδρου απόφασης και πρόβλεψη

Για την εκπαίδευση ενός δέντρου απόφασης πρώτα ενσωματώνονται τα δεδομένα στο σχεδιαστικό τμήμα της πλατφόρμας. Η εκπαίδευση πραγματοποιείται στα δεδομένα που χρησιμοποιήθηκαν για να εκπαιδευτεί το μοντέλο. Επομένως εφόσον τα δεδομένα όπως είναι ήδη γνωστό περιέχονται στην κλάση Ληξιπρόθεσμες Οφειλές γνωρίζουμε εκ των προτέρων τις απαντήσεις δηλαδή είναι ήδη γνωστές οι σωστές τιμές στην κλάση. Συνεπώς δίνεται ήδη ένα τεστ με απαντήσεις από τις απαντήσεις που ήδη έχουν δοθεί, με σκοπό να ελεγχθεί πόσο ακριβές είναι το μοντέλο που δημιουργήθηκε.

Για την επιλογή των συγκεκριμένων πεδίων για απαιτούνται για τη δημιουργία ερώτησης, χρησιμοποιείται ο διαχειριστής χαρακτηριστικού (operator select attributes) και επιλέγεται από τις παραμέτρους τη επιλογή υποσύνολο. Επειδή το πεδίο ληξιπρόθεσμες οφειλές στην τελευταία εγγραφή που προστέθηκε και αποτελεί ερώτηση

αναφοράς, είναι κενό θα πρέπει να χωριστεί το δείγμα της βάσης δεδομένων στα δύο, σε ένα τμήμα όπου όλα τα πεδία έχουν συμπληρωμένες όλες τις εγγραφές και στο δεύτερο τμήμα όπου υπάρχει κενή εγγραφή, αυτή η εγγραφή που ζητείται πρόβλεψη από το μοντέλο. Για την υλοποίηση της συγκεκριμένης διαδικασίας χρησιμοποιείται ο operator multiply για την κατασκευή αντιγράφων σε συνεργασία με το filter examples. Χρησιμοποιώντας τη μέθοδο του δένδρου απόφασης και εφαρμόζοντας το μοντέλο μέσω του διαχειριστή (operator), εφαρμογή μοντέλου (apply model) και εκπαιδεύω και εφαρμόζω το μοντέλο.

Για την κατασκευή μοντέλου ενός δένδρου απόφασης με κλάση τις Ληξιπρόθεσμες Οφειλές με σκοπό τη δημιουργία ερωτήσεων σχετικά με χαρακτηριστικά που δόθηκαν από εμάς των άλλων πεδίων και μόνο μη συμπληρωμένο πεδίο τις Ληξιπρόθεσμες Οφειλές όπου και την απάντηση θα δώσει το μοντέλο και θα υπολογιστεί και η σχετική ακρίβεια της απάντησης. Για τον σκοπό αυτόν εισάγεται η βάση δεδομένων στο σχεδιαστικό τμήμα της πλατφόρμας προσθέτοντας στον data editor της πλατφόρμας μια εγγραφή όπου συμπληρώνονται τα χαρακτηριστικά που θα μελετηθούν, αφήνοντας κενό το πεδίο των ληξιπρόθεσμων οφειλών όπου και ζητάται από το μοντέλο απάντηση. Στην έρευνα μας θεωρούμε ως δεδομένα τα πεδία με τα δεδομένα που θέλουμε να τροφοδοτήσουμε το μοντέλο τα οποία είναι :Άνδρας, ηλικίας 43 ετών με 2 τέκνα όπου δεν είναι ενοικιαστής στον τόπο διαμονής του και τα έσοδα του είναι 1000-2000 ευρώ.

Για να επιλεχθούν τα συγκεκριμένα πεδία που χρειάζονται για την ερώτηση, χρησιμοποιώ τον operator select attributes και στη συνέχεια επιλέγονται από τις παραμέτρους η επιλογή υποσύνολο. Επειδή το πεδίο ληξιπρόθεσμες οφειλές στην τελευταία εγγραφή που προστέθηκε είναι κενό, θα πρέπει να χωριστεί το δείγμα της βάσης δεδομένων σε δύο τμήματα, ένα τμήμα όπου όλα τα πεδία έχουν συμπληρωμένες όλες τις εγγραφές και στο δεύτερο τμήμα όπου υπάρχει κενή εγγραφή, αυτή η εγγραφή που ζητώ πρόβλεψη από το μοντέλο. Για την υλοποίηση της συγκεκριμένης διαδικασίας χρησιμοποιείται ο τελεστής multiply για την κατασκευή ακριβών αντιγράφων και σε συνεργασία με το filter examples. Χρησιμοποιώντας τη μέθοδο του δένδρου απόφασης και εφαρμόζοντας το μοντέλο μέσω του διαχειριστή (operator) εφαρμογή μοντέλου (apply model) εκπαιδεύω και εφαρμόζω το μοντέλο.

Η απάντηση στην ερώτηση που τέθηκε στο μοντέλο, ήταν αν Άνδρας, ηλικίας 43 ετών με 2 τέκνα όπου δεν είναι ενοικιαστής στον τόπο διαμονής του και τα έσοδα του είναι 1000-2000 ευρώ ποια η πιθανότητα ύπαρξης ληξιπρόθεσμων οφειλών. Η απάντηση

είναι πώς δεν θα έχει Ληξιπρόθεσμες Οφειλές με πιθανότητα επαλήθευσης του αλγορίθμου 0.938 ή 93,8 %.

Row No.	ΛΗΞΙΠΡΟΘ...	prediction(...	confidence(OXI)	confidence(NAI)	ΗΛΙΚΙΑ	ΦΥΛΟ	ΑΡΙΘΜΟΣ Τ...	ΕΝΟΙΚΙΑΣΗ	ΕΣΟΔΑ
1	?	OXI	0.938	0.062	43	A	2	OXI	1000-2000

Εικόνα 11: Ακρίβεια επαλήθευσης του μοντέλου

4.9. Cross Validation

Σκοπός είναι η εκπαίδευση του μοντέλου μέσω της μεθόδου cross validation μιας ειδικής τεχνικής που χωρίζει το δείγμα σε υποσύνολα και χρησιμοποιεί ένα μέρος των υποσυνόλων για εκπαίδευση και ένα άλλο μέρος υποσυνόλου του δείγματος χρησιμοποιείται για την πραγματοποίηση των ερωτήσεων. Η διασταυρούμενη επικύρωση ή εκτίμηση περιστροφής, είναι μια μέθοδος επαναδειγματοληψίας που χρησιμοποιεί διαφορετικά τμήματα των δεδομένων για να δοκιμάσει και να εκπαιδεύσει ένα μοντέλο σε διαφορετικές επαναλήψεις. Χρησιμοποιείται κυρίως σε περιβάλλοντα όπου ο στόχος είναι η πρόβλεψη και κάποιος θέλει να εκτιμήσει με πόση ακρίβεια θα αποδώσει ένα μοντέλο πρόβλεψης στην πράξη. Όταν πραγματοποιείται μια τέτοιου είδους κατάτμηση του δείγματος, οι λαμβανόμενες μετρήσεις για την ακρίβεια της μεθόδου που χρησιμοποιούμε είναι επαρκείς. Πρώτα εισάγουμε το σύνολο δεδομένων στο σχεδιαστικό τμήμα της πλατφόρμας, τα δεδομένα αυτά είναι πλήρη χωρίς κενές εγγραφές, και στη συνέχεια εισάγοντας το select attributes για την επιλογή των επιθυμητών πεδίων. Το cross validation είναι ο τελεστής που θα εφαρμόσει αυτόματα τη διαδικασία. Στο cross validation περιλαμβάνεται μια υποδιαδικασία μέσα στη συνολική διαδικασία, όπου στην αριστερή πλευρά βρίσκεται ο έλεγχος της διαδικασίας και δεξιά η εκπαίδευση. Θέτουμε το δέντρο απόφασης στη διαδικασία ελέγχου και στην διαδικασία εκπαίδευσης θέτουμε τον τελεστή apply model πραγματοποιώντας τις αντίστοιχες συνδέσεις, (ενώνοντας στον τελεστή apply model την είσοδο testing στο unl) ώστε η διαδικασία να χωρίσει το δείγμα σε υπό δείγματα και να πραγματοποιηθούν αρκετές φορές εκπαίδευση και ερωτήσεις με τον τελεστή performance ώστε να μετρηθεί η ακρίβεια του μοντέλου. Θέτοντας στις παραμέτρους του cross validation των αριθμό των τμημάτων που θα «κοπεί» το αρχικό δείγμα έτσι ώστε το ένα τμήμα να χρησιμοποιηθεί για ερώτηση και τα υπόλοιπα εννέα για την εκπαίδευση του μοντέλου. Εν συνεχεία ένα άλλο τμήμα από τα δέκα θα χρησιμοποιηθεί για ερώτηση και τα άλλα εννιά για την

εκπαίδευση του μοντέλου κ.ο.κ.. Οι κύκλοι αυτοί θα συνεχίσουν μέχρις ότου εξαντληθούν και τα δέκα κομμάτια.

accuracy: 90.60%

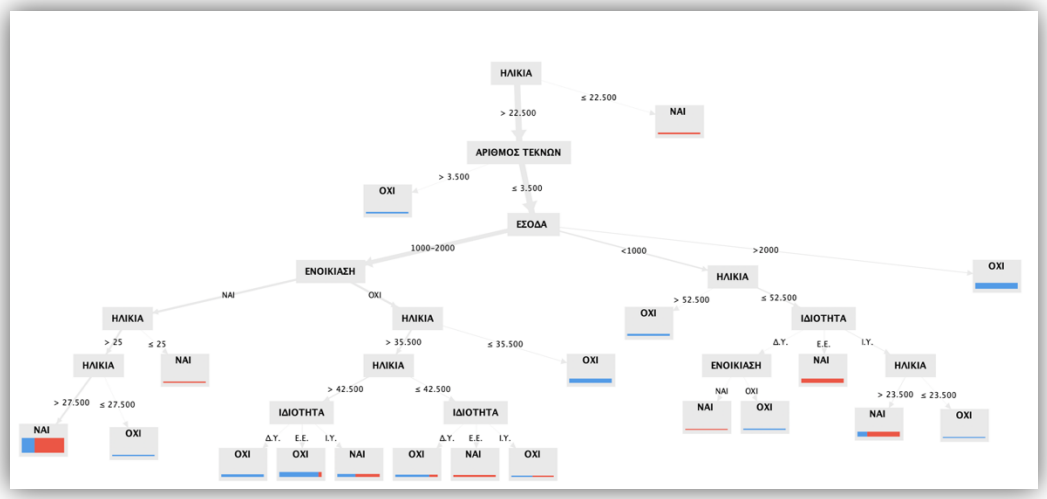
	true OXI	true NAI	class precision
pred. OXI	136	2	98.55%
pred. NAI	26	134	83.75%
class recall	83.95%	98.53%	

Εικόνα 12: Ακρίβεια του μοντέλου μετά από εκπαίδευση

Από τα αποτελέσματα παρατηρείται πως η ακρίβεια είναι 90.6% κάτι που σημαίνει πως το μοντέλο επαληθεύεται σε περισσότερες από τα 9/10 των απαντήσεων του δείγματος. Η ακρίβεια των αποτελεσμάτων στις περιπτώσεις όπου το μοντέλο προέβλεψε όχι στις Ληξιπρόθεσμες Οφειλές είναι 83,95 % ενώ οι περιπτώσεις όπου το μοντέλο πρόβλεψε NAI είναι 98,53%.

4.10 Δένδρο Απόφασης (Decision Tree)

Για την κατασκευή και οπτικοποίηση του δένδρου απόφασης παρεμφερής διαδικασία με τη διαφορά πως κατόπιν της εισαγωγής της βάσης δεδομένων και τον καθορισμό ρόλου στο σχεδιαστικό τμήμα της πλατφόρμας, Διαχωρίζουμε το σύνολο δεδομένων μέσω του διαχωρισμού των δεδομένων (split data) όπου το 70% θα χρησιμοποιηθεί για την κατασκευή του μοντέλου, ενώ το 30% θα χρησιμοποιηθεί για τον έλεγχο του μοντέλου. Ακολουθούν οι συνδέσεις προς τον operator του Δένδρου Απόφασης και προς την εφαρμογή μοντέλου. Στη συνέχεια πραγματοποιείται η σύνδεση και με τον τελεστή ελέγχου της επίδοσης του μοντέλου προς την έξοδο αλλά και με την εφαρμογή μοντέλου προς την έξοδο με σκοπό την οπτικοποίηση του Δένδρου Απόφασης. Η οπτικοποίηση του δένδρου απόφασης είναι η εξής:



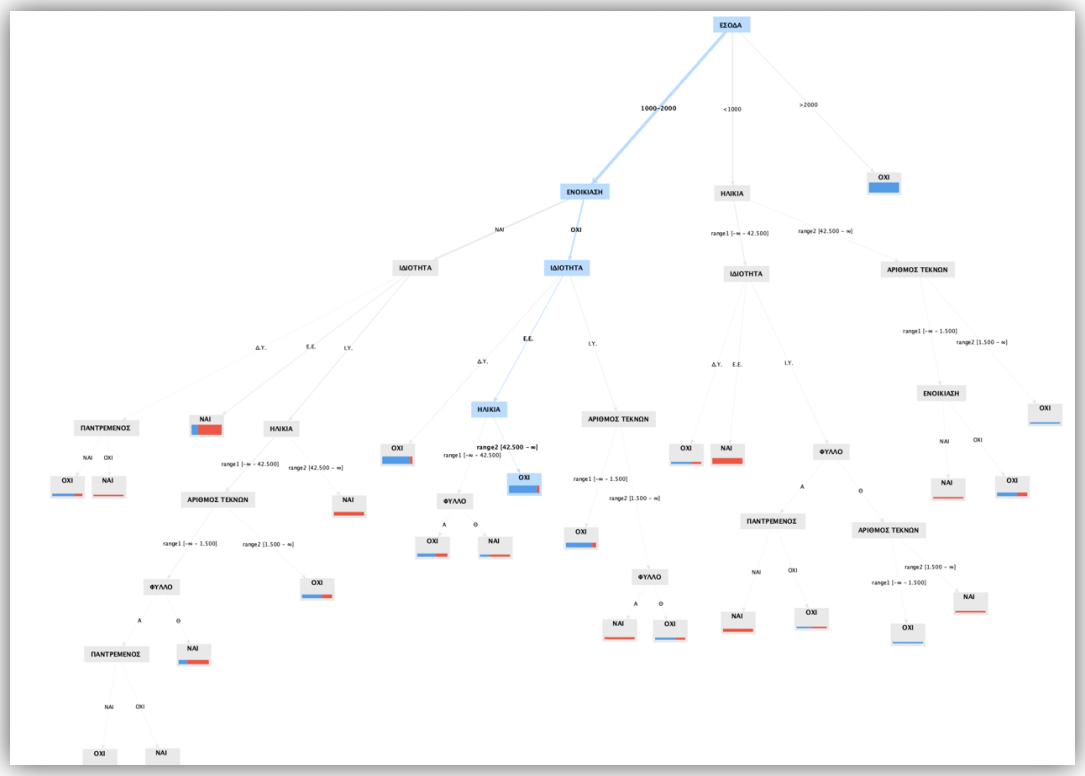
Εικόνα 13: Οπτικοποίηση του δένδρου απόφασης

Παρατηρείται πως η ηλικία συναρτήσει των Ληξιπρόθεσμων Οφειλών είναι καθοριστικός παράγοντας για τη δημιουργία των πρώτων κλάδων. Στην περίπτωση που η Ηλικία είναι μικρότερη ή ίση με τα 22 έτη, τότε σίγουρα υπάρχουν Ληξιπρόθεσμες οφειλές (4% του συνόλου του δείγματος). Εφόσον ηλικία είναι μεγαλύτερη από τα 22,5 έτη ελέγχουμε τον αριθμό των τέκνων. Παρατηρείται πως δημιουργούνται δύο νέα κλαδιά του δέντρου όπου όταν ο αριθμός των τέκνων είναι μεγαλύτερος από το 3,5 δεν υπάρχουν ληξιπρόθεσμες οφειλές. Σε αντίθετη περίπτωση εφόσον ο αριθμός των τέκνων είναι μικρότερος του 3,5 τότε δημιουργείται νέος κόμβος και ελέγχονται τα έσοδα. Στην περίπτωση αυτή όταν τα έσοδα είναι μεταξύ των 1000 και 2000 € δημιουργείται ένα νέο κλαδί του δέντρου όπου ελέγχεται ξανά η ηλικία. Επίσης, όταν τα έσοδα είναι μικρότερα των 1000 € πάλι ελέγχεται ηλικία, όμως όταν τα έσοδα είναι μεγαλύτερα των 2000 € δεν υπάρχουν ληξιπρόθεσμες οφειλές και αυτό αναφέρεται στο 11,54% του συνόλου του δείγματος δηλαδή σε 24 άτομα του σετ δεδομένων. Εφόσον τα έσοδα είναι μεταξύ των 1000 και 2000 € ελέγχεται ξανά ηλικία αυτή τη φορά με διαφορετικά κλαδιά τα 25 έτη. Στην περίπτωση που ηλικία είναι μικρότερη από 25 έτη υπάρχουν σίγουρα ληξιπρόθεσμες οφειλές ενώ σε περίπτωση ηλικία είναι μεγαλύτερη των 25 έτη δημιουργείται ένας νέος κλάδος όπου υπάρχουν ληξιπρόθεσμες οφειλές για 39 άτομα του συνόλου δεδομένων, ενώ όταν ηλικία είναι μικρότερη των 27 ετών δεν υπάρχουν ληξιπρόθεσμες οφειλές. Αντίστοιχα για ηλικία μεγαλύτερη των 52 ετών δεν υπάρχουν ληξιπρόθεσμες οφειλές ενώ για ηλικία μικρότερη των 52 ετών ελέγχεται η επαγγελματική ιδιότητα. Στην περίπτωση που επαγγελματική ιδιότητα είναι δημόσιος υπάλληλος δεν υπάρχουν ληξιπρόθεσμες οφειλές στο μεγαλύτερο μέρος του δείγματος, ενώ αντιθέτως

όταν είναι ελεύθερος επαγγελματίας υπάρχουν ληξιπρόθεσμες οφειλές για το 8,17% του συνόλου του δείγματος. Αντιθέτως όταν η ιδιότητα είναι ιδιωτικός υπάλληλος ελέγχεται ξανά ηλικία όπου όταν είναι μικρότερη των 23,5 ετών δεν υπάρχουν ληξιπρόθεσμες οφειλές αντιθέτως όταν ηλικία είναι μεγαλύτερη από 23,5 έτη εξετάζουμε το φίλο όπου σε περίπτωση που είναι άντρας υπάρχουν σίγουρα ληξιπρόθεσμες οφειλές ενώ σε περίπτωση που είναι γυναίκα επανεξετάζουμε την ηλικία.

4.11 Αυτόματος ανιχνευτής αλληλεπίδρασης (CHAID)

Ο αυτόματος ανιχνευτής αλληλεπίδρασης Chi-square automatic interaction detection (CHAID) είναι μια τεχνική που χρησιμοποιείται για την ανακάλυψη της σχέσης μεταξύ μεταβλητών. Η ανάλυση CHAID δημιουργεί ένα προγνωστικό μοντέλο, ή δέντρο, για να βοηθήσει στον προσδιορισμό του τρόπου με τον οποίο οι μεταβλητές συγχωνεύονται καλύτερα για να εξηγήσουν το αποτέλεσμα στη δεδομένη εξαρτημένη μεταβλητή. Στην ανάλυση CHAID, μπορούν να χρησιμοποιηθούν ονομαστικά, τακτικά και συνεχή δεδομένα, όπου οι συνεχείς προβλέψεις χωρίζονται σε κατηγορίες με περίπου ίσο αριθμό παρατηρήσεων. Ο CHAID δημιουργεί όλους τους πιθανούς διασταυρούμενους πίνακες για κάθε κατηγορηματικό προγνωστικό παράγοντα έως ότου επιτευχθεί το καλύτερο αποτέλεσμα και δεν μπορεί να πραγματοποιηθεί περαιτέρω διαχωρισμός. Στην τεχνική CHAID, μπορούμε να δούμε οπτικά τις σχέσεις μεταξύ των μεταβλητών διαχωρισμού και του σχετικού σχετικού παράγοντα μέσα στο δέντρο.



Εικόνα 14: Οπτικοποίηση CHAID

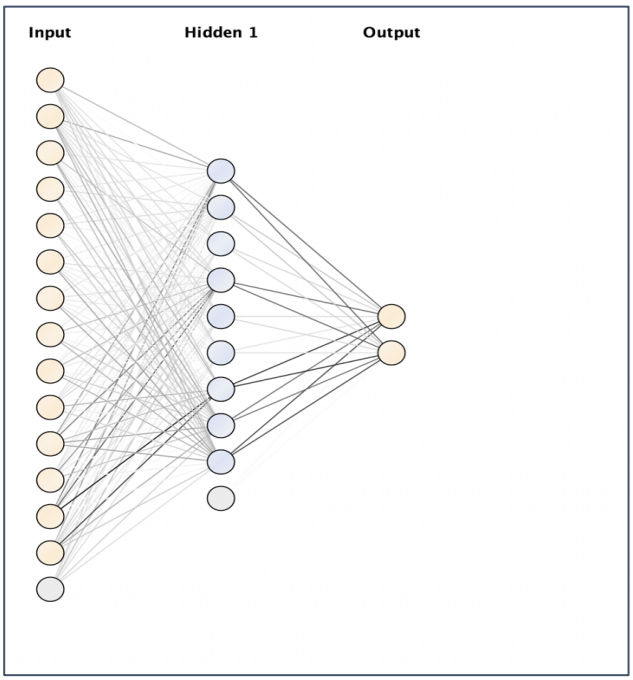
Η ανάπτυξη της απόφασης, ή του δέντρου ταξινόμησης, ξεκινά με τον προσδιορισμό της μεταβλητής στόχου ή της εξαρτημένης μεταβλητής. που θα θεωρούνταν η ρίζα. Η ανάλυση CHAID χωρίζει τον στόχο σε δύο ή περισσότερες κατηγορίες που ονομάζονται αρχικοί ή μητρικοί κόμβοι και στη συνέχεια οι κόμβοι χωρίζονται χρησιμοποιώντας στατιστικούς αλγόριθμους σε θυγατρικούς κόμβους. Σε αντίθεση με την ανάλυση παλινδρόμησης, η τεχνική CHAID δεν απαιτεί κανονική διανομή των δεδομένων.

Πρώτος παράγοντάς για τον αυτόματό ανιχνευτή αλληλεπίδρασης αποτελούν τα έσοδα όπου στην κατηγορία 1000 έως 2000 € καθοριστικό ρόλο έχει η ενοικίαση, στην κατηγορία μικρότερης των 1000 € η ηλικία, ενώ στην κατηγορία μεγαλύτερη των 2000 € δεν υπάρχουν ληξιπρόθεσμες οφειλές όπως και στο δέντρο απόφασης. Σε περίπτωση που υπάρχει ή και δεν υπάρχει ενοικίαση εξετάζεται η ιδιότητα. Σε περίπτωση που ηλικία είναι μέχρι 42 έτη εξετάζεται η ιδιότητα ενώ σε περίπτωση που η ηλικία είναι μεγαλύτερη των 42 ετών εξετάζεται ο αριθμός των τέκνων. Όταν η ιδιότητα είναι Δημόσιος Υπάλληλος εξετάζεται αν είναι έγγαμος/η ή όχι, ενώ όταν η επαγγελματική ιδιότητα είναι ελεύθερος επαγγελματίας σε ποσοστό 85% του υπάρχουν ληξιπρόθεσμες οφειλές ενώ στην περίπτωση που είναι ιδιωτικός υπάλληλος εξετάζουμε την ηλικία. Το ίδιο ισχύει και

για την κατηγορία ενοικιάσεις όπου εξετάζεται ιδιότητα όπου όταν η επαγγελματική ιδιότητα είναι δημόσιος υπάλληλος τότε δεν υπάρχουν ληξιπρόθεσμες οφειλές ενώ για επαγγελματική ιδιότητα ελεύθερου επαγγελματία, επανεξετάζεται η ηλικία ενώ όταν στην περίπτωση ιδιωτικού υπάλληλου εξετάζεται ο αριθμός των τέκνων. Παρατηρούμε πως ο αυτόματος ανιχνευτής αλληλεπιδράσεις δημιουργεί περισσότερα κλαδιά και ομάδες ταξινόμησης σε σχέση με το δέντρο απόφασης με αποτέλεσμα να υπάρχει μεγαλύτερη διάσπαση των ομάδων που δημιουργούνται για την ύπαρξη ή μη ληξιπρόθεσμων οφειλών.

4.12 Νευρωνικό Δίκτυο (Neural Network)

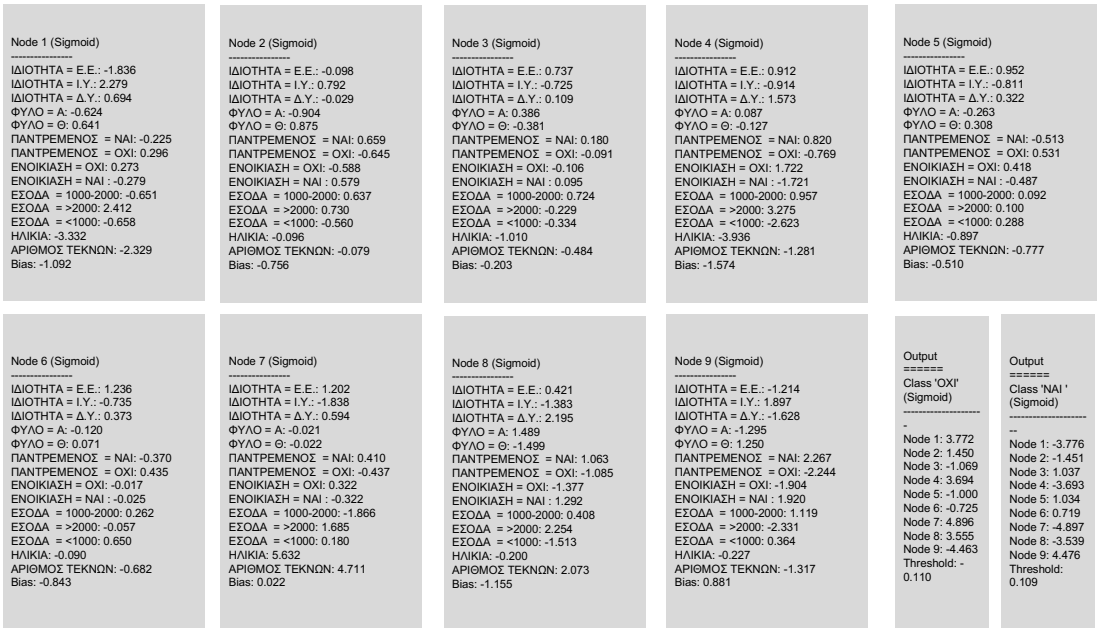
Δημιουργήθηκε νευρωνικό δίκτυο με επίβλεψη, με δεκατέσσερις εισόδους (input) που αναπαριστούν τις κατηγορίες με τις απαντήσεις από τη βάση δεδομένων, ένα κρυφό επίπεδο εννέα βρόγχων, και δύο έξοδοι όπου έχει τεθεί η ύπαρξη ληξιπρόθεσμων ή μη οφειλών. Οι είσοδοι του δικτύου αποτελούν το πρωταρχικό σήμα εισόδου και κάθε είσοδος πολλαπλασιάζεται με το συναπτικό βάρος που αντιστοιχεί. Στη συνέχεια τα αποτελέσματα αθροίζονται και προστέθηκε και ένας εξωτερικός παράγοντας (πόλωση).



Εικόνα 15: Δημιουργία ΤΝΔ πρόσθιας τροφοδότησης

Το βασικό μοντέλο τεχνητού νευρώνα περιλαμβάνει ένα σύνολο προσαρμοζόμενων παραμέτρων, που ονομάζονται βάρη, όπως στη γραμμική και τη λογιστική παλινδρόμηση. Όπως ακριβώς και στην παλινδρόμηση, τα βάρη αυτά χρησιμοποιούνται ως πολλαπλασιαστές των εισόδων του νευρώνα, και αθροίζονται. Το

άθροισμα των βαρών επί τις εισόδους ονομάζεται γραμμικός συνδυασμός των εισόδων. Τα βάρη στα νευρωνικά δίκτυα είναι ο πιο καθοριστικός παράγοντας για τη λειτουργία τους που σε συνεργασία με τη διαδικασία εκπαίδευσης που ακολουθήθηκε οδηγηθήκαμε στον προσδιορισμό των βαρών.



Εικόνα 16: Αποτελέσματα βαρών

Στον πίνακα αναφέρονται τα βάρη κάθε νευρώνα βάσει των τιμών εισόδου καθώς και τα βάρη των τιμών εξόδου καθορίζοντας διακόσιους εκπαιδευτικούς κύκλους. Βάσει των βαρών τεκμαίρεται πως για την μη ύπαρξη ληξιπρόθεσμων οφειλών επίδραση με φθίνουσα σειρά βαρών έχουν οι βρόγχοι 1, 7 και 9 που κυριαρχούν τα βάρη της ηλικίας, του εισοδήματος, της επαγγελματικής ιδιότητας και του αριθμού τέκνων και του όπου τα μεγαλύτερα από 2000 ευρώ έσοδα οδηγούν στη μη ύπαρξη ληξιπρόθεσμων οφειλών. Αντιθέτως για την ύπαρξη ληξιπρόθεσμων οφειλών οι βρόγχοι 7, 8 και 9 με κύρια χαρακτηριστικά το εισόδημα μικρότερο των 1000 ευρώ, η ενοικίαση κατοικίας και η οικογενειακή κατάσταση.

4.13 Συμπεράσματα.

Στην παρούσα εργασία πραγματοποιήθηκε διεξοδική αναφορά της διαδικασίας της εξόρυξης δεδομένων και των τεχνικών εφαρμογής της με σκοπό την υποστήριξη λήψης αποφάσεων στη Δημόσια Διοίκηση. Ο πλούτος πληροφοριών που είναι κρυμμένος στα Πληροφοριακά Συστήματα και τις βάσεις δεδομένων με τις κατάλληλες τεχνικές εξόρυξης και ανακάλυψης γνώσης, αποτελεί απαραίτητο και πολύτιμο βοήθημα όταν οι αποφάσεις που η δημόσια διοίκηση οφείλει να προτείνει πρέπει να είναι σε βάθος και

στοχευμένες. Η Δημόσια Διοίκηση για την αξιοποίηση των Δημοσίων Πολιτικών οφείλει πέραν της εφαρμογής των διοικητικών διαδικασιών να είναι σε θέση να γνωρίζει, να αναλύει και να προτείνει κατευθύνσεις βασισμένες στην «κρυφή πληροφορία» που συγκεντρώνεται στις βάσεις δεδομένων με σκοπό την υποστήριξη σύνθετων αποφάσεων με σκοπό την επίλυση πολύπλοκων προβλημάτων.

Το όφελος από την ανάλυση μεγάλου όγκου δεδομένων από τη Δημόσια Διοίκηση με στόχο την ανακάλυψη γνώσης για λήψη των κατάλληλων αποφάσεων που θα οδηγήσουν σε μεγαλύτερη αποτελεσματικότητα του κράτους έχει άμεσο αντίκτυπο στην καθημερινότητα και ποιότητα ζωής των πολιτών μιας και προϋποθέτει άμεση γνώση των αναγκών της κοινωνίας, ενώ ταυτόχρονα θα απαιτούνται λιγότεροι πόροι για την ικανοποίηση αυτών των αναγκών.

Επιχειρήθηκε να παρουσιαστεί μια διαδικασία εξόρυξης δεδομένων για την αντιμετώπιση ενός σύνθετου προβλήματος αυτό της ύπαρξης ληξιπρόθεσμων οφειλών από τα νοικοκυριά. Η γνώση των παραγόντων που επιδρούν στη δυνατότητα των πολιτών να ανταποκριθούν στις μηνιαίες υποχρεώσεις τους, ωφελεί στη διατύπωση και διαμόρφωση των κατάλληλων πολιτικών που σκοπό έχουν την υποστήριξη των πολιτών στους τομείς που δημιουργούν αδυναμία ανταπόκρισης ως προς τις υποχρεώσεις τους. Οι τεχνικές που αποκαλύπτουν τα προβλήματα αντιμετώπισης των αναγκών μέσα από τη βάση δεδομένων που δημιουργήθηκε για τις ανάγκες της εργασίας εξαιτίας της έλλειψης αντίστοιχων πληροφοριών στις βάσεις ανοιχτών δεδομένων, υπέδειξε παράγοντες που η δημόσια διοίκηση οφείλει να λάβει υπόψιν για την αντιμετώπιση αυτών των προβλημάτων.

Στο μέλλον, προτείνεται η μελέτη και η εφαρμογή τεχνικών εξόρυξης να εμπλουτιστεί σε δεδομένα που θα μπορούσαν να αντληθούν από κατάλληλες και με πολλούς και διαφορετικούς τομείς ανοικτές βάσεις δεδομένων. Για το σκοπό αυτό προτείνεται η μελέτη δεδομένων σε διαφορετικούς τομείς, χρησιμοποιώντας επιπρόσθετες τεχνικές εξόρυξης, με σκοπό την ανακάλυψη γνώσης η οποία και θα συμβάλει στη διαδικασία λήψης αποφάσεων στη δημόσια διοίκηση.

Βιβλιογραφία

Βιβλία & Δημοσιεύσεις

1. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., 1996. Advances in Knowledge Discovery and Data Mining. AAAI Press
2. Piatetsky-Shapiro, G., 1991. Knowledge Discovery in Databases: An Overview. PKP
3. Han, J., Kamber, M., Pei, J., 2012. Data Mining Concepts and Techniques. Morgan Kaufman
4. Hand, D., Mannila, H, and Smyth, P., 2001. Principles of Data Mining (Adaptive Computation and Machine Learning). Bradford Book
5. Frawley, J., Piatetsky-Shapiro, G., and Matheus, C., 1991. Knowledge Discovery in Databases: An Overview. AI Magazine
6. Merrill, D., & Tennyson, D. 1977. Concept teaching: An instructional design guide. Educational Technology
7. Collins, G., 2020. Data Science from scratch. Independently Published, 2020
8. Ackoff, R. L. 1989. From data to wisdom. Journal of Applied Systems Analysis
9. Howe, J., 2006. The rise of crowdsourcing. Wired Magazine
10. Agrawal, R., Imielinski, T. and Swami, A.N. 1993. Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data
11. Fayyad, U., Haussler, D., and Stolroz, P. 1996. Minning Scientific Data. Communications of the ACM.
12. Jiawei, H., 2001.Data Mining, Concepts and Techniques. San Francisco Morgan Kaufmann Publishers
13. Phyu, T.N., 2009 Survey of Classification Techniques in Data Mining. Proceedings of the 2009 International Multi Conference of Engineers and Computer Scientists, 18-20 March 2009, Hong Kong
14. Roiger, R., Geatz, M., 2007. Data Mining a Tutorial Based Primer. Pearson.
15. Tan, P., Steinbach, M., Kumar, V., 2005. Introduction to Data Mining. Pearson
16. Achtert E., Böhm, C., Kriegel, H. P., Kröger,P., Müller-Gorman,I., Zimek, A. 2007. Detection and visualization of subspace cluster hierarchies. Springer

17. Rossi, F., Tsoukias, A., 2009. Algorithmic Decision Theory. Springe
18. Rokach, L., Maimon, O., 2014. Data Mining with Decision Trees. World Scientific
19. Shavlik, J., Mooney, R., and Towell, G. in press. 1990. Symbolic and neural net learning algorithms. Machine Learning. Forthcoming
20. Chester, M., 1993. Neural Networks: A Tutorial. Prentice Hall
21. Civco, D.L., 1993. Artificial Neural Network for Land Cover and Mapping, International Journal of Geographical Information System
22. Brian D., Ripley, N., L., 1996. Pattern Recognition and Neural Networks. Cambridge University Press
23. Schmidhuber, J., 2015. Deep Learning in Neural Networks: An Overview. Neural Networks, Vol 61
24. Jaynes, E.T., 2003. Probability theory. Cambridge University Press
25. McGrayne, S. B., 2011. The Theory That Would Not Die. Yale University Press
26. Cohen, J., Cohen, P., West, S.G. & Aiken L.S. (2003). Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences. 3rd ed. Hillsdale, N.J.: Lawrence Erlbaum Associates.
27. Bates, D.M. & Watts, D.G. (1988). Nonlinear Regression Analysis & Its Applications. New York: John Wiley & Sons
28. Rawlings, J.O., Pantula, S.G. & Dickey, D.A. (1998). Applied Regression Analysis. A Research Tool (Second Edition). New York: Springer-Verlag, Inc
29. Henninger, Maureen (2013) The Value and Challenges of Public Sector Information, Cosmopolitan Civil Societies: An Interdisciplinary Journal
30. Jordan, Sara R. (2014) Beneficence and the Expert Bureaucracy, Public Integrity
31. Alves, David & Martinez, Luis M. & Viegas, José M. (2012) Retrieving Real-time Information to users in Public Transport Networks: An Application to the Lisbon Bus System, Procedia - Social and Behavioral Sciences
32. Arribas-Bel, Daniel (2014) Accidental, open and everywhere: Emerging data sources for the understanding of cities, Applied Geography

33. Annoni, Paola & Ferrari, PierAlda & Salini, Silvia (2006) Data Mining Analysis on Italian Family Preferences and Expenditures in P. Perner (Ed.), Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining, Springer Berlin Heidelberg

34. Corallo, Angelo & Fortunato, Laura & Matera, Marco & Alessi, Marco & Camillò, Alessio & Chetta, Valentina, Storelli, Davide (2015) Sentiment Analysis for Government: An Optimized Approach in P. Perner (Ed.), Ma- chine Learning and Data Mining in Pattern Recognition, Springer International Publishing

35. Lev-On, Azi & Steinfeld, Nili (2015) Local engagement online: Municipal Facebook pages as hubs of interaction, Government Information Quarterly

36. Kamel Boulos, Maged N. & Al-Shorbaji, Najeeb M. (2014) On the Internet of Things, smart cities and the WHO Healthy Cities, International Journal of Health Geographics

37. Cao, Longbing (2012). Social Security and Social Welfare Data Mining: An Overview, IEEE Transactions on Systems, Man & Cybernetics: Part C - Applications & Reviews

38. Raad, Elie & Al Bouna, Bechara & Chbeir, Richard (2015) Preventing sensitive relationships disclosure for better social media preservation, International Journal of Information Security

39. Pandey, Rajiv & Dhoundiyal, Manoj (2015) Quantitative Evaluation of Big Data Categorical Variables through R, Proceedings of the International Conference on Information and Communication Technologies, ICICT 2014

40. Ohemeng, Frank L. K. & Ofosu-Adarkwa, Kwaku (2015) One way traffic: The open data initiative project and the need for an effective demand side initiative in Ghana, Government Information Quarterly

41. Hansen, J.V., Nelson, R.D. 1997. Neural networks and traditional time series methods: a synergistic combination in state economic forecasts. IEEE Transactions on Neural Networks

42. Makulowich, J. Government Data Mining Systems Defy Definition. Washington Technology

43. Sund, R. Utilization of Administrative Registering using Statistical Knowledge Discovery. National Research and Development Centre for Welfare and Health, 2002; Working Paper

44. Zvarova, J., Pribik, V. Information society in Czech healthcare `starting point' to prognosis for the year 2013, 2002; International Journal of Medical Informatics

45. DSS Consulting. IAURIF – Traffic flow prediction in the Paris region, 2003. http://www.datamining.hu/angol/alk_koz.html

46. Klosgen, W., May, M. 2003. Census Data Mining: An Application.

47. Siebes, A. Data Mining for Professional Statisticians.

48. Luan, J. 2001.Data Mining as Driven by Knowledge Management in Higher Education

49. Κιόχος, Π., 1993. Στατιστική. Interbooks

50. Κουρής, Γ., 2006. Εφαρμογή Τεχνικών Data Mining σε Συστήματα Ηλεκτρονικού εμπορίου

51. Βερύκιος, Β., Καγκλής, Β., Σταυρόπουλος, Η., 2015. Η επιστήμη των δεδομένων μέσα από τη γλώσσα R. Κάλλιππος

52. Βαζιργιάννης, Μ., Χαλκίδη, Μ., 2003. Εξόρυξη Γνώσης Από Βάσεις Δεδομένων. Πρωτοπορία

53. Μ. Δενδρινός, Μ., Δ. Κουής, Δ., 2015 Βασικές Αρχές και Τεχνολογίες στην Επιστήμη της Πληροφόρησης, Κάλλιπος

Παράρτημα

Δημογραφικά Στοιχεία Δείγματος.

ΙΔΙΟΤΗΤΑ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ΕΕ	132	44,3	44,3	44,3
	ΙΥ	114	38,3	38,3	82,6
	ΔΥ	52	17,4	17,4	100,0
	Total	298	100,0	100,0	

ΦΥΛΟ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Α	150	50,3	50,3	50,3
	Θ	148	49,7	49,7	100,0
	Total	298	100,0	100,0	

ΠΑΝΤΡΕΜΕΝΟΣ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ΟΧΙ	100	33,6	33,6	33,6
	ΝΑΙ	198	66,4	66,4	100,0
	Total	298	100,0	100,0	

ΕΝΟΙΚΙΑΣΗ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ΟΧΙ	152	51,0	51,0	51,0
	ΝΑΙ	146	49,0	49,0	100,0
	Total	298	100,0	100,0	

ΕΣΟΔΑ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<1000	68	22,8	22,8	22,8
	1000-2000	194	65,1	65,1	87,9
	>2000	36	12,1	12,1	100,0
	Total	298	100,0	100,0	

ΛΗΞΙΠΡΟΘΕΣΜΕΣΟΦΕΙΛΕΣ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	OXI	162	54,4	54,4	54,4
	NAI	136	45,6	45,6	100,0
	Total	298	100,0	100,0	

Dependent Variable Encoding

Original Value	Internal Value
OXI	0
NAI	1

Στατιστικές Αναλύσεις

Categorical Variables Codings

		Frequency	Parameter coding	
			(1)	(2)
ΕΣΟΔΑ	<1000	68	1,000	,000
	1000-2000	194	,000	1,000

	2	36	,000	,000
	EE	132	1,000	,000
ΙΔΙΟΤΗΤΑ	ΙΥ	114	,000	1,000
	2	52	,000	,000
	ΟΧΙ	152	1,000	
ΕΝΟΙΚΙΑΣΗ	ΝΑΙ	146	,000	
	Α	150	1,000	
ΦΥΛΟ	Θ	148	,000	
	ΟΧΙ	100	1,000	
ΠΑΝΤΡΕΜΕΝΟΣ	ΝΑΙ	198	,000	

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	295,520[a]	,321	,429

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	16,123	8	,041

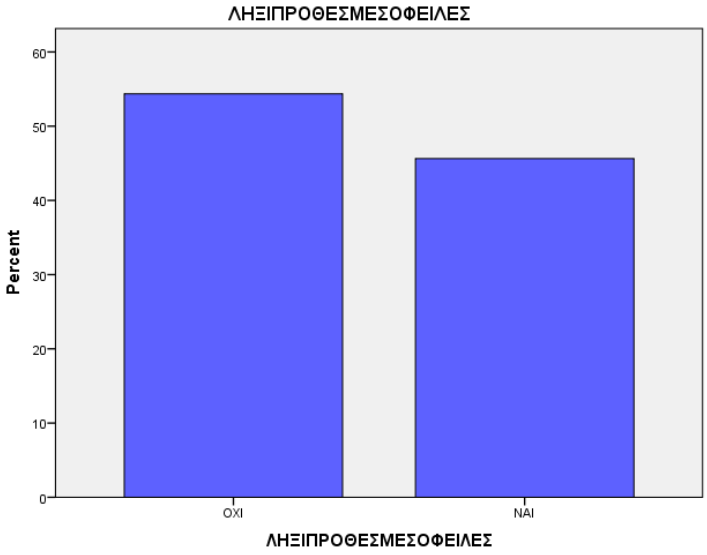
Variables in the Equation

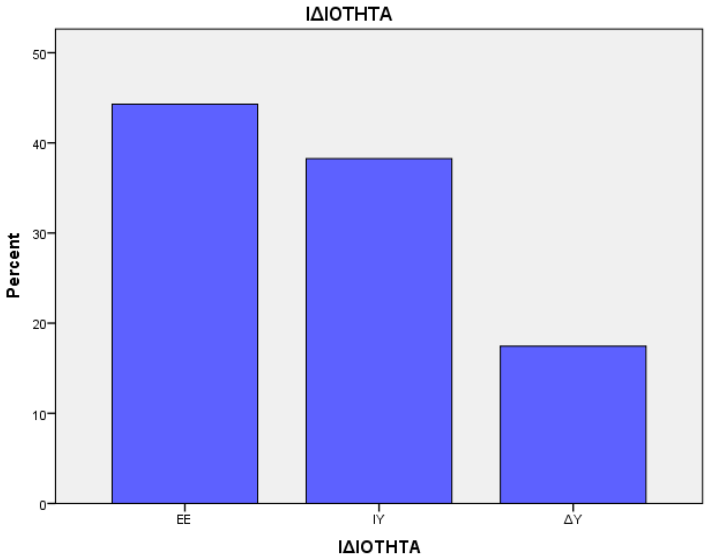
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	,000	,014	,000	1	,982	1,000
	JOB			9,707	2	,008	
	JOB(1)	1,365	,464	8,670	1	,003	3,916
	JOB(2)	1,353	,463	8,550	1	,003	3,870

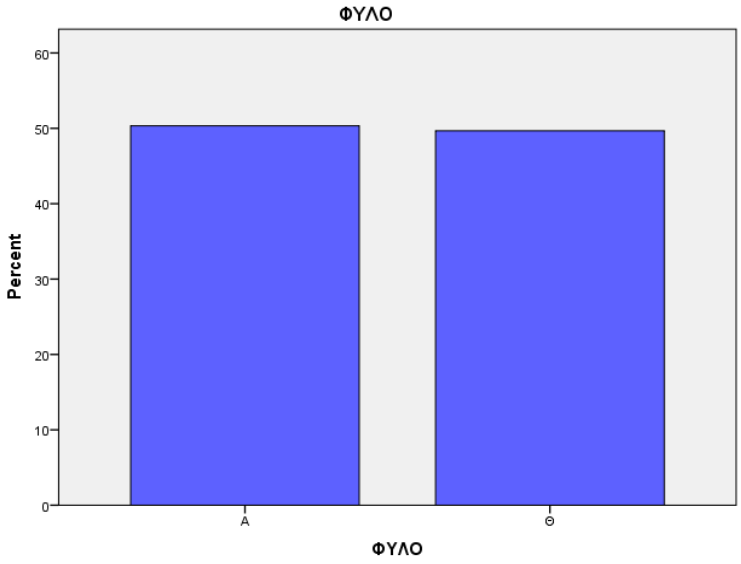
	GENDER(1)	-,364	,296	1,512	1	,219	,695
	MARITAL(1)	-,382	,493	,599	1	,439	,683
	CHILDREN	-,337	,240	1,959	1	,162	,714
	RENTAL(1)	-1,634	,316	26,811	1	,000	,195
	INCOME			8,623	2	,013	
	INCOME(1)	22,202	6195,658	,000	1	,997	4386468497,083
	INCOME(2)	21,190	6195,658	,000	1	,997	1595358591,790
	Constant	-20,956	6195,658	,000	1	,997	,000

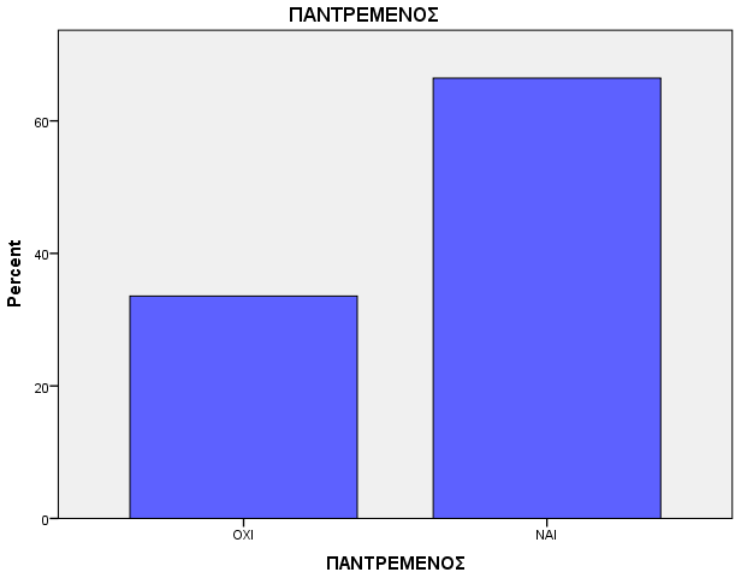
a. Variable(s) entered on step 1: AGE, JOB, GENDER, MARITAL, CHILDREN, RENTAL, INCOME.

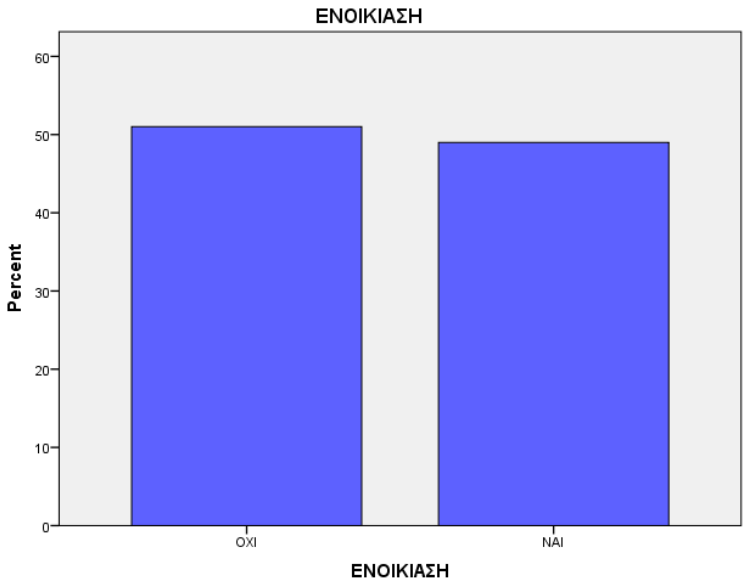
Γραφήματα

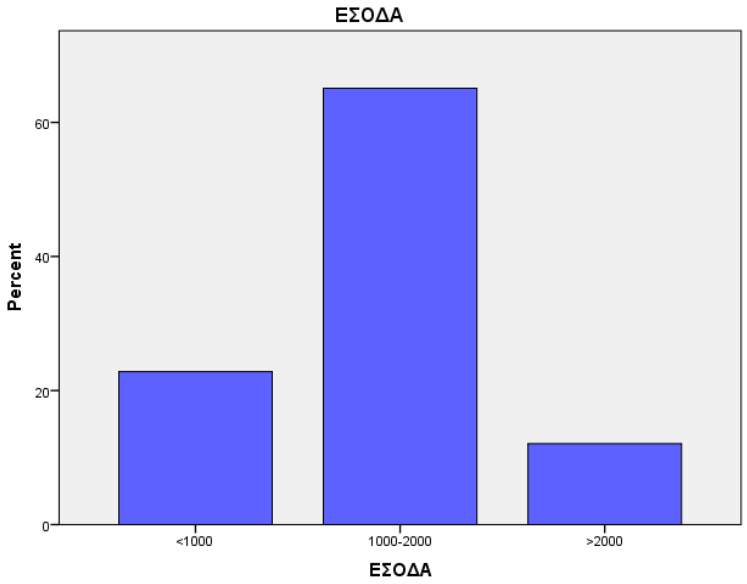

ΛΗΞΙΠΡΟΘΕΣΜΕΣΟΦΕΙΛΕΣ
Percent
60
50
40
30
20
10
0
ΟΧΙ
ΝΑΙ
ΛΗΞΙΠΡΟΘΕΣΜΕΣΟΦΕΙΛΕΣ

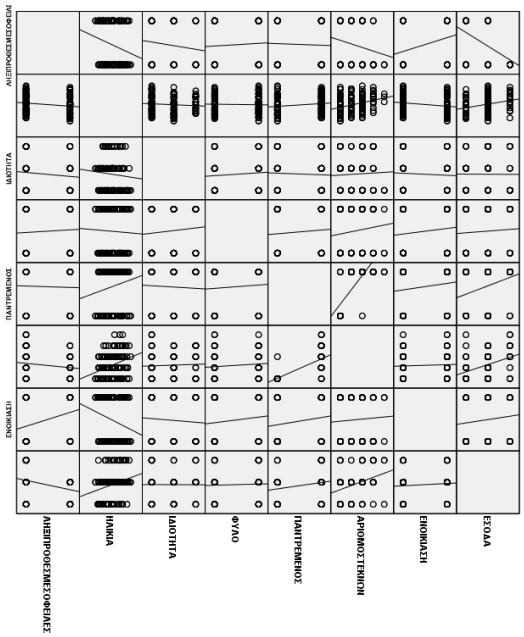

ΙΔΙΟΤΗΤΑ
Percent
50
40
30
20
10
0
ΕΕ
ΙΥ
ΔΥ
ΙΔΙΟΤΗΤΑ

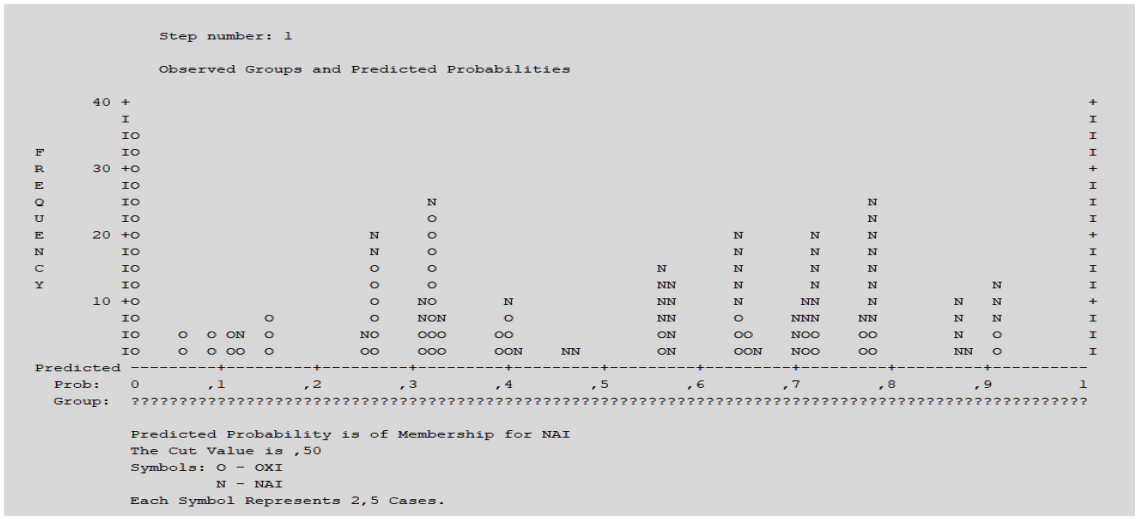

ΦΥΛΟ
Percent
60
50
40
30
20
10
0
Α
Θ
ΦΥΛΟ


ΠΑΝΤΡΕΜΕΝΟΣ
Percent
60
40
20
0
ΟΧΙ
ΝΑΙ
ΠΑΝΤΡΕΜΕΝΟΣ


ΕΝΟΙΚΙΑΣΗ
Percent
60
50
40
30
20
10
0
ΟΧΙ
ΝΑΙ
ΕΝΟΙΚΙΑΣΗ


ΕΣΟΔΑ
Percent
60
40
20
0
<1000
1000-2000
>2000
ΕΣΟΔΑ


ΑΝΕΞΑΡΤΗΣΙΑ ΜΕΣΟ ΟΦΕΛΕΙΑΣ
ΗΛΙΚΙΑ
ΙΔΙΟΤΗΤΑ
ΦΥΛΟ
ΠΛΗΓΩΜΕΝΟΣ
ΑΡΙΘΜΟΣ ΤΕΚΝΩΝ
ΕΝΟΙΚΙΑΣΗ
ΕΣΟΔΑ


Step number: 1
Observed Groups and Predicted Probabilities
FREQUENCY
Predicted
Prob: 0 ,1 ,2 ,3 ,4 ,5 ,6 ,7 ,8 ,9 1
Group: ??
Predicted Probability is of Membership for NAI
The Cut Value is ,50
Symbols: O - OXI
N - NAI
Each Symbol Represents 2,5 Cases.

Ερωτήσεις Απαντήσεις 1 Ρυθμίσεις

ερωτηματολόγιο ύπαρξης ληξιπρόθεσμων οφειλών

Ερωτηματολόγιο στο πλαίσιο ερευνητικής εργασίας

Ιδιότητα *

- Ελεύθερος Επαγγελματίας
- Ιδιωτικός Υπάλληλος
- Δημόσιος Υπάλληλος

ΗΛΙΚΙΑ *

Κείμενο σύντομης απάντησης

Φύλο *

- Άνδρας
- Γυναίκα

Παντρεμένος/η

- ΝΑΙ
- ΟΧΙ

Αριθμός Τέκνων

Κείμενο σύντομης απάντησης

Αριθμός Τέκνων

Κείμενο σύντομης απάντησης

Ενοικίαση (Σε περίπτωση που στον τόπο διαμονής είστε ενοικιαστής και όχι ιδιοκτήτης) *

- ΝΑΙ
- ΟΧΙ

Μηνιαίο Εισόδημα *

- <1000
- 1000-2000
- >2000

Ύπαρξη ληξιπρόθεσμων οφειλών *

- ΝΑΙ
- ΟΧΙ

Εθνική Σχολή Δημόσιας Διοίκησης και Αυτοδιοίκησης (Ε.Σ.Δ.Δ.Α.)

Πειραιώς 211, ΤΚ 177 78, Ταύρος

τηλ: 2131306349 , fax: 2131306479

www.ekdd.gr