ΑΛΕΞΑΝΔΡΕΙΟ ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ ΔΙΟΙΚΗΣΗ, ΛΟΓΙΣΤΙΚΗ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Εξόρυξη δεδομένων και απεικόνιση σε γεωγραφικό σύστημα πληροφοριών της εγκληματικότητας στην Ελλάδα

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΓΕΩΡΓΙΟΥ ΧΑΡΙΤΟΠΟΥΛΟΥ

Επιβλέπων : Ευστάθιος Κύρκος
Αναπληρωτής Καθηγητής, Τμήμα Λογιστικής και Χρηματοοικονομικής

Θεσσαλονίκη, 06/2017

Η σελίδα αυτή είναι σκόπιμα λευκή.

ΑΛΕΞΑΝΔΡΕΙΟ ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ ΔΙΟΙΚΗΣΗ, ΛΟΓΙΣΤΙΚΗ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Εξόρυξη δεδομένων και απεικόνιση σε γεωγραφικό σύστημα πληροφοριών της εγκληματικότητας στην Ελλάδα

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

ΓΕΩΡΓΙΟΥ ΧΑΡΙΤΟΠΟΥΛΟΥ

Επιβλέπων : Ευστάθιος Κύρκος
Αναπληρωτής Καθηγητής, Τμήμα Λογιστικής και Χρηματοοικονομικής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις ΧΧ/ΧΧ/ΧΧΧ.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Όνομα Επώνυμο Βαθμίδα, Ίδρυμα	Όνομα Επώνυμο Βαθμίδα, Ίδρυμα	Όνομα Επώνυμο Βαθμίδα, Ίδρυμα

Θεσσαλονίκη, ……

(Υπογραφή)

.....

Χαριτόπουλος, Γεώργιος

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του Προγράμματος Μεταπτυχιακών Σπουδών του Τμήματος Λογιστικής και Χρηματοοικονομικής, του Αλεξάνδρειου Τεχνολογικού Εκπαιδευτικού Ιδρύματος Θεσσαλονίκης, με ειδίκευση στη Χρηματοοικονομική Διοίκηση, τη Λογιστική και τα Πληροφοριακά Συστήματα.

Με την ολοκλήρωση της συγγραφικής μου εργασίας θα ήθελα να ευχαριστήσω τον κ. Κύρκο Ευστάθιο, επιβλέποντα καθηγητή, για τις συνεχείς και ουσιαστικές επιστημονικές συμβουλές του, καθ' όλη τη διάρκεια της εκπόνησης της παρούσας διπλωματικής εργασίας, καθώς και την μητέρα μου για την αμέριστη ηθική και πνευματική της συμπαράσταση στην παρούσα προσπάθεια.

Στη μνήμη της Ντόρας

Περίληψη

Η διασφάλιση της ορθής λειτουργίας των κοινωνικών δομών ενός κράτους προϋποθέτει, εκτός των άλλων, τη δημόσια ασφάλεια και προστασία των πολιτών. Στον σκοπό αυτό συμβάλλει η τεχνολογική ανάπτυξη και η αφθονία δεδομένων που με την ηλεκτρονική τους μορφή μπορούν να χρησιμοποιηθούν για τον χαρακτηρισμό και τη πρόβλεψη εγκλημάτων, μέσω των μεθόδων εξόρυξης δεδομένων (data mining), καθώς και την απεικόνισή τους με γεωγραφικό σύστημα πληροφοριών (Geographical Information System - G.I.S.).

Στη παρούσα εργασία με τη χρήση τεχνικών ανάλυσης εξόρυξης δεδομένων εντοπίζονται τα χαρακτηριστικά των εγκλημάτων και εξετάζεται η επιρροή κοινωνικο-δημογραφικών παραγόντων στη διάπραξή τους. Χρησιμοποιείται βάση δεδομένων εγκλημάτων που διαπράχθηκαν στην Ελλάδα το έτος 2016. Ανάλογα με το τύπο των εγκλημάτων, αυτά διακρίνονται σε ομάδες (κατά των περιουσιακών δικαιωμάτων, κατά της ιδιοκτησίας κτλ.) και επιλέγονται τα κατάλληλα για την διεξαγωγή της έρευνας. Δημιουργείται το αναγκαίο μοντέλο, όπου με τις ορθές μεθόδους, τεχνικές εξόρυξης δεδομένων και απεικόνιση σε γεωγραφικό σύστημα πληροφοριών, μπορούν να εξαχθούν συμπεράσματα για την κατανόηση πτυχών της εγκληματικότητας, εντοπισμού σημείων ¨hotspot¨ και επαναλαμβανόμενων μοτίβων εγκλημάτων σε συγκεκριμένο τόπο και χρόνο. Μελετάται η εγκληματικότητα στον ελλαδικό χώρο και εντοπίζονται τα εγκλήματα που πρωταγωνιστούν, όπου είναι οι κλοπές, οι ληστείες και οι απάτες, καθώς και τα χαρακτηριστικά τους (π.χ. τελεσμένο και ανεξιχνίαστο έγκλημα, πλημμέλημα, ημεδαπός, άνδρας, ηλικίας 18-34, άγνωστα αίτια). Τα αποτελέσματα αποτελούν πεδίο προβληματισμού και έρευνας για τις αρμόδιες αρχές (π.χ. Αστυνομία και Δικαστικές Αρχές). Η ανάλυση των εξαγόμενων πληροφοριών μπορεί να αξιοποιηθεί από την Αστυνομία για τον στρατηγικό σχεδιασμό αντιμετώπισης της εγκληματικότητας, τη πρόληψη (πρόβλεψη εγκληματικών ενεργειών), τη καταστολή (μείωση εγκλημάτων που εμφανίζονται σε μεγάλο βαθμό) και εξιχνίαση των εγκλημάτων, καθώς και την επ' αυτοφώρω σύλληψη των δραστών.

Λέξεις Κλειδιά: Εξόρυξη Δεδομένων, WEKA, Γεωγραφικό Συστήματα Πληροφοριών, Γ.Σ.Π., CrimeView, Εγκληματικότητα.

Η σελίδα αυτή είναι σκόπιμα λευκή.

Abstract

Ensuring the proper functioning of the social structures of a state presupposes, among other things, public safety and the protection of citizens. What contributes to that end is the technological development and data abundance, which in their electronic form can be used to characterize and predict crimes, through data mining methods as well as their depiction of geographic information Systems (G.I.S.).

In this thesis, by using data mining techniques, the characteristics of crimes are identified and the influence on their perception of their socio-demographic factors is examined. A database of crimes committed in Greece in the year 2016 is used. Depending on the type of crime, they are divided into groups (against property rights, property, etc.) and the suitable ones for conducting the research are chosen. The necessary model is developed where, with the right methods, data mining techniques and display in a geographical information system, conclusions can be drawn for the understanding of the aspects of crime, the identification of hotspots and the repetitive patterns of crime at a specific place and time. Criminal activity is investigated in Greece and the mainstay crimes that are thefts, robberies and frauds are detected, as well as their characteristics (e.g. ended and unsolved crime, misdemeanor, native, male, aged 18-34, unknown causes). The results are an area of reflection and research for the competent authorities (e.g. Police and Judicial Authorities). The analysis of the extracted information can be exploited by the Police for its strategic planning of tackling crime, prevention (crime prediction), repression (the reduction of large extended crimes) and detection of crimes, as well as the perpetrators in the act arrest.

Keywords: Data mining, WEKA, Geographic Information Systems, G.I.S., CrimeView, Crime.

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

Ευρετήριο Σχημάτων

Ευρετήριο Πινάκων

1 Εισαγωγή

1.1 Πρόβλημα – Σημαντικότητα του θέματος

Τα τελευταία 50 χρόνια, έχουν γίνει πολλές έρευνες για την ανάλυση της εγκληματικότητας και ιδίως για τον εντοπισμό των οικονομικών αιτιών της, με ποικιλία στην κύρια μεταβλητή επιρροής της, που άλλοτε μπορεί να ήταν η αγορά εργασίας και άλλοτε το εισόδημα των πολιτών. Ο βραβευμένος με Νόμπελ οικονομικών επιστημών Becker Gary προσέγγισε το θέμα της επιρροής των οικονομικών παραγόντων στην εγκληματικότητα από την οπτική πλευρά του δράστη της εγκληματικής ενέργειας. Έδωσε έμφαση στο κέρδος που μπορεί να έχει ο δράστης από την ενέργειά του και στο κόστος που θα έχει η πράξη του σε σχέση με την τιμωρία του εάν συλληφθεί. Ακόμη, ερευνήθηκε κατά πόσο θα έπραττε ένας πολίτης μία εγκληματική ενέργεια σε περίπτωση απολαβής υψηλά αμειβόμενων μισθών, παρέχοντάς του οικονομική ευημερία. Διάφορες έρευνες για την επιρροή της εγκληματικότητας από το σύνολο των εγκληματικών ενεργειών δεν έδιναν σε μεγάλο βαθμό αξιόπιστα αποτελέσματα. Το μοντέλο του Becker Gary αναδεικνύει πως η κατηγοριοποίηση των εγκλημάτων (κατά της ιδιοκτησίας, κατά της ζωής κτλ.) παρέχει αξιόλογα αποτελέσματα, συνδέοντας τις κατηγορίες αυτές με κοινωνικούς και δημογραφικούς παράγοντες. Επομένως, εάν για παράδειγμα ο δείκτης ακαθάριστου εγχώριου προϊόντος κατά κεφαλή είναι πάρα πολύ μικρός και υπάρχει μεγάλη ανεργία θα έχει επίδραση σε διάπραξη εγκλημάτων κατά της ιδιοκτησίας και των περιουσιακών δικαιωμάτων και όχι κατά της πολιτειακής εξουσίας (απείθεια) και συνεπώς σε όλα τα εγκλήματα στο σύνολό τους (Becker, 1968).

Οι παράγοντες που επηρεάζουν το έγκλημα ποικίλουν και μπορεί να οφείλονται σε οικονομική, κοινωνική και ανθρώπινη κρίση, καθώς και στην κρίση αξιών. Εάν θελήσουμε να κατηγοριοποιήσουμε τους παράγοντες επιρροής των εγκλημάτων σε γενικές κατηγορίες μπορούμε να πούμε, βάσει βιβλιογραφίας, ότι υπάρχουν οι κοινωνικοί (μορφωτικό επίπεδο πολιτών, εκπαιδευτικά συστήματα), οικονομικοί (ανεργία, πληθωρισμός, ΑΕΠ κατά κεφαλήν) και δημογραφικοί (πληθυσμός ανά περιφέρεια, ποσοστό ανδρών-γυναικών-παιδιών κτλ.) παράγοντες. Επιπροσθέτως, κατά Becker Gary, υπάρχει και η κατηγορία της ορθής λειτουργίας των κρατικών

μηχανισμών, όπως ύπαρξη συστήματος δικαιοσύνης, πρόληψης και καταστολής από την Αστυνομία και αυστηρότητα επιβολής των νόμων.

Η τεχνολογική εξέλιξη συμβάλει στην αντιμετώπιση της εγκληματικότητας με ποικίλους τρόπους. Καθημερινά αποθηκεύονται από τις διωκτικές αρχές (Αστυνομία, Λιμενικό, Δικαστικές Αρχές κτλ.) δεδομένα που αφορούν εγκλήματα. Κρυμμένη γνώση στους τεράστιους όγκους δεδομένων εγκλημάτων που αποθηκεύει για μεγάλο χρονικό διάστημα η Αστυνομία, μπορεί να εξορυχτεί και να βοηθήσει στην καταπολέμηση της εγκληματικότητας. Με την κατάλληλη επεξεργασία και ανάλυση των δεδομένων μπορούν να προσφερθούν χρήσιμες και πολύτιμες πληροφορίες για την πρόληψη και καταστολή του εγκλήματος. Πληροφοριακά συστήματα και λογισμικά είναι στην διάθεση των Διωκτικών Αρχών και με την εφαρμογή των κατάλληλων τεχνικών εξόρυξης δεδομένων, αναλύονται μεγάλου όγκου συσσωρευμένα δεδομένα εγκλημάτων και επιτυγχάνεται ανίχνευση κρυφών επαναλαμβανόμενων μοτίβων και σχέσεων, πρόβλεψη μελλοντικών τάσεων και συμπεριφορών, ανακάλυψη νέων προτύπων και επίλυση προβλημάτων. Με την χρήση της εξόρυξης δεδομένων η Αστυνομία μπορεί να προβλέψει και να εκτιμήσει καταστάσεις, να εντοπίσει την εγκληματική τάση και τους παράγοντες που επηρεάζουν το έγκλημα και να σκιαγραφήσει την εγκληματική συμπεριφορά του δράστη. Οι τεχνικές εξόρυξης δεδομένων αναπτύχθηκαν στο βάθος του χρόνου, ύστερα από μακρά έρευνα και εξελίσσονται συνεχώς. Ανάλογα με το τι επιζητείται η Αστυνομία κάθε φορά, χρησιμοποιούνται τα κατάλληλα εργαλεία (αλγόριθμοι) και εξάγονται κανόνες συσχέτισης εγκλημάτων και των χαρακτηριστικών τους, καθώς και ομαδοποιήσεις και κατηγοριοποιήσεις αυτών.

Η ύπαρξη γεωχωρικών γεγονότων και δεδομένων και η δυνατότητα ανάλυσης τους με την χρήση γεωγραφικών συστημάτων πληροφοριών, παρέχει την χαρτογράφηση της εγκληματικότητας και αφορά πληροφορίες σχετικές με τον χώρο, καθώς και δραστηριότητες που εκτελούνται σε αυτόν. Η χρήση χαρτών και χρονικών δεδομένων και ο προσδιορισμός χωρικών και χρονικών προτύπων, βοηθάει στην ανάπτυξη πρακτικών πρόληψης εγκλημάτων σε οριοθετημένο τόπο και χρόνο. Παρέχονται πληροφορίες, όπως για το τι έγκλημα έγινε, πότε (ημερομηνία και ώρα) και που (τόπος), καθώς και ενημέρωση για επαναλαμβανόμενα μοτίβα εγκλημάτων και κατανοείται το που, πότε και γιατί συγκεκριμένα εγκλήματα πιθανόν να συμβούν. Επομένως, επιτυγχάνεται η πρόβλεψη της εγκληματικότητας και εντοπίζονται προβληματικές περιοχές (hotspot) ώστε να παρθούν τα ανάλογα αστυνομικά μέτρα. Δίνεται η δυνατότητα προχωρημένης χωρικής ανάλυσης εγκλημάτων και επόπτευσής τους, μέσω

του γεωγραφικού συστήματος πληροφοριών και μπορούν να γίνουν ερωτήματα σε πολλαπλά γεωγραφικά επίπεδα πληροφορίας.

Η δημόσια ασφάλεια αποτελεί μέλημα όλων των πολιτών και της εκάστοτε κοινωνίας για την επίτευξη της εύρυθμης λειτουργίας της και την προστασία των πολιτών της. Η ύπαρξη όμως του φαινομένου της εγκληματικότητας και η διάπραξη καθημερινά εγκλημάτων δημιουργεί πρόβλημα στην ομαλή συμβίωση των πολιτών και χρήζει, από τις αρμόδιες αρχές, άμεσης και αποτελεσματικής αντιμετώπισης. Η ανάλυση του εγκλήματος αποτελεί σημαντικό έργο, καθώς τα εξαγόμενα αποτελέσματα μπορούν να βοηθήσουν στην αντιμετώπιση του. Για το σκοπό αυτό χρησιμοποιούνται μέθοδοι και τεχνικές εξόρυξης δεδομένων και ανάλυση γεωγραφικής γνώσης. Η χρήση των δύο αυτών τρόπων, παρέχει στην Αστυνομία μία ολοκληρωμένη προσέγγιση πρόβλεψης, εξιχνίασης και καταστολής του εγκλήματος με την εξαγωγή γνώσης για την εγκληματικότητα και λύση συγκεκριμένων προβλημάτων, καθώς και της συμβολής στον στρατηγικό σχεδιασμό αντιμετώπισής της. Η σωστή ερμηνεία των αποτελεσμάτων, καθώς και η επιλογή των τεχνικών και μεθόδων εξαρτάται από τον αναλυτή και την ορθότητα της βάσης δεδομένων που χρησιμοποιείται.

1.2 Σκοπός – Στόχοι

Ο σκοπός της παρούσας διπλωματικής εργασίας είναι η εξόρυξη γνώσης και πληροφορίας από μεγάλο όγκο δεδομένων και αφορούν εγκλήματα που διαπράχθηκαν στην Ελλάδα το έτος 2016. Ακόμη, αναδεικνύεται ο τρόπος απεικόνισης εγκλημάτων σε γεωγραφικό σύστημα πληροφοριών και η αξία συνδυασμού των δύο αυτών τρόπων άντλησης πληροφορίας για την αντιμετώπιση της εγκληματικότητας. Η συγκεκριμένη διπλωματική εργασία επιμερίζεται στους εξής στόχους:

- Βιβλιογραφική επισκόπηση για την εννοιολογική οριοθέτηση της εγκληματικότητας.
- Αναζήτηση βιβλιογραφίας για την εξόρυξη γνώσης και πληροφορίας από βάσεις δεδομένων.
- Εξέταση μεθοδολογίας πρόβλεψης εγκλημάτων και χαρακτηριστικών τους.
- Εφαρμογή τεχνολογιών εξόρυξης δεδομένων για την ανάπτυξη μοντέλου εξόρυξης γνώσης και πληροφορίας.

- Απεικόνιση χωρικών δεδομένων εγκλημάτων με γεωγραφικό σύστημα πληροφοριών.

1.3 Συνεισφορά

Η παρούσα διπλωματική εργασία συνεισφέρει με τις παρακάτω ενέργειες:

- Παρουσιάζονται έννοιες και ορισμοί για την εγκληματικότητα.
- Γίνεται διάκριση των εγκλημάτων σε κατηγορίες.
- Αναφέρονται οι εργασίες και τεχνικές της εξόρυξης δεδομένων.
- Γίνεται επεξήγηση καθαρισμού, επεξεργασίας και μετασχηματισμού δεδομένων για την δημιουργία της χρησιμοποιηθείσας βάσης δεδομένων.
- Περιγράφονται τρόποι χρήσης τεχνολογιών εξόρυξης δεδομένων και απεικόνισης χωρικών δεδομένων.
- Διεξάγεται εκπαίδευση μοντέλου εξόρυξης δεδομένων στα δεδομένα εκπαίδευσης.
- Επιλέγονται τεχνικές και μέθοδοι εξόρυξης δεδομένων για εξαγωγή συμπερασμάτων.
- Συγκρίνονται τα αποτελέσματα των διαφόρων τεχνικών και μεθόδων εξόρυξης δεδομένων που χρησιμοποιήθηκαν και αξιολογούνται για την καταλληλότητα και συνεισφορά τους στην επίτευξη του σκοπού της εργασίας.
- Παρουσιάζονται τα αποτελέσματα σε διαγράμματα.
- Εντοπίζονται επαναλαμβανόμενα μοτίβα εγκλημάτων με συγκεκριμένα χαρακτηριστικά και οι συσχετισμοί τους, καθώς και αυτά που χρήζουν ιδιαίτερης προσοχής.
- Εξάγεται κρυφή γνώση και νέα πληροφορία από την βάση δεδομένων που πρωτύτερα ήταν άγνωστη, όπως σκιαγράφηση εγκληματικού προφίλ δραστών.

1.4 Διάρθρωση της μελέτης

Η δομή της εργασίας αποτελείται από 6 κεφάλαια και για την ευκολία κατανόησης περιγράφονται ανά κεφάλαιο. Στο 1° Κεφάλαιο γίνεται αναφορά για το

φαινόμενο της εγκληματικότητας και το επικουρικό έργο της τεχνολογίας στην καταπολέμησή της, μέσω της εξόρυξης γνώσης και πληροφορίας από μεγάλο όγκο δεδομένων που αφορούν εγκλήματα. Αναλύεται ο σκοπός και στόχος της διπλωματικής και εν συνεχεία παρατίθεται η συνεισφορά της στους αναγνώστες.

Στο 2° Κεφάλαιο παρουσιάζεται το θεωρητικό υπόβαθρο της εγκληματικότητας με αναφορά σε έννοιες και ορισμούς, καθώς και περεταίρω πληροφορίες για την κατηγοριοποίηση των εγκλημάτων. Γίνεται βιβλιογραφική επισκόπηση για την εξόρυξη γνώσης και πληροφορίας από βάσεις δεδομένων, περιγράφοντας τις εργασίες κατηγοριοποίησης, συσταδιοποίησης, κανόνων συσχέτισης, παλινδρόμησης, ανίχνευσης ανωμαλιών, ανάλυσης χρονολογικών σειρών, προτύπων ακολουθιών και μείωσης των διαστάσεων. Επίσης, αναφέρονται οι τεχνικές εξόρυξης δεδομένων, καθώς και ξεχωριστά για την εξόρυξη χωρικών δεδομένων.

Στο 3° Κεφάλαιο αναπτύσσεται η μεθοδολογία πρόβλεψης εγκλημάτων, όπου εμπεριέχει τον καθορισμό του προβλήματος, τις διαδικασίες προετοιμασίας της βάσης δεδομένων με τον καθαρισμό, επεξεργασία και μετασχηματισμό των δεδομένων και τέλος, την υλοποίηση των τελικών πινάκων.

Στο 4° Κεφάλαιο παρατίθεται το ερευνητικό μέρος της διπλωματικής με την εφαρμογή στην δημιουργηθείσα βάση δεδομένων των τεχνικών εξόρυξης με το πρόγραμμα Weka και την απεικόνιση εγκλημάτων σε γεωγραφικό σύστημα πληροφοριών (G.I.S.).

Στο 5° Κεφάλαιο περιγράφονται και αξιολογούνται τα συμπεράσματα της έρευνας και αναφέρεται η χρησιμότητά τους για την Αστυνομία.

Στο τελευταίο 6° Κεφάλαιο περιέχονται η σύνοψη, τα αποτελέσματα και τα συμπεράσματα της διπλωματικής, καθώς και οι περιορισμοί και μελλοντικές επεκτάσεις για περαιτέρω έρευνα της διπλωματικής εργασίας.

2 Βιβλιογραφική Επισκόπηση – Θεωρητικό Υπόβαθρο

2.1 Εννοιολογική οριοθέτηση της Εγκληματικότητας

Το φαινόμενο της εγκληματικότητας εμφανίζεται πολύ νωρίς, από τις πρώτες κιόλας μορφές κοινωνίας και συνεχίζεται η ύπαρξή του έως και σήμερα. Η αρχέγονη καταγωγή του φαινομένου αναφέρεται και στην βίβλο με τη δολοφονία του Άβελ από τον αδερφό του Κάιν. Το έγκλημα αποτελεί αναπόσπαστο στοιχείο της οργανωμένης κοινωνίας και παρατηρείται, σε μεγαλύτερο βαθμό, σε πόλεις με αυξημένο πληθυσμιακό αριθμό συγκέντρωσης πολιτών (μεγάλα αστικά κέντρα), καθώς και σε χώρες με έντονες κοινωνικοοικονομικές διακυμάνσεις και αισθητό βαθμό έλλειψης της δημοκρατικής έννοιας στο πολίτευμά τους.

2.1.1 Εγκληματικότητα-Έννοιες και Ορισμοί

Η εγκληματολογική επιστήμη μελετά την κοινωνική και οντολογική υπόσταση του εγκλήματος. Τα ατομικά και κοινωνικά αίτια οδηγούν τους ανθρώπους σε αξιόποινες πράξεις και ο εντοπισμός και ανάλυσή τους, από την συγκεκριμένη επιστήμη, επιφέρει αλλαγές στο σύστημα δικαίου, ώστε να επιβληθούν οι αντίστοιχες θεσμικές ποινές. Παρακλάδια της εγκληματολογικής επιστήμης είναι η Εγκληματολογία (γενική ή ειδική, ατομική ή κοινωνική κτλ.), η Θυματολογία (ερμηνεία προβλημάτων ποινικού δικαίου από την πλευρά του θύματος) και η Ανακριτική (έρευνα διαλεύκανσης τελεσθέντος εγκλήματος και εντοπισμός των ατομικών στοιχείων του δράστη). Ενδεικτικά, άλλες επικουρικές επιστήμες σε αυτήν της εγκληματικότητας είναι η Δικαστική Ψυχολογία, η Κοινωνιολογία, η Γραφολογία και η Ιατροδικαστική (Κωστάρας, 2001).

Ο αριθμός (μέγεθος) των εγκλημάτων φανερώνει την ποσοτική υπόσταση της εγκληματικότητας ενώ η βαρύτητα του εγκλήματος και οι επιπτώσεις που έχει στο κοινωνικό περιβάλλον αναδεικνύουν την ποιοτική της υπόσταση. Οι αιτίες που οδηγούν κάποιον στην διάπραξη εγκλήματος ποικίλουν και οι βασικές, εξ αυτών, είναι για λόγους οικονομικούς, ηθικούς, οικογενειακούς, βιολογικούς, ψυχολογικούς και μιμητισμού. Σημαντικοί παράγοντες όπως το περιβάλλον, οι κοινωνικές, οι οικονομικές και οι πολιτικές συνθήκες που επικρατούν επηρεάζουν διαφορετικά τον κάθε, εν δυνάμει,

δράστη στην κοινωνική του συμπεριφορά. Επομένως, ως αιτία στη διαμόρφωση του εγκλήματος μπορεί να θεωρηθεί και η χωρική επιρροή. Τα βασικά κίνητρα εγκληματικής ενέργειας είναι οικονομικά, συναισθηματικά και ηθικά.

Η έννοια του δικαίου αναφέρεται στην τάξη που πρέπει να επικρατεί στην κοινωνία και επιτυγχάνεται με ένα σύνολο θεσπισμένων κανόνων (νόμοι) από το κράτος ή εθίμων από την κοινωνία. Ο σκοπός της ύπαρξης δικαίου είναι η δημιουργία πλαισίου συμπεριφοράς των ανθρώπων μεταξύ τους, ώστε να επιτύχουν αρμονική συμβίωση με επικουρική βοήθεια των κανόνων ηθικής και εθιμοτυπίας, οι οποίοι δεν είναι υποχρεωτικοί από το νόμο (Ραφτόπουλος, 1996).

Οι κανόνες δικαίου διακρίνονται σε δημόσιους και ιδιωτικούς:

- Το Δημόσιο Δίκαιο περιλαμβάνει αρκετούς κλάδους και ένας εξ αυτών είναι το ποινικό δίκαιο, το οποίο έχει ως σκοπό την προστασία της κοινωνίας. Το ποινικό δίκαιο διαχωρίζεται σε ουσιαστικό και δικονομικό. Στο ουσιαστικό δίκαιο καθορίζονται ποια είναι τα εγκλήματα που προκύπτουν από τις πράξεις ή παραλείψεις των πολιτών και με τις ποινές που προβλέπονται. Το δικονομικό δίκαιο καθορίζει τα κρατικά όργανα και τον διαδικαστικό τρόπο λειτουργίας τους, με τον οποίο ερευνούν και βεβαιώνουν τα εγκλήματα, καθώς και με τον οποίο βρίσκουν και επιβάλλουν την ποινή στους δράστες. Ενδεικτικά, οι υπόλοιποι κλάδοι του δημοσίου δικαίου είναι το Συνταγματικό Δίκαιο, το Διοικητικό Δίκαιο και η Διοικητική Δικονομία, το Εκκλησιαστικό Δίκαιο και η Εκκλησιαστική Δικονομία, το Εργατικό Δίκαιο, το Ιδιωτικό Διεθνές Δίκαιο, το Δίκαιο Κοινωνικής Ασφάλισης και η Πολιτική Δικονομία.
- Το Ιδιωτικό Δίκαιο ρυθμίζει τις σχέσεις των πολιτών μεταξύ τους ή με το κράτος και ομαδοποιούνται σε αστικές και εμπορικές. Ενδεικτικά, οι κλάδοι στους οποίους διακρίνεται το Ιδιωτικό Δίκαιο είναι το Αστικό και Εμπορικό Δίκαιο. Το Αστικό Δίκαιο σχετίζεται με τις ιδιωτικές σχέσεις των πολιτών και διακρίνεται στις Γενικές Αρχές, το Ενοχικό Δίκαιο, το Εμπράγματο Δίκαιο, το Οικογενειακό Δίκαιο και το Κληρονομικό Δίκαιο. Το Εμπορικό Δίκαιο σχετίζεται με το εμπόριο και διακρίνεται στο Γενικό Μέρος, το Δίκαιο των Εμπορικών Εταιρειών, το Δίκαιο των Αξιόγραφων, το Πτωχευτικό Δίκαιο, το Ναυτικό Δίκαιο και το Ασφαλιστικό Δίκαιο.

(Ραφτόπουλος, 1996)

Στη παρούσα εργασία μας απασχολεί το Ουσιαστικό Ποινικό Δίκαιο, το οποίο καθορίζει ποια είναι τα εγκλήματα που προκύπτουν από τις πράξεις ή παραλείψεις των

πολιτών. Το Ποινικό Δίκαιο βασίζεται στην αρχή που λέει «κανένα έγκλημα, καμία ποινή χωρίς νόμο» (nullum crimen, nulla poena sine lege). Ουσιαστικά και σύμφωνα με το άρθρο 1 του Ποινικού Κώδικα (Π.Κ.) δεν μπορεί να επιβληθεί καμία ποινή παρά μόνο για εκείνες τις πράξεις για τις οποίες ο νόμος την είχε ρητά ορίζει πριν από την τέλεσή τους.

Το άρθρο 14 του Ποινικού Κώδικα ορίζει ότι:

«Έγκλημα είναι πράξη άδικη και καταλογιστεί στον δράστη της, η οποία τιμωρείται από το νόμο. Στις διατάξεις των ποινικών νόμων ο όρος «πράξη» περιλαμβάνει και τις παραλείψεις».

Επομένως, για να προκύπτει έγκλημα θα πρέπει να υπάρχει «πράξη», δηλαδή ανθρώπινη ενέργεια ή παράλειψη και η πράξη αυτή να είναι «άδικη» και «καταλογιστή» στον δράστη της, αλλά και να περιγράφεται σε κάποιον νόμο με ορισμένη «ποινή».

Η πράξη είναι εκούσια όταν γίνεται με τη θέληση του δράστη (ηθελημένη συμπεριφορά) με ενέργεια ή παράλειψη και με την οποία επέρχεται εξωτερική μεταβολή στο περιβάλλον. Η πράξη δεν είναι εκούσια και ποινικώς αξιόλογη όταν η ενεργητικότητα ενός ανθρώπου προκύπτει μετά από ακαταμάχητη σωματική βία που του ασκείται από άλλο πρόσωπο. Όταν ο άνθρωπος δεν έχει θέληση να δράσει, τότε η σωματική του συμπεριφορά δεν θεωρείται εκούσια, διότι είναι μη ενσυνείδητη και δεν αποτελεί έγκλημα. Ακόμη, για να υπάρξει έγκλημα πρέπει η συμπεριφορά του δράστη να έχει μεταβολή στο περιβάλλον. Τέλος, η πράξη του δράστη θεωρείται θετική εκτελώντας ενέργεια και αρνητική με παράλειψη (Ραφτόπουλος, 1996).

Ο άδικος χαρακτήρας της πράξης υφίσταται όταν η πράξη δεν συνάδει με τον απαγορευτικό ή επιτακτικό κανόνα δικαίου και δεν αποκλείεται για άλλους λόγους το άδικο. Μία πράξη είναι άδικη όταν ανάγεται σε αυτές του νόμου που προσδίδουν τον χαρακτηρισμό του εγκλήματος και όταν η πράξη προσβάλλει το έννομο ατομικό ή κοινωνικό αγαθό. Οι λόγοι που αποκλείουν τον άδικο χαρακτήρα της πράξης που αντιτίθεται στον κανόνα δικαίου, είναι αυτοί που σύμφωνα με άλλους κανόνες οι πράξεις αυτές είναι ανεκτές και δε προσβάλουν την ειρηνική συμβίωση. Δηλαδή, εάν υπάρχει ανθρωποκτονία που τελέστηκε σε άμυνα δεν είναι άδικη. Το αναγκαίο μέτρο της άμυνας κρίνεται από τον βαθμό επικινδυνότητας της επίθεσης, από το είδος της βλάβης που απειλούσε, από τον τρόπο και την ένταση της επίθεσης και από τις λοιπές περιστάσεις. Θα πρέπει να διευκρινιστεί πως αυτός που πράττει χωρίς υπαιτιότητα και ο ανίκανος για καταλογισμό πράττουν αδίκως και δεν αποτελούν προϋποθέσεις για το άδικο της πράξης (Ραφτόπουλος, 1996).

Ο καταλογισμός της άδικης πράξης γίνεται στον δράστη εφόσον είναι ικανός προς καταλογισμό (βιολογική και ψυχολογική ικανότητα αξιολόγησης), είναι υπαίτιος και μπορεί να συμμορφωθεί προς το κανόνα δικαίου. Δηλαδή, να αντιλαμβάνεται τις πράξεις τις οποίες έπραξε (υπό ομαλές συνθήκες), να είναι ηλικιακά ώριμος, ψυχικά και πνευματικά υγιής και να ενεργεί υπαιτίως (Ραφτόπουλος, 1996).

Έγκλημα δεν υπάρχει όταν ο νόμος δεν προβλέπει ποινή. Η πράξη έχει μόνο αστικά επακόλουθα ή επιβάλλεται μόνο διοικητικοί ή πειθαρχική ποινή (Ραφτόπουλος, 1996).

2.1.2 Διαίρεση των αξιόποινων πράξεων (εγκλημάτων)

Σύμφωνα με το άρθρο 18 του Ποινικού Κώδικα κάθε πράξη που τιμωρείται με τη ποινή του θανάτου ή της κάθειρξης είναι κακούργημα, κάθε πράξη που τιμωρείται με φυλάκιση ή με χρηματική ποινή ή με περιορισμό σε ειδικό κατάστημα κράτησης νέων είναι πλημμέλημα και κάθε πράξη που τιμωρείται με κράτηση ή πρόστιμο είναι πταίσμα.

Τα πταίσματα είναι τα ελαφρότερα εγκλήματα και τιμωρούνται, βάσει νόμου, με την ποινή της κράτησης ή το πρόστιμο, όπως π.χ. η ρύπανση (428 Π.Κ.), η παραμέληση της αναγγελίας ανεύρεσης νεκρού (442 Π.Κ.), η άρνηση αποδοχής νομισμάτων (452 Π.Κ.) και η διατάραξη ησυχίας (417 Π.Κ.).

Τα πλημμελήματα είναι μεσαίας βαρύτητας εγκλήματα και τιμωρούνται, βάσει νόμου, με φυλάκιση και χρηματική ποινή, όπως π.χ. η απειλή (333 Π.Κ.), η πλαστογραφία (216 Π.Κ.), η δωροδοκία (235 Π.Κ.), η κλοπή (372 Π.Κ.), η εξύβριση (361 Π.Κ.) και η φθορά ξένης ιδιοκτησίας (381 Π.Κ.), καθώς και με περιορισμό σε σωφρονιστικό κατάστημα όπως π.χ. τα εγκλήματα εφήβων ηλικίας από 13 έως 17 χρόνων.

Τα κακουργήματα είναι εξαιρετικής βαρύτητας εγκλήματα και τιμωρούνται, βάσει νόμου, με ποινή ισόβιας ή πρόσκαιρης κάθειρξης, όπως π.χ. η ανθρωποκτονία με πρόθεση (381 Π.Κ.), ο βιασμός (336 Π.Κ.), η έσχατη προδοσία (134 Π.Κ.) και η ληστεία (380 Π.Κ.). Η τιμωρία της θανατικής ποινής ίσχυε παλαιότερα και καταργήθηκε το 1993.

Ο Ποινικός Κώδικας χωρίζεται σε γενικό και ειδικό μέρος. Το γενικό μέρος αποτελεί την βασική γνώση για τη θεμελίωση των εγκλημάτων, ενώ στο ειδικό μέρος ταξινομούνται τα εγκλήματα σε κατηγορίες, σύμφωνα με τη προσβολή του έννομου αγαθού.

Ενδεικτικά, οι ομάδες διαχωρισμού των εγκλημάτων στο ειδικό μέρος είναι:

- Προσβολές του πολιτεύματος, π.χ. εσχάτη προδοσία (134 Π.Κ.).
- Προδοσία της χώρας, π.χ. στρατιωτική υπηρεσία στον εχθρό (143 Π.Κ.) και κατασκοπεία (148 Π.Κ.).
- Εγκλήματα κατά ξένων κρατών, π.χ. προσβολή κατά ξένου κράτους και του αρχηγού του (153 Π.Κ.).
- Εγκλήματα κατά της ελεύθερης άσκησης των πολιτικών δικαιωμάτων, π.χ. δωροδοκία (159 Π.Κ.) και νόθευση της εκλογής (164 Π.Κ.).
- Προσβολές κατά της πολιτειακής εξουσίας, π.χ. αντίσταση (167 Π.Κ.) και απείθεια (169 Π.Κ.).
- Επιβουλή της δημόσιας τάξης, π.χ. εγκληματική οργάνωση (187 Π.Κ.) και διατάραξη της κοινής ειρήνης (189 Π.Κ.).
- Επιβουλή της θρησκευτικής ειρήνης, π.χ. κακόβουλη βλασφημία (198 Π.Κ.) και καθύβριση θρησκευμάτων (199 Π.Κ.).
- Εγκλήματα που ανάγονται στην στρατιωτική υπηρεσία και στην υποχρέωση για στράτευση, π.χ. διέγερση αυτών που έχουν υποχρέωση στρατιωτικής υπηρεσίας (202 Π.Κ.) και παράνομη αποδημία (205 Π.Κ.).
- Εγκλήματα σχετικά με το νόμισμα, π.χ. παραχάραξη (207 Π.Κ.), κυκλοφορία παραχαραγμένων νομισμάτων (208 Π.Κ.) και παράνομη έκδοση ανώνυμων ομολογιών (215 Π.Κ.).
- Εγκλήματα σχετικά με τα υπομνήματα, π.χ. πλαστογραφία (216α Π.Κ.), πλαστογραφία και κατάχρηση ενσήμων (218 Π.Κ.) και υφαρπαγή ψευδούς βεβαίωσης (220β Π.Κ.).
- Εγκλήματα σχετικά με την απονομή της δικαιοσύνης, π.χ. ψευδορκία (224 Π.Κ.), παραπλάνηση σε ψευδορκία (228 Π.Κ.) και ψευδής καταμήνυση (229 Π.Κ.).
- Εγκλήματα σχετικά με την υπηρεσία, π.χ. παθητική δωροδοκία (235 Π.Κ.), ενεργητική δωροδοκία (236 Π.Κ.), κατάχρηση εξουσίας (239 Π.Κ.) και παραβίαση οικιακού ασύλου (241 Π.Κ.).
- Κοινώς επικίνδυνα εγκλήματα, π.χ. εμπρησμός (264 Π.Κ.), πλημμύρα (268 Π.Κ.), έκρηξη (270 Π.Κ.) και πρόκληση ναυαγίου (277 Π.Κ.).
- Εγκλήματα κατά της ασφάλειας των συγκοινωνιών, των τηλεφωνικών επικοινωνιών και κατά των κοινωφελών εγκαταστάσεων, π.χ. διατάραξη της ασφάλειας των συγκοινωνιών (290 Π.Κ.), παρακώλυση συγκοινωνιών (292 Π.Κ.) και παύση εργασίας (294 Π.Κ.).

- Εγκλήματα κατά της ζωής, π.χ. ανθρωποκτονία με πρόθεση (299 Π.Κ.), ανθρωποκτονία με συναίνεση (300 Π.Κ.), παιδοκτονία (303 Π.Κ.) και έκθεση (306 Π.Κ.).
- Σωματικές βλάβες, π.χ. απλή σωματική βλάβη (308 Π.Κ.), επικίνδυνη σωματική βλάβη (309 Π.Κ.) και συμπλοκή (313 Π.Κ.).
- Μονομαχία, π.χ. πρόκληση σε μονομαχία (316 Π.Κ.) και διέγερση σε μονομαχία (320 Π.Κ.).
- Εγκλήματα κατά της προσωπικής ελευθερίας, π.χ. αρπαγή (322 Π.Κ.), εμπόριο δούλων (323 Π.Κ.), ακούσια απαγωγή (327 Π.Κ.), εκούσια απαγωγή (328 Π.Κ.), παράνομη βία (330 Π.Κ.), αυτοδικία (331 Π.Κ.), απειλή (333 Π.Κ.) και διατάραξη οικιακής ειρήνης (334 Π.Κ.).
- Εγκλήματα κατά της γενετήσιας ελευθερίας και εγκλήματα οικονομικής εκμετάλλευσης της γενετήσιας ζωής, π.χ. βιασμός (336 Π.Κ.), έγκληση (344 Π.Κ.), αιμομιξία (345 Π.Κ.), πορνογραφία ανηλίκων (348α Π.Κ.), μαστροπεία (349 Π.Κ.) και σωματεμπορία (351 Π.Κ.).
- Εγκλήματα σχετικά με το γάμο και την οικογένεια, π.χ. απάτη σχετική με γάμο (355 Π.Κ.) και μοιχεία (357 Π.Κ.).
- Εγκλήματα κατά της τιμής, π.χ. εξύβριση (361 Π.Κ.) και δυσφήμηση (362 Π.Κ.).
- Παραβίαση του απορρήτου, π.χ. παραβίαση του απορρήτου των επιστολών (370 Π.Κ.)
- Εγκλήματα κατά της ιδιοκτησίας, π.χ. κλοπή (372 Π.Κ.), υπεξαίρεση (375 Π.Κ.), ληστεία (380 Π.Κ.), φθορά ξένης ιδιοκτησίας (381 Π.Κ.) και αφαιρέσεις (378 Π.Κ.).
- Εγκλήματα κατά των περιουσιακών δικαιωμάτων, π.χ. εκβίαση (385 Π.Κ.), απάτη (386 Π.Κ.), απάτη ευτελούς αξίας (387 Π.Κ.), απιστία (390 Π.Κ.) και δόλια αποδοχή παροχών (392 Π.Κ.).
- Επαιτεία και αλητεία, π.χ. επαιτεία (407 Π.Κ.) και αλητεία (408 Π.Κ.).
- Πταίσματα, π.χ. ευθύνη για πταίσματα άλλων (411 Π.Κ.), παράνομη άσκηση επαγγέλματος (414 Π.Κ.), πρόκληση ανησυχίας (416 Π.Κ.), διατάραξη ησυχίας (417 Π.Κ.) και υπέρβαση νυχτερινή ώρας (418 Π.Κ.).
- Τελικές διατάξεις, π.χ. παραβάσεις διοικητικών διατάξεων (458 Π.Κ.) και παράβαση αστυνομικών διατάξεων (459 Π.Κ.).

(Ραφτόπουλος, 1996)

Η σύνθεση μίας εγκληματικής πράξης μπορεί να αποτελείται από μία μεγάλη γκάμα ενεργειών και ποικίλει ανάλογα με την βαρύτητα της (ρύπανση περιβάλλοντος,

ανθρωποκτονία) και με την παγκόσμια επιρροή (διεθνές έγκλημα). Στον πίνακα 2-1 συνοψίζονται οι διάφοροι τύποι εγκλημάτων και η επιρροή τους σε διεθνές και εγχώριο βαθμό, έχοντας υπόψη τους διεθνείς και ανά χώρα επικρατώντας νόμους.

Πίνακας 2-1: Τύποι εγκλημάτων σε διαφορετικά επίπεδα.

Τύπος	Τοπικό επίπεδο επιβολής του νόμου	Επίπεδο εθνικού επιπέδου ασφαλείας
Παραβάσεις της κυκλοφορίας	Οδήγηση υπό την επήρεια ουσιών, θανατηφόρο / προσωπικό τραυματισμό / υλική ζημιά τροχαίο ατύχημα, οδικός βαθμός	
Σεξουαλικό έγκλημα	Σεξουαλικά αδικήματα, σεξουαλικές επιθέσεις, κακοποίηση παιδιών	Οργανωμένη πορνεία
Κλοπή	Ληστεία, διάρρηξη, κλοπή, κλοπή αυτοκινήτου, κλεμμένη ιδιοκτησία	Κλοπή εθνικών μυστικών ή πληροφοριών για όπλα
Απάτη	Πλαστογραφία και παραποίηση, απάτες, υπεξαίρεση, εξαπάτηση ταυτότητας	Διακρατική νομιμοποίηση εσόδων από παράνομες δραστηριότητες, απάτη ταυτότητας, διακρατική οικονομική απάτη
Εμπρησμός	Πυρκαγιά σε κτίρια, διαμερίσματα	
Αδικήματα συμμοριών / ναρκωτικών	Αδικήματα ναρκωτικών (πωλήσεις ή κατοχή)	Διακρατική διακίνηση ναρκωτικών
Βίαιη εγκληματικότητα	Ποινική ανθρωποκτονία, ένοπλη ληστεία, επιθετική επίθεση, άλλες επιθέσεις	Τρομοκρατία (βιο-τρομοκρατία, βομβιστική επίθεση, αεροπειρατεία κ.λπ.)
Ηλεκτρονικό έγκλημα	Διαδικτυακές απάτες, παράνομες συναλλαγές, διείσδυση σε δίκτυο / πειρατεία, διάδοση των ιών, εγκλήματα μίσους, πειρατεία στον κυβερνοχώρο, πορνογραφία στον κυβερνοχώρο, κυβερνο-τρομοκρατία, κλοπή εμπιστευτικών πληροφοριών	

Πηγή: Chen, 2003.

2.2 Εξόρυξη γνώσης και πληροφορίας από Βάσεις Δεδομένων

2.2.1 Εισαγωγή

Η τεχνολογική εξέλιξη έχει συντελέσει στην εύκολη συλλογή, αποθήκευση και ανάλυση μεγάλου όγκου δεδομένων. Πληροφορίες και δεδομένα αποθηκεύονται καθημερινά σε αποθήκες πληροφοριών από όλους τους κλάδους της επιστήμης και του επιχειρηματικού κόσμου, ώστε μετέπειτα να μπορέσουν με την επεξεργασία τους να αποκτήσουν αξία για τον σκοπό που συλλέχθηκαν. Με τη πάροδο του χρόνου, ο όγκος των δεδομένων πολλαπλασιάζεται και η χρήση παραδοσιακών τεχνικών και εργαλείων ανάλυσης καθίσταται δύσχρηστη και πολλές φορές αδύνατη να εκπληρώσει τον στόχο του χρήστη. Στη προκειμένη περίπτωση, εμφανίζονται καινούριοι και σύγχρονοι αλγόριθμοι επεξεργασίας μεγάλου όγκου δεδομένων και σε συνδυασμό τόσο με τις παραδοσιακές μεθόδους ανάλυσης όσο και με τις νέες, μας δίνουν αξιόλογα αποτελέσματα.

Οι εφαρμογές που βρίσκει χρήση η εξόρυξη δεδομένων ποικίλλει. Η Διαχείριση της Αγοράς (Market Management) είναι ένας τομέας όπου χρησιμοποιείται το data mining για κατευθυνόμενες διαφημιστικές εκστρατείες και πώλησης. Αναλύοντας τα δεδομένα, μπορούν να δημιουργηθούν μοντέλα πελατών με κοινά χαρακτηριστικά ενδιαφέροντα και καταναλωτικές συνήθειες (τμηματοποίηση και κατηγοριοποίηση πελατών). Επομένως, μπορούν να εξαχθούν συμπεράσματα για την συχνότητα αγοράς και εμπιστοσύνης του πελάτη στην επιχείρηση και την αγορά διασταυρωμένων προϊόντων (η πώληση ενός προϊόντος μπορεί συνδυάζεται με κάποιο άλλο) και το πώς συμπεριφέρονται οι καταναλωτές (προφίλ) στην αγορά προϊόντων σε σχέση με το τι επιλέγουν να αγοράσουν, πότε αλλά και με ποιον τρόπο, επιτυγχάνοντας στοχευμένη προσέγγιση πελατών.

Εκτός από τις συμβατικές επιχειρήσεις υπάρχουν και αυτές του ηλεκτρονικού εμπορίου, όπου η χρήση του data mining γίνεται σε μεγαλύτερο όγκο δεδομένων και βρίσκει εφαρμογή σε περισσότερες περιπτώσεις. Μπορούν να εξαχθούν πληροφορίες που έχουν σχέση με το διαδικτυακό προφίλ του πελάτη, όπως για τις προτιμήσεις στην επισκεψιμότητα ιστοσελίδων, το ιστορικό αναζήτησης, τη διαδρομή περιήγησης και επιλογής ενός προϊόντος, καθώς και άλλα που βοηθούν στην εξατομίκευση του πελάτη.

Άλλος τομέας εφαρμογής του data mining είναι η Διαχείριση του Ρίσκου (Risk Management) που σχετίζεται με την ασφάλεια των επενδύσεων μιας επιχείρησης και τις απειλές που αντιμετωπίζει από τους κινδύνους που εμφανίζονται στο εσωτερικό και εξωτερικό της περιβάλλον. Οι τράπεζες χρησιμοποιούν την εξόρυξη δεδομένων για διαφήμιση, προώθηση πωλήσεων, εντοπισμό απάτης, αλλά και για τη διαχείριση του κινδύνου. Διατηρούνται αναλυτικά στοιχεία και πληροφορίες των πελατών της κάθε τράπεζας, ώστε να ελέγχεται, κατά το δυνατόν, ο πιστωτικός κίνδυνος (π.χ. ο δανειολήπτης αδυνατεί να αποπληρώσει τα χρέη του). Επίσης, οι τράπεζες αντιμετωπίζουν και άλλους κινδύνους, όπως η διακύμανση επιτοκίων και ισοτιμιών των νομισμάτων, καθώς και η αδυναμία ρευστότητας. Εκτός των τραπεζικών οργανισμών, χρήση των αποτελεσμάτων της εξόρυξης δεδομένων για εντοπισμό του κινδύνου γίνεται και από τις ασφαλιστικές εταιρίες, σχετικά με τις αποζημιώσεις πελατών, τον καθορισμό ύψους αποζημιώσεων και ασφαλίστρων, αλλά κι από τα χρηματιστήρια, όσον αφορά τις διακυμάνσεις των χρηματιστηριακών δεικτών και των τιμών των μετοχών.

Η εξόρυξη δεδομένων εφαρμόζεται και για τη Διαχείριση της Απάτης (Fraud Management) όπου βρίσκει ανταπόκριση από τις τράπεζες, τις ασφαλιστικές εταιρείες, την ελεγκτική (λογιστική) και τις τηλεπικοινωνίες, ώστε να αποφευχθεί και αντιμετωπιστεί επιτυχώς η απάτη. Γενικότερα, γίνεται η προσπάθεια μέσα από την εξόρυξη δεδομένων της δημιουργίας μοντέλων συμπεριφοράς για εντοπισμό αυτών που έχουν δόλιο σκοπό για οικονομικά εγκλήματα, καθώς και παρόμοιων παλαιότερων συμπεριφορών. Επομένως, εάν παρεκκλίνει κάποια ενέργεια από το πρότυπο συναλλαγών του πελάτη, τότε διενεργείται έλεγχος.

Ένας άλλος χώρος εφαρμογής των εξελιγμένων τεχνικών data mining είναι οι βάσεις κειμένων μεγάλου όγκου αδόμητων δεδομένων και ονομάζεται text mining. Η χρήση των παραδοσιακών μεθόδων data mining στην αναζήτηση σε βάσεις δομημένων δεδομένων περιορίζεται στη ταυτοποίηση των λέξεων, ενώ με την χρήση του text mining εντοπίζονται οι σύνδεσμοι που υφίστανται σε σχέση με την στοχευμένη αναζήτηση και παρέχεται περισσότερη πληροφορία. Χρησιμοποιώντας το text mining, δίνεται η δυνατότητα αξιοποίησης των αδόμητων δεδομένων, τα οποία είναι σε όγκο πολύ περισσότερα από τα δομημένα και της απόκτησης χρήσιμης πληροφορίας.

Η ανάπτυξη των κοινωνικών μέσων στον παγκόσμιο ιστό, τα οποία αποτελούν το βασικό συστατικό της κοινωνικής δικτύωσης, έδωσε διαφορετική κατεύθυνση στη χρήση του data mining. Τα κοινωνικά μέσα καθημερινά χρησιμοποιούνται όλο και περισσότερο και αυξάνονται, παρέχοντας τεράστιο όγκο δεδομένων. Αυτή η ραγδαία

ανάπτυξη των κοινωνικών μέσων και γενικότερα του παγκόσμιου ιστού που παρέχει πληθώρα δεδομένων και πληροφοριών καθιστά, πολλές φορές, τον αναζητητή ανίκανο να οργανώσει και να δομήσει τα αποτελέσματα της αναζήτησής του.

Ενδεικτικά αναφέρονται κάποια κοινωνικά δίκτυα:

- Τα κοινωνικά νέα και συστάσεις (social news and recommendations) π.χ. Digg, Flipboard και Reddit.
- Οι καταγεγραμμένοι κοινωνικοί ιστότοποι (social bookmarking sites) π.χ. Slashdot, Reddit, Squidoo, Stumbleupon, Delicious και Digg.
- Οι μικρο-blogging υπηρεσίες (micro blogging services) π.χ. το Twitter και Tumblr.
- Τα συστήματα blogging (blogging systems): π.χ. το WordPress, το Blogger, το Tumblr, το Medium, το Quora, το Google+, το Facebook Notes, το Ghost, το Squarespace και Typepad.
- Τα κοινωνικά δίκτυα (social networks) π.χ. το Facebook, το Myspace και το LinkedIn.
- Οι κοινότητες διαμοιρασμού (social sharing): π.χ. το YouTube, το Meetup, το Lastfm και το Flickr.
- Τα wikis: (ο όρος "wiki" σημαίνει "γρήγορα" στα Χαβανέζικα) π.χ. το MediaWiki, το Meta-Wiki, το Wikibooks, το Wikidata, το Wikinews, το Wikipedia, το Docuwiki και Pdwiki.

Η ανάγκη εξόρυξης δεδομένων και γνώσης από τον Παγκόσμιο Ιστό δημιούργησε το web analytics και συγκεκριμένα από την χρήση των κοινωνικών μέσων την Ανάλυση Κοινωνικών Δικτύων (Social Network Analysis, SNA). Με τη χρήση SNA δίνεται η δυνατότητα καλύτερης διαχείρισης των δεδομένων περιεχομένου δομής, χρήσης και προφίλ του χρήστη, που βρίσκονται στον παγκόσμιο ιστό. Τα δεδομένα περιεχομένου είναι κατάλληλα δομημένα (π.χ. δομημένα δεδομένα, εικόνες, κείμενα), τα δεδομένα δομής σκιαγραφούν τον γράφο σύνδεσης ιστοσελίδων με υπερσυνδέσεις, τα δεδομένα χρήσης αναφέρονται στο τρόπο χρήσης ενός δικτυακού τόπου (π.χ. IP, πρωτύτερη επίσκεψη άλλων δικτυακών τόπων, συνολική διαδρομή, ώρα και μέρα αναζήτησης) και τα δεδομένα προφίλ του χρήστη (π.χ. δημογραφικά στοιχεία, προτιμήσεις και ενδιαφέροντα χρήστη) παρέχουν πληροφορίες για τον χρήστη του διαδικτυακού τόπου (Χαλκίδη & Βαζιργιάννης, 2005).

Η εξόρυξη δεδομένων βρίσκει εφαρμογή σε πολλές περιπτώσεις της καθημερινότητας και βοηθάει τον άνθρωπο να πάρει τις κατάλληλες αποφάσεις. Εκτός από τη χρήση του data mining στον εμπορικό και επιστημονικό κόσμο, το συναντάμε και σε διάφορους φορείς και οργανισμούς του κράτους (Αστυνομία, Λιμενικό, Στρατός,

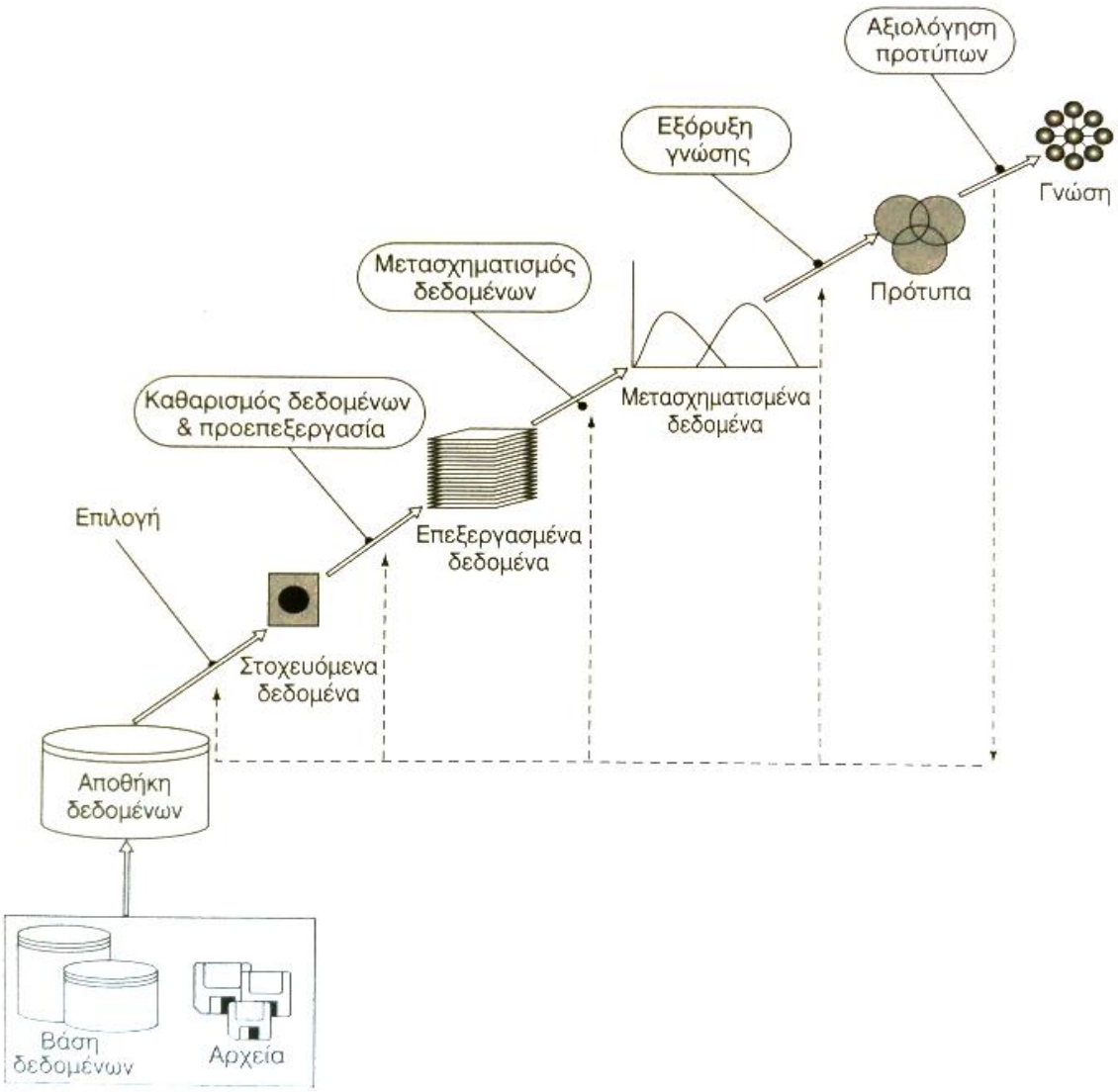
Περιφέρειες κτλ.) όπου τους βοηθάει να χαράξουν την πολιτική τους και να πάρουν αποφάσεις σχετικά με την δράση τους.

Η εκάστοτε Αστυνομική Αρχή διατηρεί στη κατοχή της έναν μεγάλο όγκο ποινικών δεδομένων που καταγράφει καθημερινά και μπορεί στα πλαίσια της ποινικής έρευνας να συγκεντρώσει ακόμη περισσότερα από άλλες υπηρεσίες (π.χ. Ιατροδικαστική, Δικαστική Αρχή κτλ.). Ο όγκος αυτός των δεδομένων περιέχει πολλές πληροφορίες που είναι χρήσιμες για τη πρόληψη και καταστολή του εγκλήματος. Η ανάλυση των εγκληματικών δεδομένων επιφέρει συνδυασμούς κοινωνικο-οικονομικών και κοινωνικο-δημογραφικών μοτίβων σε απεικονίσεις (γραφήματα), επιτρέποντας τη χρήση των εξαγόμενων πληροφοριών. Η εφαρμογή της εξόρυξης δεδομένων στην ανάλυση των εγκληματικών αδικημάτων από την αστυνομία, βοηθάει στην ανακάλυψη μοτίβων εγκληματικής συμπεριφοράς για το που (τόπο), πότε (χρόνο) και γιατί (αίτιο) συγκεκριμένα εγκλήματα πιθανόν να συμβούν. Οι κανόνες συσχέτισης, κατά την εξόρυξη δεδομένων, παρέχουν πληροφορίες σχέσεων μεταξύ των εγκληματικών χαρακτηριστικών και μπορούν να καθοδηγήσουν την αστυνομία να αφουγκραστεί την εγκληματική τάση και να προσδιορίσει περιοχές εμφάνισης εγκληματικότητας, ώστε να την οδηγήσει στο πιθανό εντοπισμό μελλοντικών εγκληματικών ενεργειών.

2.2.2 Ορισμοί και Έννοιες

Η χρήση μεγάλων αποθηκών δεδομένων για την εξόρυξη γνώσης και γενικότερα η διαδικασία που ακολουθείται ονομάζεται «Ανακάλυψη Γνώσης από Βάσεις Δεδομένων» (Knowledge Discovery in Databases, KDD) με συνώνυμη χρήση του όρου «Εξόρυξη Δεδομένων» (Data Mining), ο οποίος, κάποιες φορές, χρησιμοποιείται για το προσδιορισμό των τεχνικών ανάλυσης και εξόρυξης γνώσης από τις βάσεις δεδομένων.

Ο ορισμός που δόθηκε από τους Frawley, Piatesky-Shaphiro & Matheus το 1991 για την ανακάλυψη γνώσης από βάσεις δεδομένων είναι ότι: «Η ανακάλυψη γνώσης από βάσεις δεδομένων είναι η ντετερμινιστική διαδικασία αναγνώρισης έγκυρων καινοτόμων, ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων στα δεδομένα» (Χαλκίδη & Βαζιργιάννης, 2005).



Σχήμα 2-1: Τα Βήματα της διαδικασίας Ανακάλυψης Γνώσης από Βάσεις Δεδομένων.
Πηγή: Χαλκίδη & Βαζιργιάννης, 2005.

«Ο όρος εξόρυξη δεδομένων (data mining) είναι αυτός που έχει επικρατήσει και χαρακτηρίζει τη διαδικασία της εύρεσης δομών γνώσεις, οι οποίες περιγράφουν με ακρίβεια μεγάλα σύνολα πρωτογενών δεδομένων. Οι δομές αυτές αναδεικνύουν γνώση (συσχετίσεις ή κανόνες) που είναι κρυμμένη μέσα στα δεδομένα και δεν μπορεί να εξαχθεί από τον άνθρωπο-χρήστη της βάσης δεδομένων με ¨γυμνό¨ μάτι» (Χαλκίδη & Βαζιργιάννης, 2005).

Άλλοι ορισμοί που δοθήκαν για την Εξόρυξη Δεδομένων είναι:

- «Η Εξόρυξη Δεδομένων (Data Mining) ορίζεται ως η διαδικασία ανακάλυψης προτύπων μέσα από δεδομένα, δίνοντας έτσι έμφαση στη διάσταση της Μηχανικής Μάθησης (Witten & Frank, 2000) και

- Η Εξόρυξη Δεδομένων συνίσταται στην ανακάλυψη ή ¨εξόρυξη¨ γνώσης από μεγάλους όγκους δεδομένων (Han & Kamber, 2001)», (Κύρκος, 2015).

Για την καλύτερη κατανόηση των ανωτέρω ορισμών περιγράφονται βασικές έννοιες των όρων που χρησιμοποιούνται:

Δεδομένα: «Είναι:

α) πληροφορία συχνά με τη μορφή των γεγονότων ή σχημάτων τα οποία αποκτώνται από πειράματα ή έρευνες αγοράς κα ι τα οποία χρησιμοποιούνται ως βάση για υπολογισμούς ή σχεδία συμπερασμάτων.

β) πληροφορία με τη μορφή αριθμών, κειμένων, εικόνων και ήχων σε μορφή κατάλληλη για αποθήκευση ή επεξεργασία από υπολογιστές.

γ) η συλλογή ακατέργαστων δεδομένων, όπως για παράδειγμα οι απαντήσεις μιας έρευνας αγοράς. Όταν τα δεδομένα αυτά τύχουν επεξεργασίας τότε μετατρέπονται σε πληροφορία.

δ) σύμφωνα με τους Ackoff (1989) και Bellinger et. al. (2009), σύμβολα και αναπαριστούν ένα γεγονός ή αναφορά σε ένα γεγονός, χωρίς συσχέτιση με άλλα πράγματα». (Ματσατσίνης, 2010).

Πληροφορία: «Είναι:

α) συγκεκριμένη-σαφής γνώση αποκτηθείσα ή προμηθευθείσα γύρω από κάτι ή κάποιον.

β) τα συλλεγμένα γεγονότα ή δεδομένα γύρω από ένα συγκεκριμένο θέμα.

γ) η επικοινωνία γεγονότων και γνώσεις.

δ) δεδομένα υπολογιστών τα οποία έχουν οργανωθεί και παρουσιάζονται με συστηματικό τρόπο για να κάνουν ξεκάθαρο το βασικό μήνυμα.

ε) δεδομένα που έχουν επεξεργαστεί σε μορφή κατανοητή από τον δέκτη και έχουν πραγματική κατανοητή αξία για τις τρέχουσες και μελλοντικές αποφάσεις.

στ) σύμφωνα με τους Ackoff (1989) και Bellinger et. al. (2009), δεδομένα τα οποία έχουν τύχει επεξεργασίας για να γίνουν χρήσιμα και να μπορούν να δώσουν απαντήσεις σε ερωτήσεις της μορφής ¨ποιος¨ ¨τι¨ ¨που¨ και ¨πότε¨». (Ματσατσίνης, 2010).

Γνώση: «Είναι:

α) γενική ενημέρωση ή κατοχή πληροφορίας, γεγονότων, ιδεών, αληθειών ή αρχών.

β) σαφής ενημέρωση ή σαφής πληροφόρηση για μία κατάσταση ή ένα γεγονός.

γ) όλες οι πληροφορίες στα γεγονότα, οι αλήθειες και οι αρχές, οι οποίες αποκτώνται μέσω της μάθησης μέσα στο χρόνο.

δ) εξοικείωση ή κατανόηση, η οποία προστίθεται μέσω της εμπειρίας ή της μελέτης» (Ματσατσίνης, 2010).

Βάση δεδομένων: «Βάση δεδομένων ή τράπεζα πληροφοριών (Data Base ή Data Bank) είναι μία συλλογή δεδομένων αποθηκευμένων σε ηλεκτρονική μορφή, οργανωμένη κατά τέτοιο τρόπο ώστε οι διάφορες εφαρμογές να μπορούν μεν εύκολα να τη χρησιμοποιούν και να την ενημερώνουν, αλλά και που οι ίδιες δεν καθορίζουν αναγκαστικά το σχεδιασμό ή το περιεχόμενό της. Τα δεδομένα των βάσεων δεδομένων αποθηκεύονται σε δίσκους δισκέτες ή σε άλλα μαγνητικά μέσα περιλαμβάνει ακόμα έναν αριθμό από προγράμματα εφαρμογών που επεξεργάζονται τα δεδομένα» (Ματσατσίνης, 2010).

Διαχείριση δεδομένων: Η διαχείριση δεδομένων ασχολείται με την συλλογή, την επαλήθευση, την οργάνωση, την αποθήκευση, την ασφάλεια, την ανάκτηση και την συντήρηση των δεδομένων. Συγκεκριμένα, στην συλλογή γίνεται η συγκέντρωση και ανάλογη επεξεργασία της μορφής των δεδομένων για να μπορούν να είναι αξιοποιήσιμα από το σύστημα στο οποίο θα εισαχθούν και στην επαλήθευση γίνεται ο έλεγχος για την πληρότητα και ορθότητα των δεδομένων. Στην οργάνωση τα δεδομένα ετοιμάζονται για την αποθήκευση τους, με τέτοιο τρόπο, ώστε να ικανοποιούν τους χρήστες τους και στην αποθήκευση αποθηκεύονται στις περιφερειακές μαγνητικές μονάδες, ώστε να μπορούν εν συνεχεία να χρησιμοποιηθούν. Κατά την ασφάλεια γίνεται η προστασία των δεδομένων από κακόβουλα λογισμικά, καταστροφές, λανθασμένες ενέργειες ή όποια άλλη ενέργεια μη εξουσιοδοτημένων χρηστών, ενώ στην ανάκτηση μπορούν οι εξουσιοδοτημένοι χρήστες να ανακτήσουν τα δεδομένα. Τέλος, στη συντήρηση γίνεται η αναδιοργάνωση των αποθηκευμένων αρχείων, ώστε να υπάρχει τάξη και απαλλαγή μη χρήσιμων υπολειμμάτων που προκύπτουν από τις νέες εγγραφές, διαγραφές και μεταβολές των χρηστών (Ματσατσίνης, 2010).

Πρότυπο: «Είναι μια έκφραση Ε σε μία γλώσσα L η οποία περιγράφει ένα υποσύνολο δεδομένων F_E c F εκμεταλλευόμενο κοινές ιδιότητες των δεδομένων του» (Χαλκίδη & Βαζιργιάννης, 2005).

Διαδικασία KDD: Είναι διαδικασία πολλών βημάτων με την προ-επεξεργασία των δεδομένων (καθαρισμός, ολοκλήρωση, επιλογή δεδομένων εξόρυξης, μετασχηματισμός δεδομένων), την εξόρυξη των δεδομένων (αλγόριθμοι

εξόρυξης, αναζήτηση προτύπων, εύρεση πληροφορίας) και την αποτίμηση και αναπαράσταση των αποτελεσμάτων (επιλογή χρήσιμης και νέας πληροφορίας) Τα στάδια KDD φαίνονται στο σχήμα 2-1 και μπορούν να εφαρμοστούν επαναληπτικά (Νανόπουλος & Μανωλόπουλος, 2008).

Εγκυρότητα: Είναι η βεβαιότητα έως ένα βαθμό της συνέπειας σε καινούρια δεδομένα (Χαλκίδη & Βαζιργιάννης, 2005).

Ενδεχομένως χρήσιμο: Είναι η δυνατότητα να μπορούν τα εξαγόμενα πρότυπα να χρησιμοποιηθούν για κάποιο σκοπό, όπως την λήψη αποφάσεων (Χαλκίδη & Βαζιργιάννης, 2005).

Εν τέλει κατανοητό: Είναι ο προσδιορισμός των προτύπων ώστε να κατανοηθούν και να φανούν χρήσιμα στους χρήστες (Χαλκίδη & Βαζιργιάννης, 2005).

Ντετερμινιστική διαδικασία: Είναι η διαδικασία όπου γνωρίζουμε τις αρχικές συνθήκες και με μαθηματικές διαδικασίες είναι εφικτός ο καθορισμός κάθε μετέπειτα σταδίου ενός συστήματος.

Μηχανική Μάθηση: Είναι ένα μέρος της Τεχνικής Νοημοσύνης, η οποία περιλαμβάνει τεχνικές εξόρυξης γνώσης, το οποίο διερευνά τη δημιουργία προγραμμάτων που να μαθαίνουν και στην εξόρυξη δεδομένων χρησιμεύει στην πρόβλεψη και κατηγοριοποίηση (Dunham, 2004).

Η Εξόρυξη Δεδομένων είναι αποτέλεσμα συνεργασίας επιστημόνων και ερευνητών διαφορετικών επιστημονικών πεδίων. Σκοπός της συνεργασίας αυτής ήταν η εξέλιξη και δημιουργία αποτελεσματικότερων εργαλείων διαχείρισης μεγάλου όγκου δεδομένων, τα οποία μπορεί να διέφεραν ως προς τον τύπο τους. Επιστήμες όπως η στατιστική (δειγματοληψία, έλεγχος υποθέσεων, εκτίμηση), η τεχνική νοημοσύνη, η αναγνώριση προτύπων και η μηχανική μάθηση (αλγόριθμοι αναζήτησης, θεωρίες μάθησης, τεχνικές μοντελοποίησης), μέθοδοι όπως η βελτιστοποίηση, οπτικοποίηση, ανάκτηση πληροφοριών, επεξεργασία σήματος, η τεχνολογία βάσεων δεδομένων, καθώς και οι τεχνικές παράλληλων και κατανεμημένων υπολογισμών αποτέλεσαν με την συμβολή τους το μείγμα εξέλιξης της εξόρυξης δεδομένων (Tan, Steinbach & Kumar, 2015).

2.2.3 Εργασίες και Τεχνικές Εξόρυξης Δεδομένων

Η εφαρμογή της εξόρυξης δεδομένων γίνεται με διαφορετικούς τρόπους, ανάλογα με το επιθυμητό αποτέλεσμα, γι' αυτό και τα δεδομένα που χρησιμοποιούνται, κάθε φορά, διαμορφώνονται αντίστοιχα ως προς τη δομή τους και σε συμβατότητα με τον επιλεγμένο αλγόριθμο. Οι τεχνικές εξόρυξης δεδομένων που εφαρμόζονται, κατά βάση, έχουν ως στόχο τη περιγραφή (περιγραφικές εργασίες) και τη πρόβλεψη (προγνωστικές εργασίες) από μεγάλο όγκο δεδομένων. Στην διαδικασία της περιγραφής διερευνούνται οι σχέσεις των δεδομένων και εξάγονται συμπεράσματα από τα αποτελέσματα, ενώ στη πρόβλεψη επιτυγχάνεται η τιμή ενός χαρακτηριστικού (εξαρτημένη μεταβλητή) που είναι απόρροια από την χρήση των υπαρχόντων άλλων (ανεξάρτητες μεταβλητές). Στην συνέχεια περιγράφονται οι βασικές εργασίες εξόρυξης δεδομένων, καθώς και τεχνικές (αλγορίθμους) που χρησιμοποιούν.

2.2.3.1 Κατηγοριοποίηση

Μία από τις πιο γνωστές και βασικές εργασίες της εξόρυξης δεδομένων είναι η κατηγοριοποίηση, όπου συνήθως χρησιμοποιείται για πρόβλεψη και εκτίμηση. Με τη βοήθεια της στατιστικής των προτύπων και της μηχανικής μάθησης γίνεται προσπάθεια ανάκτησης και εξαγωγής πληροφορίας από μεγάλο όγκο δεδομένων. Σκοπός της κατηγοριοποίησης είναι η αντιστοίχιση ενός χαρακτηριστικού σε ένα εκ των υπαρχόντων συνόλων χαρακτηριστικών. Κατά τη διαδικασία της κατηγοριοποίησης δημιουργείται ένα μοντέλο από διάφορες κατηγορίες δεδομένων και τη χρήση ενός επιλεγμένου αλγορίθμου, ανάλογα με τον επιδιωκόμενο σκοπό. Τα αποτελέσματα της εκμάθησης του μοντέλου (κατηγοριοποιητής) από γνωστή κατηγορία δεδομένων αναπαρίστανται ως δέντρα αποφάσεων, μαθηματικοί τύποι ή κανόνες κατηγοριοποίησης. Κατόπιν της δημιουργίας του μοντέλου, γίνεται δοκιμή με δεδομένα εκπαίδευσης για να υπολογιστεί η ακρίβεια του, όπου συγκρίνονται τα αποτελέσματα κατηγοριοποίησης των δοκιμαστικών δεδομένων με τη πρόβλεψη κατηγορίας του μοντέλου. Εάν το ποσοστό σωστής κατηγοριοποίησης των δοκιμαστικών δεδομένων είναι αποδεκτό τότε το μοντέλο μπορεί να χρησιμοποιηθεί σε μετέπειτα δεδομένα άγνωστης κατηγοριοποίησης (Χαλκίδη & Βαζιργιάννης, 2005).

Υπάρχουν διάφοροι αλγόριθμοι που χρησιμοποιούνται κατά τη κατηγοριοποίηση και οι οποίοι ομαδοποιούνται ανάλογα με το τύπο τεχνικής που εφαρμόζεται.

<u>Αλγόριθμοι θεμελιωμένοι στην Στατιστική</u>:

❖ Παλινδρόμηση:

Η παλινδρόμηση εφαρμόζεται σε διάφορες περιπτώσεις, όπως για κατηγοριοποίηση και πρόβλεψη και σχετίζεται με την εκτίμηση των εξερχόμενων τιμών των χαρακτηριστικών (κατηγορίες), χρησιμοποιώντας τις τιμές εισόδου (βάση δεδομένων). Διαδικαστικά, κατά τη παλινδρόμηση ένα σύνολο δεδομένων αντιστοιχείται σε μία εξίσωση.

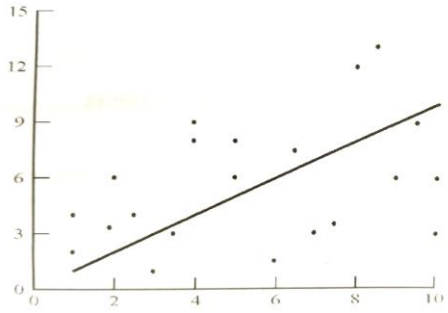
Στην γραμμική παλινδρόμηση χρησιμοποιείται ο τύπος:

$$y = c_0 + c_1 x_1 + \ldots + c_n x_n$$

Σχήμα 2-2: Τύπος γραμμικής παλινδρόμησης.
Πηγή: Dunham, 2004.

Όπου: C_0, C_1,…, C_n = συντελεστές παλινδρόμησης, X_1,… X_n = παράμετροι εισόδου και y = παράμετρος εξόδου.

Η εκτίμηση των εξερχόμενων τιμών γίνεται, ως συνέχεια της εξίσωσης, με τη κατάταξη των δεδομένων σε δύο περιοχές με μία ευθεία γραμμή (χώρος δύο διαστάσεων) που είναι το σημείο εξισορρόπησης ή διαίρεσης των δύο κατηγοριών. Υπάρχει όμως η περίπτωση τα δεδομένα να μην μπορούν να αναπαριστούν ένα γραμμικό μοντέλο ή αν αυτό συμβαίνει να είναι ανεπαρκές με λανθασμένα δεδομένα (θόρυβος) ή να μην είναι κάποια, εξ αυτών, συνηθισμένα (ακραίες τιμές, τιμές>1 & <0).



Σχήμα 2-3: Ανεπαρκής γραμμική παλινδρόμηση.
Πηγή: Dunham, 2004.

Η παλινδρόμηση χρησιμοποιείται στην κατηγοριοποίηση για την κατάταξη των κατηγοριοποιημένων δεδομένων σε περιοχές και για πρόβλεψη της εξερχόμενης τιμής κάποιας κατηγορίας.

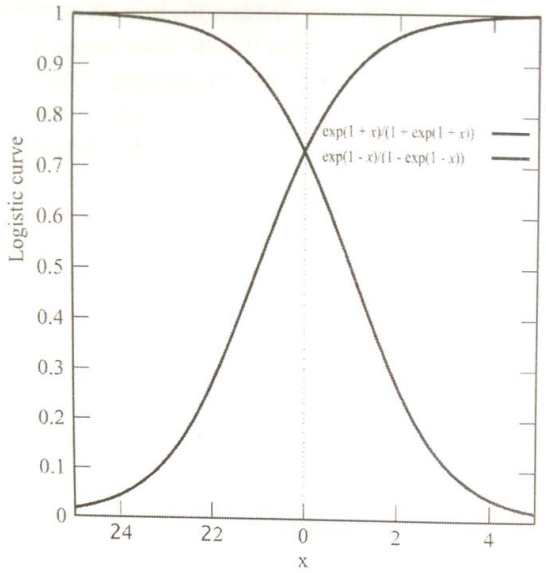
Διαφορετική τεχνική παλινδρόμησης από την γραμμική αποτελεί η λογιστική παλινδρόμηση που χρησιμοποιεί τον τύπο:

$$p = \frac{e^{(c_0 + c_1 x_1)}}{1 + e^{(c_0 + c_1 x_1)}}$$

Σχήμα 2-4: Τύπος λογιστικής καμπύλης μιας μεταβλητής.
Πηγή: Dunham, 2004.

Όπου: C_0, C_1,…, C_n = συντελεστές παλινδρόμησης, X_1,… X_n = παράμετροι εισόδου και p = πιθανότητα να ανήκει σε κάποια κατηγορία.

Όπως φαίνεται στο σχήμα 2-5 η λογιστική παλινδρόμηση δε χρησιμοποιεί ευθεία γραμμή αλλά καμπύλη τιμών μεταξύ 0 και 1, ώστε να υπολογιστεί η πιθανότητα συμμετοχής των δεδομένων σε κάποια εκ των δύο κατηγοριών, όπως και στην γραμμική παλινδρόμηση.
(Dunham, 2004)



Σχήμα 2-5: Λογιστική καμπύλη.
Πηγή: Dunham, 2004.

❖ Bayesian κατηγοριοποίηση

Στο κανόνα του Bayes για την πιθανότητα ικανοποίησης της εκάστοτε υπόθεσης από καθορισμένη πλειάδα (υπό συνθήκη πιθανότητα), βασίζεται μια άλλη τεχνική κατηγοριοποίησης που ονομάζεται Bayesian κατηγοριοποίηση. Στην προκειμένη περίπτωση, αναλύεται η συνεισφορά όλων των ανεξάρτητων χαρακτηριστικών και της συνέπειας τους στη πρόβλεψη.

Ο τύπος που χρησιμοποιείται σύμφωνα με την θεωρία του Bayes είναι ο εξής:

$$P(H|X) = \frac{P(H) * P(X|H)}{P(X)}$$

Σχήμα 2-6: Τύπος Πιθανότητας κατά Bayes.
Πηγή: Κύρκος, 2015.

Όπου P(H|X) = πιθανότητα επαλήθευσης H, ισχύοντας το γεγονός X, P(H) = εκ των προτέρων πιθανότητα ισχύς της υπόθεσης H, P(X) = εκ των προτέρων πιθανότητα να συμβεί το γεγονός X και P(H|X) = πιθανότητα να συμβεί το X, ισχύοντας η υπόθεση H (Κύρκος, 2015).

Το αποτέλεσμα των υπό συνθήκη πιθανοτήτων που προέκυψε από την συνολική εκπαίδευση με τον συνδυασμό των εξαγόμενων τιμών των χαρακτηριστικών της πλειάδας χρησιμοποιείται για να γίνει η πρόβλεψη.

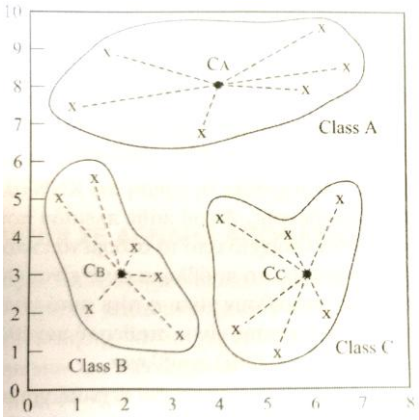
Πλεονεκτήματα της κατηγοριοποίησης αυτής είναι η ευκολία χρήσης και εκπαίδευσης (μία φορά), η θεμελίωση στην στατιστική, η χρήση ελλιπών δεδομένων (παράλειψη πιθανότητας), η χρήση αριθμητικών και ονομαστικών μεταβλητών, η παροχή εγκυρότητας και ακρίβειας στα αποτελέσματα. Αντιθέτως, μειονέκτημα είναι ότι τα γνωρίσματα δεν είναι ανεξάρτητα και γίνεται χρήση υποσυνόλων των χαρακτηριστικών, αγνοώντας τα εξαρτημένα εξ αυτών με άλλα. Ακόμη, ότι δεν υπάρχει καθιερωμένος τρόπος εξαγωγής γράφου και ότι υπάρχει αδυναμία χειρισμού συνεχών δεδομένων, όπου η διαίρεση τους σε διαστήματα ίσως επιφέρει λανθασμένα αποτελέσματα.
(Dunham, 2004)

<u>Αλγόριθμοι θεμελιωμένοι στην Απόσταση</u>:

Οι αλγόριθμοι θεμελιωμένοι στην απόσταση έχουν ως βάση ότι τα χαρακτηρίστηκα της ίδιας κατηγορίας είναι πιο κοντά μεταξύ τους από αυτά άλλων κατηγοριών και έχοντας ως μέτρο ομοιότητας την απόσταση μπορούν να οριστούν τα χαρακτηριστικά διαφορετικών μεταξύ τους κατηγοριών της βάσης δεδομένων (ανάκτηση πληροφοριών).

❖ Απλή προσέγγιση

«Δεδομένης μιας βάσης δεδομένων $D=\{t_1,t_{2,...}t_n\}$, από πλειάδες όπου κάθε πλειάδα $t_{i=}$ $\{t_{i1},t_{i2,...}t_{ik}\}$ περιέχει αριθμητικές τιμές, και ενός συνόλου από κατηγορίες $C=\{C_{1,...}C_m\}$, όπου κάθε κατηγορία $C_j=\{C_{j1},C_{j2...}C_{jk}\}$ έχει αριθμητικές τιμές, το πρόβλημα της κατηγοριοποίησης έγκειται στο να χωρίσουμε κάθε μία t_i στην κατηγορία C_j έτσι ώστε $sim(t_i, C_j) \geq sim(t_i, C_j)$ $\forall$ $C_\iota \in C$ όπου $C_\iota \neq C_j$» (Dunham, 2004).

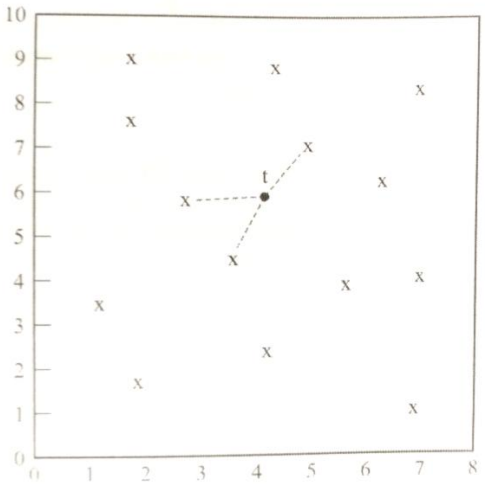


Σχήμα 2-7: Κατηγοριοποίηση με απλό αλγόριθμο απόστασης.
Πηγή: Dunham, 2004.

❖ Κ-πλησιέστεροι γείτονες (k-nearest neighbors-KNN)

Σύμφωνα με τη μέθοδο των Κ-πλησιέστερων γειτόνων, βρίσκονται όλα τα δείγματα εκπαίδευσης που έχουν ομοιότητες με το δείγμα ελέγχου ως προς τα χαρακτηριστικά του, βάση της απόστασης. Συγκεκριμένα, τα δείγματα εκπαίδευσης περιλαμβάνουν τα δεδομένα και την επιθυμητή τους κατηγοριοποίηση. Όταν κατηγοριοποιείται ένα καινούριο στοιχείο από το δείγμα ελέγχου, τότε λαμβάνεται υπόψη η απόστασή του από τα Κ στοιχεία εκπαίδευσης πλησιέστερων σε απόσταση εκχωρήσεων που έχουν οριστεί. Δηλαδή όπως φαίνεται στο σχήμα 2-8 εάν ορίστηκε το

K=3 τότε το στοιχείο t του δείγματος ελέγχου θα κατανεμηθεί στη κατηγορία με τα περισσότερα στοιχεία x που ανήκουν στην ίδια κατηγορία από το σύνολο των K στοιχείων (Dunham, 2004).



Σχήμα 2-8: Κατηγοριοποίηση με KNN.
Πηγή: Dunham, 2004.

<u>Αλγόριθμοι θεμελιωμένοι σε Δέντρα αποφάσεων</u>:

Η κατηγοριοποίηση με τη χρήση δέντρων αποφάσεων είναι από τις πιο διαδεδομένες τεχνικές, η οποία χρησιμοποιείται και για την πρόβλεψη. Τα δέντρα αποφάσεων αποτελούν τις πιο ισχυρές προσεγγίσεις για την ανακάλυψη γνώσης και την εξόρυξη δεδομένων με δυνατότητα έρευνας μεγάλου και πολύπλοκου όγκου δεδομένων. Επιτρέπεται η μοντελοποίηση και εξάγονται χρήσιμα πρότυπα. Τα δέντρα αποφάσεων είναι αποτελεσματικά εργαλεία για εξόρυξη δεδομένων και κειμένου, εκμάθηση μηχανών, εξαγωγή πληροφοριών και αναγνώριση μοτίβου, βοηθώντας στην αποτελεσματικότερη, οικονομικότερη, αποδοτικότερη και με ακρίβεια διαδικασία έρευνας. Τα δέντρα αποφάσεων κατανοούνται εύκολα από τον χρήστη, χειρίζονται αριθμητικά, ονομαστικά και δεδομένα κειμένου, επιτυγχάνουν υψηλή απόδοση με λίγη προσπάθεια, εφαρμόζονται σε πολλές πλατφόρμες εξόρυξης δεδομένων και εξεργάζονται εσφαλμένα δεδομένα και ελλείπουσες τιμές (Bhargava et al., 2013).

Τα δέντρα αποφάσεων χρησιμοποιούνται ως μοντέλα κατηγοριοποίησης και δημιουργούνται ύστερα από εκπαίδευση ενός συνόλου δεδομένων, τα οποία είναι ήδη κατηγοριοποιημένα. Ο κάθε ένας εσωτερικός κόμβος του δέντρου αποφάσεων αντιπροσωπεύει ένα χαρακτηριστικό, ενώ τα κλαδιά (τόξα) του τις πιθανές τιμές του

γνωρίσματος. Τέλος, κάθε φύλλο του δέντρου αποφάσεων αντιπροσωπεύει μία από τις οριζόμενες κατηγορίες. Η εκκίνηση της κατηγοριοποίησης γίνεται από τη ρίζα του δέντρου (σύνολο δεδομένων εκπαίδευσης), όπου με τον έλεγχο των γνωρισμάτων καταλήγουμε στους εσωτερικούς κόμβους. Σε εκείνο το σημείο ελέγχεται εάν ικανοποιείται ο συγκεκριμένος κόμβος από το δείγμα και επιλέγετε το κλαδί όπου θα συνεχίσει για τον επόμενο εσωτερικό κόμβο, μέχρι την κατάληξη σε κάποιο φύλλο (κατηγορία) (Χαλκίδη & Βαζιργιάννης, 2005).

Για τη συγκεκριμένη τεχνική κατηγοριοποίησης έχουν δημιουργηθεί διάφοροι αλγόριθμοι, εκ των οποίων, μερικοί γνωστοί είναι οι ID3, C4.5 και C5, J48, CART, SLIQ και SPRINT.

❖ ID3

Σύμφωνα με τον αλγόριθμο ID3 η ρίζα του δέντρου αποφάσεων αποτελείται από ένα κόμβο (σύνολο δεδομένων εκπαίδευσης) που μετατρέπεται σε φύλλο σε περίπτωση που τα δείγματα είναι όλα της ίδιας κατηγορίας. Ο αλγόριθμος μετράει πληροφορίες για το ποσοστό αβεβαιότητας, ταχύτητας, ή έκπληξης του συνόλου δεδομένων. Η πληροφορία (εντροπία) αυτή χρησιμοποιείται για να διαχωρίσει το σύνολο σε υποσύνολα και αυτό με το μεγαλύτερο κέρδος πληροφορίας επιλέγετε σαν γνώρισμα ελέγχου. Στη συνέχεια το γνώρισμα ελέγχου (κόμβος) διαχωρίζεται και η διαδικασία ολοκληρώνεται στη περίπτωση που όλα τα γνωρίσματα ανήκουν στην ίδια κατηγορία και δεν μπορεί να γίνει περαιτέρω διαχωρισμός, δεν υπάρχουν μη κατηγοριοποιημένα γνωρίσματα του κλαδιού γνωρίσματος ελέγχου (Χαλκίδη & Βαζιργιάννης, 2005).

❖ C4.5 και C5

Ο αλγόριθμος C4.5 είναι βελτιωμένος σε σχέση με τον ID3 ως προς:

- τα ελλιπή δεδομένα που αγνοούνται,
- τα συνεχή δεδομένα, όπου χωρίζεται το δείγμα βάσει των τιμών των γνωρισμάτων,
- του κλαδέματος, όπου αντικαθίσταται ένα υποδέντρο από φύλλο σε περίπτωση σφάλματος (αντικατάσταση υποδέντρου ή από το πιο χρησιμοποιημένο υποδέντρο (ανυψώσει υποδέντρου),
- τους κανόνες, όπου γίνεται κατηγοριοποίηση με τα δέντρα αποφάσεων ή τους κανόνες που δημιουργούν και
- τη διάσπαση, όπου προτιμούν τα γνωρίσματα που διαιρούνται πολλαπλώς.

Ο αλγόριθμος C5 είναι η αναβάθμιση του C4.5 για καλύτερη χρήση σε μεγάλο όγκο δεδομένων (Dunham, 2004).

❖ J48

Ο κατηγοριοποιητής J48 χρησιμοποιείται για κατηγοριοποίηση, δημιουργώντας ένα δυαδικό δέντρο, αφού είναι να απλό δέντρο αποφάσεων C4.5. Η τεχνική των δέντρων αποφάσεων είναι η ενδεδειγμένη για προβλήματα κατηγοριοποίησης. Με την δημιουργία του δέντρου αποφάσεων εφαρμόζεται αυτό σε κάθε πλειάδα της βάσης δεδομένων και καταλήγει σε κατηγοριοποίηση της. Κατά την δημιουργία του δέντρου αποφάσεων, ο κατηγοριοποιητής J48 αγνοεί τις τιμές που λείπουν και προβλέπονται με βάση τις αντίστοιχες γνωστές τιμές των άλλων καταχωρημένων χαρακτηριστικών. Γενικώς, ο κατηγοριοποιητής J48 διαιρεί τα δεδομένα σε εύρος, βάσει των τιμών των χαρακτηριστικών για τα συγκεκριμένα στοιχεία που βρέθηκαν στο δείγμα εκπαίδευσης και επιτρέπει την κατηγοριοποίηση μέσω δέντρων αποφάσεων ή μέσω κανόνων που δημιουργούνται από αυτά (Patil & Sherekar, 2013).

❖ CART

Ο αλγόριθμος CART χρησιμοποιεί δυαδικό δέντρο αποφάσεων με την τεχνική δέντρων κατηγοριοποίησης και παλινδρόμησης, όπου επιτυγχάνεται μεγάλο βάθος αφού η διακλάδωση γίνεται κάθε φορά σε δύο υποκατηγορίες. Χρησιμοποιείται ο βαθμός πληροφορίας (εντροπία) για την επιλογή του καλύτερου χαρακτηριστικού διάσπασης και διαχειρίζεται η ελλιπής πληροφορία με την μέθοδο της αγνόησης (Dunham, 2004).

❖ SLIQ

Ο αλγόριθμος SLIQ επιτυγχάνει ακρίβεια και βελτίωση χρόνου εκτέλεσης του δέντρου αποφάσεων με την προ-ταξινόμηση και την κατά πλάτος ανάπτυξη, πετυχαίνοντας επεξεργασία μεγάλου όγκου δεδομένων. Οι αριθμητικές ιδιότητες ταξινομούνται σε μια λίστα ιδιοτήτων με την οποία επιτυγχάνεται η εξάλειψη της ανάγκης να γίνεται πάλι σε κάθε κόμβο η ταξινόμησή τους. Ακόμη, ταξινομούνται η συμβολικές ιδιότητες ως προς τις τιμές τους και δημιουργείται η λίστα κλάσεως με το φύλλο του δέντρου αποφάσεων και την κλάση των εγγραφών (Νανόπουλος & Μανωλόπουλος, 2008).

❖ SPRINT

Σε αντίθεση με τον αλγόριθμο SLIQ, ο SPRINT δεν δημιουργεί χωρική πολυπλοκότητα των εγγραφών, ώστε να απαιτεί μεγάλη μνήμη. Α και ο αλγόριθμος SPRINT βασίζεται στον SLIQ χρησιμοποιεί διαφορετικές λίστες ιδιοτήτων και σε αριθμό έναν περισσότερο από τον SLIQ. Ουσιαστικά χρησιμοποιείται λίστα ιδιοτήτων για την τιμή, τον κωδικό και την κλάση της εγγραφής (Νανόπουλος & Μανωλόπουλος, 2008).

<u>Αλγόριθμοι θεμελιωμένοι σε Νευρωνικά Δίκτυα</u>:

Η δημιουργία μοντέλου κατηγοριοποίησης ή πρόβλεψης μπορεί να γίνει με τη χρήση των νευρωνικών δικτύων. Η προσέγγιση αυτού του τρόπου κατηγοριοποίησης βασίζεται στον ανθρώπινο εγκέφαλο και συγκεκριμένα στη δομή του, ως προς τους νευρώνες. Όπως και στα δέντρα αποφάσεων έτσι και στα νευρωνικά δίκτυα το μοντέλο κατηγοριοποιεί το σύνολο της βάσης δεδομένων. Αυτό γίνεται με την είσοδο συγκεκριμένων τιμών των χαρακτηριστικών μιας εγγραφής στους αντίστοιχους κόμβους του γράφου. Η δημιουργηθείσα τιμή εξόδου δείχνει το ποσοστό πιθανότητας αντιστοιχίσεις της εγγραφής στην κατηγορία και στη συνέχεια η εγγραφή τοποθετείται στην κατηγορία με το υψηλότερο ποσοστό πιθανότητας. Η διαδικασία εκπαίδευσης επαναλαμβάνεται εωσότου όλα τα δεδομένα κατηγοριοποιηθούν σε ικανοποιητικό βαθμό.

Κατά τη κατηγοριοποίηση με νευρωνικά δίκτυα καθορίζεται ο αριθμός των κόμβων εξόδου, τα χαρακτηριστικά εισόδου και εισέρχονται όλες οι πλειάδες στο δίκτυο για ορθότερη κατηγοριοποίηση. Κατασκευάζεται η τοπολογία του δικτύου και αξιολογούνται τα αποτελέσματα εξόδου, ώστε να προσαρμοστεί ανάλογα το βάρος που θα έχει η κατηγορία εξόδου των αποτελεσμάτων για την επανάληψη της διαδικασίας. (Dunham, 2004)

❖ Διάδοση

Σύμφωνα με τη διάδοση δίνεται μία τιμή σε κάθε κόμβο εισόδου από την πλειάδα βάσεις δεδομένων και εξάγεται σε κάθε τόξο εξόδου του κόμβου μία νέα τιμή ύστερα από επεξεργασία. Η ίδια διαδικασία συνεχίζεται στους επομένους κόμβους με τελικό αποτέλεσμα μία πλειάδα από τιμές εξόδου των κόμβων εξόδου (Dunham, 2004).

❖ Εποπτευόμενη μάθηση του νευρωνικού δικτύου

Ονομάζεται εποπτευόμενοι μάθηση διότι το επιθυμητό αποτέλεσμα εξόδου είναι πρωτύτερα γνωστό και εκτελείται ανάδραση στα δεδομένα του συνόλου εκπαίδευσης. Αντιθέτως, στην μη εποπτευόμενοι μάθηση οι τιμές της εξόδου δεν είναι γνωστές. Στη διαδικασία αυτή της κατηγοριοποίησης δίνετε βάση σε μία πλειάδα των αρχικών δεδομένων εκπαίδευσης και την προσαρμογή στα βάρη των τόξων τους για την μετέπειτα εκπαίδευση (Dunham, 2004).

❖ Νευρωνικά δίκτυα Perceptron

Το perceptron είναι νευρώνας με πολλές εισόδους και μία έξοδο. Είναι γνωστός για την απλότητα λειτουργίας του και μπορεί να χρησιμοποιηθεί για κατηγοριοποίηση σε δύο κατηγορίες (Dunham, 2004).

Αλγόριθμοι θεμελιωμένοι σε Κανόνες:

Οι αλγόριθμοι αυτοί βασίζονται σε κανόνες if-then (εάν-τότε) που δημιουργούνται για τη κατηγοριοποίηση. Επομένως, από το σύνολο της πλειάδας της βάσης δεδομένων γίνεται αξιολόγηση του τύπου «αληθές-ψευδές» για ξεκαθάρισμα των δεδομένων στα δέντρα αποφάσεων. Ακολουθείται κάποια σειρά διάσπασης και εξετάζονται όλες οι κατηγορίες κατά την δημιουργία τους, ενώ στους κανόνες δεν ακολουθείται σειρά και εξετάζεται μία κατηγορία την κάθε φορά (Dunham, 2004).

❖ Δημιουργία κανόνων από ένα δέντρο αποφάσεων

Στη διαδικασία της δημιουργίας κανόνων από ένα δέντρο αποφάσεων, ο αλγόριθμος δημιουργεί κανόνα για κάθε φύλλο (οριζόμενη κατηγορία) του δέντρου.

Άλλοι αλγόριθμοι που χρησιμοποιούνται βασισμένοι σε κανόνες είναι αυτοί που δημιουργούνται από ένα νευρωνικό δίκτυο ή δεν απορρέουν από κάποιο δέντρο αποφάσεων ή νευρωνικό δίκτυο (Dunham, 2004).

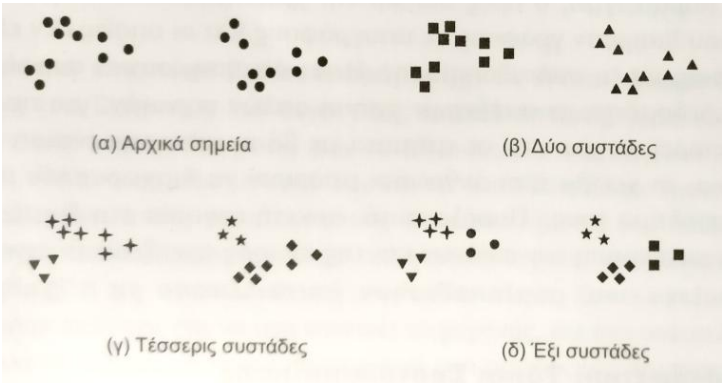
Συνδυαστικές τεχνικές:

Ο συνδυασμός των τεχνικών παρέχει τη δυνατότητα αναζήτησης τρόπων για καλύτερα αποτελέσματα μία εκ των τεχνικών για συνδυασμό είναι η ενίσχυση η τεχνική αυτή χρησιμοποιείται είτε ως σύνθεση τεχνικών, όπου τα αποτελέσματα της μίας

χρησιμοποιούνται ως είσοδος στην άλλη είτε ως ανεξάρτητες τεχνικές, όπου τα αποτελέσματα διαφόρων τεχνικών για μία κατηγορία στη συνέχεια συνδυάζονται (Dunham, 2004).

2.2.3.2 Συσταδιοποίηση

Η συσταδιοποίηση ομαδοποιεί τα στοιχεία της βάσης δεδομένων, ανάλογα με τις ομοιότητες χαρακτηριστικών τους και μοιάζει αρκετά με την κατηγοριοποίηση. Αρκετοί θεωρούν την συσταδιοποίηση ως μορφή της κατηγοριοποίησης, αφού προσδιορίζονται τα δεδομένα με κατηγορίες. Στη συσταδιοποίηση οι ομάδες (συστάδες) δεν είναι προκαθορισμένες όπως στη κατηγοριοποίηση, αλλά δημιουργούνται και όσο πιο ομογενοποιημένες είναι ως ομάδες τόσο μεγαλύτερη διαφορά υπάρχει μεταξύ τους. Ακόμη, στη συσταδιοποίηση μπορεί να μην υπάρχει εξαρχής γνώση για τις συστάδες και τον αριθμό τους αλλά και τα αποτελέσματα να είναι δυναμικά, σε αντίθεση με την κατηγοριοποίηση (Tan, Steinbach & Kumar, 2015).



Σχήμα 2-9: Διαφορετικοί τρόποι συσταδιοποίησης ίδιου συνόλου στοιχείων.
Πηγή: Dunham, 2004.

Η συσταδιοποίηση, σε αντίθεση με την κατηγοριοποίηση (εποπτευόμενη), κάποιες φορές αναφέρεται και ως μη εποπτευόμενη κατηγοριοποίηση, όπου τα νέα δεδομένα προσδιορίζονται σε υπάρχουσες κατηγορίες.

Υπάρχουν διαφορετικοί τύποι συσταδιοποίησης:

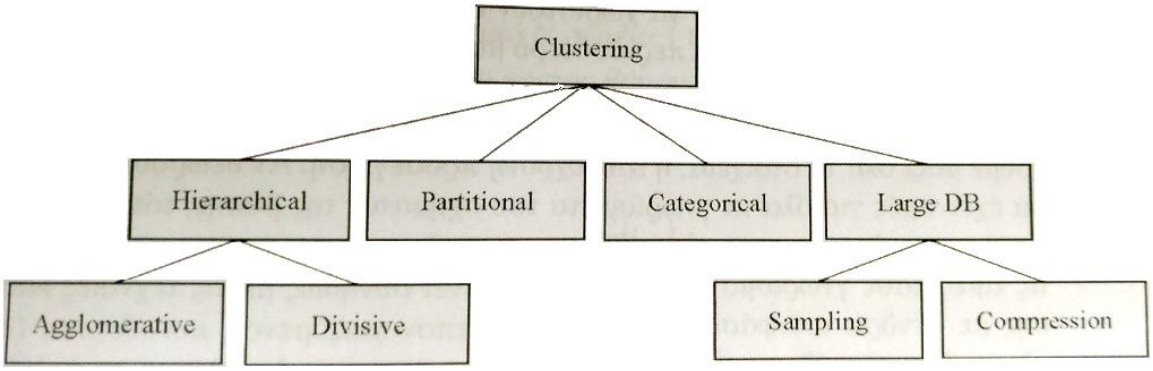
Ιεραρχική ή Διαμεριστική. Στη διαμεριστική συσταδιοποίηση αρχικά γίνεται διαίρεση των στοιχείων σε συστάδες, ώστε αυτά να ανήκουν σε ένα υποσύνολο, ενώ στην ιεραρχική συσταδιοποίηση υπάρχει φωλιασμένο σύνολο συστάδων, όπου υπάρχουν

υποσυστάδες οργανωμένες σαν δέντρο. Κάθε συστάδα (κόμβος), εκτός από τα φύλλα, έχει υποσυστάδες και η ρίζα του δέντρου περιέχει όλα τα στοιχεία.

Αποκλειστική ή Επικαλυπτόμενη ή Ασαφείς. Στην αποκλειστική συσταδιοποίηση κάθε στοιχείο της βάσης δεδομένων αποδίδεται σε μία συστάδα, ενώ στην επικαλυπτόμενη συσταδιοποίηση μπορεί να ανήκει σε περισσότερες κατηγορίες (ομάδες). Στην ασαφή συσταδιοποίηση κάθε στοιχείο ανήκει σε όλες τις κατηγορίες με σταθμισμένη ιδιότητα απόδοσης συμμετοχής, όπου έχει κλίμακα από το «0» (δεν ανήκει) έως και το «1» (ανήκει πλήρως).

Πλήρης ή Μερική. Στη πλήρη συσταδιοποίηση κάθε στοιχείο ανήκει σε μία συστάδα σε αντίθεση με την μερική συσταδιοποίηση, όπου κάποια στοιχεία μπορεί να ανήκουν σε σύνολα ως «θόρυβος», «ακραίες τιμές» ή «αδιάφορο υπόβαθρο».
(Tan, Steinbach & Kumar, 2015)

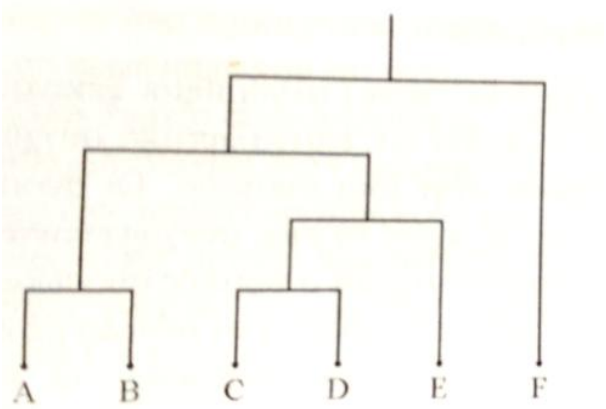
Υπάρχουν διάφοροι τύποι αλγορίθμων για την συσταδιοποίηση και ένας τρόπος κατηγοριοποίησης είναι να χωριστούν σε ιεραρχικούς, διαμεριστικούς, σε μεγάλες βάσεις δεδομένων και σύμφωνα με κατηγορικά γνωρίσματα (Dunham, 2004).



Σχήμα 2-10: Κατηγοριοποίηση αλγορίθμων συσταδιοποίησης.
Πηγή: Dunham, 2004.

Ιεραρχικοί Αλγόριθμοι:

Οι ιεραρχικοί αλγόριθμοι συσταδιοποίησης δημιουργούν σύνολα συστάδων με τη δομή των δεδομένων ως δένδρο (δενδρόγραμμα). Η ρίζα του δέντρου αποτελεί την αρχική συστάδα με όλα τα στοιχεία και οι εσωτερικοί κόμβοι τις υποσυστάδες. Το μήκος μεταξύ των υποσυστάδων (παιδιά) ενός κόμβου δεν πρέπει να είναι μεγαλύτερο από αυτό που δημιουργείται προηγούμενα μεταξύ του κόμβου αυτού με άλλο κόμβο ως παιδιά ενός άλλου κόμβου (Dunham, 2004).



Σχήμα 2-11: Δενδρόγραμμα.
Πηγή: Dunham, 2004.

Οι ιεραρχικοί αλγόριθμοι αποτελούνται από τους συσσωρευτικούς, όπου κάθε στοιχείο ανήκει στη συστάδα του και με επαναληπτικές συγχωνεύσεις υπάρχει καταληκτική συστάδα, καθώς και οι διαιρετικοί, όπου τα στοιχεία ξεκινούν από μία συστάδα και διασπώνται σε δύο επιμέρους, έως ότου τα στοιχεία να ανήκουν στη δική τους συστάδα (Dunham, 2004).

<u>Διαμεριστικοί Αλγόριθμοι</u>:

Οι διαμεριστικοί αλγόριθμοι, αφού πρωτίστως οριστεί ο επιθυμητός αριθμός συστάδων K των στοιχείων, αποδίδουν ως έξοδο ένα μόνο σύνολο συστάδων. Ένας πολύ γνωστός αλγόριθμος που χρησιμοποιείται στην διαμεριστική συσταδιοποίηση είναι ο K-means.

❖ Συσταδιοποίηση K-means

Στην συσταδιοποίηση k-means καθορίζετε εκ των προτέρων ο επιθυμητός αριθμός συστάδων (παράμετρος K). Με τυχαία επιλογή των σημείων k ως κέντρα συστάδων (κέντρα βάρους) και σύμφωνα με την Ευκλείδεια μετρική αποστάσεων, όλες οι περιπτώσεις ανατίθενται στο πλησιέστερο κέντρο της συστάδας. Κατόπιν, υπολογίζεται η τιμή ή ο μέσος (κέντρο βάρους) κάθε συστάδας. Μετέπειτα το κέντρο βάρους κάθε συστάδας ενημερώνεται για την νέα τιμή που παίρνει από τα στοιχεία που αποδίδονται στην συστάδα με επανάληψη εωσότου σταθεροποιηθεί η τιμή τους (Witten & Frank, 2000).

❖ Αλγόριθμος Πλησιέστερου Γείτονα

Στην περίπτωση χρήσης του αλγόριθμου πλησιέστερου γείτονα γίνεται επαναληπτική συγχώνευση των στοιχείων των συστάδων στις πλησιέστερες μεταξύ τους, όπου υπάρχει περίπτωση να δημιουργηθούν και καινούριες συστάδες.

Ενδεικτικά κάποιοι άλλοι αλγόριθμοι που χρησιμοποιούνται είναι:

- το Δένδρο Ελάχιστης Ζεύξης (Minimum Spanning Tree, MST),
- ο Αλγόριθμος Συσταδιοποίησης Τετραγωνικού Σφάλματος (Squared Error Clustering Algorithm, SECA),
- ο Αλγόριθμος PAM (Partitioning Around Medoids-Διαμερισμός γύρω από Medoids, Medoid =σύνολο στοιχείων συστάδας με την ελάχιστη μέση ανομοιότητα των στοιχείων),
- ο Αλγόριθμος Ενέργειας Δεσμού (Bond Energy Algorithm-BEA),
- Συσταδιοποίηση με Γενετικούς Αλγορίθμους και
- Συσταδιοποίηση με Νευρωνικά Δίκτυα.

(Dunham, 2004)

<u>Συσταδιοποίηση σε μεγάλες βάσεις δεδομένων</u>:

Οι κλασικοί αλγόριθμοι συσταδιοποίησης πιθανόν να είναι ακατάλληλοι για επεξεργασία δυναμικών βάσεων δεδομένων, λόγω της πολυπλοκότητας τους όπου χρειάζονται επαρκής μνήμη για την επεξεργασία των δεδομένων. Η ύπαρξη και χρήση μεγάλων βάσεων δεδομένων και η ανάγκη επαναληπτικών ενεργειών των αλγορίθμων, δημιούργησε την ανάγκη για μεγαλύτερες δυνατότητες της κυρίας μνήμης ή χρήση αλγορίθμων που να λειτουργούν με διαφορετική φιλοσοφία.

Η αποδοτικότητα των αλγορίθμων σε μεγάλες βάσεις δεδομένων εξαρτάται από:

- την σάρωση της βάσης δεδομένων που πρέπει να γίνει το περισσότερο μία φορά,
- την ¨online¨ ικανότητα παροχής πληροφοριών της κατάστασης και του καλύτερου αποτελέσματος, καθώς και την ενημέρωση των αποτελεσμάτων αυξητικά, με τη πρόσθεση ή αφαίρεση δεδομένων,
- την ύπαρξη δυνατότητας προσωρινής διακοπής και συνέχειας ή οριστικής διακοπής,
- τη δυνατότητα χρήσης διαφορετικών τεχνικών σάρωσης της βάσης δεδομένων,

- την λειτουργία με λίγη διαθέσιμη κυρία μνήμη και
- την λιγότερη δυνατή εξεργασία των πλειάδων.

(Dunham, 2004)

❖ Αλγόριθμος DBSCAN

Η χρήση του αλγόριθμου DBSCAN (Density-Based Spatial Clustering of Applications with Noise) οδηγεί σε συστάδες με ελάχιστο μέγεθος αλλά και πυκνότητα, η οποία είναι ο ελάχιστος αριθμός σημείων συγκεκριμένης μεταξύ τους αποστάσεις. Με τη μέθοδο αυτή αποφεύγεται η δυνατότητα συστάδων από απομονωμένα σημεία.

Ενδεικτικά άλλοι αλγόριθμοι μεγάλων βάσεων δεδομένων είναι ο BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) και ο CURE (Clustering using Representatives-Συσταδιοποίηση με χρήση Αντιπροσώπων).

(Dunham, 2004)

<u>Συσταδιοποίηση με κατηγορικά γνωρίσματα</u>:

Οι παραδοσιακοί αλγόριθμοι συσταδιοποίησης, σε κάποιες εκ των περιπτώσεων χρήσης τους, αντιμετωπίζουν προβλήματα κατά την επεξεργασία μη αριθμητικών (κατηγορικών) δεδομένων. Για τον λόγο αυτό δημιουργήθηκαν οι αλγόριθμοι με κατηγορικά γνωρίσματα.

❖ Αλγόριθμος ROCK

Ο Αλγόριθμος ROCK (Robust Clustering using links-εύρωστη συσταδιοποίηση με χρήση συνδέσμων) χρησιμοποιεί την δυαδική τιμή (Boolean) και τα κατηγορικά γνωρίσματα. Η ομοιότητα μεταξύ των στοιχείων βασίζεται στο πλήθος των συνδέσμων τους και όταν αυτή υπερβαίνει ένα οριζόμενο κατώφλι, καλούνται γείτονες. Ο αλγόριθμος ROCK δημιουργεί ομάδες με τα στοιχεία που έχουν τους περισσότερους κοινούς γείτονες (συνδέσμους).

(Dunham, 2004)

2.2.3.3 Κανόνες Συσχέτισης

Οι κανόνες συσχέτισης είναι ένας τρόπος ανάδειξης των συσχετίσεων μεταξύ των δεδομένων μίας βάσης δεδομένων, όπου υπάρχουν και δεν είναι εμφανείς, δεν αποτελούνται από συναρτησιακές εξαρτήσεις, αλλά ούτε συνδέονται με την σχέση αιτιότητας-συσχέτισης. Σύμφωνα με τους κανόνες αυτούς, εντοπίζονται οι συνήθεις ταυτόχρονες χρήσεις των στοιχείων που υπάρχει μεταξύ τους κάποια σύνδεση, όπως για παράδειγμα η αγορά ενός προϊόντος από κατάστημα λιανικής πώλησης να συνοδεύεται από την αγορά κάποιου άλλου προϊόντος. Το ποσοστό εμφάνισης συγκεκριμένου στοιχείου/ων στις συναλλαγές που έγιναν κατά την αγορά προϊόντων από το κατάστημα, ονομάζεται υποστήριξη (support). Εμπιστοσύνη, στους κανόνες συσχέτισης, ονομάζεται η συχνότητα εμφάνισης του στοιχείου ενός στοιχειοσυνόλου Y σε συναλλαγές με το στοιχειοσύνολο X. Τα δύο αυτά χαρακτηριστικά (υποστήριξη και εμπιστοσύνη) φανερώνουν την σημαντικότητα των συσχετίσεων, ώστε να επιλεγούν αυτές που μας ενδιαφέρουν (Dunham, 2004).

Ανάλογα με την μορφή των δεδομένων χρησιμοποιούνται και οι αντίστοιχοι αλγόριθμοι για τους κανόνες συσχέτισης ή διαμορφώνονται τα δεδομένα. Υπάρχουν δυαδικά (κατηγορικά και συμμετρικά) και ονομαστικά χαρακτηριστικά, καθώς και συνεχή χαρακτηριστικά. Όταν διαχειριζόμαστε συνεχή χαρακτηριστικά σε κανόνες συσχέτισης, αυτοί ονομάζονται ποσοτικοί κανόνες συσχέτισης (quantitative association rules) και γενικός υπάρχουν τρείς σημαντική τύποι μεθόδων. Οι μέθοδοι αυτοί βασίζονται στην διακριτοποίηση των δεδομένων, στην στατιστική και στην μη διακριτοποίηση των δεδομένων.

Οι ορισμοί που μπορούν να δοθούν για τους κανόνες συσχέτισης, την υποστήριξη και εμπιστοσύνη είναι:

- « Με δεδομένο ένα σύνολο από στοιχεία $I=\{I_1, I_2,\ldots I_m\}$ και μια βάση δεδομένων από συναλλαγές $D=\{t_1,t_2,\ldots t_n\}$ όπου $t_1=\{I_{i1}, I_{i2},\ldots I_{ik}\}$ και $I_{ij} \in I$, ένας κανόνας συσχέτισης (association rule) είναι ένα επαγωγικό συμπέρασμα της μορφής $X \Rightarrow Y$, όπου $X,Y \subset I$ είναι σύνολα στοιχείων που ονομάζονται στοιχειοσύνολα και $X \cup Y = \emptyset$ » (Dunham, 2004).
- «Η υποστήριξη (support-s) για ένα κανόνα συσχέτισης $X \Rightarrow Y$ είναι το ποσοστό των συναλλαγών στη βάση δεδομένων που περιέχουν το $X \cup Y$» (Dunham, 2004).

- «Η εμπιστοσύνη ή η ισχύς (confidence, strength - α) για ένα κανόνα συσχέτισης $X \Rightarrow Y$, είναι το κλάσμα του αριθμού των συναλλαγών που περιέχουν το $X \cup Y$ προς τον αριθμό των συναλλαγών που περιέχουν το X» (Dunham, 2004).

❖ Αλγόριθμος Apriori

Ο πρώτος εκ των αλγορίθμων που χρησιμοποιήθηκε για την εξόρυξη κανόνων συσχέτισης είναι ο αλγόριθμος Apriori, με τον οποίο δημιουργούνται στοιχειοσύνολα με καθορισμένο μέγεθος και ελέγχεται με σάρωση της βάσης δεδομένων εάν αυτά έχουν κάποια σχετική συχνότητα εμφάνισης. Ουσιαστικά, ελέγχεται εάν τα καινούρια στοιχειοσύνολα είναι συχνά (κλειστά), όπου θα ισχύει και η ιδιότητα ότι τα υποσύνολα του καθενός από αυτά θα είναι επίσης συχνά. Αντίστροφα, όταν ένα στοιχειοσύνολο δεν είναι συχνό, τότε ούτε και τα δημιουργηθέντα υπερσύνολα του είναι συχνά (Dunham, 2004).

❖ Αλγόριθμος της Δειγματοληψίας

Ο συγκεκριμένος αλγόριθμος χρησιμοποιείται για την μείωση της σάρωσης του μεγάλου όγκου των βάσεων δεδομένων σε ένα ή δύο το πολύ περάσματα. Με την διαδικασία αυτή γίνεται δειγματοληψία στην βάση δεδομένων σε τέτοιο βαθμό, ώστε να χωράει στην μνήμη κατά την επεξεργασία. Κατόπιν, χρησιμοποιείται κάποιος άλλος αλγόριθμος, όπως ο Apriori, για την εύρεση των συχνών στοιχειοσυνόλων (Dunham, 2004).

❖ Αλγόριθμος της Διαμέρισης

Ο αλγόριθμος της διαμέρισης διαχωρίζει την βάση δεδομένων σε τμήματα και ταυτόχρονα ενώ σαρώνει την βάση δεδομένων, βρίσκει τα συχνά στοιχειοσύνολα των τμημάτων που έχει σαρώσει με την σειρά, ένα κάθε φορά. Η βάση δεδομένων σαρώνεται συνολικά δύο φορές και στο δεύτερο πέρασμα χρησιμοποιούνται τα στοιχειοσύνολα που είναι συχνά, έστω και σε μία μόνο διαμέριση για να βρεθεί εάν ισχύει το ίδιο σε όλη την βάση δεδομένων (Dunham, 2004).

2.2.3.4 Παλινδρόμηση

Η εργασία της εξόρυξης δεδομένων με την παλινδρόμηση σχετίζεται με τη προγνωστική μοντελοποίηση και τη σχέση που υπάρχει μεταξύ των μεταβλητών. Η στοχευμένη μεταβλητή (μεταβλητή απόκρισης) που είναι εξαρτημένη και συνεχής, συνδέεται με άλλη ή άλλες μεταβλητές (ανεξάρτητες), οι οποίες την επηρεάζουν. Η παλινδρόμηση έχει εφαρμογή σε αρκετές περιπτώσεις, όπως στη πρόβλεψη χρηματιστηριακών δεικτών με την χρήση άλλων οικονομικών δεικτών και των πωλήσεων ενός προϊόντος σε σχέση με την διαφήμιση του (Tan, Steinbach & Kumar, 2015).

Ενδεικτικά, τεχνικές ανάλυσης παλινδρόμησης είναι η Απλή Γραμμική Παλινδρόμηση (μέθοδος των ελάχιστων τετραγώνων, ανάλυση των σφαλμάτων παλινδρόμησης, ανάλυση του αποδοτικότητας προσαρμογής), Πολλαπλή Γραμμική Παλινδρόμηση, η Πολυωνυμική Παλινδρόμηση και η Λογιστική (ή Λογαριθμική) Παλινδρόμηση (Κύρκος, 2015).

2.2.3.5 Ανίχνευση Ανωμαλιών

Η ανίχνευση ανωμαλιών είναι μία από τις εργασίες της εξόρυξης δεδομένων, όπου επιδιώκεται ο εντοπισμός στοιχείων με διαφορετικότητα, έναντι του συνόλου των στοιχείων της βάσης δεδομένων. Τα διαφορετικά στοιχεία είναι γνωστά ως «ακραίες τιμές», λόγω της απομακρυσμένης τους απόστασης από τα υπόλοιπα σημεία κατά την απεικόνισή τους σε ένα διάγραμμα. Ακόμη, συναντάμε τον όρο «ανίχνευση αποκλίσεων», διότι οι τιμές των χαρακτηριστικών των διαφορετικών στοιχείων αποκλίνουν από τις συνηθισμένες και αναμενόμενες. Τέλος, μπορεί να χρησιμοποιηθεί ο όρος «εξόρυξη εξαιρέσεων» για τα διαφορετικά στοιχεία, λόγω των ασυνήθιστων ανωμαλιών τους.

Παραδείγματα στα οποία εφαρμόζεται η ανίχνευση ανωμαλιών είναι ο εντοπισμός της απάτης (χρήση κλεμμένης πιστωτικής κάρτα με ασυνήθιστες αγορές), η ανίχνευση εισβολών σε υπολογιστές (εντοπισμός ασυνήθιστης συμπεριφοράς), η ορθή λειτουργία της δημόσιας υγείας (εντοπισμός περιστατικών μη δικαιολογημένων, σύμφωνα με τις στατιστικές αναφορές των νοσοκομείων και ιατρικών κλινικών) και οι

διαταραχές στο περιβάλλον (πρόβλεψη σπάνιων καιρικών συμβάντων, όπως ξηρασία και πλημμύρες).

Οι τεχνικές που χρησιμοποιούνται για τις εργασίες εντοπισμού ανωμαλιών είναι:

- Οι στατιστικές προσεγγίσεις στις οποίες δημιουργούνται διάφορα μοντέλα, συνήθως κατανομής πιθανότητας, βασισμένα στα δεδομένα της βάσης. Στη συνέχεια, εκτιμούνται τα νέα στοιχεία σύμφωνα με τα μοντέλα για την προσαρμοστικότητα τους ή την πιθανότητα να ανήκουν σε αυτά.
- Η ανίχνευση ακραίων τιμών βάσει εγγύτητας, στην οποία εντοπίζονται τα απομακρυσμένα στοιχεία σε απόσταση από το σύνολο των σημείων. Συνήθης μέθοδος που χρησιμοποιείται είναι αυτή των Κ πλησιέστερων γειτόνων.
- Η ανίχνευση ακραίων τιμών βάσει πυκνότητας, στην οποία εντοπίζονται τα στοιχεία με χαμηλή πυκνότητα και σχετίζεται με αυτή της εγγύτητας. Ένας τρόπος υπολογισμού της πυκνότητας είναι να ισούται με το αντίστροφο της μέσης απόστασης από τους Κ πλησιέστερους γείτονες, βάσει του αλγορίθμου πλησιέστερου γείτονα ή του αλγορίθμου DBSCAN (συσταδιοποίηση).
- Η ανάλυση συστάδων, όπου κατά βάση χρησιμοποιείται για τον εντοπισμό ισχυρών σχέσεων μεταξύ στοιχείων και ουσιαστικά το αντίθετο με τις ακραίες τιμές. Στην προκειμένη περίπτωση, μπορούν να απομονωθούν οι απομακρυσμένες μικρές συστάδες ή να υπολογιστή ο βαθμός που ανήκουν τα συσταδιοποιημένα στοιχεία στην συστάδα τους.

(Tan, Steinbach & Kumar, 2015)

2.2.3.6 Ανάλυση Χρονολογικών Σειρών

Η ανάλυση χρονολογικών σειρών βοηθάει στην εύρεση ακολουθιών με τιμές ενός χαρακτηριστικού, στηριζόμενη στην βάση δεδομένων χρονοσειρών. Επομένως, επιζητάτε από μία χρονολογική σειρά τιμών (αριθμών) με ίσα χρονικά διαστήματα (ημερησίως, ετησίως κτλ.) την μετέπειτα νέα ακολουθία του. Οι βασικές λειτουργίες της ανάλυσης χρονοσειρών είναι η χρήση μονάδων μέτρησης αποστάσεις για εντοπισμό ομοιοτήτων σε διαφορετικές χρονοσειρές, ο καθορισμός της συμπεριφοράς τους από τη δομή τους και η πρόβλεψη τιμών με χρήση διαγραμμάτων (Dunham, 2004).

Για να επιτευχτεί ο καθορισμός της ομοιότητας των χρονοσειρών, καθορίζεται μία ρεαλιστική για τον χρήστη συνάρτηση απόστασης. Μια τεχνική υπολογισμού της

απόστασης ανάμεσα σε δύο ακολουθίες είναι η χαρτογράφησή τους και η χρήση μιας νόρμας (μέτρο απόστασης διανύσματος). Ακόμη, μπορούν να εξαχθούν κάποια χαρακτηριστικά γνωρίσματα για χρήση του καθορισμού ομοιότητας ή να βρεθεί η πιο μακρινή κοινή υπο-ακολουθία των ακολουθιών, χρησιμοποιώντας το μήκος της για τον καθορισμό της απόστασης (Χαλκίδη & Βαζιργιάννης, 2005).

2.2.3.7 Πρότυπα Ακολουθιών

Οι εργασίες των προτύπων ακολουθιών γίνονται στα δεδομένα μιας βάσης για να καθορίσουν σειριακά πρότυπα, τα οποία στηρίζονται σε μία χρονική ακολουθία ενεργειών και τα εξαγόμενα δεδομένα συσχετίζονται βάση του χρόνου ή άλλων ακολουθιών. Τα πρότυπα ακολουθιών χρησιμεύουν στην εύρεση στοιχείων και εξήγηση φαινομένων που επαναλαμβάνονται, στη πρόγνωση των εισερχομένων δραστηριοτήτων και τον υπολογισμό των ομοιοτήτων. Ο τρόπος με τον οποίο δημιουργούνται τα πρότυπα συνήθως επικεντρώνεται στα συμβολικά πρότυπα και η σκέψη στην οποία βασίζονται είναι:

«Λαμβάνοντας υπόψη ένα ενδεχομένως μεγάλο πρότυπο (συμβολοσειρά) S, ενδιαφερόμαστε για τα πρότυπα ακολουθιών της μορφής a→b, όπου τα a, b, ab είναι υποσυμβολοσειρές μέσα στο S, τέτοιες ώστε η συχνότητα του ab να μην είναι μικρότερη από κάποια ελάχιστη υποστήριξη και η πιθανότητα ότι το a ακολουθείται αμέσως από το b να μην είναι μικρότερη από την ελάχιστη εμπιστοσύνη» (Χαλκίδη & Βαζιργιάννης, 2005).

Ακόμη, τα εξαγόμενα αποτελέσματα των προτύπων ακολουθιών μπορούν να έχουν περιορισμούς με προσχέδια πρότυπα (σειριακά επεισόδια, παράλληλα επεισόδια και κανονικές εκφράσεις). Τα σειριακά επεισόδια εμφανίζονται ως σύνολα στοιχείων σε συνολική κατάταξη, ενώ τα παράλληλα ως ασήμαντα καταταγμένα σύνολα γεγονότων. Τέλος, στις κανονικές εκφράσεις μπορούμε να έχουμε ένα προσχέδιο (A|B)C*(D|E) όπου ζητάμε πρότυπα που ικανοποιούν την πρωτίστως τα στοιχεία A και B, μετά το C και κατόπιν το D και E ανεξαρτήτου προτεραιότητας.
(Χαλκίδη & Βαζιργιάννης, 2005).

2.2.3.8 Μείωση των Διαστάσεων

Στις εργασίες μείωσης των διαστάσεων επιδιώκεται η αντιπροσώπευση της βάσης δεδομένων από λιγότερα σύνολα, προσπαθώντας να διατηρηθεί η αρχική δομή και έχοντας υπόψη ότι χάνονται αρκετές πληροφορίες. Στην μείωση των διαστάσεων βασικός στόχος είναι η μείωση του χώρου, το οποίο επιτυγχάνεται με τη προβολή n-διάστατων συνόλων κάθε στοιχείου της βάσης δεδομένων σε k-διάστατο χώρο, ισχύοντας $k<<n$.

Οι βασικές δύο μέθοδοι που ακολουθούνται για την μείωση των διαστάσεων είναι η τοπική ή σχηματική συντήρηση και η σφαιρική ή τοπολογική διατήρηση. Στη πρώτη μέθοδο δεν χρησιμοποιούνται οι γενικές ιδιότητες που έχει ένα σύνολο δεδομένων αλλά απλοποιούνται οι ακολουθίες δεδομένων, χωρίς να υπάρχει σχέση με το υπόλοιπο σύνολο. Ουσιαστικά εντοπίζονται και διατηρούνται τα πιο σημαντικά χαρακτηριστικά γνωρίσματα k που περιέχουν την περισσότερη πληροφορία, ώστε να μην αλλοιωθεί αισθητά η αρχική δομή.

Στη σφαιρική τοπολογική διατήρηση γίνεται συνήθως απεικόνιση για εντοπισμό της μικρότερης δυνατής αντιπροσώπευσης των στοιχείων της βάσης δεδομένων, εντοπίζοντας τα χαρακτηριστικά k με τα οποία δίνεται η σφαιρική και πιο φειδωλή αντικειμενική συνάρτηση.
(Χαλκίδη & Βαζιργιάννης, 2005).

2.3 Χωρικά Δεδομένα

Τα δεδομένα κατατάσσονται στα χωρικά σύμφωνα με την χωρική συνιστώσα τους, από την οποία δίνεται πληροφορία για την θέση τους. Οι βάσεις χωρικών δεδομένων περιέχουν χωρικές αλλά και μη χωρικές πληροφορίες. Τα χωρικά δεδομένα μπορούν να τοποθετηθούν σε φυσικό χώρο με χαρακτηριστικά όπως το γεωγραφικό πλάτος και μήκος, τη διεύθυνση ή τη διαμέριση τους κατά θέση. Ακόμη τα χωρικά δεδομένα μπορούν να χαρακτηριστούν στον χώρο (κοντινά, μακρινά, ανατολικά, δυτικά, γειτονικά κτλ.) και να παρέχουν πληροφορία απόστασης ή τοπολογίας. Ο χαρακτηρισμός αυτός των χωρικών δεδομένων βοηθάει στην εύκολη ευρετηρίαση και την χρήση συγκεκριμένων δομών δεδομένων. Στην εξόρυξη δεδομένων η γνώση της απόστασης

βοηθάει στις μετρήσεις ομοιότητας μεταξύ των στοιχείων της βάσης δεδομένων (Dunham, 2004).

2.3.1 Εξόρυξη Χωρικών Δεδομένων

Κάποιες τεχνικές που αναφέρθηκαν στην ενότητα 2.2.3 χρησιμοποιούνται για την εξόρυξη χωρικών δεδομένων, όμως για το συγκεκριμένο σκοπό δημιουργήθηκαν καινούργιες τεχνικές και αλγόριθμοι, καθώς και κανόνες.

Η δημιουργία χωρικών κανόνων εστιάζει στη περιγραφή της δομής και των συσχετίσεων των χωρικών στοιχείων. Υπάρχουν οι κανόνες χωρικών χαρακτηριστικών οι οποίοι περιγράφουν τα δεδομένα μιας βάσης ή τμήμα της. Οι κανόνες χωρικών διαχωρισμών εστιάζουν στην περιγραφή των διαφορετικών χαρακτηριστικών βάσει των οποίων διαφοροποιούνται οι κλάσεις των δεδομένων. Οι κανόνες χωρικών συσχετίσεων είναι αποτελέσματα ενός υποσυνόλου δεδομένων από κάποιο άλλο και ουσιαστικά, χρησιμοποιούνται κανόνες συσχέτισης σε στοιχεία χωρικών δεδομένων (Dunham, 2004).

Στη συνέχεια παρατίθεται παράδειγμα για καλύτερη κατανόηση.

- Κανόνας χωρικού χαρακτηριστικού: το ακαθάριστο εγχώριο προϊόν κατά κεφαλή στην κεντρική Μακεδονία είναι 12.000 ευρώ.
- Κανόνας χωρικού διαχωρισμού: το ακαθάριστο εγχώριο προϊόν κατά κεφαλή στην κεντρική Μακεδονία είναι 12.000 ευρώ ενώ στη Πελοπόννησο είναι 14.000 ευρώ.
- Κανόνας χωρικής συσχέτισης: το ακαθάριστο εγχώριο προϊόν κατά κεφαλή στην κεντρική Μακεδονία και συγκεκριμένα στην Θεσσαλονίκη είναι 13.000 ευρώ.

❖ Αλγόριθμοι χωρικής κατηγοριοποίησης

Κατά την χωρική κατηγοριοποίηση γίνεται η διαμέριση των συνόλων χωρικών στοιχείων βάσει χωρικών ή/και μη χωρικών χαρακτηριστικών. Τεχνικές που χρησιμοποιούνται είναι:

- Η επέκταση του αλγορίθμου ID3 με τον οποίο γίνεται εισαγωγή γράφων γειτνίασης που δημιουργούνται από στοιχεία στο χώρο και κάθε ένα τους αποτελεί κόμβο στον γράφο. Οι ακμές που συνδέονται οι κόμβοι θα πρέπει να ενώνουν γειτονικούς κόμβους, ώστε να λαμβάνονται υπόψη τα κοντινά από το δοθέν στοιχεία.

- Το δέντρο χωρικής απόφασης που δημιουργεί δέντρα αποφάσεων με την χρήση στοιχείων, κοντινών στα χωρικά που μας ενδιαφέρουν και τη περιγραφή των κλάσεων από το σύνολο των κατηγορημάτων τους. Επιλέγονται τα κατηγορήματα που είναι πιο σχετικά ώστε να δημιουργηθούν ακριβή μικρότερα δέντρα αποφάσεων.

(Dunham, 2004).

❖ Αλγόριθμοι χωρικής συσταδιοποίησης

Οι αλγόριθμοι χωρικής συσταδιοποίησης χρησιμοποιούνται για μεγάλες βάσεις δεδομένων και οι εξ αυτών καλοί θα πρέπει να μπορούν να ξεχωρίζουν τις συστάδες από τα σχήματα που δημιουργούνται. Δηλαδή, ο αλγόριθμος θα πρέπει να ξεχωρίζει τα σημεία της κάθε συστάδας, όπως όταν για παράδειγμα σε ένα δισδιάστατο χώρο υπάρχουν τέσσερις συστάδες όπως φαίνεται στο σχήμα 2-12 και κάποια σημεία της μίας είναι πιο κοντά σε σημεία άλλης συστάδας. Επίσης, δεν θα πρέπει να υπάρχει επιρροή από ακραία σημεία, τα οποία δεν περιλαμβάνονται κατά την δημιουργία των συστάδων, καθώς και να μην έχει σημασία η σειρά σάρωσης των σημείων (Dunham, 2004).

Σχήμα 2-12: Σχήματα χωρικών συστάδων.
Πηγή: Dunham, 2004.

Ενδεικτικά κάποιες τεχνικές που χρησιμοποιούνται κατά την χορήγηση συσταδιοποίησης είναι:

- Η χρήση του αλγόριθμου CLARANS χωρικής τάξης [SD(CLARANS)-Spatial Dominant(CLustering Algorithm based on RANdomized Search)] με τον οποίο συσταδοποιούνται οι χωρικές συνιστώσες και χρησιμοποιούνται κατόπιν τα μη χωρικά χαρακτηριστικά της κάθε συστάδας για περιγραφή.

- Ο αλγόριθμος DBCLASD (Distribution Based Clustering of LArge Spatial Databases) βασίζεται σε κατανομές και συσταδοποιεί μεγάλες βάσεις χωρικών δεδομένων. Ο DBCLASD προσδιορίζει τη κατανομή με βάση τις αποστάσεις ανάμεσα στους πλησιέστερους γείτονες.

(Dunham, 2004).

2.3.2 Γεωγραφικά Συστήματα Πληροφοριών

Η διαδικασία της έρευνας των χαρακτηριστικών της γης με την επικουρική συμμετοχή της τηλεσκόπησης και σε συνδυασμό με την τεχνολογική έξαρση, οδήγησε στα Γεωγραφικά Συστήματα Πληροφοριών (Geographical Information Systems-GIS). Τα συστήματα GIS χρησιμοποιούν χωρικά δεδομένα, τα οποία παρέχουν πληροφορίες σχετικές με την γεωγραφική θέση των στοιχείων (Ansari & Kale, 2014).

Η ύπαρξη, τα τελευταία χρόνια, αρκετών εμπορικών GIS και η ανάπτυξη των λειτουργιών τους, έδωσε την δυνατότητα στους χρήστες να επιτύχουν ακριβέστερα και πιο αξιόπιστα αποτελέσματα. Η απόκτηση χωρικής ανάλυσης από την εισαγωγή περιβαλλοντικών ακατέργαστων δεδομένων σε σύστημα GIS έδωσε σημαντικές πληροφορίες και γνώσεις για πληθώρα εφαρμογών στην καθημερινότητα. Τα GIS είναι συστήματα εφαρμογής (εργαλεία λογισμικών), όπου επιτρέπεται η δημιουργία, συλλογή, αποθήκευση, ανάλυση, μετασχηματισμός, διαχείριση και απεικόνιση χωρικών δεδομένων και των σχετικών χαρακτηριστικών γνωρισμάτων τους από τον πραγματικό περιβάλλον. Κάποιες εκ των εφαρμογών του GIS είναι για δημιουργία πολεοδομικών σχεδίων, αξιολόγηση περιβαλλοντικών επιπτώσεων, χαρτογράφηση, κατανομές του πληθυσμού, διαχείριση πόρων και περιουσιακών στοιχείων, επιστημονικών ερευνών, καθώς και για διαχείριση κρίσεων, όπως αντιμετώπιση άμεσων προβλημάτων π.χ. απομόνωση περιοχής με σχεδιασμό διαδρομών (Leu & Wang, 2006).

Ένα GIS μπορεί να αποθηκεύσει σε ψηφιακή μορφή στοιχεία του πραγματικού κόσμου (τόπους, δρόμους, κτίρια) και να αποδίδονται σε ορθογώνιο σχέδιο παράλληλων γραμμών σάρωσης που ακολουθείται από δέσμη ηλεκτρονίων σε μία οθόνη, παρέχοντας εικόνα (raster images). Οι εικόνες αυτές αποτελούνται από σειρές και στήλες κυψελών με κάθε κελί να έχει μία τιμή, η οποία είναι διακριτή, συνεχής ή μηδενική. Η πληροφορία απεικονίζεται με εικονοστοιχεία στην εικόνα (πλέγμα σημείων με πληροφορία). Άλλη μέθοδος αποθήκευσης δεδομένων σε GIS είναι με την μέθοδο

διανύσματος, όπου χρησιμοποιούνται σημεία, γραμμές, ή σχήματα από γραμμές (πολύγωνα) για την αναπαραγωγή των αντικειμένων. Τα δεδομένα διανύσματος εμφανίζονται ως γραφικά με ξεκάθαρα όρια, που χρησιμοποιούνται σε παραδοσιακούς χάρτες ενώ τα δεδομένα raster εμφανίζονται ως εικόνα με ίσως πολύ εμφάνιση των ορίων των αντικειμένων (Leu & Wang, 2006).

Το σύστημα GIS έχει την ικανότητα να χειρίζεται και να εκτελεί λειτουργίες σε γεωχωρικά δεδομένα, όπως η τοποθεσία (χωρικά δεδομένα) και τα χαρακτηριστικά της (δεδομένα χαρακτηριστικών). Στις μέρες μας, έχει ενισχυθεί η αποτελεσματικότητα της αστυνομίας στην έρευνα εγκλημάτων, όπου η ανάλυση και πρόβλεψή τους γίνεται χρησιμοποιώντας, κατά κόρον, συστήματα GIS. Η κατανόηση του «πού» «πότε» και «γιατί» έγινε ένα έγκλημα και η γραφική απεικόνιση του σε χάρτες βοηθάει στην καταπολέμηση της εγκληματικότητας (Ansari & Kale, 2014).

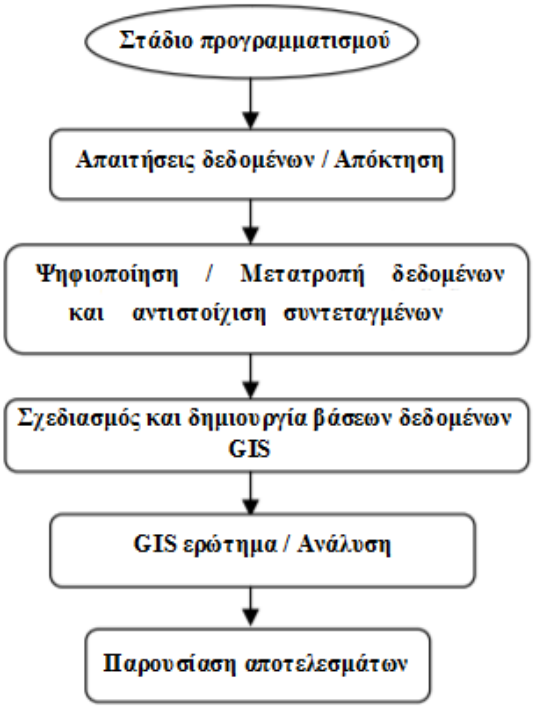
Η χρήση ψηφιακών χαρτών οπτικοποιεί με ταχύ ρυθμό τις τάσεις της εγκληματικότητας και παρέχει χωρικά πρότυπα εγκληματικών συμβάντων. Η προβολή σε χάρτες των τόπων εγκληματικών γεγονότων, των συλλήψεων που έγιναν, των τόπων διαμονής των δραστών, καθώς και άλλων πληροφοριών παρέχει εύκολη μέθοδο απεικόνισης δραστηριοτήτων και γεγονότων. Ο τρόπος αυτός απεικόνισης παρέχει αμεσότητα εύρεσης πληροφοριών και ιδιαιτέρως περισσότερων του ενός είδους πληροφορίες ταυτόχρονα. Οι θέσεις του εγκλήματος στους χάρτες ενός συστήματος GIS μπορούν να παρουσιάζουν πληροφορίες, όπως για την ημέρα και το είδος του εγκλήματος, την συχνότητα συγκεκριμένου εγκλήματος, το τρόπο που λειτουργίας δράστη εγκληματικής ενέργειας (modus operandi) (Johnson, C.P., 2000).

Μεγάλος όγκος δεδομένων που αφορούν εγκλήματα επεξεργάζονται με διάφορες τεχνικές, τόσο στην χωρική αλλά και χρονική τους πληροφορία. Μια τεχνική ανάλυσης είναι αυτή των Hotspots όπου αντιπροσωπεύουν περιοχές με συγκέντρωση εγκλημάτων. Υπάρχουν διάφορες κατηγορίες εύρεσης Hotspot, όπως είναι η διερευνητική και γεωστατική ανάλυση των δεδομένων, καθώς και η χωρική αυτοσυσχέτιση για εντοπισμό Hotspot εγκληματικότητας. Επομένως, με την χρήση των μεθόδων εύρεσης Hotspot παρέχονται χάρτες που απεικονίζουν την θέση του εγκλήματος, την υψηλή συγκέντρωση εγκληματικότητας, καθώς και τον διαχωρισμό του χάρτη σε περιοχές όπου παρατηρούνται συγκεκριμένα εγκλήματα (π.χ. πώληση ναρκωτικών). Ένα Hotspot σημείο φανερώνει μια κατάσταση σύνθεσης συστάδων σε μια χωρική κατανομή, όπου αποτελεί γεωγραφική περιοχή με συγκεκριμένο αριθμό εγκληματικών περιστατικών και φανερώνει ιδιαίτερο ενδιαφέρον. Η αποτύπωση σε χάρτη τάσεων εγκληματικής

δραστηριότητας, περιοχών με υψηλή πυκνότητα εγκλημάτων, καθώς και με χρονικές πληροφορίες μπορεί να φανεί πολύ χρήσιμη στην αστυνομία για την χάραξη της πολιτικής της (Ansari & Kale, 2014).

Υπάρχουν διάφορες μεθοδολογίες που ακολουθούνται κατά την χρήση GIS μία εκ των οποίων που προτείνεται, αποτελείται από τα στάδια που φαίνονται στο σχήμα 2-13 και είναι τα εξής:

- Στάδιο προγραμματισμού
- Απαιτήσεις δεδομένων / Απόκτηση
- Ψηφιοποίηση / Μετατροπή δεδομένων και αντιστοίχιση συντεταγμένων
- Σχεδιασμός και δημιουργία βάσεων δεδομένων GIS
- GIS ερώτημα / Ανάλυση
- Παρουσίαση αποτελεσμάτων



Σχήμα 2-13: Διάγραμμα ροής μεθοδολογίας GIS.

Πηγή: Ojiako et al., 2016.

Η σχεδίαση μιας δομημένης ψηφιακής βάσης δεδομένων που θα χρησιμοποιηθεί για τη χρήση σε ένα σύστημα GIS είναι σημαντική και πολύπλοκη διαδικασία. Η δημιουργία ενός λεπτομερούς μοντέλου δεδομένων αποτελείται από τρία επίπεδα που είναι:

1. Εννοιολογικός Σχεδιασμός

Αποτελεί το αρχικό βήμα του σχεδιασμού της βάσης δεδομένων, όπου προσδιορίζεται και περιγράφεται το περιεχόμενο της. Καθορίζονται τα βασικά εδαφικά αντικείμενα και η χωρική σχέση που έχουν μεταξύ τους. Το μοντέλο δεδομένων εστιάζει στον άνθρωπο και είναι συνήθως μερικά δομημένο από επιλεγμένα αντικείμενα και διεργασίες, σχετιζόμενα με κάποιο τομέα προβλημάτων. Ο συλληφθέν σχεδιασμός πραγματοποιείται χωρίς τη χρήση λογισμικού ή συστημάτων όπως συμβαίνει στην υλοποίηση της βάσης δεδομένων.

2. Λογική σχεδίαση

Το δεύτερο στάδιο διαμορφώνει τις οντότητες του πραγματικού περιβάλλοντος και τις μοντελοποιεί σύμφωνα με τον λογικό σχεδιασμό του πραγματικού κόσμου. Χρησιμοποιείται μια σχεσιακή βάση δεδομένων για την καταγραφή των δεδομένων του εννοιολογικού σχεδιασμού στο σύστημα GIS. Οι οντότητες, τα χαρακτηριστικά τους και οι σχέσεις τους αντιπροσωπεύονται με ενιαίο τρόπο ενημέρωσης ώστε να μην υπάρχουν επαναλαμβανόμενες ίδιες καταγραφές που αλληλεπικαλύπτονται, καθώς και απώλεια πληροφοριών.

3. Φυσικός σχεδιασμός

Το τρίτο στάδιο περιλαμβάνει την απόδοση των οντοτήτων του πραγματικού κόσμου στο σύστημα GIS βάση του δομημένου μοντέλου που επιλέχθηκε, όπως είναι το σχεσιακό, γεωσχεσιακό, δικτυακό ή ιεραρχικό. Τα γεωχωρικά δεδομένα και τα χαρακτηριστικά τους είναι κατάλληλα δομημένα, ώστε να γίνουν αποδεκτά από το λογισμικό και το υλικό του GIS που χρησιμοποιείται. Με αυτό τον τρόπο επιτυγχάνεται η ορθή απόδοση της εικόνας (σημεία, γραμμές, πολύγωνα) των οντοτήτων και η σωστή λειτουργία των ρυθμισμένων απαιτήσεων. Τέτοιες ρυθμίσεις μπορεί να είναι οι εξής:

- Οι αποθηκευμένες πληροφορίες να είναι προσπελάσιμες
- Οι αποθηκευμένες πληροφορίες να μπορούν να ανακτηθούν σε μεταγενέστερη ημερομηνία,
- Να γίνεται ενημέρωση της βάσης δεδομένων κατά διαστήματα,
- Να μπορούν να εκτελεσθούν αναλυτικές λειτουργίες,
- Να μπορούν να απαντηθούν γενικές ερωτήσεις που αφορούν το τομέα προβλημάτων.

(Ojiako et al., 2016)

3 Μεθοδολογία Πρόβλεψης Εγκλημάτων

3.1 Καθορισμός Προβλήματος

Η εγκληματική συμπεριφορά εξαρτάται από περιστασιακούς και ποικίλους παράγοντες και διευκολύνεται από εγκληματικές ευκαιρίες και περιβαλλοντικά χαρακτηριστικά. Οι πολίτες που ρέπουν προς την διάπραξη εγκλημάτων είναι άνθρωποι συνήθειας και τείνουν σε επαναλαμβανόμενα μοτίβα συμπεριφοράς.

Η παρουσία εγκλημάτων σε μια οριοθετημένη περιοχή αποτελεί ύπαρξη γεωχωρικών γεγονότων και δεδομένων με θεματική και χρονική συγγένεια, ικανών προς έρευνα. Η γνώση των συνηθειών των εγκληματιών, καθώς και των περιοχών δράσης τους, επιφέρει αποτελεσματικότερες ενέργειες αντιμετώπισης (πρόληψης και καταστολής) του εγκλήματος, από τα όργανα επιβολής του νόμου. Μια τεχνική ανάλυση δεδομένων εγκληματικότητας που διαμορφώνει γεωχωρικές κατανομές μπορεί καλύτερα να εντοπίσει και να περιγράψει τα μοτίβα εγκληματικότητας. Μοντελοποιούνται συγκεντρωτικές επιφάνειες συνόλων δεδομένων (γραφικές παραστάσεις), όπου αποθηκεύονται γεωχωρικής κατανομής εγκλήματα, εντός των καθορισμένων περιοχών. Τα γραφήματα χρησιμοποιούνται για εντοπισμό συνόλου δεδομένων με παρόμοια γεωχωρικής κατανομής χαρακτηριστικά (Phillips & Lee, 2012).

Τα τελευταία 50 χρόνια, έχουν γίνει πολλές έρευνες για την οικονομική ανάλυση της εγκληματικότητας και εντοπισμό των οικονομικών αιτιών της εγκληματικότητας, με ποικιλία της κύριας μεταβλητής επιρροής της, που άλλοτε μπορεί να ήταν η αγορά εργασίας και άλλοτε το εισόδημα των πολιτών. Ο βραβευμένος με Νόμπελ οικονομικών επιστημών Becker Gary προσέγγισε το θέμα της επιρροής των οικονομικών παραγόντων στην εγκληματικότητα από την οπτική πλευρά του δράστη της εγκληματικής ενέργειας. Έδωσε έμφαση στο κέρδος που μπορεί να έχει ο δράστης από την ενέργειά του και στο κόστος που θα έχει η πράξη του σε σχέση με την τιμωρία του εάν συλληφθεί. Ακόμη, ερευνήθηκε κατά πόσο θα έπραττε ένας πολίτης μία εγκληματική ενέργεια σε περίπτωση απολαβής υψηλά αμειβόμενων μισθών, παρέχοντάς του οικονομική ευημερία. Διάφορες έρευνες για την επιρροή της εγκληματικότητας από το σύνολο των εγκληματικών ενεργειών δεν έδιναν σε μεγάλο βαθμό αξιόπιστα αποτελέσματα. Το μοντέλο του Becker Gary αναδεικνύει πως η κατηγοριοποίηση των εγκλημάτων (κατά της ιδιοκτησίας, κατά

της ζωής κτλ.) παρέχει αξιόλογα αποτελέσματα, συνδέοντας τις κατηγορίες αυτές με κοινωνικούς και δημογραφικούς παράγοντες. Επομένως, εάν για παράδειγμα ο δείκτης ακαθάριστου εγχώριου προϊόντος κατά κεφαλή είναι πάρα πολύ μικρός και υπάρχει μεγάλη ανεργία θα έχει επίδραση σε διάπραξη εγκλημάτων κατά της ιδιοκτησίας και των περιουσιακών δικαιωμάτων και όχι κατά της πολιτειακής εξουσίας (απείθεια) και συνεπώς σε όλα τα εγκλήματα στο σύνολό τους (Becker, 1968).

Εάν θελήσουμε να κατηγοριοποιήσουμε τους παράγοντες επιρροής των εγκλημάτων σε γενικές κατηγορίες μπορούμε να πούμε, βάσει βιβλιογραφίας, ότι υπάρχουν οι κοινωνικοί (μορφωτικό επίπεδο πολιτών, εκπαιδευτικά συστήματα), οικονομικοί (ανεργία, πληθωρισμός, ΑΕΠ κατά κεφαλήν) και δημογραφικοί (πληθυσμός ανά περιφέρεια, ποσοστό ανδρών-γυναικών-παιδιών κτλ.) παράγοντες. Επιπροσθέτως, κατά Becker Gary, υπάρχει και η κατηγορία της ορθής λειτουργίας των κρατικών μηχανισμών, όπως ύπαρξη συστήματος δικαιοσύνης, πρόληψης και καταστολής από την αστυνομία, αυστηρότητα επιβολής των νόμων και άλλα.

Ο συνδυασμός τεχνικών της εξόρυξης δεδομένων εγκλημάτων σε σχέση με κοινωνικούς και δημογραφικούς παράγοντες, καθώς και η απεικόνιση και ανάλυση των δεδομένων σε σύστημα GIS, δίνει την δυνατότητα για μεγαλύτερη κατανόηση της δυναμικής των παράνομων δραστηριοτήτων. Παρέχεται η ευκαιρία για ανακάλυψη των μοτίβων εγκληματολογικής συμπεριφοράς και εντοπίζονται οι προβληματικές περιοχές, βοηθώντας στην καταπολέμηση του εγκλήματος. Κατανοείται το που, πότε και γιατί συγκεκριμένα εγκλήματα (σχετικά με το νόμισμα, κατά της ιδιοκτησίας, κατά των περιουσιακών δικαιωμάτων, κατά της ζωής, κτλ.) είναι πιθανό να συμβούν. Η πληροφορία αυτή μπορεί να χρησιμοποιηθεί για περαιτέρω έρευνα αιτιών και στοχευμένες προσπάθειες πρόληψης εγκλήματος (αύξηση περιπολιών, φωτισμός περιοχών, εγκατάσταση φωτογραφικών μηχανών και καμερών, χρήση προηγμένης τεχνολογίας για εντοπισμό παραχαραγμένων χαρτονομισμάτων, καθώς και διαδικτυακής απάτης κτλ.).

3.2 Καθαρισμός, Επεξεργασία και Μετασχηματισμός Δεδομένων

Για την παρούσα έρευνα χρησιμοποιήθηκαν δεδομένα τα οποία πάρθηκαν από την επίσημη ηλεκτρονική διεύθυνση της Ελληνικής Αστυνομίας (www.astynomia.gr) και της Ελληνικής Στατιστικής Αρχής (www.statistics.gr), καθώς επίσης και από τον

διαδικτυακό τόπο (web site) «http://data.gov.gr». Το data.gov.gr είναι ο κεντρικός κατάλογος των δημόσιων δεδομένων που παρέχει πρόσβαση σε βάσεις δεδομένων (datasets) των φορέων της ελληνικής κυβέρνησης. Ο σκοπός του data.gov.gr είναι να αυξηθεί η πρόσβαση σε υψηλής αξίας, μηχανικά αναγνώσιμα σύνολα δεδομένων με την παροχή ενιαίων υπηρεσιών καταλογογράφησης, ευρετηρίασης, αποθήκευσης, αναζήτησης και διαθεσιμότητας των δεδομένων και των πληροφοριών δημόσιου τομέα, καθώς και διαδικτυακές υπηρεσίες προς τους πολίτες και τρίτα συστήματα πληροφοριών (data.gov.gr, 2016).

Η αρχική βάση δεδομένων που χρησιμοποιήθηκε για την έρευνα, περιέχει σύνολο δεδομένων (πίνακες σε αρχείο Excel) που συλλέχθηκαν από τις πηγές που προαναφέρθηκαν και κυρίως από το έντυπο «Στατιστική Επετηρίδα Ελληνικής Αστυνομίας 2016», το οποίο δημιουργήθηκε από την Διεύθυνση Πληροφορικής της Ελληνικής Αστυνομίας. Αρχικώς, τα δεδομένα συλλέχθηκαν σε ένα πίνακα αρχείου Excel (.xlsx) και περιλάμβαναν αδικήματα του ποινικού κώδικα (π.χ. ανθρωποκτονία από αμέλεια, ανθρωποκτονία με πρόθεση, κιβδηλεία, παραχάραξη, πλαστογραφία, εξύβριση, εμπρησμός, αυτοδικία, απόδραση κρατουμένου, απείθεια, αντίσταση) και κατά ιδιοκτησίας και περιουσιακών δικαιωμάτων (διακεκριμένη κλοπή με διάρρηξη, άλλη διακεκριμένη κλοπή, κλοπή με αρπάγη, κλοπή μεταφορικού μέσου για χρήση για πολύ μικρό χρονικό διάστημα, άλλη κλοπή με διάρρηξη, άλλη κλοπή, κλοπές και υπεξαιρέσεις ευτελούς αξίας, ληστεία με αρπάγη, άλλη ληστεία, εκβίαση, απάτη).

Τα εγκλήματα που συλλέχθηκαν κατονομάζονται ειδικά, εάν εξιχνιάστηκαν ή όχι, αναφέρεται ο βαθμός εγκλήματος (κακουργήματα, πλημμελήματα, πταίσματα), η κατάσταση (απόπειρα, τελεσμένο) και κάποια δημογραφικά στοιχεία, όπως η εθνικότητα δράστη (ημεδαπός, αλλοδαπός), η τάξη ηλικίας του δράστη (π.χ. 07-17, 18-24) και το φύλο του δράστη (άνδρας, γυναίκα). Επιπλέον, στα εγκλήματα έχει προστεθεί και η αυτοκτονία που αν και δεν είναι έγκλημα και δεν καταλογίζεται στον δράστη σε περίπτωση απόπειρας. Πρέπει να σημειωθεί ότι νομικά, υπάρχουν περιπτώσεις που συντρέχει ποινή από τον νόμο στη περίπτωση του άρθρου 301 του Π.Κ. "Συμμετοχή σε αυτοκτονία" (εγκλήματα κατά της ζωής) ή όποια ύπαρξη τέλεσης εγκληματικής ενέργειας, καθώς και ηθικά είναι καταδικαστέα από πολλές θρησκείες, διότι ως πράξη δηλώνει ασέβεια προς τον Θεό. Οι αυτοκτονίες και απόπειρες αυτοκτονιών χρησιμοποιούνται στην βάση δεδομένων με επιπλέον στοιχεία όπως φύλο, ηλικία και αίτιο-λόγος (π.χ. οικονομικοί, οικογενειακοί), διότι μπορούν να εξαχθούν πολύτιμες

πληροφορίες σχετικά με τους λόγους αυτοχειρίας και την επιρροή της οικονομικής κρίσης στην Ελλάδα.

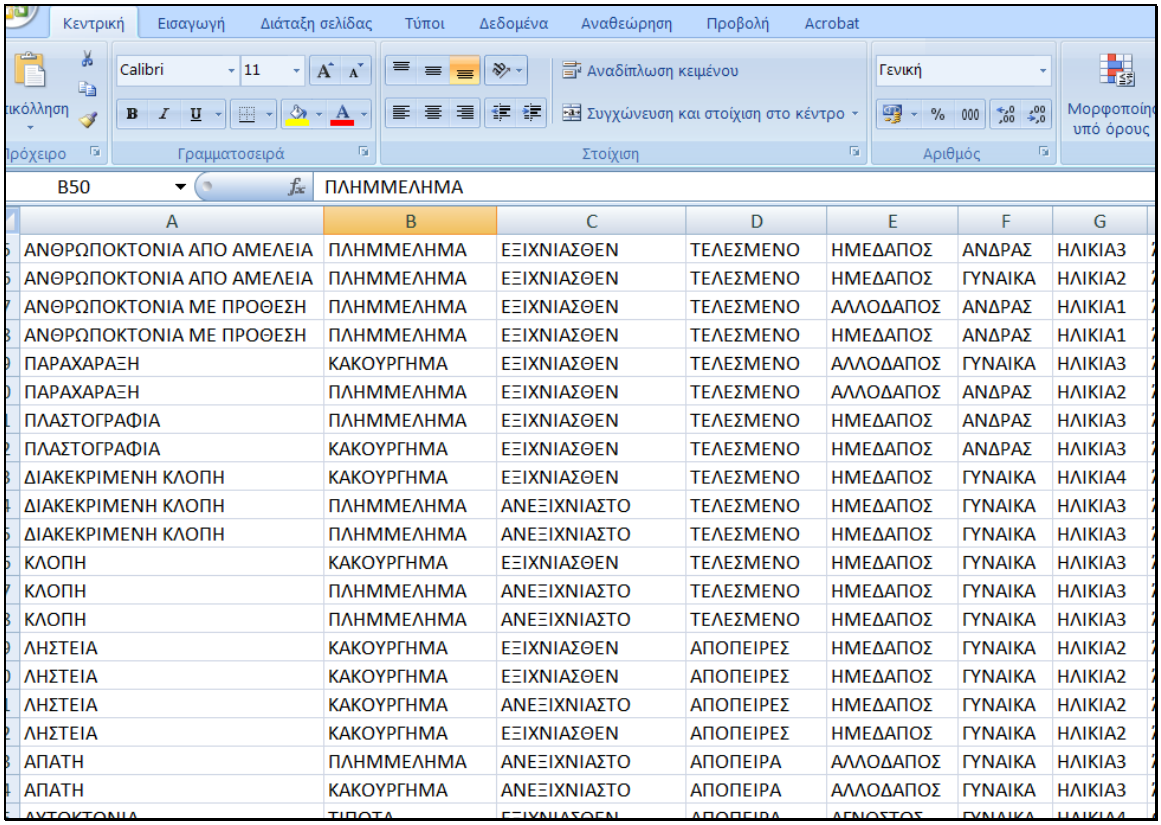
Στην συνέχεια, για την ποιότητα του συνόλου των δεδομένων έγιναν περεταίρω ενέργειες όπως καθαρισμός, μετασχηματισμός και διακριτοποίηση των δεδομένων προς αποφυγή των προβληματικών (θόρυβος, ελλιπή δεδομένα, άχρηστα δεδομένα, λανθασμένες και κατεστραμμένες τιμές, ακραίες τιμές, ασυμβατότητα τιμών), καθώς και για μείωση των δεδομένων και ευκολότερη ομαδοποίηση (διακριτοποίηση σε μεγαλύτερα διαστήματα των γειτονικών τιμών του συνεχούς χαρακτηριστικού, π.χ. ηλικία 07-17,18-34). Εγκλήματα όπως η εξύβριση, ο εμπρησμός, η αυτοδικία, η απόδραση κρατουμένου, η απείθεια και η αντίσταση, αφαιρέθηκαν από τα δεδομένα, διότι δεν προσφέρουν όφελος για τον συγκεκριμένο σκοπό της έρευνας. Ακόμη, αφαιρέθηκαν εγκλήματα που δεν επηρεάζουν το αποτέλεσμα της έρευνας, με ιδιαίτερα πάρα πολύ μικρές τιμές (π.χ. "Κιβδηλεία" με 3 περιπτώσεις σε όλο το έτος 2016, σε αντίθεση με "Άλλη κλοπή" που είναι 83.833 περιπτώσεις) και κάποια άλλα συγχωνεύτηκαν (π.χ. "κλοπή με αρπαγή" και "άλλη κλοπή") για διευκόλυνση της έρευνας.

Τα δεδομένα που συγκεντρώθηκαν περιέχουν ποσοτικά συνεχή χαρακτηριστικά όπως φαίνονται στο σχήμα 3.1 και στην συνέχεια με την προ-επεξεργασία όλων των δεδομένων, δημιουργήθηκε πίνακας σε αρχείο Excel με το καθένα έγκλημα και τα ονομαστικά (nominal) χαρακτηριστικά του (8 στήλες και 123.631 γραμμές), όπως φαίνεται στο σχήμα 3.2.

ΑΔΙΚΗΜΑΤΑ ΚΑΤΑ ΙΔΙΟΚΤΗΣΙΑΣ ΚΑΙ ΠΕΡΙΟΥΣΙΑΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ ΠΟΥ ΕΙΔΙΚΑ ΚΑΤΟΝΟΜΑΖΟΝΤΑΙ	ΑΔΙΚΗΜΑΤΑ ΚΑΤΑ ΒΑΘΜΟ ΕΞΙΧΝΙΑΣΘΕΝΤΑ ΚΑΙ ΜΗ ΕΞΙΧΝΙΑΣΘΕΝΤΑ			ΑΔΙΚΗΜΑΤΑ ΚΑΤΑ ΒΑΘΜΟ ΕΞΙΧΝΙΑΣΘΕΝΤΑ			ΕΞΙΧ
	ΚΑΚΟΥΡΓΗΜΑΤΑ	ΠΛΗΜΜΕΛΗΜΑΤΑ	ΣΥΝΟΛΟ	ΚΑΚΟΥΡΓΗΜΑΤΑ	ΠΛΗΜΜΕΛΗΜΑΤΑ	ΣΥΝΟΛΟ	ΚΑΚΟΥΡ
ΔΙΑΚΕΚΡΙΜΕΝΗ ΚΛΟΠΗ ΜΕ ΔΙΑΡΡΗΞΗ	1012	180	1192	712	80	792	3
ΑΛΛΗ ΔΙΑΚΕΚΡΙΜΕΝΗ ΚΛΟΠΗ	983	350	1333	573	107	680	2
ΚΛΟΠΗ ΜΕ ΑΡΠΑΓΗ	6	1258	1264	1	110	111	
ΚΛΟΠΗ ΜΕΤΑΦΟΡΙΚΟΥ ΜΕΣΟΥ ΓΙΑ ΧΡΗΣΗ ΓΙΑ ΠΟΛΥ ΜΙΚΡΟ ΧΡΟΝΙΚΟ ΔΙΑΣΤΗΜΑ	1	19	20	0	4	4	
ΑΛΛΗ ΚΛΟΠΗ ΜΕ ΔΙΑΡΡΗΞΗ	65	23579	23644	3	1643	1646	
ΑΛΛΗ ΚΛΟΠΗ	148	83685	83833	21	4763	4784	
ΚΛΟΠΕΣ ΚΑΙ ΥΠΕΞΑΙΡΕΣΕΙΣ ΕΥΤΕΛΟΥΣ ΑΞΙΑΣ	0	24	24	0	12	12	
ΛΗΣΤΕΙΑ ΜΕ ΑΡΠΑΓΗ	215	30	245	28	4	32	1
ΑΛΛΗ ΛΗΣΤΕΙΑ	4445	444	4889	764	38	802	2
ΕΚΒΙΑΣΗ	60	139	199	18	42	60	
ΑΠΑΤΗ	259	4242	4501	149	728	877	3

Σχήμα 3-1: Αδικήματα κατά ιδιοκτησίας και περιουσιακών δικαιωμάτων.

Πηγή: Astynomia.gr, 2016.

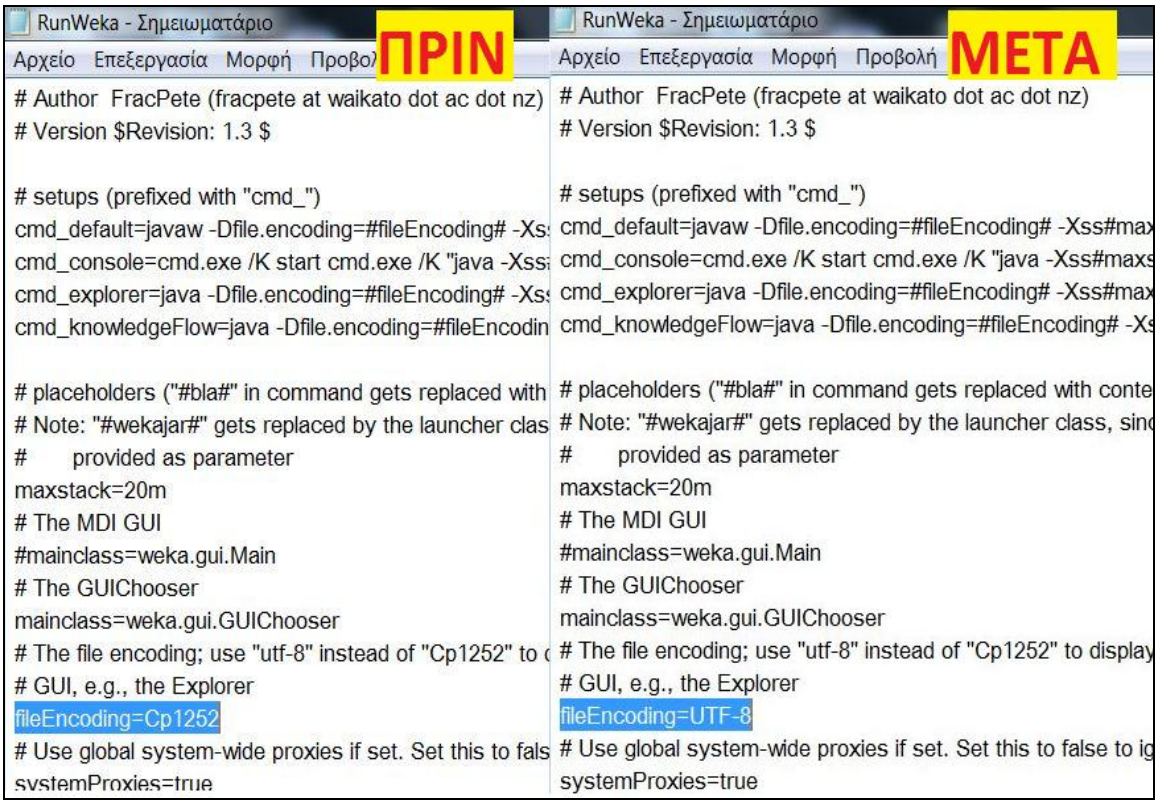


	A	B	C	D	E	F	G
5	ΑΝΘΡΩΠΟΚΤΟΝΙΑ ΑΠΟ ΑΜΕΛΕΙΑ	ΠΛΗΜΜΕΛΗΜΑ	ΕΞΙΧΝΙΑΣΘΕΝ	ΤΕΛΕΣΜΕΝΟ	ΗΜΕΔΑΠΟΣ	ΑΝΔΡΑΣ	ΗΛΙΚΙΑ3
6	ΑΝΘΡΩΠΟΚΤΟΝΙΑ ΑΠΟ ΑΜΕΛΕΙΑ	ΠΛΗΜΜΕΛΗΜΑ	ΕΞΙΧΝΙΑΣΘΕΝ	ΤΕΛΕΣΜΕΝΟ	ΗΜΕΔΑΠΟΣ	ΓΥΝΑΙΚΑ	ΗΛΙΚΙΑ2
7	ΑΝΘΡΩΠΟΚΤΟΝΙΑ ΜΕ ΠΡΟΘΕΣΗ	ΠΛΗΜΜΕΛΗΜΑ	ΕΞΙΧΝΙΑΣΘΕΝ	ΤΕΛΕΣΜΕΝΟ	ΑΛΛΟΔΑΠΟΣ	ΑΝΔΡΑΣ	ΗΛΙΚΙΑ1
8	ΑΝΘΡΩΠΟΚΤΟΝΙΑ ΜΕ ΠΡΟΘΕΣΗ	ΠΛΗΜΜΕΛΗΜΑ	ΕΞΙΧΝΙΑΣΘΕΝ	ΤΕΛΕΣΜΕΝΟ	ΗΜΕΔΑΠΟΣ	ΑΝΔΡΑΣ	ΗΛΙΚΙΑ1
9	ΠΑΡΑΧΑΡΑΞΗ	ΚΑΚΟΥΡΓΗΜΑ	ΕΞΙΧΝΙΑΣΘΕΝ	ΤΕΛΕΣΜΕΝΟ	ΑΛΛΟΔΑΠΟΣ	ΓΥΝΑΙΚΑ	ΗΛΙΚΙΑ3
10	ΠΑΡΑΧΑΡΑΞΗ	ΠΛΗΜΜΕΛΗΜΑ	ΕΞΙΧΝΙΑΣΘΕΝ	ΤΕΛΕΣΜΕΝΟ	ΑΛΛΟΔΑΠΟΣ	ΑΝΔΡΑΣ	ΗΛΙΚΙΑ2
11	ΠΛΑΣΤΟΓΡΑΦΙΑ	ΠΛΗΜΜΕΛΗΜΑ	ΕΞΙΧΝΙΑΣΘΕΝ	ΤΕΛΕΣΜΕΝΟ	ΗΜΕΔΑΠΟΣ	ΑΝΔΡΑΣ	ΗΛΙΚΙΑ3
12	ΠΛΑΣΤΟΓΡΑΦΙΑ	ΚΑΚΟΥΡΓΗΜΑ	ΕΞΙΧΝΙΑΣΘΕΝ	ΤΕΛΕΣΜΕΝΟ	ΗΜΕΔΑΠΟΣ	ΑΝΔΡΑΣ	ΗΛΙΚΙΑ3
13	ΔΙΑΚΕΚΡΙΜΕΝΗ ΚΛΟΠΗ	ΚΑΚΟΥΡΓΗΜΑ	ΕΞΙΧΝΙΑΣΘΕΝ	ΤΕΛΕΣΜΕΝΟ	ΗΜΕΔΑΠΟΣ	ΓΥΝΑΙΚΑ	ΗΛΙΚΙΑ4
14	ΔΙΑΚΕΚΡΙΜΕΝΗ ΚΛΟΠΗ	ΠΛΗΜΜΕΛΗΜΑ	ΑΝΕΞΙΧΝΙΑΣΤΟ	ΤΕΛΕΣΜΕΝΟ	ΗΜΕΔΑΠΟΣ	ΓΥΝΑΙΚΑ	ΗΛΙΚΙΑ3
15	ΔΙΑΚΕΚΡΙΜΕΝΗ ΚΛΟΠΗ	ΠΛΗΜΜΕΛΗΜΑ	ΑΝΕΞΙΧΝΙΑΣΤΟ	ΤΕΛΕΣΜΕΝΟ	ΗΜΕΔΑΠΟΣ	ΓΥΝΑΙΚΑ	ΗΛΙΚΙΑ3
16	ΚΛΟΠΗ	ΚΑΚΟΥΡΓΗΜΑ	ΕΞΙΧΝΙΑΣΘΕΝ	ΤΕΛΕΣΜΕΝΟ	ΗΜΕΔΑΠΟΣ	ΓΥΝΑΙΚΑ	ΗΛΙΚΙΑ3
17	ΚΛΟΠΗ	ΠΛΗΜΜΕΛΗΜΑ	ΑΝΕΞΙΧΝΙΑΣΤΟ	ΤΕΛΕΣΜΕΝΟ	ΗΜΕΔΑΠΟΣ	ΓΥΝΑΙΚΑ	ΗΛΙΚΙΑ3
18	ΚΛΟΠΗ	ΠΛΗΜΜΕΛΗΜΑ	ΑΝΕΞΙΧΝΙΑΣΤΟ	ΤΕΛΕΣΜΕΝΟ	ΗΜΕΔΑΠΟΣ	ΓΥΝΑΙΚΑ	ΗΛΙΚΙΑ3
19	ΛΗΣΤΕΙΑ	ΚΑΚΟΥΡΓΗΜΑ	ΕΞΙΧΝΙΑΣΘΕΝ	ΑΠΟΠΕΙΡΕΣ	ΗΜΕΔΑΠΟΣ	ΓΥΝΑΙΚΑ	ΗΛΙΚΙΑ2
20	ΛΗΣΤΕΙΑ	ΚΑΚΟΥΡΓΗΜΑ	ΕΞΙΧΝΙΑΣΘΕΝ	ΑΠΟΠΕΙΡΕΣ	ΗΜΕΔΑΠΟΣ	ΓΥΝΑΙΚΑ	ΗΛΙΚΙΑ2
21	ΛΗΣΤΕΙΑ	ΚΑΚΟΥΡΓΗΜΑ	ΑΝΕΞΙΧΝΙΑΣΤΟ	ΑΠΟΠΕΙΡΕΣ	ΗΜΕΔΑΠΟΣ	ΓΥΝΑΙΚΑ	ΗΛΙΚΙΑ2
22	ΛΗΣΤΕΙΑ	ΚΑΚΟΥΡΓΗΜΑ	ΕΞΙΧΝΙΑΣΘΕΝ	ΑΠΟΠΕΙΡΕΣ	ΗΜΕΔΑΠΟΣ	ΓΥΝΑΙΚΑ	ΗΛΙΚΙΑ2
23	ΑΠΑΤΗ	ΠΛΗΜΜΕΛΗΜΑ	ΑΝΕΞΙΧΝΙΑΣΤΟ	ΑΠΟΠΕΙΡΑ	ΑΛΛΟΔΑΠΟΣ	ΓΥΝΑΙΚΑ	ΗΛΙΚΙΑ3
24	ΑΠΑΤΗ	ΚΑΚΟΥΡΓΗΜΑ	ΑΝΕΞΙΧΝΙΑΣΤΟ	ΑΠΟΠΕΙΡΑ	ΑΛΛΟΔΑΠΟΣ	ΓΥΝΑΙΚΑ	ΗΛΙΚΙΑ3
25	ΑΥΤΟΚΤΟΝΙΑ	ΤΙΠΟΤΑ	ΕΞΙΧΝΙΑΣΘΕΝ	ΑΠΟΠΕΙΡΑ	ΑΓΝΩΣΤΟΣ	ΓΥΝΑΙΚΑ	ΗΛΙΚΙΑ4

Σχήμα 3-2: Βάση δεδομένων σε πίνακα Excel.

Ο δημιουργηθέν πίνακας δεδομένων (Excel), θα χρησιμοποιηθεί για εξόρυξη δεδομένων στο λογισμικό WEKA (για περισσότερες πληροφορίες σχετικά με το λογισμικό WEKA βλέπε παράρτημα 1). Για τον λόγο αυτό, θα πρέπει η βάση δεδομένων να είναι σε μορφή επεξεργάσιμη για το λογισμικό και να παρέχει αναλύσιμα αποτελέσματα. Το πρόγραμμα WEKA χρησιμοποιεί αλγορίθμους, κάποιοι εκ των οποίον λειτουργούν καλύτερα με ονομαστικά δεδομένα και άλλοι με δυαδικά. Επίσης, δεν αναγνωρίζει την ελληνική γλώσσα και μπορούν να μετατραπούν τα χαρακτηριστικά σε αγγλικές διακριτές ονομασίες ή να παραμετροποιηθεί το πρόγραμμα. Στην δεύτερη περίπτωση εξάγεται η βάση δεδομένων σε αρχείο .csv (comma separated values), κάτι το οποίο υπάρχει ως επιλογή από το λογισμικό MS Excel και μπορεί να αναγνωριστεί από το λογισμικό WEKA. Στην συνέχεια, ανοίγεται το αρχείο .csv από κάποιο πρόγραμμα ανάγνωσης και επεξεργασίας κειμένου (έγγραφο κειμένου, notepad) μετατρέπονται τα ¨;¨ (ερωτηματικά) του κειμένου σε ¨,¨ (κόμμα), γιατί τα αρχεία .csv χρησιμοποιούν ορισμένους διαφορετικούς χαρακτήρες για συμβολισμούς (κενό, αλλαγή γραμμής κτλ.) και αποθηκεύεται σε κωδικοποίηση (encoding) “UTF-8”. Το λογισμικό WEKA χρησιμοποιεί αρχεία με κατάληξη .arff (Attribute-Relation File Format) και μπορεί να μετατρέψει σε αυτήν την μορφή τα αρχεία .csv. Τέλος, για να μπορεί το πρόγραμμα

WEKA να διαβάζει τα αρχεία με κωδικοποίηση “UTF-8”, θα πρέπει να αλλαχθεί στο αρχείο RunWeka.ini (βρίσκεται στους φακέλους που είναι εγκατεστημένο το πρόγραμμα WEKA στον υπολογιστή) η κωδικοποίηση από “Cp1252” σε “UTF-8” όπως φαίνεται στο σχήμα 3.3.



Σχήμα 3-3: Αρχείο RunWeka.ini.

Στην παρούσα εργασία θα χρησιμοποιηθεί η πρώτη μέθοδος της μετονομασίας της βάσης δεδομένων με αγγλικές διακριτές ονομασίες, ώστε να αποφευχθεί η δεύτερη μέθοδος, η οποία είναι περισσότερο πολύπλοκη, καθώς και για διευκόλυνση της διαδικασίας και αποφυγή οποιουδήποτε λάθους.

3.3 Υλοποίηση Πινάκων Δεδομένων

Η δημιουργία της βάσης δεδομένων σε αρχείο Excel έγινε με την συλλογή τον καθαρισμό, μετασχηματισμό και διακριτοποίηση των δεδομένων και χρησιμοποιήθηκαν με αγγλική διακριτή ονομασία τα εξής χαρακτηριστικά:

1 CRIME=ΕΓΚΛΗΜΑ

2 RANK=ΒΑΘΜΟΣ

3 SOLVING=ΕΞΙΧΝΙΑΣΗ

4 CONDITION=ΚΑΤΑΣΤΑΣΗ

5 NATIONALITY=ΕΘΝΙΚΟΤΗΤΑ

6 SEX=ΦΥΛΟ

7 AGE=ΗΛΙΚΙΑ

8 CAUSE=ΑΙΤΙΑ

Η επιλογή των εγκλημάτων έγινε επιλεκτικά από τις ομάδες διαχωρισμού τους, με κριτήριο την σοβαρότητά τους και το αντίκτυπο της κάθε ενέργειας στην κοινωνία, καθώς και κατά πόσο βοηθούν την συγκεκριμένη έρευνα. Η ανθρωποκτονία από αμέλεια και ανθρωποκτονία με πρόθεση ανήκουν στα εγκλήματα κατά της ζωής, η ληστεία, κλοπή και διακεκριμένη κλοπή ανήκουν στα εγκλήματα κατά της ιδιοκτησίας, η απάτη υπάγεται στα εγκλήματα κατά των περιουσιακών δικαιωμάτων και η παραχάραξη και πλαστογραφία, αντιστοίχως, στα εγκλήματα σχετικά με το νόμισμα και τα υπομνήματα. Η διακεκριμένη κλοπή αναφέρεται σε αντικείμενα προορισμένα για θρησκευτική λατρεία, επιστημονικής, καλλιτεχνικής, αρχαιολογικής ή ιστορικής σημασίας που εκτίθενται δημοσίως ή σε μεταφερόμενα αντικείμενα με δημόσια συγκοινωνία και από ταξιδιώτη. Ακόμη, διακεκριμένη κλοπή διαπράττεται όταν δύο ή περισσότεροι ενώθηκαν για τον σκοπό αυτό και εάν πρόσωπο τελεί κλοπές κατ' επάγγελμα ή κατά συνήθεια ή το αντικείμενο υπερβαίνει τις 73.000 ευρώ. Τέλος, χρησιμοποιείται στην βάση δεδομένων και η αυτοκτονία, ώστε να εξαχθούν πολύτιμα συμπεράσματα σχετικά με τους λόγους αυτοχειρίας και την επιρροή της από την δυσχερή επικρατούσα οικονομική κατάσταση, τα τελευταία χρόνια, στην Ελλάδα.

Πίνακας 3-1: CRIME=ΈΓΚΛΗΜΑ.

Διακριτή ονομασία	Τιμή
MANSL (MANSLAUGHTER)	ΑΝΘΡΩΠΟΚΤΟΝΙΑ ΑΠΟ ΑΜΕΛΕΙΑ
HOM (HOMICIDE)	ΑΝΘΡΩΠΟΚΤΟΝΙΑ ΜΕ ΠΡΟΘΕΣΗ
ROB (ROBBERY)	ΛΗΣΤΕΙΑ
THEFT	ΚΛΟΠΗ
DIS-TH (DISTINGUISHED THEFT)	ΔΙΑΚΕΚΡΙΜΕΝΗ ΚΛΟΠΗ
FRAUD	ΑΠΑΤΗ
COUN (COUNTERFEITING)	ΠΑΡΑΧΑΡΑΞΗ
FORG (FORGERY)	ΠΛΑΣΤΟΓΡΑΦΙΑ
SUIC (SUICIDE)	ΑΥΤΟΚΤΟΝΙΑ

Σύμφωνα με το άρθρο 18 του Ποινικού Κώδικα τα εγκλήματα διαχωρίζονται σε κακουργήματα, πλημμελήματα και πταίσματα. Τα κακουργήματα είναι εξαιρετικής βαρύτητας εγκλήματα, τα πλημμελήματα είναι μεσαίας βαρύτητας εγκλήματα και τα πταίσματα είναι τα ελαφρότερα εγκλήματα και τιμωρούνται με κράτηση (όχι φυλάκιση) ή πρόστιμο. Οι τιμές για τον βαθμό εγκλήματος που επιλέχθηκαν στην βάση δεδομένων είναι το "κακούργημα" και "πλημμέλημα" ή "τίποτα" για την περίπτωση της αυτοκτονίας. Εξαιρούνται τα πταίσματα, τα οποία δεν εμπεριέχονταν στις πηγές δεδομένων και σε αντίθετη περίπτωση θα προσαύξαναν τα δεδομένα κατά πολύ μεγάλο βαθμό, δυσκολεύοντας την έρευνα και πιθανόν αποδίδοντας μη αξιόπιστα αποτελέσματα, αφού δεν αντιπροσωπεύουν δεδομένα μεγάλης βαρύτητας και τα εγκλήματα στα οποία συνήθως αναφέρονται δεν είναι (έχουν καθαριστεί) στην τελική βάση δεδομένων.

Πίνακας 3-2: RANK=ΒΑΘΜΟΣ

Διακριτή ονομασία	Τιμή
MISD (MISDEMEANOR)	ΠΛΗΜΜΕΛΗΜΑ
FEL (FELONY)	ΚΑΚΟΥΡΓΗΜΑ
NOTH (NOTHING)	ΤΙΠΟΤΑ

Η εξιχνίαση του εγκλήματος έχει σημασία στην απόδοση της δικαιοσύνης και την ατιμωρησία του δράστη σε περίπτωση ανεξιχνίαστων εγκλημάτων. Ακόμη, βοηθάει στην ορθή λειτουργία των κρατικών μηχανισμών δικαιοσύνης, καθώς και στην επιρροή που ασκείται στους εγκληματίες για ροπή σε καθ' επάγγελμα και καθ' εξακολούθηση διάπραξης εγκλημάτων. Οι τιμές που επιλέχθηκαν για το χαρακτηριστικό της εξιχνίασης είναι το ¨εξιχνιασθέν¨ και το ¨ανεξιχνίαστο¨.

Πίνακας 3-3: SOLVING=ΕΞΙΧΝΙΑΣΗ

Διακριτή ονομασία	Τιμή
SOL (SOLVED)	ΕΞΙΧΝΙΑΣΘΕΝ
UNSOL (UNSOLVED)	ΑΝΕΞΙΧΝΙΑΣΤΟ

Η κατάσταση στην οποία ένα έγκλημα βρίσκεται, δηλαδή αν έχει ολοκληρωθεί ή όχι, επηρεάζει την ποινή (κακούργημα, πλημμέλημα) μιας και το αποτέλεσμα της πράξης είναι διαφορετικό. Οι τιμές για την εξιχνίαση εγκλήματος είναι το ¨τελεσμένο¨ και το ¨απόπειρα¨.

Πίνακας 3-4: CONDITION=ΚΑΤΑΣΤΑΣΗ

Διακριτή ονομασία	Τιμή
ENDED	ΤΕΛΕΣΜΕΝΟ
ATT (ATTEMPT)	ΑΠΟΠΕΙΡΑ

Οι τιμές του χαρακτηριστικού ¨εθνικότητα¨, σε περίπτωση Έλληνα είναι ¨ημεδαπός¨, ειδάλλως για όποια άλλη εθνικότητα είναι ¨αλλοδαπός¨. Τέλος, εάν δεν προκύπτει η εθνικότητα του δράστη από της αρχικές πηγές δεδομένων τότε παίρνει την τιμή ¨άγνωστος¨.

Πίνακας 3-5: NATIONALITY=ΕΘΝΙΚΟΤΗΤΑ

Διακριτή ονομασία	Τιμή
NAT (NATIVE)	ΗΜΕΔΑΠΟΣ
FOREI (FOREIGN)	ΑΛΛΟΔΑΠΟΣ
UNKN (UNKNOWN)	ΆΓΝΩΣΤΟΣ

Σχετικά με το φύλλο του δράστη της εγκληματικής ενέργειας, οι τιμές που αποδίδονται στο χαρακτηριστικό ¨φύλλο¨ είναι ¨άνδρας¨ και ¨γυναίκα¨.

Πίνακας 3-6: SEX=ΦΥΛΟ

Διακριτή ονομασία	Τιμή
MAN	ΑΝΔΡΑΣ
WOMAN	ΓΥΝΑΙΚΑ

Η ηλικία των δραστών έχει διακριτοποιηθεί σε 4 ομάδες που είναι ηλικίες από 07 έως και 17 με τιμή ¨ηλικία1¨, από 18 έως και 34 με τιμή ¨ηλικία2¨, από 35 έως και 59 με τιμή ¨ηλικία3¨ και από 60 και άνω με τιμή ¨ηλικία4¨.

Πίνακας 3-7: AGE=ΗΛΙΚΙΑ

Διακριτή ονομασία	Τιμή
AGE1	ΗΛΙΚΙΑ1=07-17
AGE2	ΗΛΙΚΙΑ2=18-34
AGE3	ΗΛΙΚΙΑ3=35-59
AGE4	ΗΛΙΚΙΑ4=60-

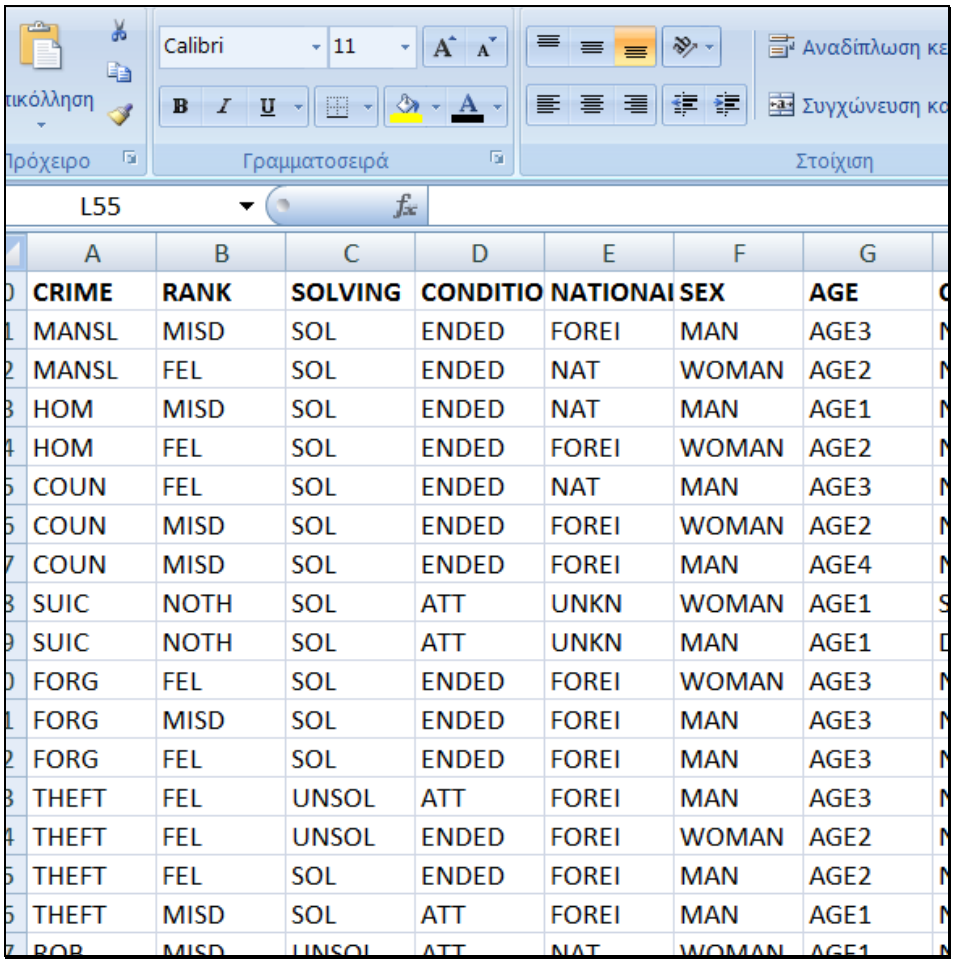
Η αιτία των εγκλημάτων παίρνει τιμές ¨οικονομικοί¨, ¨οικογενειακοί¨, ¨ασθένεια¨ ¨αισθηματικοί¨ και ¨άγνωστοι λόγοι¨ , αναλόγως με τον λόγο για τον οποίο έγιναν και συνάδει με την έννοια των τιμών.

Πίνακας 3-8: CAUSE=ΑΙΤΙΑ

Διακριτή ονομασία	Τιμή
NOT-KN (NOT KNOWN)	ΆΓΝΩΣΤΟΙ ΛΟΓΟΙ
SENS (SENSITIVE)	ΑΙΣΘΗΜΑΤΙΚΟΙ
DISE (DISEASE)	ΑΣΘΕΝΕΙΑ
FAM (FAMILY)	ΟΙΚΟΓΕΝΕΙΑΚΟΙ
ECON (ECONOMIC)	ΟΙΚΟΝΟΜΙΚΟΙ

Εφόσον αντικατασταθούν τα χαρακτηριστικά και οι τιμές τους από την ελληνική στην αγγλική διακριτή ονομασία για να μπορεί το λογισμικό Weka να αναγνωρίζει τους

χαρακτήρες (η ελληνική γλώσσα δεν εμπεριέχεται ως προεπιλογή), δημιουργείται ο τελικός πίνακας Excel (μπορούν να χρησιμοποιηθούν διάφορα φίλτρα για απόκτηση πληροφοριών), όπως φαίνεται στο σχήμα 3.4.



	A	B	C	D	E	F	G
0	**CRIME**	**RANK**	**SOLVING**	**CONDITIO**	**NATIONAI**	**SEX**	**AGE**
1	MANSL	MISD	SOL	ENDED	FOREI	MAN	AGE3
2	MANSL	FEL	SOL	ENDED	NAT	WOMAN	AGE2
3	HOM	MISD	SOL	ENDED	NAT	MAN	AGE1
4	HOM	FEL	SOL	ENDED	FOREI	WOMAN	AGE2
5	COUN	FEL	SOL	ENDED	NAT	MAN	AGE3
6	COUN	MISD	SOL	ENDED	FOREI	WOMAN	AGE2
7	COUN	MISD	SOL	ENDED	FOREI	MAN	AGE4
8	SUIC	NOTH	SOL	ATT	UNKN	WOMAN	AGE1
9	SUIC	NOTH	SOL	ATT	UNKN	MAN	AGE1
10	FORG	FEL	SOL	ENDED	FOREI	WOMAN	AGE3
11	FORG	MISD	SOL	ENDED	FOREI	MAN	AGE3
12	FORG	FEL	SOL	ENDED	FOREI	MAN	AGE3
13	THEFT	FEL	UNSOL	ATT	FOREI	MAN	AGE3
14	THEFT	FEL	UNSOL	ENDED	FOREI	WOMAN	AGE2
15	THEFT	FEL	SOL	ENDED	FOREI	MAN	AGE2
16	THEFT	MISD	SOL	ATT	FOREI	MAN	AGE1
17	ROB	MISD	UNSOL	ATT	NAT	WOMAN	AGE1

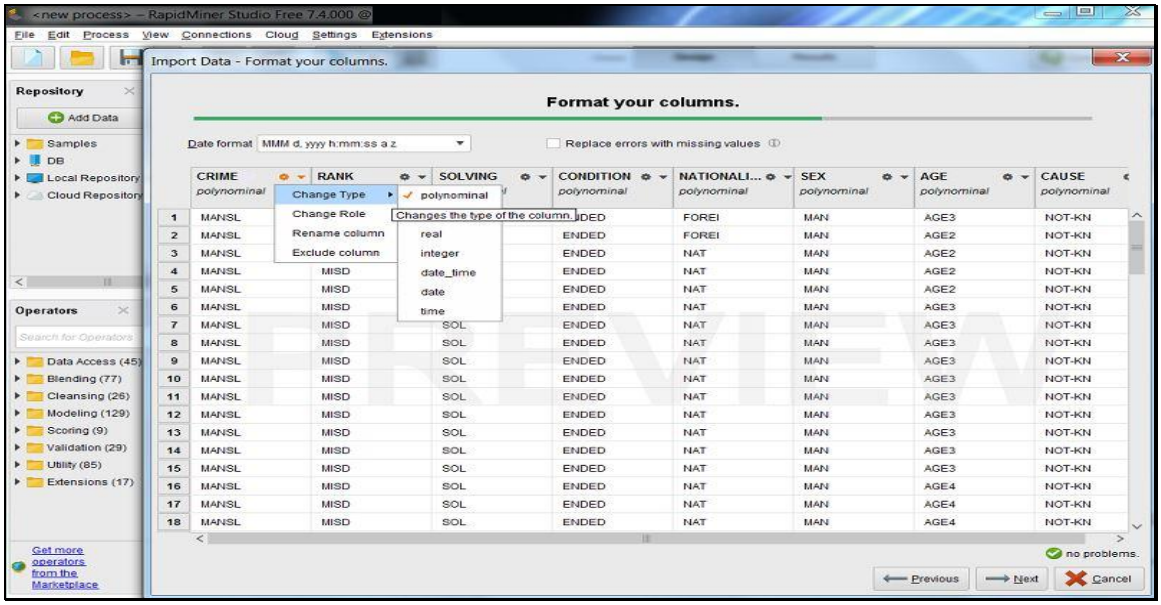
Σχήμα 3-4: Βάση δεδομένων σε πίνακα Excel με αγγλικές διακριτές ονομασίες.

4 Εφαρμογή τεχνικών Εξόρυξης Δεδομένων με το πρόγραμμα Weka και απεικόνιση εγκλημάτων σε Γεωγραφικό Σύστημα Πληροφοριών (GIS)

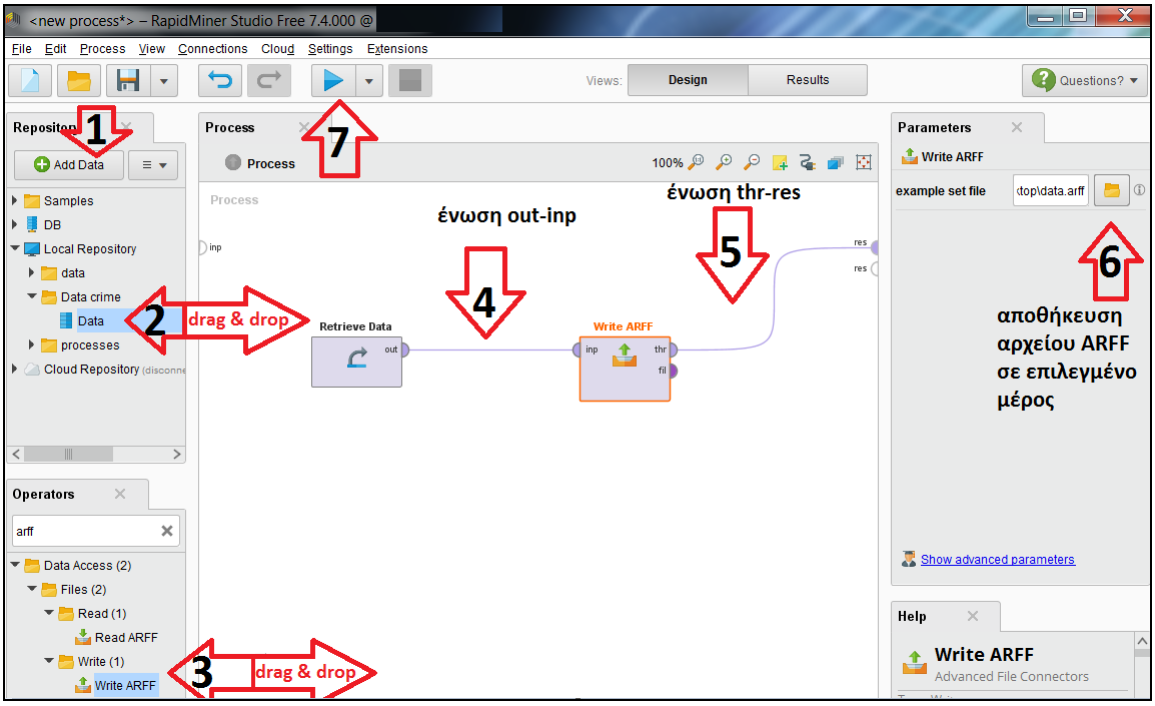
4.1 Εισαγωγή βάσης δεδομένων στο πρόγραμμα WEKA

Η δημιουργηθείσα βάση δεδομένων Excel (.xlsx) μετατρέπεται σε μορφή αρχείου με κατάληξη ¨.arff ¨ για να εισαχθεί στην συνέχεια στο πρόγραμμα Weka και να γίνει η επεξεργασία και εξόρυξη δεδομένων. Υπάρχουν διάφοροι τρόποι για την μετατροπή του πίνακα Excel και στην παρούσα περίπτωση θα χρησιμοποιηθεί το πρόγραμμα RapidMiner, το οποίο είναι ένα πάρα πολύ χρήσιμο εργαλείο για τους ερευνητές και αναλυτές. Με το συγκεκριμένο λογισμικό δίνεται η δυνατότητα για προετοιμασία των δεδομένων, μηχανική και βαθιά μάθηση, εξόρυξη κειμένου και παροχή προγνωστικών analytics (για περισσότερες πληροφορίες σχετικά με το λογισμικό RapidMinner βλέπε παράρτημα 1).

Εγκαθίσταται το λογισμικό RapidMiner, γίνεται η εκκίνησή του και εισάγεται (drag & drop) ο πίνακας Excel (βάση δεδομένων) για εξαγωγή του αρχείου σε κατάληξη .arff, όπως φαίνεται στα σχήματα 4.1 και 4.2. Κατά την εισαγωγή του πίνακα Excel, εάν είναι επιθυμητό, μπορούν να γίνουν τροποποιήσεις στα δεδομένα.

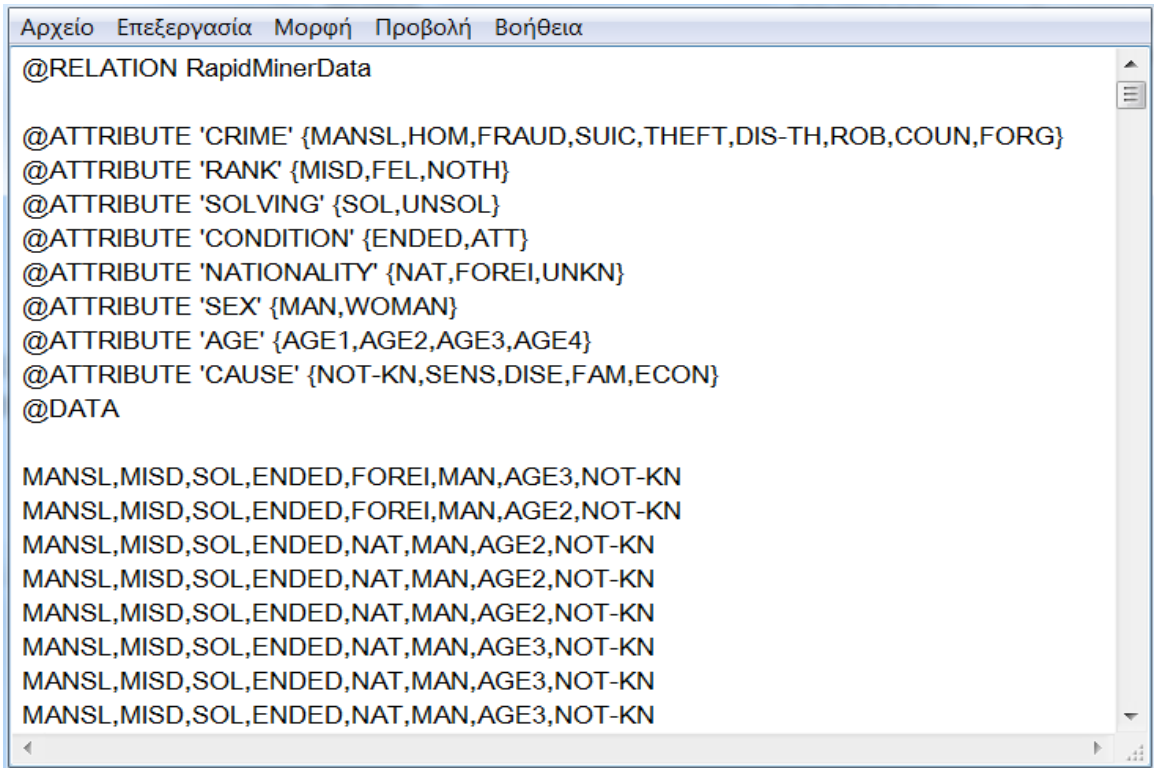


Σχήμα 4-1: Εισαγωγή πίνακα Excel στο RapidMiner.



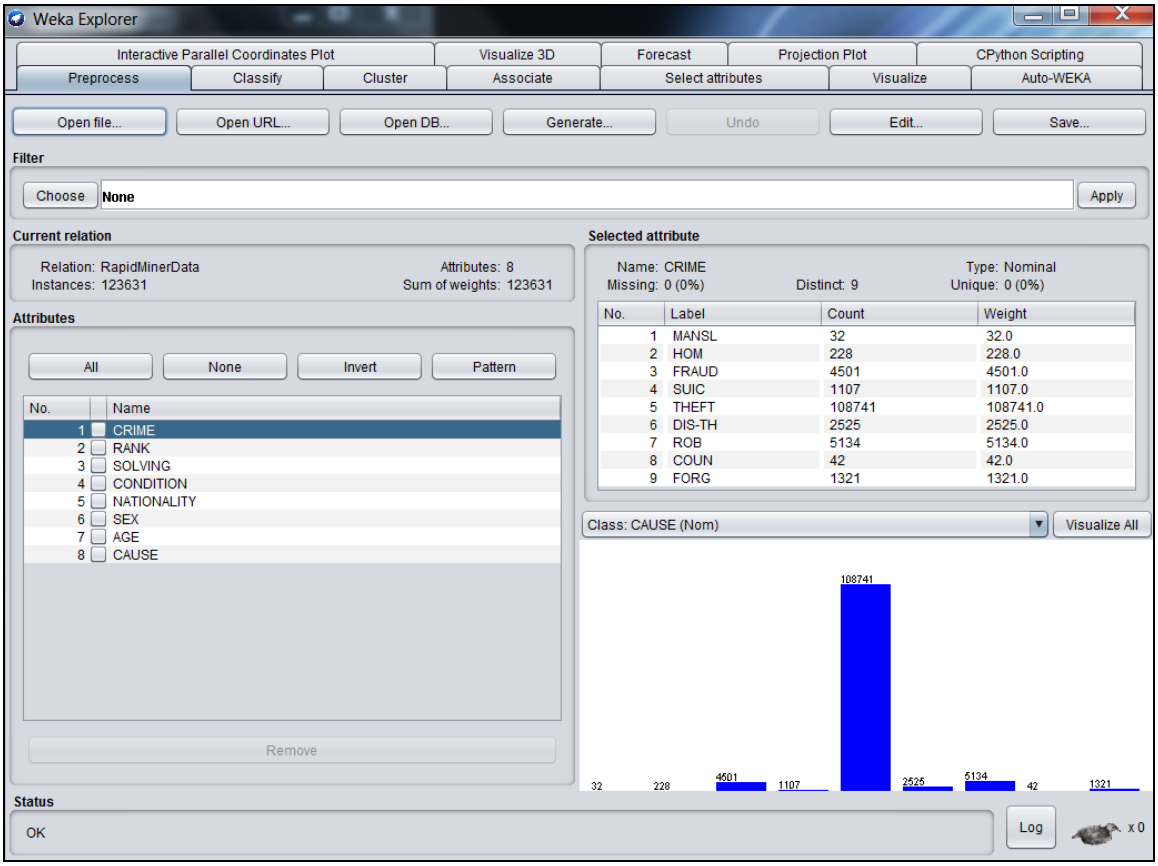
Σχήμα 4-2: Διαδικασία εξαγωγής αρχείου ARFF.

Κατόπιν, το εξαγόμενο αρχείο με κατάληξη ¨.arff ¨ έχει μορφή όπως φαίνεται στο σχήμα 4.3.



Σχήμα 4-3: Βάση δεδομένων σε αρχείο ARFF.

Στην συνέχεια ¨φορτώνεται¨ το αρχείο (ARFF) των δεδομένων στο πρόγραμμα WEKA, όπως φαίνεται στο σχήμα 4.4.



Σχήμα 4-4: Βάση δεδομένων σε αρχείο ARFF.

Στο αρχικό παράθυρο του λογισμικού δίνονται διάφορες σημαντικές πληροφορίες για τα δεδομένα. Συγκεκριμένα, η βάση δεδομένων περιλαμβάνει 123.631 περιστατικά εγκλημάτων από τα οποία το κάθε ένα έχει 8 χαρακτηριστικά (Crime, Rank, Solving κτλ.). Ενδεικτικά, το χαρακτηριστικό ¨Crime¨ είναι ονομαστικού (nominal) τύπου, δεν υπάρχει κάποια απουσία καταχωρήσεων (missing=0, 0%), έχει 9 διακριτές τιμές (distinct=9) με διαφορετικές τιμές και βαρύτητα. Στο ραβδωτό γράφημα απεικονίζεται η κατανομή των 123.631 περιστατικών στις ονομαστικές τιμές του χαρακτηριστικού ¨Crime¨ με κατηγοριοποίηση την αιτία. Ανάλογα με την επιλογή του χαρακτηριστικού από το πλαίσιο ¨Attributes¨ εμφανίζονται οι αντίστοιχες πληροφορίες. Από το πτυσσόμενο κουμπί ¨Class¨ επιλέγεται το ¨Crime¨ και στην συνέχεια το ¨Visualize All¨ και εξάγεται μία γενική εικόνα όλων των χαρακτηριστικών με κατηγοριοποίηση το έγκλημα, όπως φαίνεται στο σχήμα 4-5.

Σχήμα 4-4: Ραβδωτά διαγράμματα των χαρακτηριστικών της βάσης δεδομένων.

Στο διάγραμμα ¨Crime¨ του σχήματος 4-4 φαίνεται το μέγεθος του είδους εγκλήματος, όπου διαπιστώνεται ότι υπερτερεί το έγκλημα της κλοπής με πολύ μεγάλη διαφορά έναντι των άλλων, ενώ στο διάγραμμα ¨Rank¨ ο βαθμός των εγκλημάτων που ξεχωρίζει είναι τα πλημμελήματα. Ακόμη, παρατηρείται ότι τα εγκλήματα στον μεγαλύτερο αριθμό είναι ανεξιχνίαστα (διάγραμμα ¨Solving¨) και τελεσμένα (διάγραμμα ¨Condition¨). Οι δράστες των εγκλημάτων στο σύνολό τους είναι ημεδαποί (διάγραμμα ¨Nationality¨) και άνδρες (διάγραμμα ¨Sex¨). Σύμφωνα με την διακριτοποίηση των ηλικιών των δραστών, οι ηλικίες που ξεχωρίζουν για την συχνή εμφάνισή τους είναι από 18 έως 34 ετών (διάγραμμα ¨Age¨), ενώ στις αιτίες (διάγραμμα ¨Cause¨) οι περισσότεροι λόγοι διάπραξης εγκλήματος είναι άγνωστοι. Συμπερασματικά, από τα διαγράμματα προκύπτει ότι υπερτερούν σε αριθμό στην βάση δεδομένων οι ανεξιχνίαστες τελεσμένες πλημμελληματικές κλοπές, οι οποίες διαπράχθηκαν για άγνωστους λόγους από ημεδαπούς άνδρες, ηλικίας μεταξύ 18 έως 34 ετών.

Ανάλογα με την επιλογή του χαρακτηριστικού από το πλαίσιο ¨Attributes¨ εμφανίζονται οι αντίστοιχες πληροφορίες και διαγράμματα και μπορούν να πραγματοποιηθούν διάφορες επιθυμητές εργασίες, όπως η επιλογή να διαγραφεί κάποιο χαρακτηριστικό. Επιλέγοντας το κουμπί ¨Edit¨ μπορούν να τροποποιηθούν τα περιστατικά και χαρακτηριστικά της βάσης δεδομένων και με το κουμπί ¨Choose¨ από το πλαίσιο ¨Filter¨ διαλέγεται το ανάλογο φίλτρο για εφαρμογή. Παραδείγματος χάριν, εάν

χρησιμοποιηθεί το φίλτρο ¨NominalToBinary¨ (Weka.filters.unsupervised.attribute. NominalToBinary) μετατρέπονται όλα τα ονομαστικά χαρακτηριστικά σε δυαδικά αριθμητικά χαρακτηριστικά, ενώ το φίλτρο ¨Add¨ (Weka.filters.unsupervised.attribute. Add) είναι ένα φίλτρο εμφάνισης που προσθέτει ένα νέο χαρακτηριστικό στο σύνολο δεδομένων και το φίλτρο ¨Discretize¨ (Weka.filters.unsupervised.attribute.Discretize) είναι ένα φίλτρο περιστατικού με το οποίο επιτυγχάνεται διακριτοποίηση ενός εύρους αριθμητικών χαρακτηριστικών στο σύνολο δεδομένων σε ονομαστικά χαρακτηριστικά. Η επιλογή φίλτρων βοηθάει για την συντόμευση του χρόνου υλοποίησης του μοντέλου σε μεγάλες βάσης δεδομένων, καθώς και στην απλοποίησή του και την διαλογή των χρήσιμων μεταβλητών.

4.2 Εφαρμογή τεχνικών και Αποτελέσματα της Εξόρυξης Δεδομένων

❖ Κατηγοριοποίηση

Για τη πραγματοποίηση της εξαγωγής πληροφορίας από την δημιουργηθείσα βάση δεδομένων, εφαρμόζεται η κατηγοριοποίηση. Συνήθως, η συγκεκριμένη εργασία χρησιμοποιείται για πρόβλεψη και εκτίμηση με σκοπό την αντιστοίχιση ενός χαρακτηριστικού σε ένα εκ των υπαρχόντων συνόλων χαρακτηριστικών. Αρχικά, η τεχνική που ακολουθείται είναι αυτή των δέντρων αποφάσεων και η χρήση του αλγορίθμου J48 (για πληροφορίες σχετικά με την τεχνική και τον αλγόριθμο J48 ανατρέξτε στην βιβλιογραφική επισκόπηση).

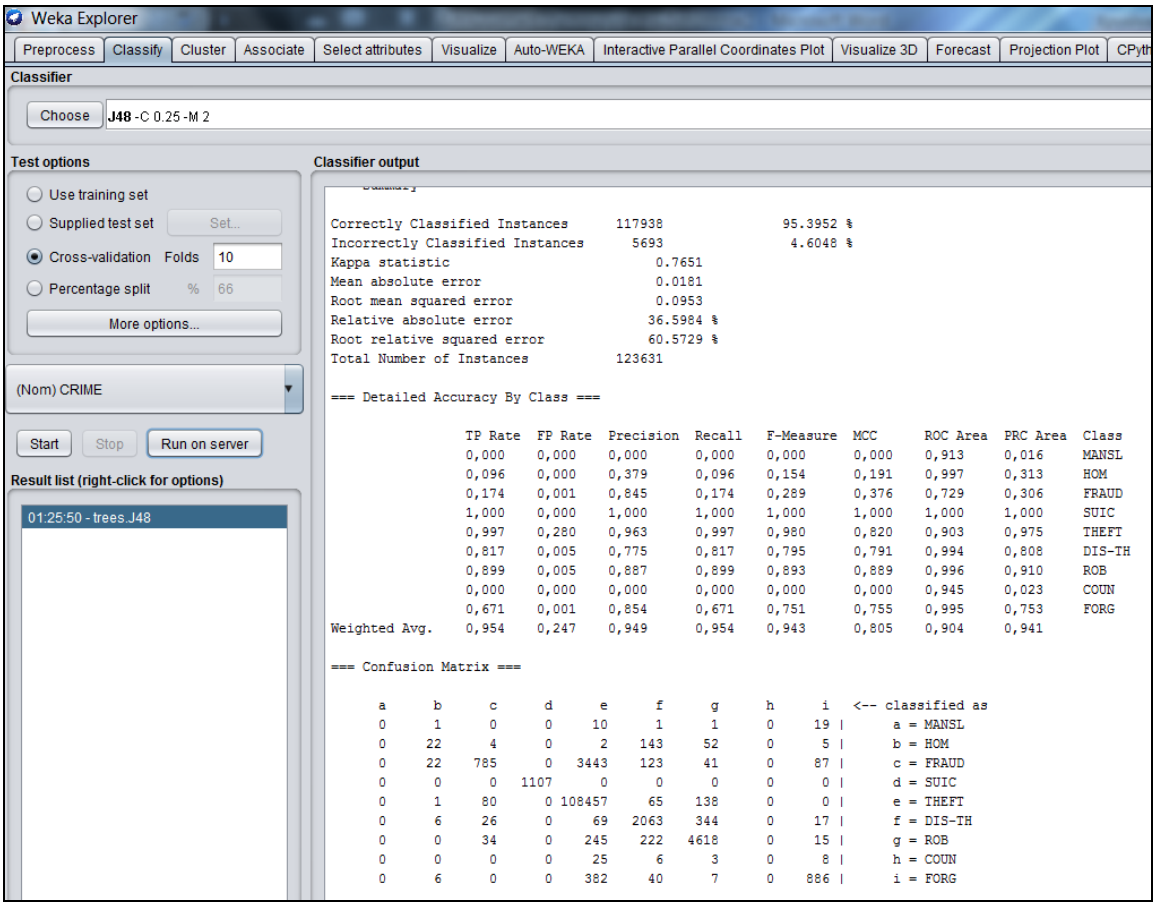
Εφόσον γίνει η ¨φόρτωση¨ των δεδομένων στο λογισμικό Weka και επιλεχθεί η καρτέλα ¨Classify¨, διαλέγουμε από το πλαίσιο ¨Classifier¨, πατώντας το κουμπί ¨Choose¨, τον αλγόριθμο J48 (Weka.classifiers.trees.J48) και ως κλάση ορίζεται το χαρακτηριστικό ¨Crime¨. Κατόπιν, επιλέγονται οι επιθυμητοί παράμετροι από την καρτέλα επιλογής που εμφανίζεται με διπλή επιλογή (¨κλικ¨) του κέρσορα πάνω στο όνομα του αλγορίθμου (J48- C 0.25-M 2), ώστε να επιτευχθούν τα καλύτερα δυνατά αποτελέσματα.

Η καρτέλα επιλογής του αλγορίθμου J48 έχει τις εξής επιλογές:

- Seed - Οι σπόροι που χρησιμοποιούνται για την τυχαία επιλογή των δεδομένων όταν χρησιμοποιείται κλάδεμα μειωμένου σφάλματος.
- Unpruned - Εάν πραγματοποιείται κλάδεμα.

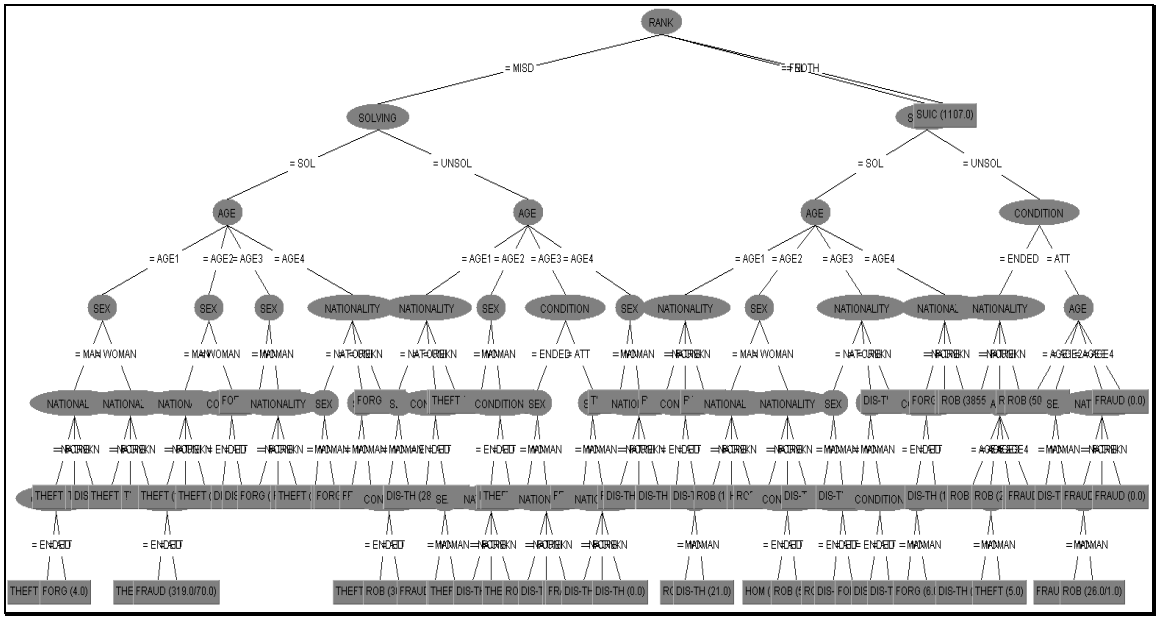
- ConfidenceFactor - Ο παράγοντας εμπιστοσύνης που χρησιμοποιείται για κλάδεμα (μικρότερες τιμές υφίστανται περισσότερο κλάδεμα).
- NumFolds - Προσδιορίζει την ποσότητα δεδομένων που χρησιμοποιείται για κλάδεμα μειωμένου σφάλματος. Ένα ¨fold¨ χρησιμοποιείται για κλάδεμα, το υπόλοιπο για την ανάπτυξη του δέντρου.
- NumDecimalPlaces - Ο αριθμός των δεκαδικών θέσεων που θα χρησιμοποιηθούν για την έξοδο αριθμών στο μοντέλο.
- BatchSize - Ο προτιμώμενος αριθμός περιπτώσεων που πρέπει να επεξεργαστούν εάν πραγματοποιείται πρόβλεψη παρτίδας. Μπορούν να παρέχονται περισσότερες ή λιγότερες περιπτώσεις, αλλά αυτό δίνει στις εφαρμογές την ευκαιρία να καθορίσουν ένα προτιμώμενο μέγεθος παρτίδας.
- ReducedErrorPruning - Εάν χρησιμοποιείται κλάδεμα με μειωμένο σφάλμα αντί για κλάδεμα C.4.5.
- UseLaplace - Το αν μετράει στα φύλλα εξομαλύνεται με βάση το Laplace.
- DoNotMakeSplitPointActualValue - Εάν είναι αληθές, το σημείο διαίρεσης δεν μετατοπίζεται σε πραγματική τιμή δεδομένων. Αυτό μπορεί να οδηγήσει σε σημαντικές επιταχύνσεις για μεγάλα σύνολα δεδομένων με αριθμητικά χαρακτηριστικά.
- Debug - Εάν έχει οριστεί αληθές, ο ταξινομητής μπορεί να παράγει πρόσθετες πληροφορίες στην κονσόλα.
- SubtreeRaising - Το κατά πόσο θα πρέπει να εξετάσετε τη λειτουργία αύξησης των υποσυνόλων κατά το κλάδεμα.
- SaveInstanceData - Να αποθηκεύσετε τα δεδομένα εκπαίδευσης για οπτικοποίηση.
- BinarySplits - Να χρησιμοποιηθούν δυαδικοί διαχωρισμοί σε ονομαστικά χαρακτηριστικά κατά την ανάπτυξη των δέντρων.
- DoNotCheckCapabilities - Αν οριστεί, οι ιδιότητες ταξινομητή δεν ελέγχονται πριν κατασκευαστεί ο ταξινομητής (Χρησιμοποιήστε με προσοχή για να μειώσετε το χρόνο εκτέλεσης).
- MinNumObj - Ο ελάχιστος αριθμός περιπτώσεων ανά φύλλο.
- UseMDLcorrection - Εάν η διόρθωση ¨MDL¨ χρησιμοποιείται όταν εντοπίζονται διαχωρίσεις σε αριθμητικά χαρακτηριστικά.
- CollapseTree - Κατάργηση εξαρτημάτων που δεν μειώνουν το σφάλμα κατά την εκπαίδευση.

Επιλέγεται στο πλαίσιο ¨Test options¨ η τεχνική Cross-validation με Folds=10, όπου διαχωρίζεται το αρχικό σύνολο δεδομένων σε 10 ισόποσα υποσύνολα με τυχαία επιλογή του ενός για εκπαίδευση του κατηγοριοποιητή. Η ίδια διαδικασία εκπαίδευσης γίνεται και για τα υπόλοιπα 9 υποσύνολα δεδομένων για το καθένα ξεχωριστά. Στην συνέχεια με την επιλογή του κουμπιού ¨Start¨ ξεκινάει η λειτουργία του αλγορίθμου και δημιουργείται το δέντρο, δίνοντας τα αποτελέσματα που φαίνονται στο σχήμα 4.5.



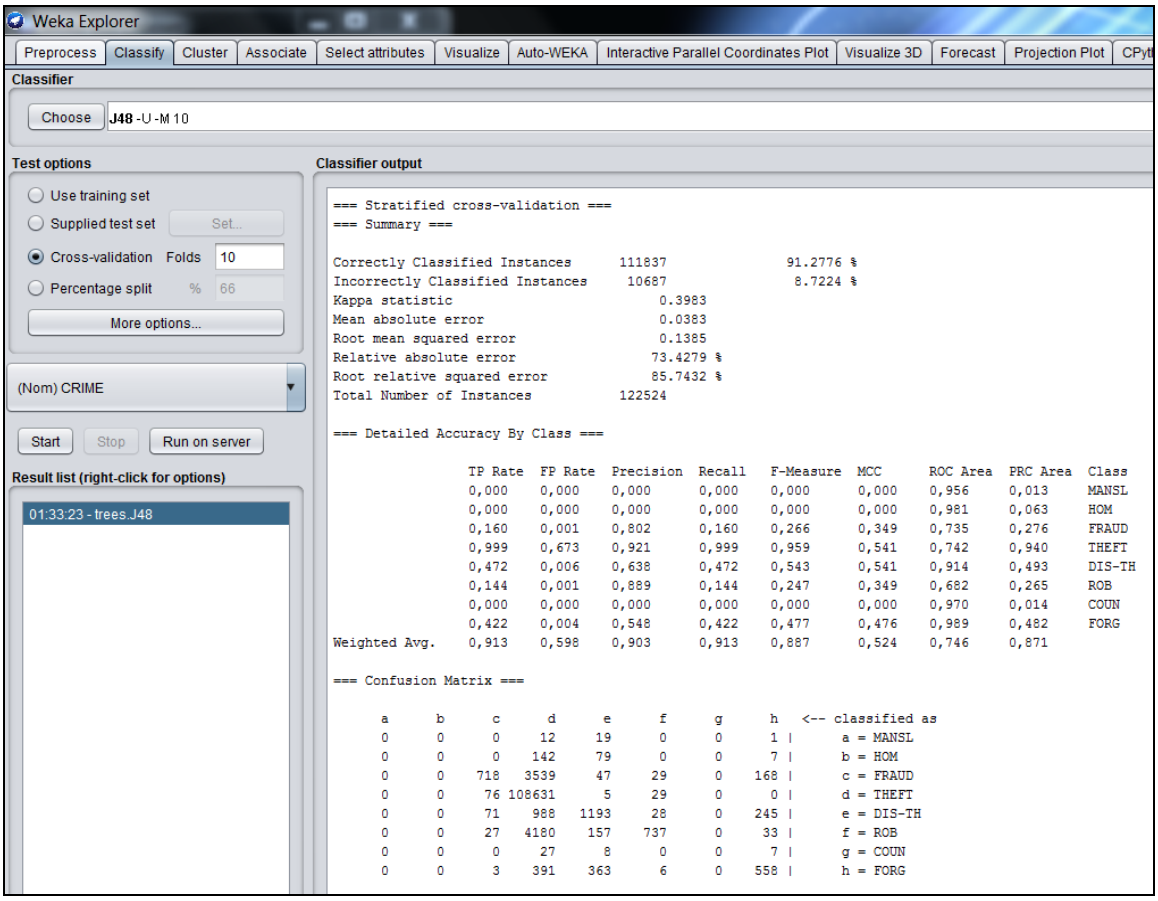
Σχήμα 4-5: Αποτελέσματα αλγορίθμου J48.

Στο σχήμα 4-6 φαίνεται το διάγραμμα του δέντρου αποφάσεων, το οποίο εμφανίζεται με την επιλογή ¨visualize tree¨ από την λίστα αποτελεσμάτων (δεξί κλικ).

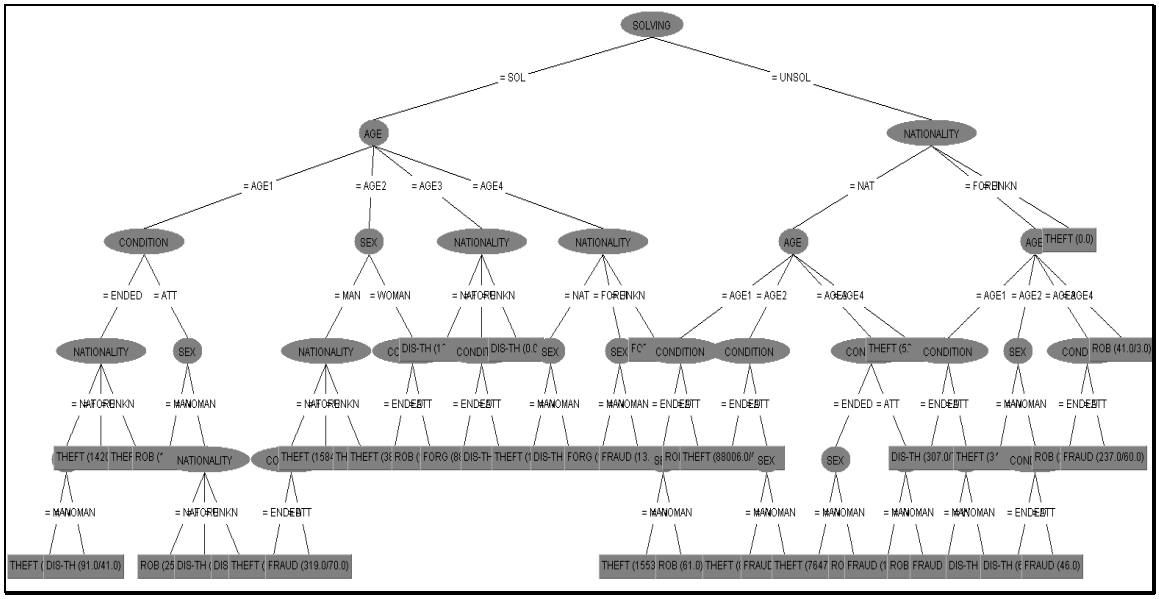


Σχήμα 4-6: Διάγραμμα αλγορίθμου J48 με κλάση ¨Crime¨

Επιλέγοντας την καρτέλα ¨Preprocess¨, αφαιρούνται χαρακτηριστικά και κάποιες τιμές από άλλα που θεωρούνται πλεονάζουσες, ώστε να περιοριστούν τα δεδομένα. Επομένως, αφαιρούνται τα χαρακτηριστικά ¨Cause¨ και ¨Rank¨ από το πλαίσιο ¨Attributes¨ και από την καρτέλα ¨Edit¨ αφαιρείται η τιμή ¨Suic¨ από το χαρακτηριστικό ¨Crime¨. Ο λόγος αφαίρεσης αυτών των χαρακτηριστικών είναι η δοκιμή με λιγότερα δεδομένα και ο έλεγχος των αποτελεσμάτων εάν έχουν μεγάλη απόκλιση και επηρεάζονται σημαντικά από την αφαίρεση κάποιον χαρακτηριστικών και τιμών. Για τον σκοπό αυτό επιλέχθηκαν δύο χαρακτηριστικά με μεγάλο αριθμό συμμετοχής, καθώς και η τιμή ¨Suic¨ η οποία δεν αποτελεί έγκλημα. Έπειτα, από την καρτέλα επιλογής του αλγορίθμου J48 επιλέγεται ως ελάχιστος αριθμός περιπτώσεων ανά φύλλο (MinNumObj) ίσος με 10 και στην πραγματοποίηση κλαδέματος (Unpruned) επιλέγεται η ένδειξη ¨True¨. Οι υπόλοιπες επιλογές παραμένουν ως έχουν και ξαναεκτελείται ο κατηγοριοποιητής, δίνοντας τα αποτελέσματα που φαίνονται στο σχήμα 4-7 και το διάγραμμα δέντρου στο σχήμα 4-8.



Σχήμα 4-7: Αποτελέσματα αλγορίθμου J48 με μείωση χαρακτηριστικών και κλάση ¨Crime¨

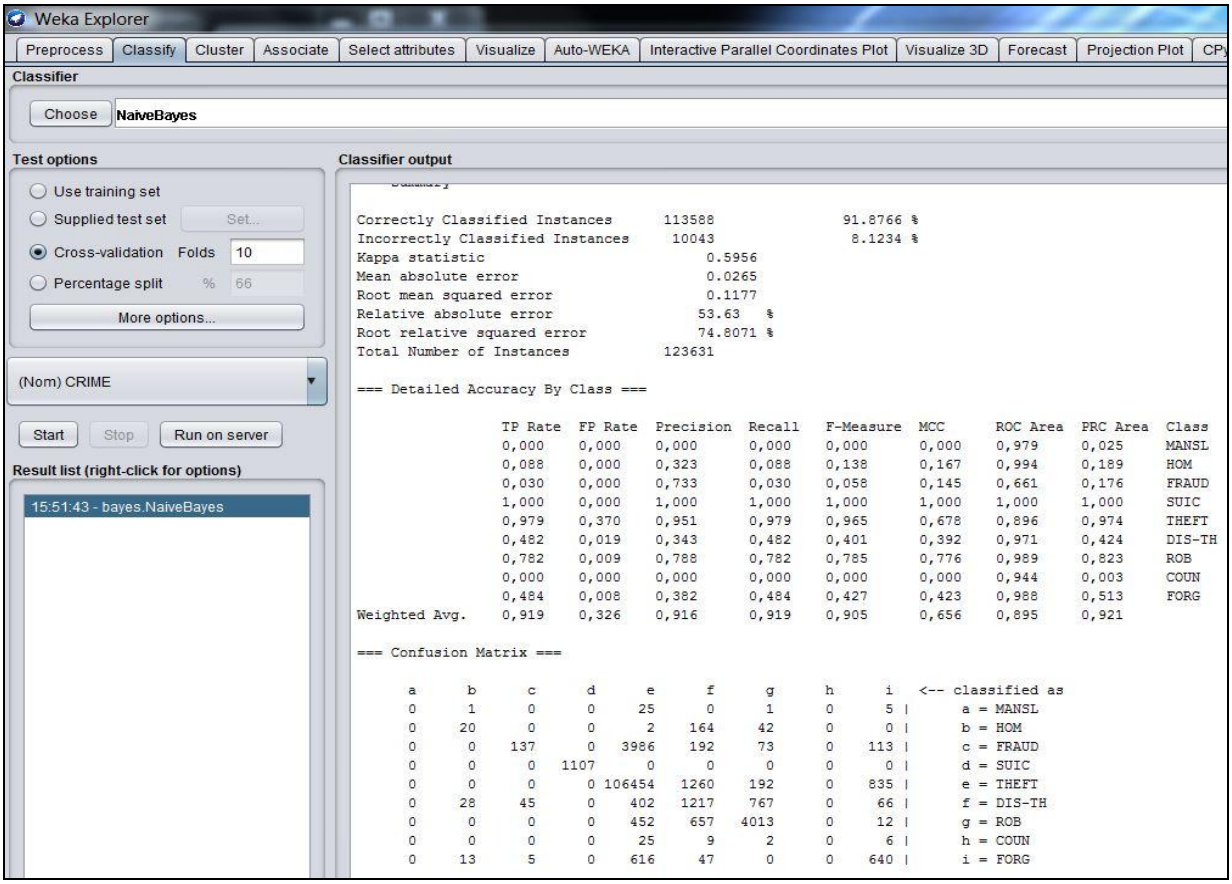


Σχήμα 4-8: Διάγραμμα αλγορίθμου J48 με μείωση χαρακτηριστικών και κλάση ¨Crime¨.

Όπως παρατηρείται με πρώτη ματιά στα αρχικά αποτελέσματα το ¨Correctly Classified Instances¨ (σωστά κατηγοριοποιημένες περιπτώσεις) είναι με ποσοστό 95,395% και στην δεύτερη περίπτωση με 91,277%. Η διαφορά είναι πολύ μικρή και

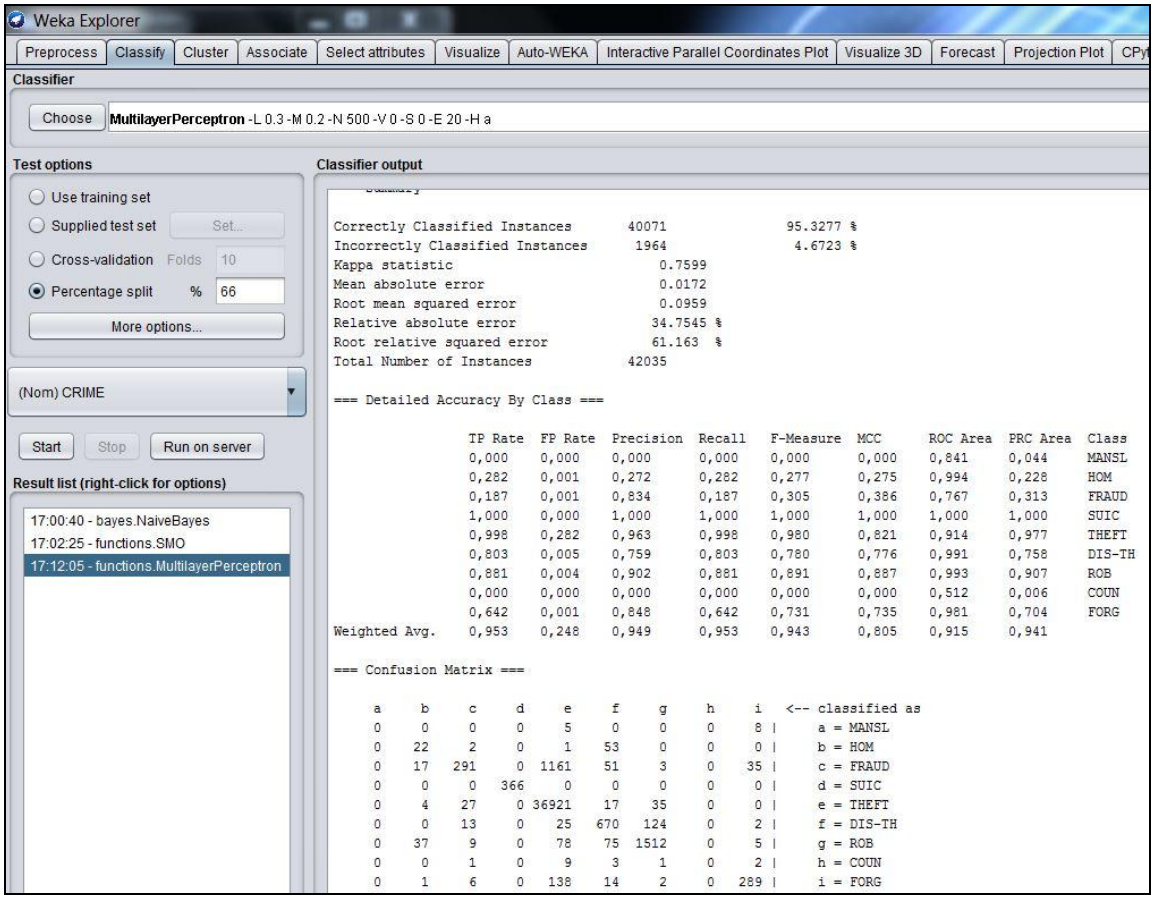
επίσης και τα 2 αποτελέσματα έχουν πολύ υψηλό ποσοστό σωστών κατηγοριοποιημένων περιπτώσεων. Ελέγχοντας όμως τους δείκτες, παρατηρούνται αισθητές διαφορές, όπως στον δείκτη TP Rate και συγκεκριμένα στις τιμές ¨ROB¨, ¨COUNT¨ και ¨FORG¨, καθώς και στον ROC Rate στις τιμές ¨THEFT¨ και ¨ROB¨. Συμπερασματικά, υπάρχουν διαφορές και επηρεάζονται τα αποτελέσματα από τις αλλαγές που έγιναν.

Ο κατηγοριοποιητής NaïveBayes εξάγει υποθέσεις πιθανοτήτων αντί να κάνει προβλέψεις, δηλαδή αναλύει την συνεισφορά όλων των ανεξάρτητων χαρακτηριστικών και της συνέπειάς τους, υπολογίζοντας την πιθανότητα που έχει η κάθε υπόθεση να ανήκει σε μια κατηγορία. Χρησιμοποιείται η βάση δεδομένων με τον κατηγοριοποιητή NaïveBayes για τον υπολογισμό των μέσων όρων των περιπτώσεων στην κάθε ομάδα. Επιλέγεται στο πλαίσιο ¨Test options¨ η τεχνική Cross-validation με Folds=10, με κλάση το χαρακτηριστικό ¨Crime¨ και εκτελείτε ο κατηγοριοποιητής (weka.classifiers.bayes. NaiveBayes). Ο αλγόριθμος έχει σωστές κατηγοριοποιημένες περιπτώσεις σε αρκετά υψηλό ποσοστό που είναι 91,876% και χρειάστηκε 0,14 δευτερόλεπτα για την εξαγωγή των αποτελεσμάτων, τα οποία φαίνονται στο σχήμα 4-9.

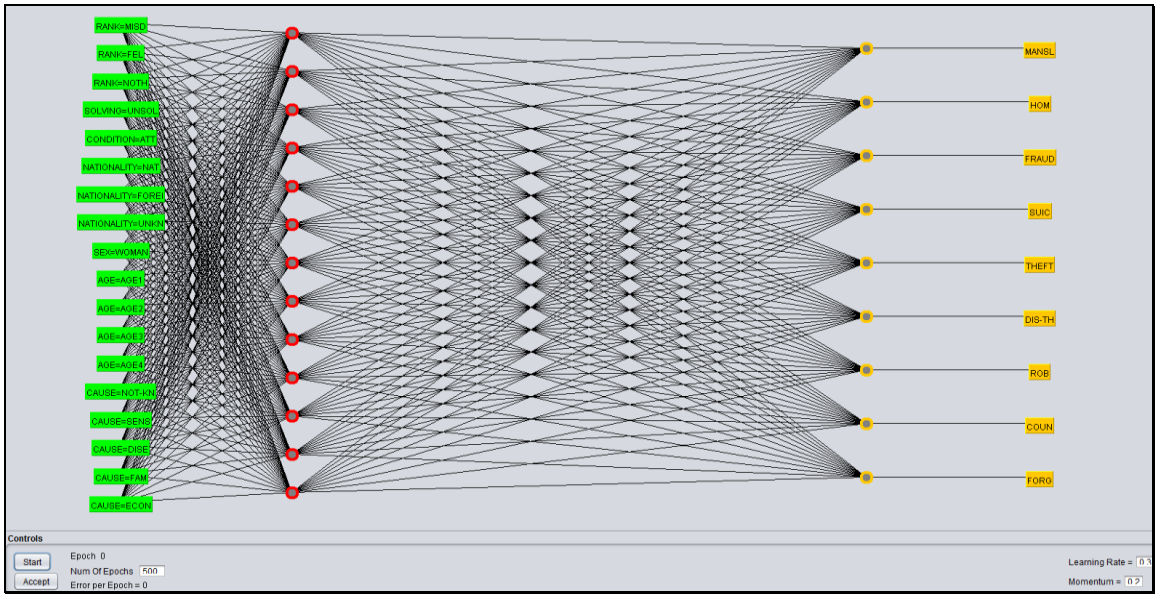


Σχήμα 4-9: Αποτελέσματα αλγορίθμου NaïveBayes με κλάση το χαρακτηριστικό ¨Crime¨.

Επόμενος κατηγοριοποιητής που χρησιμοποιείται είναι ο MultilayerPerceptron. Ο κατηγοριοποιητής αυτός είναι νευρωνικό δίκτυο με πολλές εισόδους και μία έξοδο, εμπρόσθιας διάδοσης, χωρίς να γίνεται ανατροφοδότηση στους κόμβους και φημίζεται για την απλότητα της λειτουργίας του. Με τον ίδιο τρόπο όπως και στον αλγόριθμο J48 ανοίγουμε την καρτέλα επιλογής και γίνονται οι επιθυμητές αλλαγές (π.χ. ο HiddenLayers καθορίζει τα κρυμμένα επίπεδα του νευρωνικού δικτύου, ο GUI εμφανίζει γραφική διεπαφή κτλ.), επιλέγεται στο πλαίσιο ¨Test options¨ η τεχνική Percentage split 66% (εκπαίδευση σε ποσοστό των συνολικών δεδομένων και δοκιμή στο υπόλοιπο), με κλάση το χαρακτηριστικό ¨Crime¨ και εκτελείτε ο κατηγοριοποιητής (weka.classifiers. functions.MultilayerPerceptron). Ο αλγόριθμος έχει σωστές κατηγοριοποιήσεις περιστατικών σε πολύ μεγάλο ποσοστό της τάξεως του 95,327% και χρειάστηκε 696.88 δευτερόλεπτα για την εξαγωγή των αποτελεσμάτων, τα οποία φαίνονται στο σχήμα 4-10 και η εικόνα του νευρωνικού δικτύου στο σχήμα 4-11.

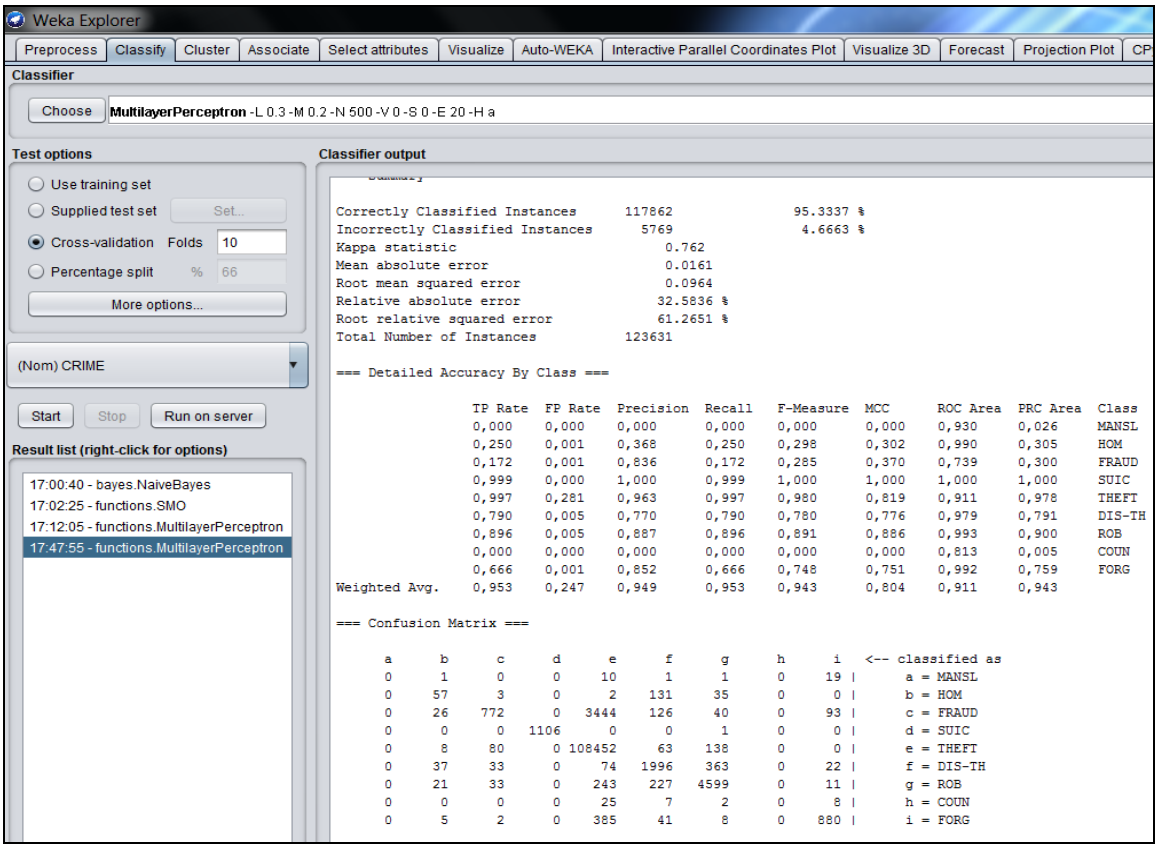


Σχήμα 4-10: Αποτελέσματα αλγορίθμου MultilayerPerceptron με τεχνική Percentage split 66%.



Σχήμα 4-11: Νευρωνικό δίκτυο MultilayerPerceptron με τεχνική Percentage split 66%.

Ο ίδιος κατηγοριοποιητής χρησιμοποιείται με την τεχνική Cross-validation με Folds=10 και με κλάση το χαρακτηριστικό ¨Crime¨. Ο αλγόριθμος έχει 95,333% σωστές κατηγοριοποίησης περιστατικών και χρειάστηκε 701.44 δευτερόλεπτα για την εξαγωγή των αποτελεσμάτων, τα οποία φαίνονται στο σχήμα 4-12.



Σχήμα 4-12: Αποτελέσματα αλγορίθμου MultilayerPerceptron με τεχνική Cross-validation με Folds=10.

Η διαφορά μεταξύ των δύο τεχνικών (Cross-validation και Percentage split) είναι μηδαμινή στο ποσοστό των σωστών κατηγοριοποιήσεων των περιστατικών με μόλις 0,006% απόκλιση, ενώ οι χρόνοι επεξεργασίας των δεδομένων που χρειάστηκε ο αλγόριθμος είναι ιδιαίτερα πολύ μεγάλος.

❖ Κανόνες Συσχέτισης

Στην συνέχεια αναζητούνται οι συσχετίσεις μεταξύ των δεδομένων του συνόλου από αυτά, οι οποίες δεν είναι εμφανείς και δεν συνδέονται με την σχέση αιτιότητας-συσχέτισης ή συναρτησιακές εξαρτήσεις. Χρησιμοποιούνται κανόνες συσχέτισης και ο αλγόριθμος Apriori για τον εντοπισμό των συνήθων ταυτόχρονων χρήσεων των στοιχείων, τα οποία έχουν κάποια σχέση. Ο αλγόριθμος Apriori σαρώνει την βάση δεδομένων και αναζητάει τα συχνά εκείνα στοιχεία του ενός στοιχειοσυνόλου (1-itemsets). Αυτά τα οποία ικανοποιούν το ελάχιστο όριο υποστήριξης (minsup) χρησιμοποιούνται για τον εντοπισμό των επόμενων στοιχειοσυνόλων στην δεύτερη σάρωση των δύο στοιχειοσυνόλων. Σε κάθε σάρωση της βάσης δεδομένων, η οποία γίνεται μία φορά για κάθε φάση, τα υποψήφια στοιχειοσύνολα βασίζονται στα προηγούμενα και αυξάνουν το χαρακτηριστικό τους κατά ένα. Επομένως, στην τρίτη σάρωση θα βρεθούν στοιχεία των τριών στοιχειοσυνόλων και θα συνεχιστεί η διαδικασία έως ότου δεν υπάρχουν σύνολα στοιχείων με ένα παραπάνω χαρακτηριστικό (k-itemsets) (για πληροφορίες σχετικά με την τεχνική και τον αλγόριθμο Apriori ανατρέξτε στην βιβλιογραφική επισκόπηση).

Η καρτέλα επιλογής του αλγορίθμου Apriori έχει τις εξής επιλογές:

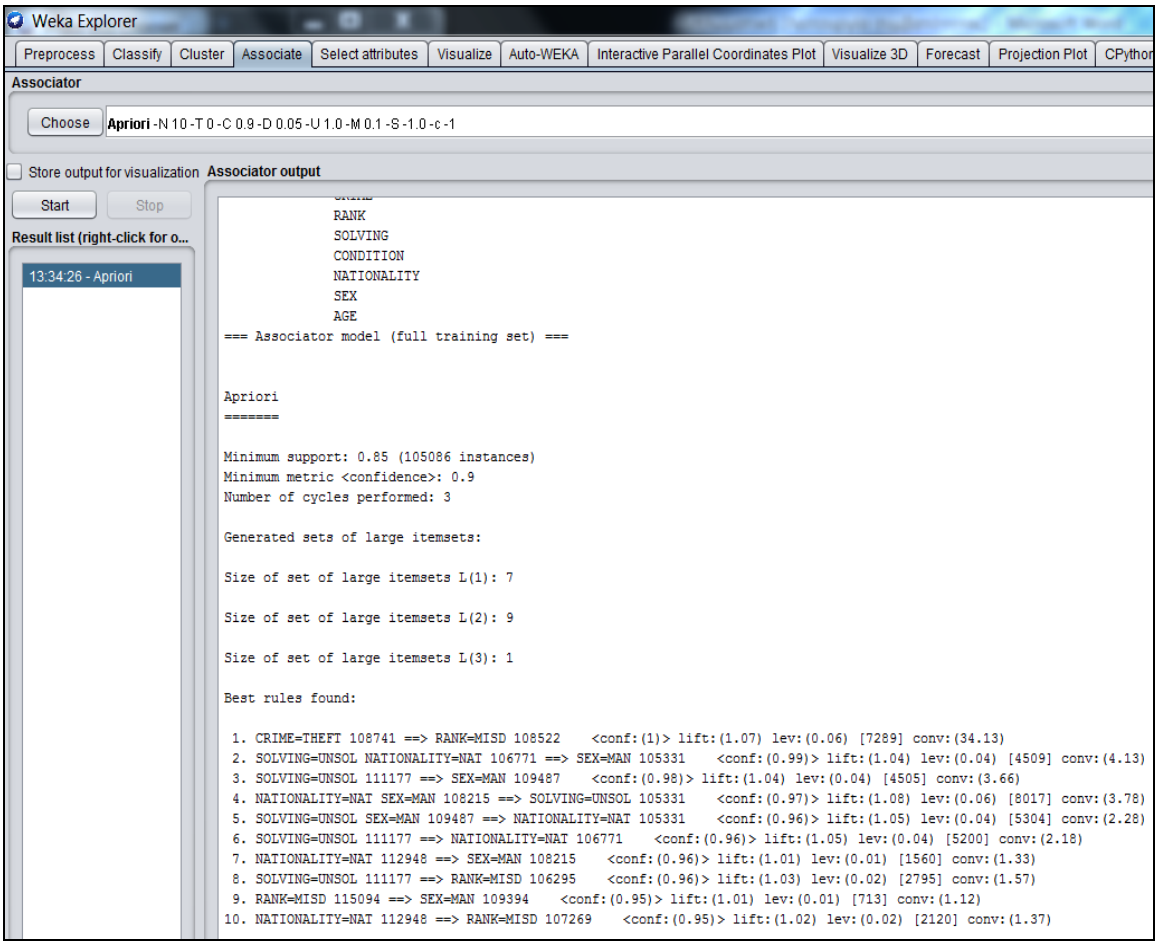
- Verbose - Εάν είναι ενεργοποιημένη, ο αλγόριθμος θα εκτελείται σε λεπτομερή λειτουργία.
- MinMetric - Ελάχιστη μετρική βαθμολογία. Εξετάζει μόνο κανόνες με βαθμολογίες υψηλότερες από αυτή την τιμή.
- NumRules - Αριθμός κανόνων που πρέπει να βρεθεί.
- LowerBoundMinSupport - Χαμηλότερο όριο για ελάχιστη υποστήριξη.
- ClassIndex - Ευρετήριο του χαρακτηριστικού κλάσης. Αν οριστεί στο -1, το τελευταίο χαρακτηριστικό λαμβάνεται ως χαρακτηριστικό κλάσης.
- OutputItemSets - Αν ενεργοποιηθεί τα στοιχειοσύνολα εξέρχονται επίσης.
- Car - Αν ενεργοποιηθεί οι κανόνες συσχέτισης τάξης εξορύσσονται αντί των (γενικών) κανόνων σύνδεσης.

- DoNotCheckCapabilities - Αν οριστεί, οι δυνατότητες του συσχετιστή δεν ελέγχονται πριν από τη δημιουργία του συσχετιστή (Χρήση με προσοχή μειώνει το χρόνο εκτέλεσης).
- RemoveAllMissingCols - Κατάργηση στηλών με όλες τις τιμές που λείπουν.
- SignificanceLevel - Σημασιακό επίπεδο. Έλεγχος σημαντικότητας (μόνο μέτρηση εμπιστοσύνης).
- TreatZeroAsMissing - Εάν είναι ενεργοποιημένη, το μηδέν (δηλαδή η πρώτη ονομαστική τιμή) αντιμετωπίζεται με τον ίδιο τρόπο όπως μια τιμή που λείπει.
- Delta - Μειώνει εξαιρετικά την υποστήριξη από αυτόν τον παράγοντα. Μειώνει την υποστήριξη έως ότου επιτευχθεί ελάχιστη υποστήριξη ή απαιτείται αριθμός κανόνων.
- MetricType - Ορίστε τον τύπο της μέτρησης βάσει της οποίας θα ταξινομηθούν οι κανόνες. Το ¨Confidence¨ (εμπιστοσύνη) είναι η αναλογία των παραδειγμάτων που καλύπτονται από την προϋπόθεση ότι επίσης καλύπτονται από την συνέπεια (οι κανόνες σύνδεσης των τάξεων μπορούν να εξορύσσονται μόνο με εμπιστοσύνη). Το ¨Lift¨ (ανύψωση) είναι εμπιστοσύνη διαιρούμενη με το ποσοστό όλων των παραδειγμάτων που καλύπτονται από την συνέπεια. Αυτό είναι ένα μέτρο της σπουδαιότητας της ένωσης που είναι ανεξάρτητη από την υποστήριξη. Το ¨Leverage¨ (μόχλευση) είναι η αναλογία πρόσθετων παραδειγμάτων που καλύπτονται τόσο από την προϋπόθεση όσο και από τις συνέπειες που υπερβαίνουν τα αναμενόμενα, εάν η αρχή και η συνέπεια ήταν ανεξάρτητες η μία από την άλλη. Ο συνολικός αριθμός των παραδειγμάτων που αντιπροσωπεύει αυτό παρουσιάζεται σε παρενθέσεις μετά τη μόχλευση. Η καταδίκη είναι ένα άλλο μέτρο απόκλισης από την ανεξαρτησία. Το ¨Conviction¨ (πεποίθηση) δίνεται από την P(προϋπόθεση)P(!Συνέπεια) / P(προϋπόθεση, !συνέπεια).
- UpperBoundMinSupport - Υψηλότερο όριο για ελάχιστη υποστήριξη.

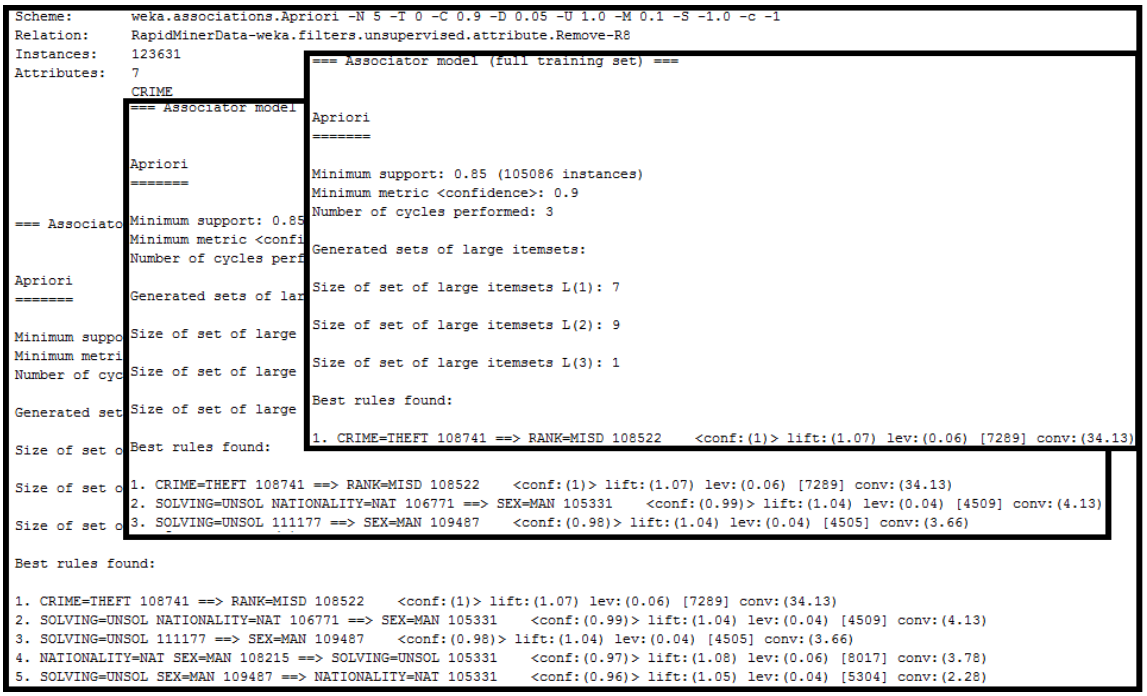
Στην συνέχεια, ο αλγόριθμος Apriori εκτελείται αρκετές φορές στα δεδομένα, αφού επιλεγούν οι επιθυμητές ρυθμίσεις και χαρακτηρίστηκα (προεπιλογή, μεταβολή αριθμού κανόνων (NumRules), Delta, Confidence (εμπιστοσύνης) και ¨lift¨ (MetricType).

Από την καρτέλα ¨Preprocess¨ βλέπουμε ότι τα αίτια που υπερισχύουν με αρκετά μεγάλη διαφορά είναι τα ¨NOT-KN¨, δηλαδή άγνωστος λόγος. Αφαιρείται το χαρακτηριστικό ¨Cause¨, διότι τα αποτελέσματα που εξάγονται με αυτό επηρεάζονται σε μεγάλο βαθμό από την τιμή ¨NOT-KN¨ και το δεύτερο σκέλος των κανόνων δίνει

αποτελέσματα μόνον με αυτή την τιμή, χωρίς να έχει ιδιαίτερη αξία η ερμηνεία των κανόνων. Οι ρυθμίσεις παραμένουν στην προεπιλογή του προγράμματος και τα αποτελέσματα του αλγορίθμου Apriori φαίνονται στο σχήμα 4-13, ενώ στο σχήμα 4-14 περιλαμβάνονται συγκεντρωτικά εξαγόμενες δοκιμές με αλλαγή του αριθμού κανόνων από 10 σε 5, 3 και 1.

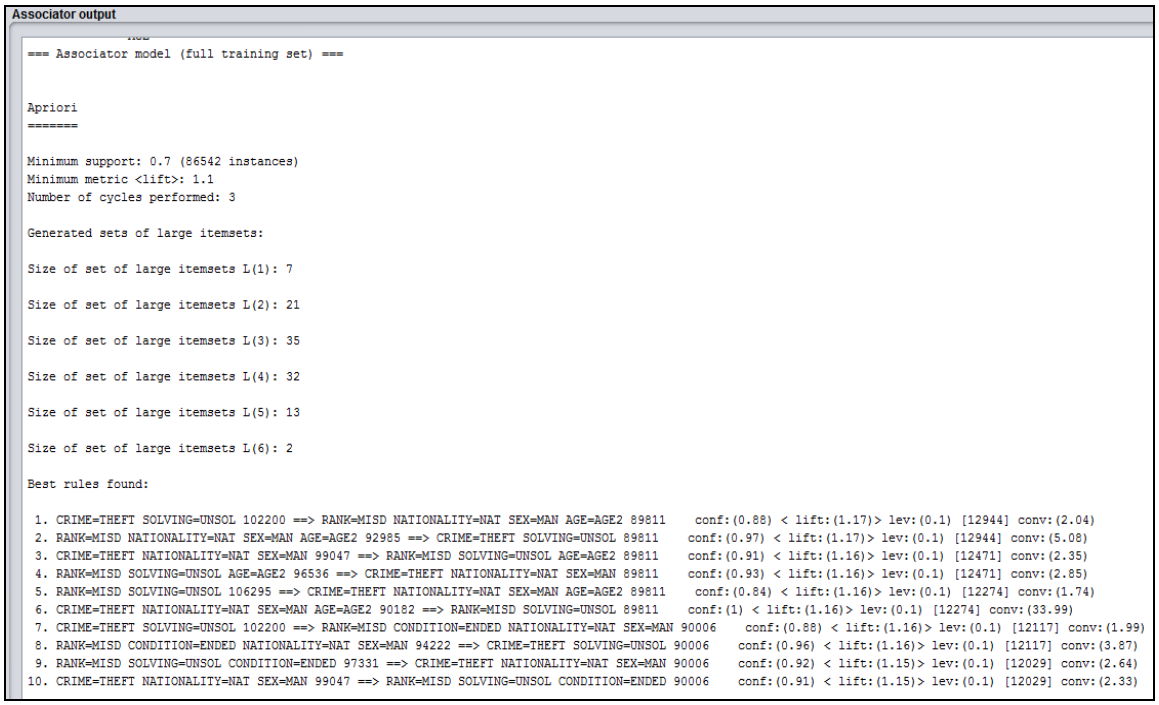


Σχήμα 4-13: Αποτελέσματα αλγορίθμου Apriori με ρυθμίσεις προεπιλογής.



Σχήμα 4-14: Αποτελέσματα αλγορίθμου Apriori με ¨NumRules¨=5, 3 και 1.

Ρυθμίζεται ο αλγόριθμος Apriori από την καρτέλα επιλογών με ¨MetricType¨ το ¨Lift¨, ελάχιστη τιμή (MinMetric) ίση με 1,1, αριθμό κανόνων ίσο με 10, ¨Delta¨ ίσο 0,1 και εκτελείται με παραγόμενα αποτελέσματα που φαίνονται στο σχήμα 4-15, ενώ με αριθμό κανόνων ίσο με 3 φαίνονται στο σχήμα 4-16.



Σχήμα 4-15: Αποτελέσματα αλγορίθμου Apriori με ¨MetricType¨=¨Lift¨ και ¨NumRules¨=10.

```
Scheme:       weka.associations.Apriori -N 3 -T 1 -C 1.1 -D 0.1 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     RapidMinerData-weka.filters.unsupervised.attribute.Remove-R8
Instances:    123631
Attributes:   7
              CRIME
              RANK
              SOLVING
              CONDITION
              NATIONALITY
              SEX
              AGE
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.8 (98905 instances)
Minimum metric <lift>: 1.1
Number of cycles performed: 2

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 19

Size of set of large itemsets L(3): 14

Size of set of large itemsets L(4): 4

Best rules found:

1. RANK=MISD SOLVING=UNSOL 106295 ==> CRIME=THEFT SEX=MAN 101006    conf:(0.95) < lift:(1.13)> lev:(0.09) [11480] conv:(3.17)
2. CRIME=THEFT SEX=MAN 104126 ==> RANK=MISD SOLVING=UNSOL 101006    conf:(0.97) < lift:(1.13)> lev:(0.09) [11480] conv:(4.68)
3. RANK=MISD SOLVING=UNSOL 106295 ==> CRIME=THEFT NATIONALITY=NAT 99006    conf:(0.93) < lift:(1.12)> lev:(0.09) [10744] conv:(2.47)
```

Σχήμα 4-16: Αποτελέσματα αλγορίθμου Apriori με ¨MetricType¨=¨Lift¨ και ¨NumRules¨=3 .

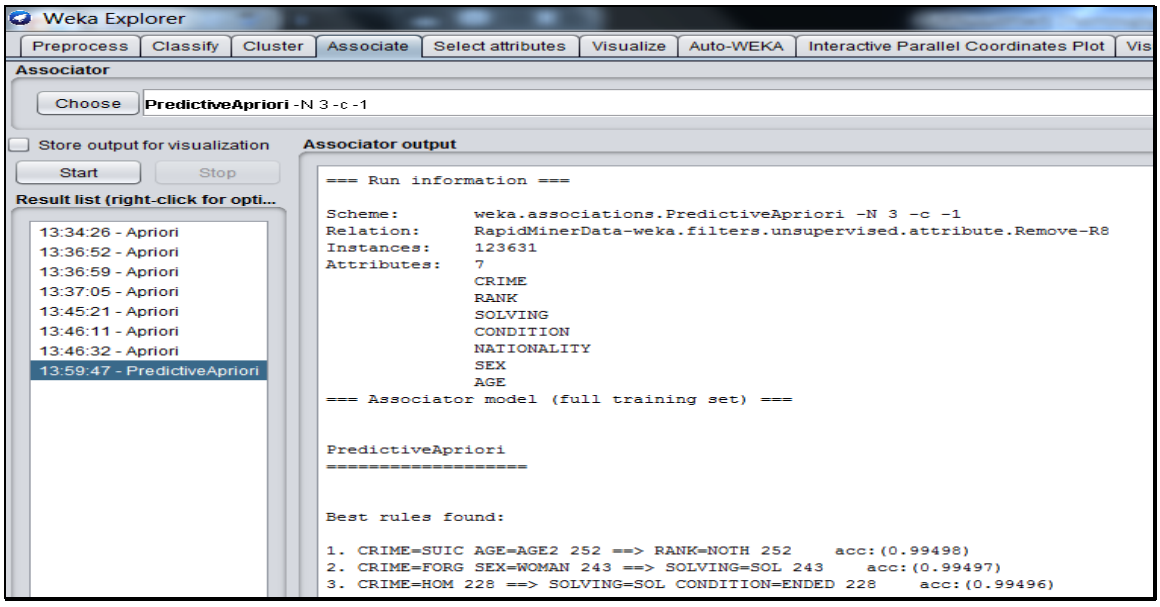
Οι αλγόριθμοι κανόνων συσχέτισης έχουν στόχο την μεγιστοποίηση της αναμενόμενης ακρίβειας πρόβλεψης και λαμβάνουν υπόψη τόσο την εμπιστοσύνη όσο και την υποστήριξη. Είναι εύκολο να βρεθεί συγκεκριμένος αριθμός κανόνων με υψηλή εμπιστοσύνη, όμως δεν θα είναι μεγάλος ο αριθμός των καταγραφών με υποστήριξη. Ακόμη, μπορούν να εξαχθούν πιο γενικοί κανόνες συσχέτισης με μεγάλη υποστήριξη, όμως με χαμηλή εμπιστοσύνη. Στον αλγόριθμο Apriori καθορίζεται το ¨minsup¨ και ¨minconf¨, όπου ουσιαστικά χρησιμοποιούνται όρια εμπιστοσύνης και υποστήριξης για επιστροφή κανόνων που βρίσκονται πάνω από τα όρια αυτά. Αντίθετα στον αλγόριθμο PredictiveApriori δεν συμβαίνει αυτό και ο χρήστης πρέπει να καθορίσει μόνο πόσους εξαγώγιμους κανόνες θέλει. Για την ανακάλυψη της γνώσης απαιτείται αξιόλογος αριθμός κανόνων που έχουν ενδιαφέρον με υψηλή υποστήριξη, όπου και η αντίστοιχη εμφανιζόμενη εμπιστοσύνη θα αναμένεται μελλοντικά. Οι κανόνες με χαμηλή υποστήριξη δεν έχουν ενδιαφέρον. Ο αλγόριθμος PredictiveApriori αναζητά τη βέλτιστη ανταλλαγή μεταξύ εμπιστοσύνης και υποστήριξης. Επιστρέφει τους κανόνες που μεγιστοποιούν την αναμενόμενη ακρίβεια πρόβλεψης, σύμφωνα με την εξίσωση του σχήματος 4-17 που προκύπτει από Μπαϋεσιανή ανάλυση της ακρίβειας πρόβλεψης, όπου βασίζεται στην υπόθεση ότι οι καταγραφές της βάσης δεδομένων είναι ανεξάρτητες και διανέμονται πανομοιότυπα και εξαλείφει τις αισιόδοξες προτιμήσεις υψηλής εμπιστοσύνης. Ο αλγόριθμος PredictiveApriori υπολογίζει την ¨προηγούμενη

εμπιστοσύνη¨ από την διαθέσιμη βάση δεδομένων και έχει καλύτερες επιδόσεις από τον Apriori στη δημιουργία κανόνων συσχέτισης, όχι όμως και στον χρόνο επεξεργασίας (Scheffer, 2001). Τα αποτελέσματα της εφαρμογής του αλγορίθμου PredictiveApriori (weka.associations. PredictiveApriori) φαίνονται στο σχήμα 4-18 για αριθμό κανόνων ίσο με 3.

$$E(c([x \Rightarrow y])|\hat{c}([x \Rightarrow y]), s(x)) \quad = \quad \frac{\int cB[c, s(x)](\hat{c}([x \Rightarrow y]))\pi(c)dc}{\int B[c, s(x)](\hat{c}([x \Rightarrow y]))\pi(c)dc}$$

Σχήμα 4-17: Εξίσωση ακρίβειας πρόβλεψης.
Πηγή: Scheffer, 2001.

«Όπου [x ⇒ y] = κανόνας συσχέτισης με τηρούμενη εμπιστοσύνη ĉ ([x ⇒ y]). Όπου π(c) = προηγούμενη ακρίβεια. Όταν δίνεται η ακρίβεια c, η εμπιστοσύνη ĉ καθορίζεται από την διωνυμική κατανομή που γράφεται ως B[c, s](ĉ). Η εξίσωση του σχήματος 4-17 ποσοτικοποιεί την αναμενόμενη εμπιστοσύνη ενός κανόνα με εμπιστοσύνη ĉ του οποίου το σώμα (η αριστερή πλευρά του κανόνα) x έχει υποστήριξη s (x). Η αναμενόμενη τιμή λαμβάνεται από δύο τυχαίες μεταβλητές: βάση δεδομένων D (διέπεται από P=ακρίβειας πρόβλεψης), και όλους τους κανόνες με s (x) και ĉ. Οι αλγόριθμοι του κανόνα σύνδεσης δεν αντλούν κανόνες τυχαίους, αλλά επιλέγουν εκείνους με προδιάθεση σε κανόνες με μεγάλη υποστήριξη και εμπιστοσύνη» (Scheffer, 2001).

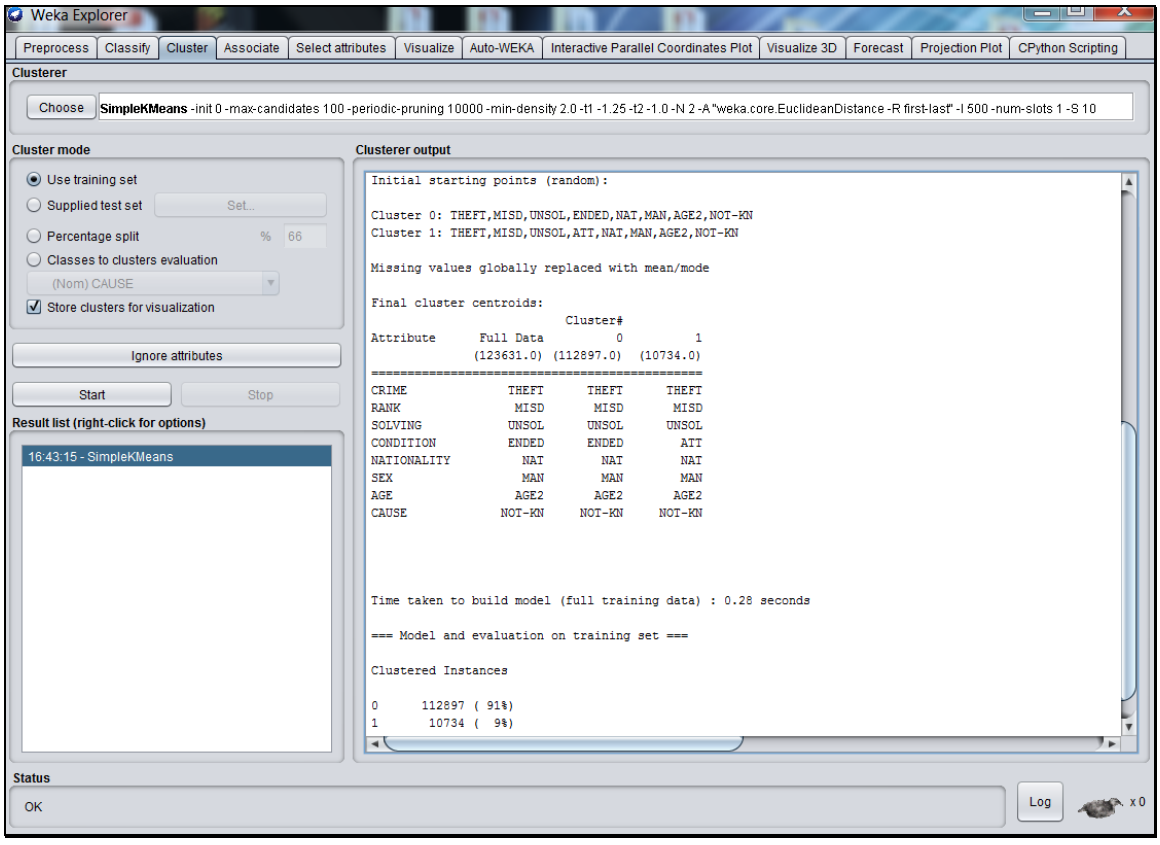


Σχήμα 4-18: Αποτελέσματα αλγορίθμου PredictiveApriori.

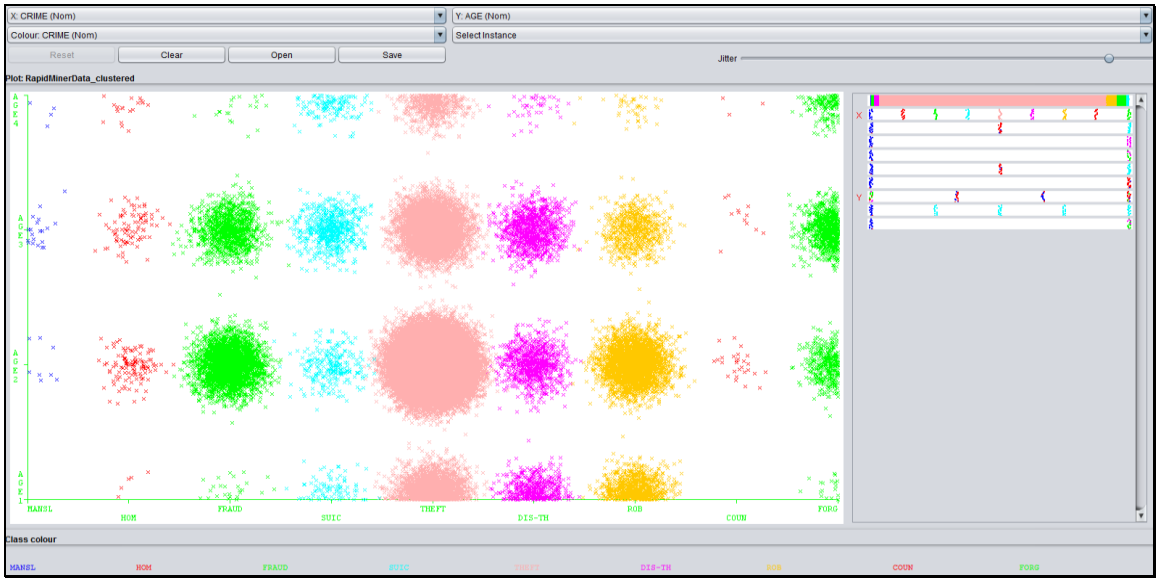
❖ Συσταδιοποίηση

Η συσταδιοποίηση ομαδοποιεί τα στοιχεία της βάσης δεδομένων, ανάλογα με τις ομοιότητες χαρακτηριστικών τους και όσο πιο ομογενοποιημένες είναι ως ομάδες τόσο μεγαλύτερη διαφορά υπάρχει μεταξύ των ομάδων. Ο πιο γνωστός αλγόριθμος στην συσταδιοποίηση είναι ο k-means (για πληροφορίες σχετικά με την συσταδιοποίηση και τον αλγόριθμο k-means ανατρέξτε στην βιβλιογραφική επισκόπηση).

Θα χρησιμοποιηθεί ο αλγόριθμος k-means για να εξαχθούν πρότυπα από την βάση δεδομένων για τις περιπτώσεις (εγκλήματα) και να καταταχθούν σε συστάδες ανάλογα με τα χαρακτηριστικά τους. Στην μη επιβλεπόμενη μάθηση δεν χρειάζεται εξαρτώμενη μεταβλητή (ετικέτες κλάσης) και προτιμάται να μαθαίνουν οι αλγόριθμοι από το σύνολο δεδομένων και να εισάγουν αντικείμενα στην κατηγορία τους, βάσει της υποκείμενης κατανομής δεδομένων. Προεπιλέγεται ο επιθυμητός αριθμός συστάδων (παράμετρος K=2) και εκτελείται ο αλγόριθμος (use training set) τα αποτελέσματα του οποίου φαίνονται στο σχήματα 4-19, καθώς και οι οπτικοποιήσεις στα σχήματα 4-20, 4-21, 4-22 και 4-23.

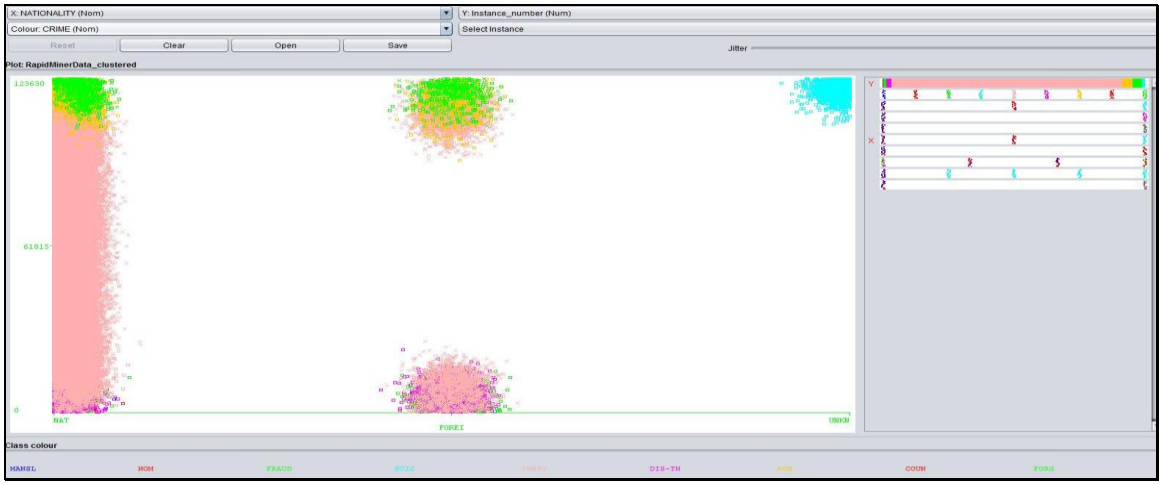


Σχήμα 4-19: Αποτελέσματα αλγορίθμου SimpleKMeans.



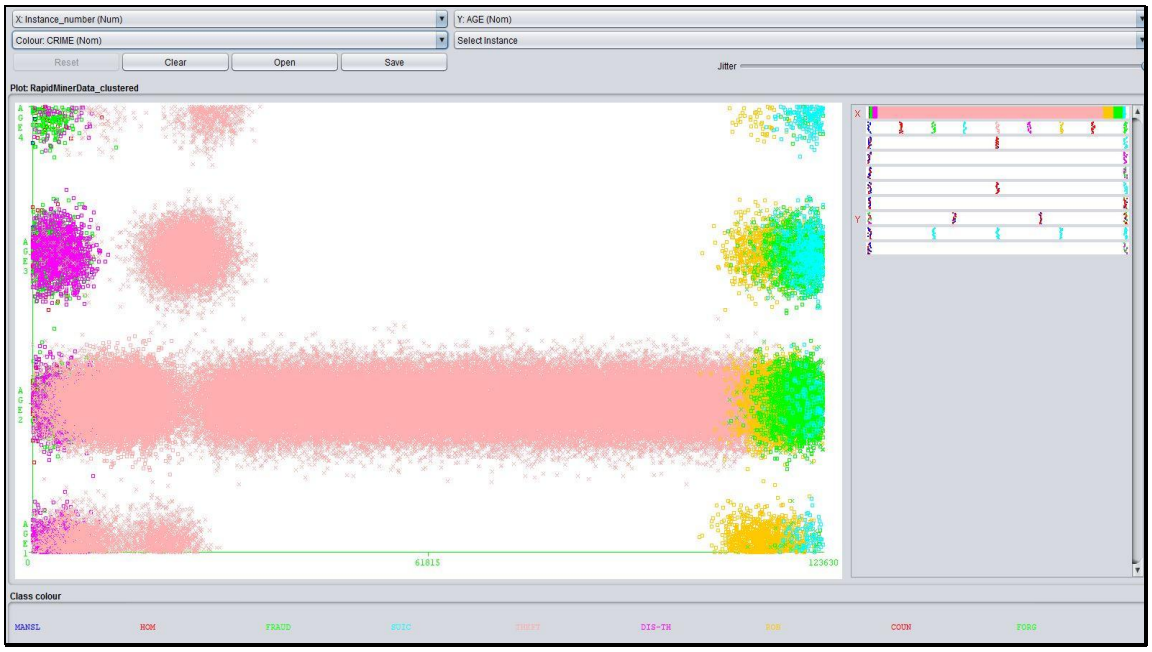
Σχήμα 4-20: Οπτικοποίηση εγκλημάτων σε σχέση με το έγκλημα και την ηλικία.

Η ομαδοποίηση των περιπτώσεων (σχήμα 4-20) με όμοια χαρακτηριστικά σε σχέση με το έγκλημα και την ηλικία, δείχνει για τις ηλικίες των 07 έως 17 χρόνων ότι υπερέχουν οι κλοπές, οι ληστείες και οι διακεκριμένες κλοπές, ενώ για τις ηλικίες των 18 έως 34 χρόνων υπάρχουν σε μεγαλύτερο βαθμό τα προηγούμενα εγκλήματα με επιπρόσθετα, σε αισθητό αριθμό σε σχέση με το σύνολό τους, τις απάτες, τις πλαστογραφίες, τις αυτοκτονίες, ανθρωποκτονίες με πρόθεση και παραχάραξη. Στις ηλικίες των 35 έως 59 ισχύει ότι και για τις ηλικίες 18-34 με ελάχιστη μείωση του όγκου, με εξαίρεση για τα εγκλήματα της ανθρωποκτονίας από αμέλεια και πλαστογραφίας που είναι περισσότερες από πριν. Τέλος, για τις ηλικίες των 60 και άνω χρόνων παρατηρείται αισθητή μείωση του αριθμού των περιπτώσεων εκτός από τις αυτοκτονίες και κλοπές.



Σχήμα 4-21: Οπτικοποίηση εγκλημάτων σε σχέση με την εθνικότητα και τον αριθμό περιπτώσεων.

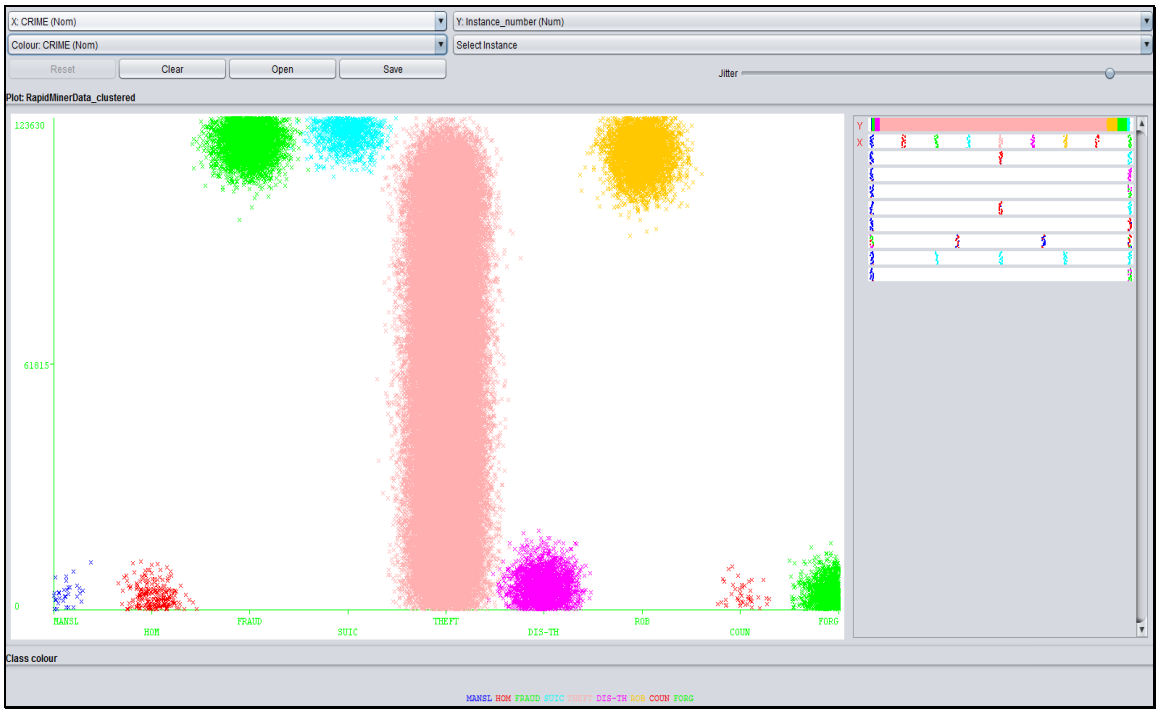
Σύμφωνα με την οπτικοποίηση του σχήματος 4-21 παρατηρείται ομαδοποίηση των περιπτώσεων με όμοια χαρακτηριστικά σε σχέση με την εθνικότητα και τον αριθμό περιπτώσεων, όπου φαίνεται ότι οι κλοπές υπερισχύουν σε αριθμό έναντι των υπολοίπων εγκλημάτων με ημεδαπούς δράστες και παρουσιάζουν ομοιότητες με τις απάτες, τις ληστείες και τις διακεκριμένες κλοπές. Στους αλλοδαπούς δράστες παρατηρούνται ομοιότητες στον αριθμό περιπτώσεων μεταξύ των κλοπών και απατών και με ελάχιστα λιγότερο αριθμό στις ληστείες. Ακόμη, υπάρχει ομαδοποίηση των κλοπών από αλλοδαπούς δράστες με τις διακεκριμένες κλοπές και την πλαστογραφία, όπου οι δύο τελευταίες υστερούν σε αριθμό περιπτώσεων έναντι των κλοπών. Επίσης, φαίνεται ότι οι περιπτώσεις αυτοκτονίας είναι λίγες, δεν σχετίζονται με άλλα εγκλήματα και οι αυτόχειρες είναι άγνωστης εθνικότητας.



Σχήμα 4-22: Οπτικοποίηση εγκλημάτων σε σχέση με τον αριθμό περιπτώσεων και την ηλικία.

Η οπτικοποίηση του σχήματος 4-22 δείχνει την ομαδοποίηση των περιπτώσεων με όμοια χαρακτηριστικά σε σχέση με τον αριθμό περιπτώσεων και την ηλικία, όπου φαίνεται ότι ο μεγαλύτερος αριθμός περιπτώσεων των κλοπών είναι ηλικίας 18-34 και ομαδοποιείται με τις απάτες και λιγότερο με τις διακεκριμένες κλοπές, τις ληστείες και την πλαστογραφία. Ομαδοποίηση περίπου ισόποσων περιπτώσεων υπάρχει στις αυτοκτονίες, απάτες, πλαστογραφίες και ληστείες, ηλικίας 35-59 χρόνων, ενώ στις ηλικίες αυτές υπάρχει μεγάλη ομογενοποίηση των περιπτώσεων των κλοπών, καθώς και των διακεκριμένων κλοπών. Στις ηλικίες των 60 και άνω φαίνεται ομαδοποίηση στις

απάτες και ανθρωποκτονίες από αμέλεια και ξεχωριστά στις αυτοκτονίες και ληστείες. Τέλος, φαίνεται ότι στις μικρές ηλικίες των 7 έως 17 ομαδοποιούνται οι ληστείες με σχεδόν διπλάσιο αριθμό από τις αυτοκτονίες και πλαστογραφίες και ξεχωριστά οι κλοπές με τις διακεκριμένες κλοπές με τις τελευταίες να είναι ελάχιστες σε αριθμό περιπτώσεων.



Σχήμα 4-23: Οπτικοποίηση εγκλημάτων σε σχέση με έγκλημα και τον αριθμό περιπτώσεων.

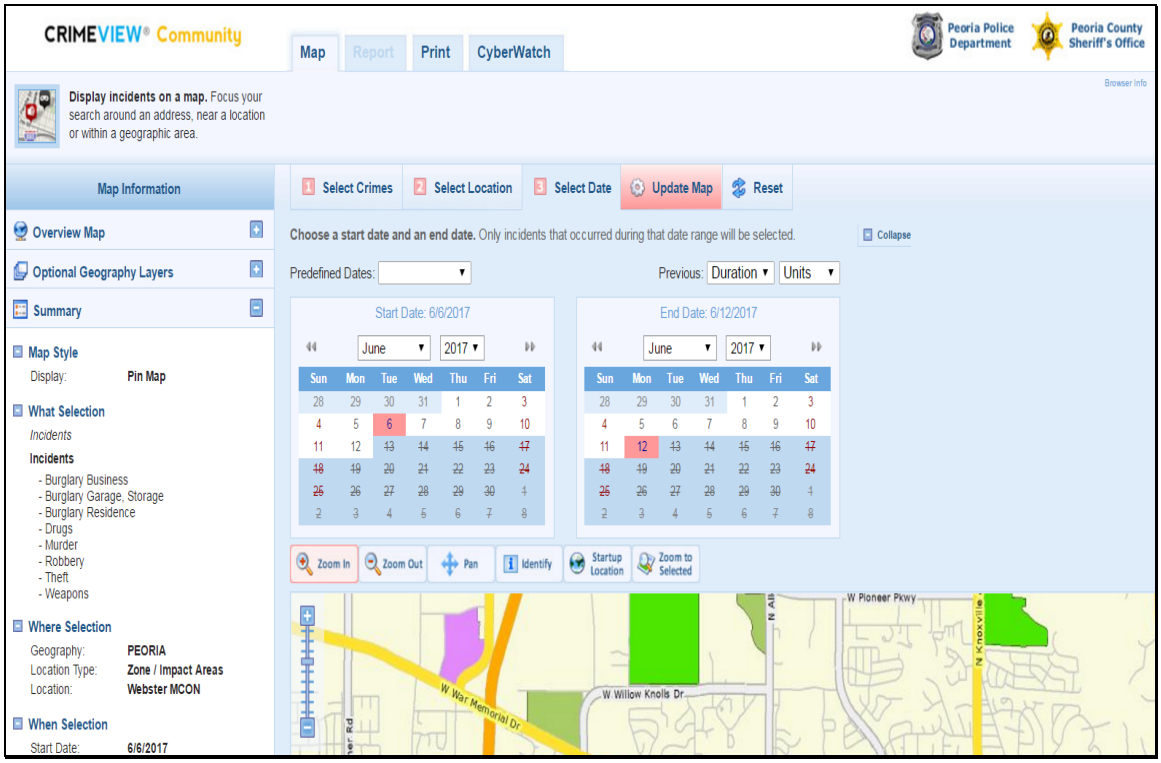
Στο σχήμα 4-23 φαίνεται ο όγκος των περιπτώσεων της βάσης δεδομένων που διαχωρίζεται στις τιμές (κλοπές, ληστείες, αυτοκτονίες κτλ.) του χαρακτηριστικού ¨Crime¨. Διακρίνεται, σε φθίνουσα αριθμητική σειρά, η πληθώρα περιπτώσεων σε κλοπές, ληστείες, απάτες, διακεκριμένες κλοπές, πλαστογραφίες και αυτοκτονίες.

Τα πλήρη αποτελέσματα όλων των χρησιμοποιηθέντων αλγορίθμων παρατίθενται στο παράρτημα 2.

4.3 Απεικόνιση εγκλημάτων σε GIS

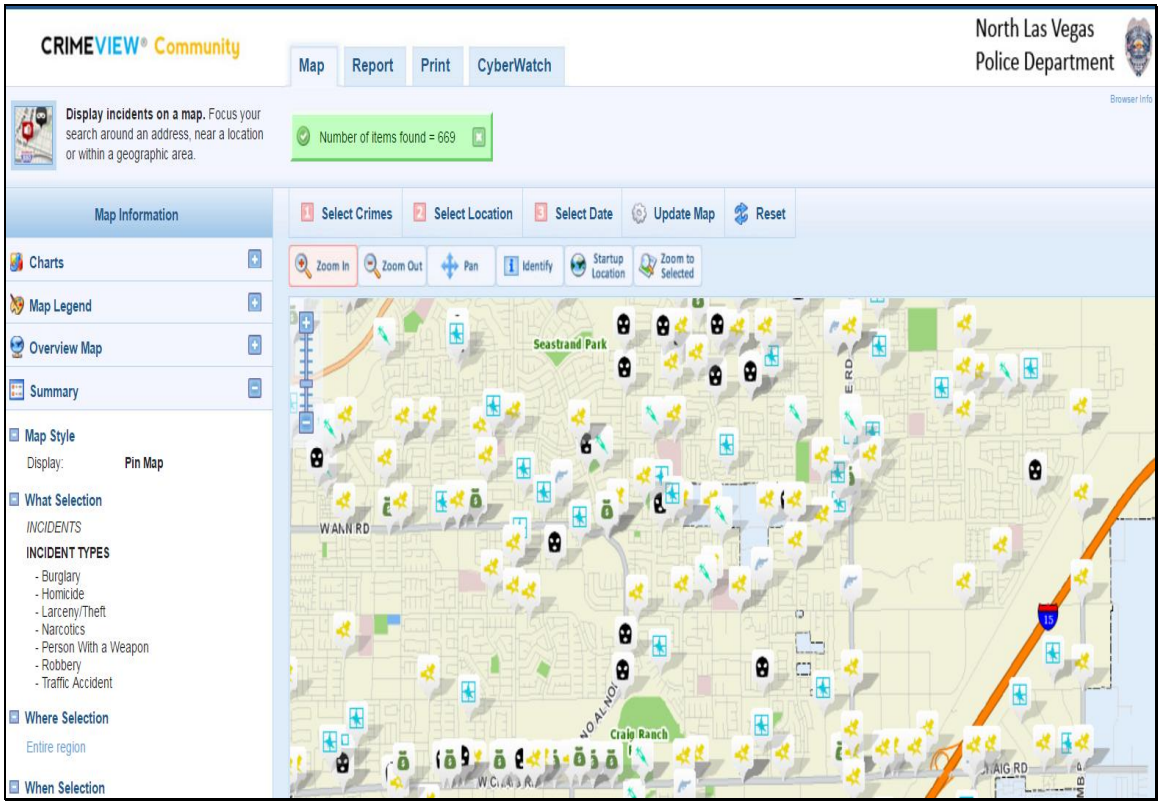
Η Ελληνική Αστυνομία και συγκεκριμένα η Διεύθυνση Πληροφορικής του Αρχηγείου της Ελληνικής Αστυνομίας λειτουργεί υποδομή Γεωγραφικού Συστήματος Πληροφοριών από το 2008 με την οποία καλύπτει της ανάγκες της σε θέματα

χαρτογράφησης της εγκληματικότητας, παραγωγής χαρτογραφικών δεδομένων και χαρτών. Το λογισμικό που χρησιμοποιείται είναι το ArcGIS 9.3 Enterprise Standard της ESRI, καθώς και η εφαρμογή γεωγραφικής απεικόνισης εγκληματικότητας "CRIMEVIEW", όπου δίνεται η δυνατότητα προχωρημένης χωρικής ανάλυσης συμβάντων και επόπτευσής τους, μέσω ενός φυλλομετρητή διαδικτύου. Η συγκεκριμένη εφαρμογή εφαρμόζεται σε πολλές χώρες, όπως και στις Ηνωμένες Πολιτείες Αμερικής και υπάρχουν δικτυακοί τόποι στους οποίους μπορεί ο χρήστης να αντλήσει πληροφορίες, όπως π.χ. για την πόλη Πεόρια της κομητείας Ιλινόις και το Λας Βέγκας της πολιτείας της Νεβάδας των Η.Π.Α. (βλέπε 4-24 και 4-25) (opengov.gr).



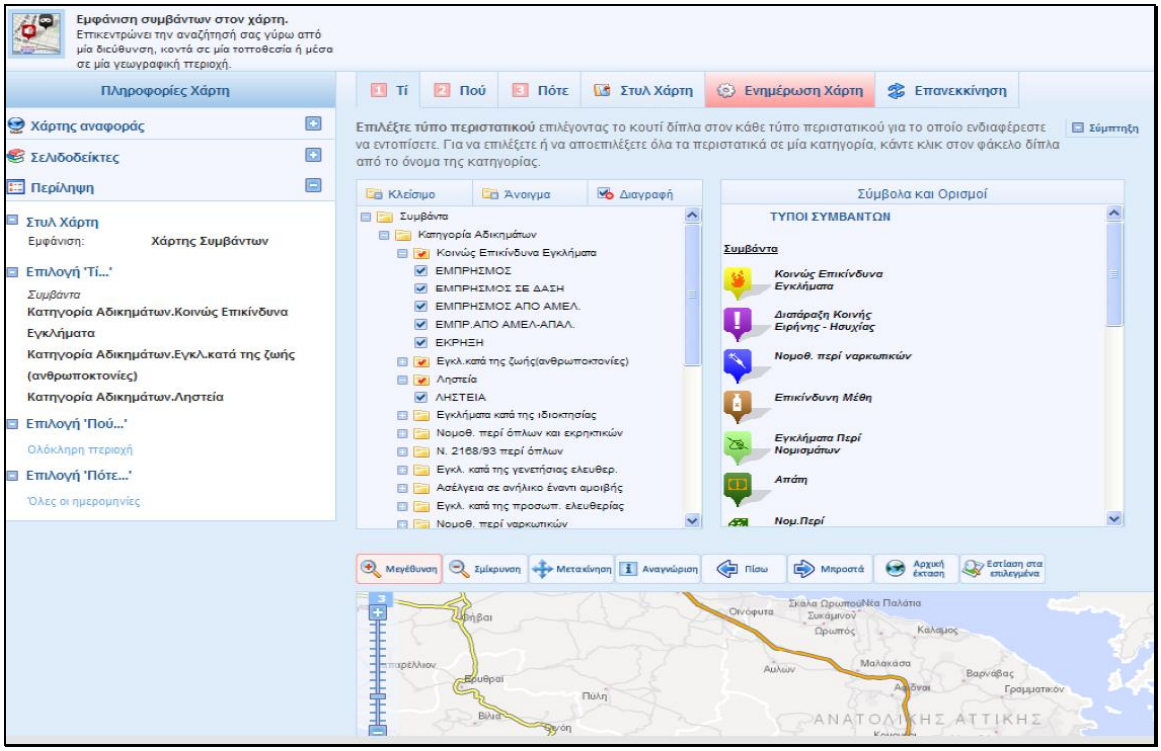
Σχήμα 4-24: Αναζήτηση με κριτήρια ΄΄Εγκλημα¨, ¨Τοποθεσία¨ και ¨Ημερομηνία¨ στην Πεόρια.

Πηγή: peoria.il, 2017.



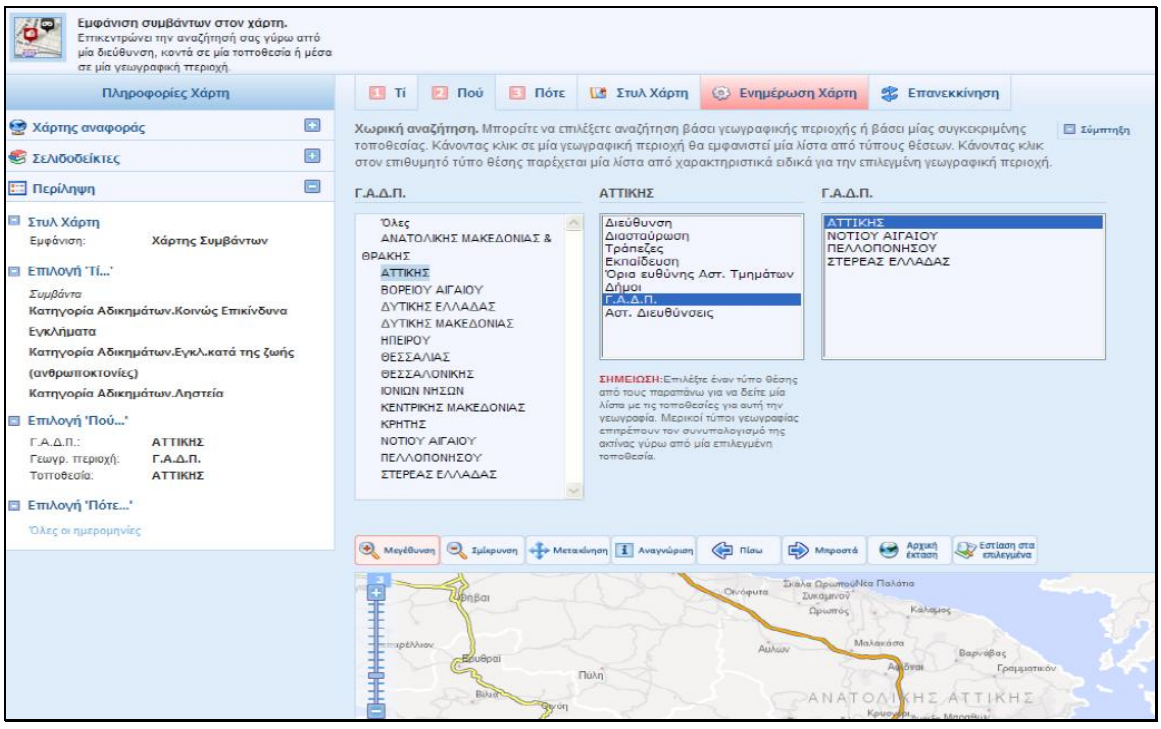
Σχήμα 4-25: Αναζήτηση με κριτήρια ¨Έγκλημα¨, ¨Τοποθεσία¨ και ¨Ημερομηνία¨ στο Β. Λας Βέγκας.
Πηγή: cityofnorthlasvegas.com, 2017.

Η εφαρμογή ¨Βιβλίο Αδικημάτων Συμβάντων¨ της Ελληνικής Αστυνομίας στην οποία καταχωρούνται τα συμβάντα και υπάρχει πρόσβαση από όλα τα Αστυνομικά Τμήματα της Ελλάδας, είναι συνδεδεμένη με την εφαρμογή "CRIMEVIEW" και παρέχετε αυτόματη ενημέρωση των συμβάντων και απεικόνιση τους σε χαρτογραφικό υπόβαθρο της τοποθεσίας τους. Ακόμη, μπορούν να γίνουν άμεσα ερωτήματα σε πολλαπλά γεωγραφικά επίπεδα πληροφορίας. Υπάρχουν φίλτρα αναζητήσεων συμβάντων ανά είδος αδικήματος (επιλογή ενός ή περισσοτέρων), αριθμό συμβάντος, μέθοδος και μέσο τέλεσης, γεωγραφική/διοικητική οντότητα (περιφέρεια, νομό, αστυνομικό τμήμα) ημερομηνία, ώρα, καθώς και άλλα τα οποία βοηθούν στην διαδικασία της αναζήτησης του τι συμβάν έγινε, πού έγινε και πότε (βλέπε σχήματα 4-26, 4-27, 4-28) (opengov.gr, 2017; astynomia.gr, marathondata.gr, 2017).



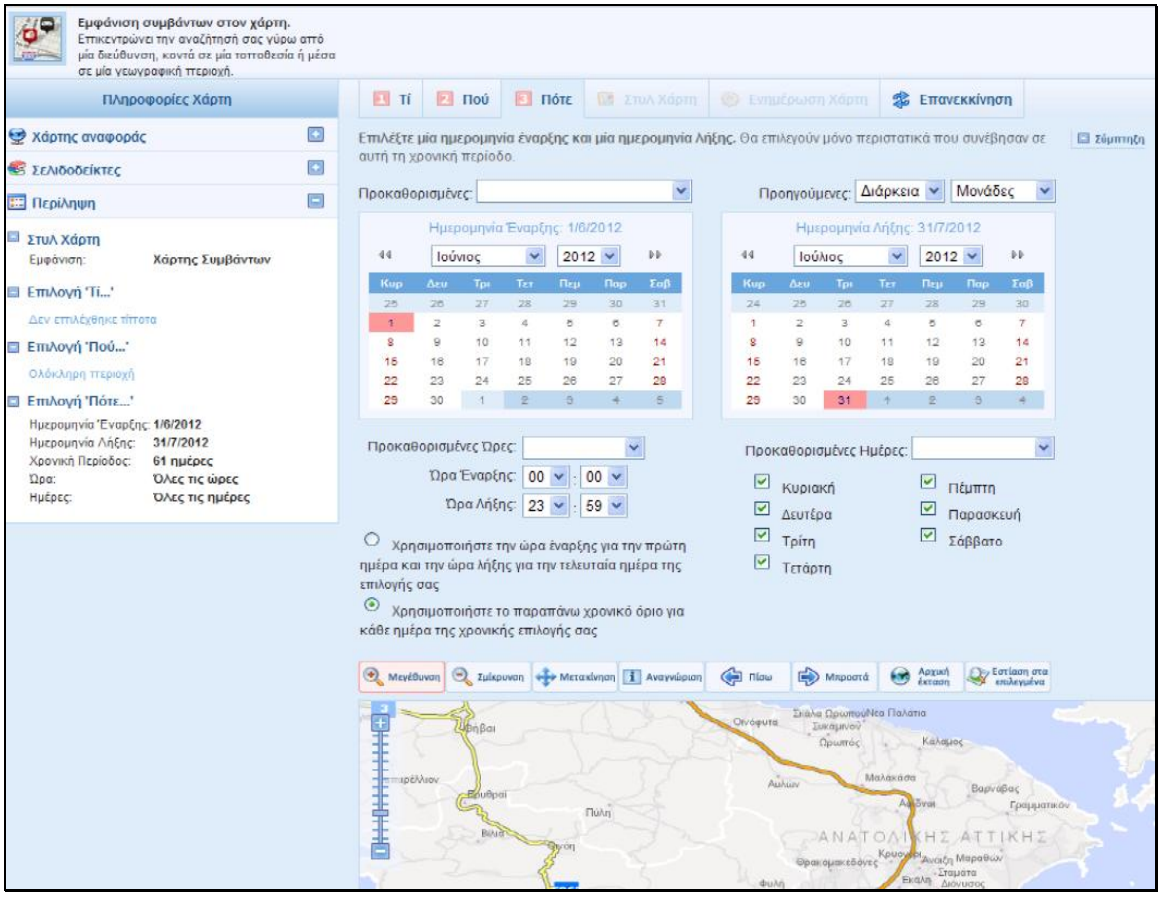
Σχήμα 4-26: Αναζήτηση με κριτήριο το ¨τι¨.

Πηγή: Διεύθυνση Πληροφορικής/Α.Ε.Α., 2017.



Σχήμα 4-27: Αναζήτηση με κριτήριο το ¨που¨.

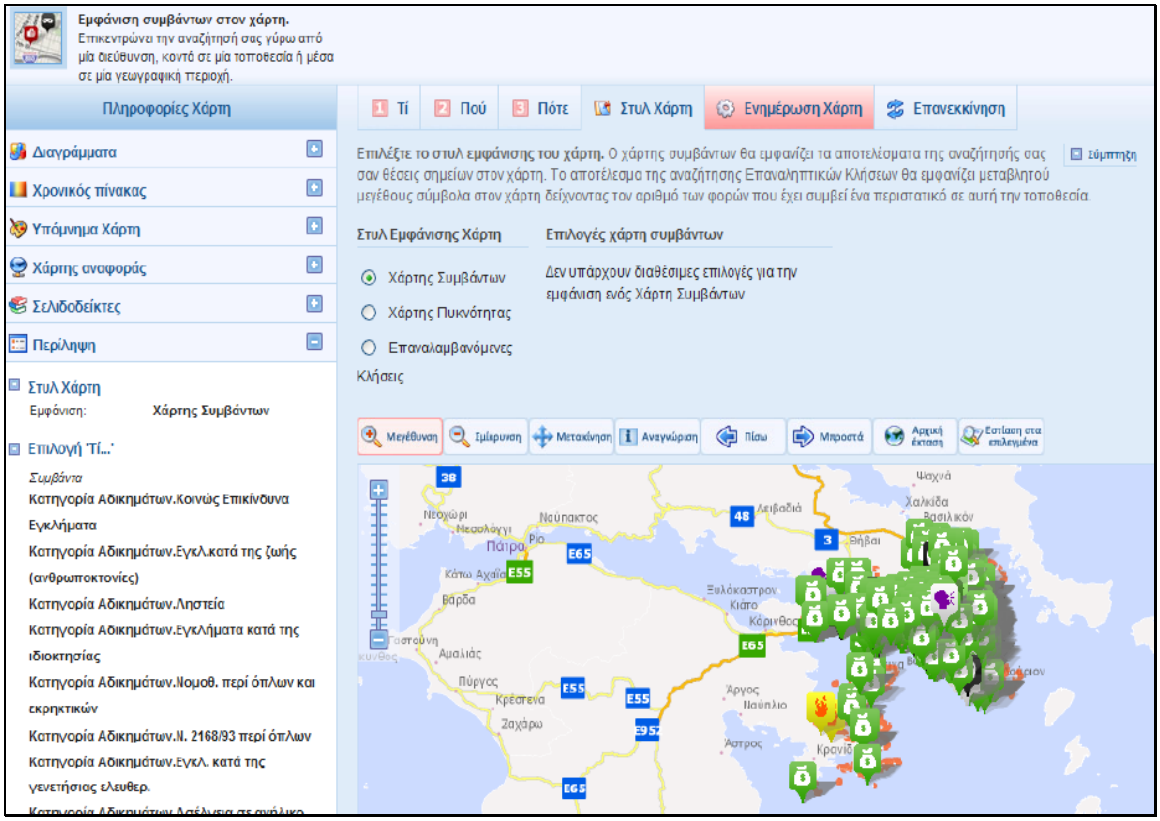
Πηγή: Διεύθυνση Πληροφορικής/Α.Ε.Α., 2017.



Σχήμα 4-28: Αναζήτηση με κριτήριο το ¨πότε¨.

Πηγή: Διεύθυνση Πληροφορικής/Α.Ε.Α., 2017.

Επίσης, υπάρχει η δυνατότητα απεικόνισης χάρτη διαγραμμάτων συμβάντων και του χρονικού τους πίνακα, επαναλαμβανόμενων κλήσεων και πυκνότητας αδικημάτων ανά γεωγραφική/διοικητική οντότητα. Ο χειριστής μπορεί να δημιουργήσει αναφορές, όπως λεπτομερειών συμβάντων και χρονικού δείκτη, δυναμικά γραφήματα σε μορφή πίτας ή ραβδογράμματος και να τα εξάγει σε μορφή xls, doc και pdf, και τέλος να εκτυπώσει αυτές και τις δημιουργηθείσες απεικονίσεις (βλέπε σχήματα 4-29, 4-30, 4-31, 4-32, 4-33 και 4-34) (opengov.gr, 2017; Διεύθυνση Πληροφορικής/Α.Ε.Α., 2017; marathondata.gr, 2017).



Σχήμα 4-29: Χάρτης συμβάντων¨.

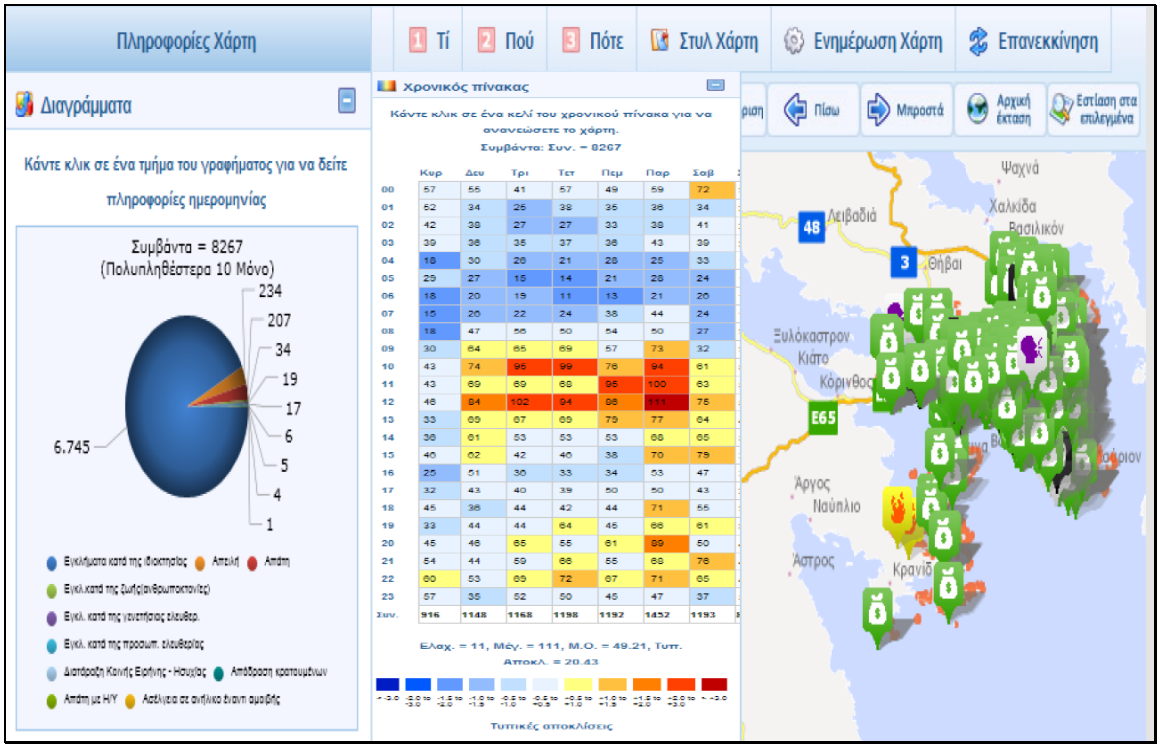
Πηγή: Διεύθυνση Πληροφορικής/Α.Ε.Α., 2017.

Σχήμα 4-30: Χάρτης Πυκνότητας αρπαγής τσάντας, κινητών, πορτοφολιών κτλ.

Πηγή: Διεύθυνση Πληροφορικής/Α.Ε.Α., 2017.

Σχήμα 4-31: Χάρτης συμβάντων¨.

Πηγή: Διεύθυνση Πληροφορικής/Α.Ε.Α., 2017.

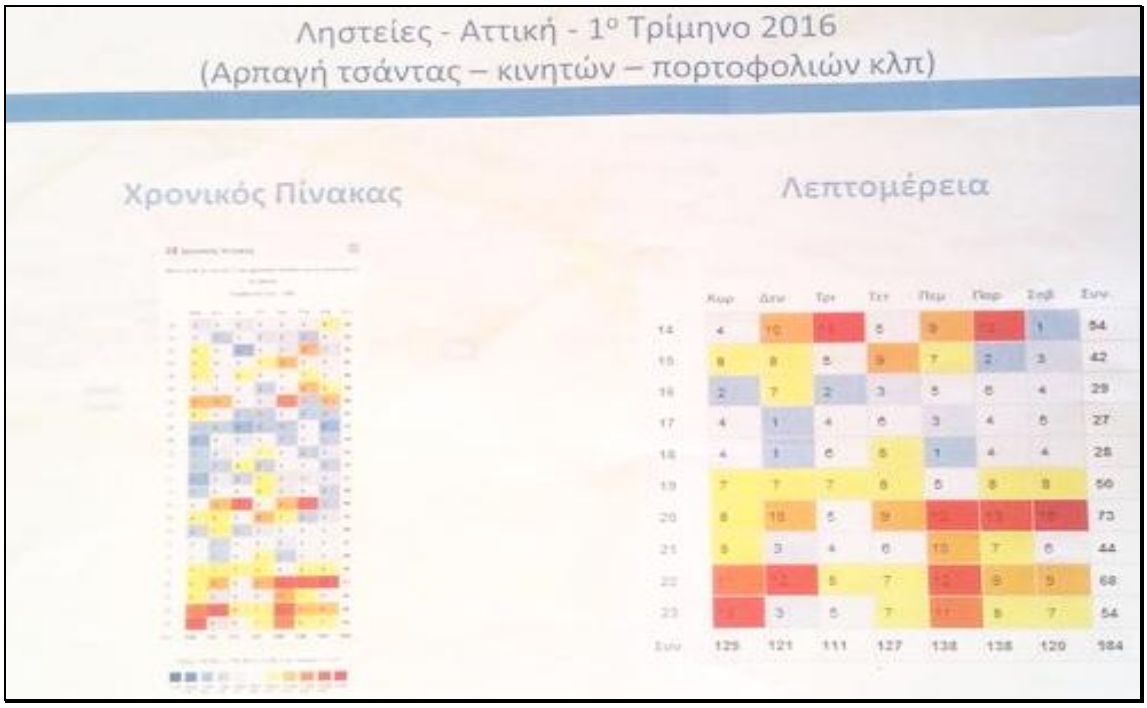


Σχήμα 4-32: Δημιουργία διαγράμματος και του χρονικού πίνακα.

Πηγή: Διεύθυνση Πληροφορικής/Α.Ε.Α., 2017.

Σχήμα 4-33: Δημιουργία διαγράμματος.

Πηγή: thetoc.gr, 2017.



Σχήμα 4-34: Δημιουργία χρονικού πίνακα αρπαγής τσάντας, κινητών κτλ. για το 1^{Ο} 3μηνο 2016.

Πηγή: thetoc.gr.com, 2017.

Τα εικονίδια που απεικονίζονται στον χάρτη αντιστοιχούν στο είδος του συμβάντος, όπου για παράδειγμα στις ληστείες εμφανίζεται μια κουκούλα κάλυψης

προσώπου σε μαύρο φόντο, ένα ζάρι για παραβάσεις του νόμου περί παιγνίων, ένας άσπρος σάκος σε πράσινο φόντο με την ένδειξη του ευρώ για εγκλήματα κατά της ιδιοκτησίας, ένα περίγραμμα ανθρώπου ξαπλωμένο σε κόκκινο φόντο για ανθρωποκτονίες, μια σύριγγα για εγκλήματα περί ναρκωτικών κτλ.. Ανάλογα με την σοβαρότητα και πυκνότητα των εγκλημάτων, οι περιοχές στον χάρτη αλλά και στο ημερολόγιο του χρονικού πίνακα έχουν διάφορα χρώματα (κόκκινο, πορτοκαλί, κίτρινο) με δυνατότητα ενημέρωσης της σοβαρότητας, χρησιμοποιώντας έντονο κόκκινο χρώμα. Επομένως, το σύστημα μπορεί να ενημερώσει τον χρήστη για ¨hotspot¨ σημεία και επαναλαμβανόμενα μοτίβα εγκλημάτων σε συγκεκριμένο τόπο και χρόνο, ώστε να παρθούν οι πρέπουσες αποφάσεις και να σχεδιαστούν και συντονιστούν οι κατάλληλες ενέργειες αντιμετώπισης.

Η εφαρμογή “CRIMEVIEW” αποτελεί ένα πολύ χρήσιμο και απαραίτητο εργαλείο για την Αστυνομία παρέχοντας χαρτογραφημένη απεικόνιση του εγκλήματος σε όλη την Ελλάδα και ρυθμίζοντας κάθε φορά τις επιλογές ανάλογα με τα επιθυμητά αποτελέσματα, αντλούνται πολύτιμες πληροφορίες για την στρατηγική αστυνόμευσης, τη καταστολή, τη πρόληψη, την εξιχνίαση του εγκλήματος και γενικώς την παροχή μεγάλης βοήθειας στο έργο της.

5 Αξιολόγηση Αποτελεσμάτων Εξόρυξης Δεδομένων – Συμπεράσματα

Για την παρούσα έρευνα χρησιμοποιήθηκε μία αρκετά μεγάλου όγκου βάση δεδομένων. Σύμφωνα με τα αποτελέσματα του κατηγοριοποιητή J48, ο οποίος επεξεργάστηκε τα δεδομένα δύο φορές και με διαφορετικές ρυθμίσεις, παρατηρήθηκε πολύ μεγάλο ποσοστό σωστών κατηγοριοποιημένων περιπτώσεων. Συγκεκριμένα, πάνω από 91% (95,395% και 91,277%) που σημαίνει ότι τα αποτελέσματα της κατηγοριοποίησης είναι πάρα πολύ ικανοποιητικά και με μεγάλη ακρίβεια. Ο αριθμός των φύλλων του δέντρου αποφάσεων στην πρώτη δοκιμή των 123.631 περιπτώσεων ήταν 86 φύλλα ενώ στην δεύτερη των 122.524 περιπτώσεων ήταν 44 φύλλα. Η τιμή του μήκους του δέντρου και ο χρόνος περάτωσης της δημιουργίας του μοντέλου για την πρώτη δοκιμή ήταν αντίστοιχα 143 (αριθμός κόμβων) και 2,79 δευτερόλεπτα ενώ για την δεύτερη 75 και 0,47 δευτερόλεπτα. Αξιοσημείωτο είναι το γεγονός ότι και στις δύο δοκιμές στο δέντρο αποφάσεων (με μορφή κειμένου), φαίνονται μετά από τους κόμβους των φύλων μέσα σε παρενθέσεις, ο αριθμός των περιπτώσεων και έπειτα των λαθών (π.χ. 1000.0/6.0), όπου τα λάθη είναι πάρα πολύ λίγα. Μετά το σημείο στίξης της άνω και κάτω τέλειας ¨:¨ δηλώνεται η τιμή της κλάσης που επιλέχθηκε ως χαρακτηριστικό (Crime). Στα αποτελέσματα και των δύο τιμών φαίνεται η ακρίβεια της κατηγοριοποίησης των περιπτώσεων σε σχέση με τις τιμές της επιλεχθείσας κλάσης. Για διευκόλυνση κατανόησης των αποτελεσμάτων επεξηγούνται οι ετικέτες:

- TP Rate (True Positives Rate): ποσοστό πραγματικών θετικών (περιπτώσεις που έχουν ταξινομηθεί σωστά ως δεδομένη κλάση).
- FP Rate (False Positives Rate): ποσοστό ψευδών θετικών (περιπτώσεις που ταξινομούνται ψευδώς ως δεδομένη κατηγορία).
- Precision: το ποσοστό των θετικών προβλέψεων που είναι πραγματικά θετικές. Είναι ένα ποσοστό TP (αναφέρεται ως θετική τιμή πρόβλεψης-PPV) και ισούται με TP / προβλεπόμενες Θετικές.
- Recall: το ποσοστό των περιπτώσεων που πραγματικά είναι θετικές και είχαν προβλεφθεί θετικά. Είναι ένα ποσοστό TP (αναφέρεται ως ευαισθησία) και ισούται με TP / πραγματικές θετικές.

- F-Measure: Συνδυασμένο μέτρο ακριβείας και ανάκλησης υπολογιζόμενο ως 2 * ακρίβεια * ανάκληση / (ακρίβεια + ανάκληση).
- MCC (Matthews Correlation Coefficient): είναι ένας συντελεστής συσχετισμού μεταξύ των παρατηρούμενων και προβλέψεων (μέτρο της ποιότητας) των δυαδικών κατηγοριοποιήσεων.
- ROC Area (Receiver Operator Characteristic Area): Η τυπική καμπύλη ROC είναι μια γραφική παράσταση του TP Rate έναντι του FP Rate. Η βέλτιστη κατηγοριοποίηση έχει ROC Area τιμή που τείνει στο 1 και με τιμή 0,5 θεωρείται ως τυχαία.
- PRC Area (Precision Recall Curve Area): Η καμπύλη ακρίβειας-ανάκλησης, όπου ακρίβεια είναι το κλάσμα των ανακτώμενων περιπτώσεων που είναι σχετικές, ενώ ανάκληση είναι το κλάσμα των σχετικών περιπτώσεων που ανακτώνται.
- Kappa statistic: Το στατιστικό στοιχείο Kappa μετρά τη συμφωνία πρόβλεψης με την πραγματική κλάση (μεταξύ των κατηγοριοποιήσεων και των αληθινών κλάσεων) - 1.0 σημαίνει ότι είναι πλήρης.
- Error Rates: Τα ποσοστά σφάλματος χρησιμοποιούνται για την αριθμητική πρόβλεψη και δεν έχουν πολύ νόημα για εργασίες κατηγοριοποίησης (στην παλινδρόμηση έχει διαφορετική σημασία).
- Confusion Matrix: Οι πρώτοι αριθμοί παρουσιάζονται στο ¨Confusion Matrix¨, με τα a, b κτλ. να αντιπροσωπεύουν τις ετικέτες κλάσης. Ο αριθμός των σωστά κατηγοριοποιημένων περιπτώσεων είναι το άθροισμα των διαγωνίων στο πλέγμα και τα υπόλοιπα είναι εσφαλμένα ταξινομημένα.

Η πρώτη δοκιμή είναι πολύ καλή με εξαίρετα αποτελέσματα και προτιμάται. Το Kappa statistic είναι 0.7651 και το ROC Area δίνει SUIC (Αυτοκτονία)=1,000, THEFT (Κλοπή)=0,975, DIS-TH (Διακεκριμένη κλοπή)=0,808, ROB (Ληστεία)=0,910 και FORG (Πλαστογραφία)=0,753 ενώ αντίθετα στην δεύτερη έχουμε Kappa statistic=0.3983, THEFT=0,940, DIS-TH=0,493, ROB =0,265 και FORG=0,482.

Όπως φαίνεται στην πρώτη δοκιμή, έγινε καλή προεπεξεργασία των δεδομένων και υπάρχει μεγάλη συμφωνία των κατηγοριοποιήσεων με τις πραγματικές κλάσεις, καθώς και υψηλές τιμές των δεικτών σε Αυτοκτονία, Κλοπή, Διακεκριμένη κλοπή, Ληστεία και Πλαστογραφία.

Στην συνέχεια παρατίθεται ο πίνακας 5-1 με τους αλγορίθμους που χρησιμοποιήθηκαν στην κατηγοριοποίηση, καθώς και με συγκεκριμένα στατιστικά στοιχεία που παρουσιάζουν ενδιαφέρον.

Πίνακας 5-1: Αλγόριθμοι Κατηγοριοποίησης.

Αλγόριθμος/ Στατιστικά	**J48** Cross-validation Folds=10	**J48** Unpruned=True Cross-validation Folds=10	**NaïveBayes** Cross-validation Folds=10	**Multilayer Perceptron** Percentage split 66%	**Multilayer Perceptron** Cross-validation Folds=10
Correctly Classified Instances	**95,395%**	**91,277%**	**91,876%**	**95.327%**	**95.333%**
Time taken to build model	**2,79¨**	**0,47¨**	**0,14¨**	**696.88¨**	**701.44¨**
Kappa statistic	**0.7651**	**0.3983**	**0.5956**	**0.7599**	**0.762**
SUIC					
ROC Area	**1,000**	**-**	**1,000**	**1,000**	**1,000**
Precision	**1,000**	**-**	**1,000**	**1,000**	**1,000**
Recall	**1,000**	**-**	**1,000**	**1,000**	**0,999**
THEFT					
ROC Area	**0,975**	**0,940**	**0,974**	**0,977**	**0,978**
Precision	**0,963**	**0,921**	**0,951**	**0,963**	**0,963**
Recall	**0,997**	**0,999**	**0,979**	**0,998**	**0,997**
DIS-TH					
ROC Area	**0,808**	**0,493**	**0,424**	**0,758**	**0,791**
Precision	**0,775**	**0,638**	**0,343**	**0,759**	**0,770**
Recall	**0,817**	**0,472**	**0,482**	**0,803**	**0,790**
ROB					
ROC Area	**0,910**	**0,265**	**0,823**	**0,907**	**0,900**
Precision	**0,887**	**0,889**	**0,788**	**0,902**	**0,887**
Recall	**0,899**	**0,144**	**0,782**	**0,881**	**0,896**
FORG					
ROC Area	**0,753**	**0,482**	**0,513**	**0,704**	**0,759**
Precision	**0,854**	**0,548**	**0,382**	**0,848**	**0,852**
Recall	**0,671**	**0,422**	**0,484**	**0,642**	**0,666**

Όπως παρατηρείται στον πίνακα 4-1 το ποσοστό σωστών κατηγοριοποιημένων περιπτώσεων με όλους τους κατηγοριοποιητές είναι πάνω από 91% και είναι ένα πάρα πολύ ικανοποιητικό ποσοστό με ασφαλή πρόβλεψη κατηγοριοποίησης, βάσει

χαρακτηριστικών. Ο χρόνος περάτωσης της δημιουργίας του μοντέλου είναι μικρός πλην του κατηγοριοποιητή MultilayerPerceptron που είναι ιδιαίτερα μεγάλος είτε με την τεχνική Percentage split 66% (696.88¨) είτε με την Cross-validation με Folds=10 (701.44¨) και δεν ενδείκνυται για πολύ μεγάλες βάσης δεδομένων, όπως στην παρούσα εργασία, λόγω του ότι χρειάζεται περισσότερη υπολογιστική ισχύ και χρόνο. Η συμφωνία πρόβλεψης των κατηγοριοποιήσεων και των αληθινών κλάσεων (Kappa statistic) για την δεύτερη περίπτωση του αλγορίθμου J48 (Unpruned=True) είναι κακή, διότι είναι κάτω από το 0,5 ενώ του NaïveBayes είναι μέτρια με τους υπολοίπους να είναι αρκετά ικανοποιητικοί. Όσον αφορά τις τιμές του ROC Area, όπου επιλέχθηκαν στον πίνακα 4-1 αυτές που είναι κοντά στο 0,5 (όχι καλός, θεωρείται τυχαίως) και πάνω, καθώς και του ¨Precision¨ και ¨Recall¨, παρατηρείται ότι στην πρώτη δοκιμή ο J48 και στις δύο δοκιμές ο MultilayerPerceptron είναι οι πιο αποτελεσματικοί κατηγοριοποιητές σε σχέση με τους υπολοίπους. Η διαφορά στα αποτελέσματα μεταξύ των δύο τεχνικών που χρησιμοποιήθηκαν στον αλγόριθμο MultilayerPerceptron είναι μηδαμινή και με μόλις 0,006% απόκλιση στο ποσοστό των σωστών κατηγοριοποιήσεων των περιστατικών. Ακόμη, αξιοπρόσεκτο είναι ότι παρά τις όποιες διαφορές μεταξύ των κατηγοριοποιητών, εκτός από το πολύ κοντινό ποσοστό σωστών κατηγοριοποιήσεων των περιστατικών που είναι σταθερά πάνω από το 91%, είναι επίσης σταθερός και ο μετρητής ROC Area, ¨Precision¨ και ¨Recall¨ για τις τιμές ¨SUIC¨ και ¨THEFT¨ της κλάσης ¨Crime¨. Για κάποιες τιμές κλάσης (ανθρωποκτονίες, απάτες και παραχάραξη) η κατηγοριοποίηση δεν ήταν επιτυχής, λόγω της ανισοκατανομής των κλάσεων. Τα αποτελέσματα παρέμειναν τα ίδια ακόμη και ύστερα από δοκιμές με τεχνικές ¨ClassBalancer¨, ¨SMOTE¨ και ¨Resample¨ και η αντιμετώπιση αυτού του προβλήματος παραμένει ως αντικείμενο περαιτέρω έρευνας.

Με την χρήση των συγκεκριμένων μοντέλων μπορεί να βοηθηθεί η αστυνομία στον εντοπισμό εγκλημάτων που χρήζουν ιδιαίτερης προσοχής και ταλανίζουν την ομαλή κοινωνική συμβίωση των πολιτών, ώστε να παρθούν τα ανάλογα μέτρα για την αποτελεσματικότερη αντιμετώπιση των συγκεκριμένων εγκλημάτων. Γίνεται εκτίμηση των εγκλημάτων με την ανάκτηση και εξαγωγή πληροφορίας από την μεγάλη βάση δεδομένων που χρησιμοποιείται. Μπορεί να διαπιστωθεί εάν ο όγκος κάποιων εγκλημάτων (π.χ. κλοπές) έχει την αντίστοιχη βαρύτητα και αντίκτυπο στην κοινωνία και επιβάλλεται μεγαλύτερη προσοχή στην αντιμετώπισή τους. Για παράδειγμα, σύμφωνα με τα αποτελέσματα των δεικτών, οι κλοπές που υπερτερούν σε αριθμό περιπτώσεων πρέπει να αντιμετωπιστούν αποτελεσματικότερα, όπως και οι

διακεκριμένες κλοπές, οι ληστείες, οι πλαστογραφίες και αυτοκτονίες. Αντιθέτως, τα αποτελέσματα για τις απάτες που είναι σε αριθμό πολύ περισσότερες (4.501) από τις αυτοκτονίες (1.107) και σχεδόν διπλάσιες από τις διακεκριμένες κλοπές (2.525), δεν έχουν την ίδια βαρύτητα.

Από την επεξεργασία των δεδομένων με τον κατηγοριοποιητή NaïveBayes, όπως φαίνεται και στο σχήμα 5-1 (για όλη την διαδικασία αποτελεσμάτων βλέπε παράρτημα 2), είναι ευδιάκριτο ότι το χαρακτηριστικό ¨CRIME¨ με τιμή ¨THEFT¨ έχει τον μεγαλύτερο αριθμό περιστατικών με υψηλότερες τις τιμές χαρακτηριστικών RANK=MISD, SOLVING=UNSOL, CONDITION=ENDED, NATIONALITY=NAT, SEX=MAN, AGE=AGE2 και CAUSE=NOT-KN. Αμέσως επόμενο είναι το ROB με υψηλότερες τις τιμές χαρακτηριστικών RANK= FEL, SOLVING=UNSOL, CONDITION=ENDED, NATIONALITY=NAT, SEX=MAN, AGE=AGE2 και CAUSE=NOT-KN και το FRAUD με τιμές χαρακτηριστικών RANK=MISD, SOLVING=UNSOL, CONDITION=ENDED, NATIONALITY=NAT, SEX=MAN, AGE=AGE2 και CAUSE=NOT-KN. Επομένως, οι ανεξιχνίαστες τελεσμένες πλημμελληματικές κλοπές, πλημμελληματικές ληστείες και κακουργηματικές απάτες ημεδαπών ανδρών δραστών με άγνωστους λόγους δράσης και ηλικίας από 18 έως 34 χρόνων, υπερτερούν έναντι των άλλων εγκλημάτων για το έτος 2016 στην Ελλάδα. Κατόπιν, ακολουθούν οι εξιχνιασμένες τελεσμένες κακουργηματικές διακεκριμένες κλοπές αλλοδαπών και ημεδαπών (σχεδόν στο ίδιο ποσοστό) ανδρών δραστών με άγνωστους λόγους δράσης και ηλικίας από 35 έως 59 χρόνων. Έπειτα, είναι οι εξιχνιασμένες τελεσμένες πλημμελληματικές πλαστογραφίες αλλοδαπών ανδρών δραστών με άγνωστους λόγους δράσης και ηλικίας από 35 έως 59 χρόνων και μετά οι απόπειρες αυτοκτονιών ανδρών ηλικίας από 35 έως 59 χρόνων. Αξιοσημείωτο είναι ότι στις αυτοκτονίες (απόπειρες και τελεσμένες) η σειρά αιτίας είναι πρωτίστως άγνωστοι λόγοι, μετά ασθένειας, έπειτα αισθηματικοί και οικογενειακοί στο ίδιο ποσοστό και τέλος οι οικονομικοί. Στον πίνακα 5-2 φαίνονται συγκεντρωτικά τα 6 πρώτα σε αριθμό περιστατικών εγκλήματα, καθώς και οι υψηλότερες τιμές των χαρακτηριστικών των αντίστοιχων εγκλημάτων. Πέραν της στατιστικής χρησιμότητας των εξαγόμενων πληροφοριών, βάσει παρατηρητικότητας, μπορεί να ωφελήσουν στην Αστυνομία για περαιτέρω προβληματισμό και έρευνα κατανόησης των πτυχών της εγκληματικότητας, όπως ποιοι είναι οι παράγοντες που οδηγούν στην αριθμητική υπεροχή συγκεκριμένων εγκλημάτων με ιδιαίτερες τιμές χαρακτηριστικών.

Naive Bayes Classifier

Attribute	Class MANSL (0)	HOM (0)	FRAUD (0.04)	SUIC (0.01)	THEFT (0.88)	DIS-TH (0.02)	ROB (0.04)	COUN (0)	FORG (0.01)
RANK									
MISD	31.0	3.0	4243.0	1.0	108523.0	531.0	475.0	32.0	1264.0
FEL	3.0	227.0	260.0	1.0	220.0	1996.0	4661.0	12.0	59.0
NOTH	1.0	1.0	1.0	1108.0	1.0	1.0	1.0	1.0	1.0
[total]	35.0	231.0	4504.0	1110.0	108744.0	2528.0	5137.0	45.0	1324.0
SOLVING									
SOL	33.0	229.0	878.0	1108.0	6542.0	1473.0	835.0	43.0	1322.0
UNSOL	1.0	1.0	3625.0	1.0	102201.0	1054.0	4301.0	1.0	1.0
[total]	34.0	230.0	4503.0	1109.0	108743.0	2527.0	5136.0	44.0	1323.0
CONDITION									
ENDED	33.0	229.0	3258.0	454.0	100587.0	2294.0	4700.0	43.0	1308.0
ATT	1.0	1.0	1245.0	655.0	8156.0	233.0	436.0	1.0	15.0
[total]	34.0	230.0	4503.0	1109.0	108743.0	2527.0	5136.0	44.0	1323.0
NATIONALITY									
NAT	31.0	144.0	3854.0	1.0	102657.0	1376.0	4460.0	31.0	403.0

Σχήμα 5-1: Επεξεργασία των δεδομένων με τον κατηγοριοποιητή NaïveBayes.

Πίνακας 5-2: Αποτελέσματα αλγορίθμου NaïveBayes.

R.	CRIME	RANK	SOLVING	CONDITION	NATIONALITY	SEX	AGE	CAUSE
1	THEFT	MISD	UNSOL	ENDED	NAT	MAN	AGE2	NOT-KN
2	ROB	FEL	UNSOL	ENDED	NAT	MAN	AGE2	NOT-KN
3	FRAUD	MISD	UNSOL	ENDED	NAT	MAN	AGE2	NOT-KN
4	DIS-TH	FEL	SOL	ENDED	FOREI-NAT	MAN	AGE3	NOT-KN
5	FORG	MISD	SOL	ENDED	FOREI	MAN	AGE3	NOT-KN
6	SUIC	-	SOL	ATT	NAT	MAN	AGE3	NOT-KN

Κατά την εξαγωγή κανόνων συσχέτισης πρέπει οι κανόνες να πληρούν τα κατώτατα όρια υποστήριξης (minsup) και εμπιστοσύνης (minconf). Με την χρήση του αλγορίθμου Apriori και PredictiveApriori αναζητούνται οι όποιοι πιθανοί συνδυασμοί

χαρακτηριστικών. Στη πρώτη δοκιμή του αλγορίθμου Apriori με τις ρυθμίσεις στην προεπιλογή (minMetric=0.9, metricType=Confidence) δημιουργείται σύνολο 7 χαρακτηριστικών εκ των οποίων προκύπτουν 9 σχέσεις με 1 συχνά συνυπάρχον χαρακτηριστικό. Οι κανόνες που εξάγονται και πληρούν τις ρυθμίσεις ορίων που επιλέχθηκαν είναι οι εξής:

1. CRIME=THEFT 108741 ==> RANK=MISD 108522 <conf:(1)> lift:(1.07) lev:(0.06) [7289] conv:(34.13)
2. SOLVING=UNSOL NATIONALITY=NAT 106771 ==> SEX=MAN 105331 <conf:(0.99)> lift:(1.04) lev:(0.04) [4509] conv:(4.13)
3. SOLVING=UNSOL 111177 ==> SEX=MAN 109487 <conf:(0.98)> lift:(1.04) lev:(0.04) [4505] conv:(3.66)
4. NATIONALITY=NAT SEX=MAN 108215 ==> SOLVING=UNSOL 105331 <conf:(0.97)> lift:(1.08) lev:(0.06) [8017] conv:(3.78)
5. SOLVING=UNSOL SEX=MAN 109487 ==> NATIONALITY=NAT 105331 <conf:(0.96)> lift:(1.05) lev:(0.04) [5304] conv:(2.28)
6. SOLVING=UNSOL 111177 ==> NATIONALITY=NAT 106771 <conf:(0.96)> lift:(1.05) lev:(0.04) [5200] conv:(2.18)
7. NATIONALITY=NAT 112948 ==> SEX=MAN 108215 <conf:(0.96)> lift:(1.01) lev:(0.01) [1560] conv:(1.33)
8. SOLVING=UNSOL 111177 ==> RANK=MISD 106295 <conf:(0.96)> lift:(1.03) lev:(0.02) [2795] conv:(1.57)
9. RANK=MISD 115094 ==> SEX=MAN 109394 <conf:(0.95)> lift:(1.01) lev:(0.01) [713] conv:(1.12)
10. NATIONALITY=NAT 112948 ==> RANK=MISD 107269 <conf:(0.95)> lift:(1.02) lev:(0.02) [2120] conv:(1.37)

Στην αριστερή μεριά του κανόνα φαίνεται πόσες φορές εμφανίστηκε το αριστερό σκέλος μίας ή συνδυασμού τιμών των χαρακτηριστικών (π.χ. 1 κανόνας: CRIME=THEFT 108741) και στο δεξιό μέρος (RANK=MISD 108522) ο αριθμός εμφανίσεων ταυτόχρονα και των δύο σκελών. Επίσης, από την δεξιά μεριά εμφανίζονται τα κριτήρια (conf:(1)> lift:(1.07) lev:(0.06) [7289] conv:(34.13)). Όσο περισσότερο το κριτήριο ¨Confidence¨ τείνει προς την μονάδα, τόσο περισσότερο ισχυρός είναι ο κανόνας. Το κριτήριο ¨Lift¨ όσο πιο κάτω από την μονάδα έχει τιμή, τόσο το αριστερό σκέλος του κανόνα επιδρά αρνητικά στην εμφάνιση του δεξιού, με τιμή 1 δεν υπάρχει

επιρροή, ενώ όσο πιο πάνω από την μονάδα είναι, τότε το αριστερό σκέλος επηρεάζει θετικά τις πιθανότητες εμφανίσεις του δεξιού σκέλους.

Όλοι οι κανόνες που εξήχθησαν έχουν πάρα πολύ υψηλό Confidence, ίσο ή μεγαλύτερο του 0,95 και τα αποτελέσματα από την μελέτη των κανόνων είναι ότι όταν το έγκλημα είναι κλοπή τότε είναι πλημμέλημα, όταν τα ανεξιχνίαστα εγκλήματα συνδυάζονται με ημεδαπούς δράστες τότε είναι άνδρες, το πλήθος των ημεδαπών δραστών διαπράττει πλημμελήματα και γενικώς στους 9 από τους 10 κανόνες γίνονται συνδυασμοί τεσσάρων συγκεκριμένα τιμών χαρακτηριστικών που είναι: το Πλημμέλημα, ο Ημεδαπός, ο Άνδρας και το Ανεξιχνίαστο έγκλημα (RANK=MISD, NATIONALITY=NAT, SEX=MAN, SOLVING=UNSOL).

Στην προκειμένη περίπτωση χρήσης κανόνων συσχέτισης στην δημιουργηθείσα βάση δεδομένων, επιδιώκεται η εξαγωγή γνώσης και χρήσιμων συμπερασμάτων. Γίνεται προσπάθεια ανάδειξης των συσχετίσεων μεταξύ των δεδομένων της βάσης δεδομένων, όπου υπάρχουν και δεν είναι εμφανείς για την Αστυνομία. Αντλούνται πληροφορίες σχέσεων μεταξύ των εγκληματικών χαρακτηριστικών, ώστε να οδηγήσουν τις εν δυνάμει έρευνες σε συγκεκριμένη κατεύθυνση και την ευκολότερη και γρηγορότερη εξιχνίαση του εγκλήματος, εντοπισμού και σύλληψης του δράστη, συμβάλλοντας αποτελεσματικότερα στην καταστολή του εγκλήματος και την εύρυθμη λειτουργία της κοινωνίας. Για παράδειγμα, σύμφωνα με τα αποτελέσματα των κανόνων συσχέτισης συνδυάζονται τέσσερις συγκεκριμένες τιμές χαρακτηριστικών, όπου αναζητείται για την διάπραξη ανεξιχνίαστου πλημμελήματος πιθανόν άνδρας δράστης ημεδαπός.

Στην δεύτερη περίπτωση χρήσης του αλγορίθμου Apriori με ¨MetricType¨ το ¨Lift¨ πήραμε τους παρακάτω 10 κανόνες:

1. CRIME=THEFT SOLVING=UNSOL 102200 ==> RANK=MISD NATIONALITY=NAT SEX=MAN AGE=AGE2 89811 conf:(0.88) < lift:(1.17)> lev:(0.1) [12944] conv:(2.04)
2. RANK=MISD NATIONALITY=NAT SEX=MAN AGE=AGE2 92985 ==> CRIME=THEFT SOLVING=UNSOL 89811 conf:(0.97) < lift:(1.17)> lev:(0.1) [12944] conv:(5.08)
3. CRIME=THEFT NATIONALITY=NAT SEX=MAN 99047 ==> RANK=MISD SOLVING=UNSOL AGE=AGE2 89811 conf:(0.91) < lift:(1.16)> lev:(0.1) [12471] conv:(2.35)

4. RANK=MISD SOLVING=UNSOL AGE=AGE2 96536 ==> CRIME=THEFT NATIONALITY=NAT SEX=MAN 89811 conf:(0.93) < lift:(1.16)> lev:(0.1) [12471] conv:(2.85)
5. RANK=MISD SOLVING=UNSOL 106295 ==> CRIME=THEFT NATIONALITY=NAT SEX=MAN AGE=AGE2 89811 conf:(0.84) < lift:(1.16)> lev:(0.1) [12274] conv:(1.74)
6. CRIME=THEFT NATIONALITY=NAT SEX=MAN AGE=AGE2 90182 ==> RANK=MISD SOLVING=UNSOL 89811 conf:(1) < lift:(1.16)> lev:(0.1) [12274] conv:(33.99)
7. CRIME=THEFT SOLVING=UNSOL 102200 ==> RANK=MISD CONDITION=ENDED NATIONALITY=NAT SEX=MAN 90006 conf:(0.88) < lift:(1.16)> lev:(0.1) [12117] conv:(1.99)
8. RANK=MISD CONDITION=ENDED NATIONALITY=NAT SEX=MAN 94222 ==> CRIME=THEFT SOLVING=UNSOL 90006 conf:(0.96) < lift:(1.16)> lev:(0.1) [12117] conv:(3.87)
9. RANK=MISD SOLVING=UNSOL CONDITION=ENDED 97331 ==> CRIME=THEFT NATIONALITY=NAT SEX=MAN 90006 conf:(0.92) < lift:(1.15)> lev:(0.1) [12029] conv:(2.64)
10. CRIME=THEFT NATIONALITY=NAT SEX=MAN 99047 ==> RANK=MISD SOLVING=UNSOL CONDITION=ENDED 90006 conf:(0.91) < lift:(1.15)> lev:(0.1) [12029] conv:(2.33)

Παρατηρείται ότι στους 10 κανόνες γίνονται συνδυασμοί των τεσσάρων προηγούμενων τιμών χαρακτηριστικών (Πλημμέλημα, Ημεδαπός, Άνδρας, Ανεξιχνίαστο) με επιπλέον έντονη την παρουσία του εγκλήματος της κλοπής, της κατάστασης του εγκλήματος που είναι τετελεσμένο και της ηλικίας μεταξύ 18 έως 34 χρόνων. Οι νέες ρυθμίσεις επέφεραν πιο χαλαρούς κανόνες αλλά με υψηλά ποσοστά, εμπεριέχοντας τους πρώτους κανόνες με επιπρόσθετες τιμές χαρακτηριστικών. Ιδιαίτερα ξεχωρίζουν ο κανόνας 1 όπου όταν έχουμε ανεξιχνίαστη κλοπή τότε τείνει, αυξάνοντας τις πιθανότητες εμφανίσεις (lift:1.17), προς διάπραξη πλημμελήματος από ημεδαπό άνδρα ηλικίας 18 έως 34 χρόνων.

Στην περίπτωση αλλαγής των ρυθμίσεων του αλγορίθμου Apriori και την χαλάρωση των κανόνων επιδιώκεται η επιστροφή περισσότερων τιμών χαρακτηριστικών στα αποτελέσματα, ώστε να βοηθηθούν και επεκταθούν οι έρευνες της Αστυνομίας με

περισσότερα πιθανά στοιχεία. Τα αποτελέσματα έδωσαν επιπλέον δύο στοιχεία, εκτός των τεσσάρων προηγούμενων γνωστών που είναι το είδος εγκλήματος και η ηλικία. Για παράδειγμα να αναζητείται για την διάπραξη ανεξιχνίαστης κλοπής που τείνει να είναι πλημμέλημα, πιθανόν άνδρας δράστης ημεδαπός με ηλικία από 18 έως 34 χρόνων. Όσο περισσότερο χαλαροί (μειώνεται η ακρίβεια) είναι οι κανόνες αυξάνονται και οι πιθανότητες αποπροσανατολισμού των ερευνών.

Όταν ζητηθούν από τον αλγόριθμο 3 κανόνες με τις ίδιες ρυθμίσεις, (MetricType=Lift) τότε εξάγονται οι εξής:

1. RANK=MISD SOLVING=UNSOL 106295 ==> CRIME=THEFT SEX=MAN 101006 conf:(0.95) < lift:(1.13)> lev:(0.09) [11480] conv:(3.17)
2. CRIME=THEFT SEX=MAN 104126 ==> RANK=MISD SOLVING=UNSOL 101006 conf:(0.97) < lift:(1.13)> lev:(0.09) [11480] conv:(4.68)
3. RANK=MISD SOLVING=UNSOL 106295 ==> CRIME=THEFT NATIONALITY=NAT 99006 conf:(0.93) < lift:(1.12)> lev:(0.09) [10744] conv:(2.47)

Οι κανόνες που προκύπτουν από τον αλγόριθμο Apriori αλληλο-επιβεβαιώνονται και οι τιμές των χαρακτηριστικών που συναντιόνται έντονα είναι η Κλοπή, το Πλημμέλημα, ο Ημεδαπός, ο Άνδρας και το Ανεξιχνίαστο έγκλημα. Παρατηρείται μια επανάληψη συγκεκριμένων τιμών χαρακτηριστικών στους κανόνες, διότι η βάση δεδομένων είναι πάρα πολύ μεγάλη και υπάρχουν πολλές εγγραφές περιστατικών με ίδιες ή παρόμοιες τιμές χαρακτηριστικών, όπου κατά την κατηγοριοποίηση από τους αλγορίθμους και την διάσπαση σε υποσυστάδες υπερτερούν και εντοπίζονται ως σημαντικά έναντι των άλλων. Στην προκειμένη περίπτωση περιορίζονται οι κανόνες συσχέτισης, δίνοντας λιγότερα στοιχεία για αναζήτηση στην Αστυνομία τα οποία όμως ήδη υπήρχαν από τους προηγούμενους κανόνες που εξήχθησαν με υψηλά σκορ εμπιστοσύνης, άρα δεν βοηθούν περαιτέρω στις έρευνες της Αστυνομίας.

Ο αλγόριθμος PredictiveApriori εκτιμά την ¨προηγούμενη εμπιστοσύνη¨ από την διαθέσιμη βάση δεδομένων και επιστρέφει τους n κανόνες που μεγιστοποιούν την αναμενόμενη ακρίβεια πρόβλεψης. Η ευνοϊκή υπολογιστική απόδοση του αλγορίθμου PredictiveApriori μπορεί να πιστωθεί στη δυναμική τεχνική κλαδέματος, όπου χρησιμοποιεί ένα ανώτερο όριο στην ακρίβεια όλων των κανόνων, σε σχέση με τα υπερσύνολα ενός συγκεκριμένου στοιχειοσυνόλου και μπορούν να εξαιρεθούν πολύ μεγάλα τμήματα του χώρου αναζήτησης (Scheffer, 2004).

Στην συνέχεια αξιοσημείωτο είναι ότι ο αλγόριθμος PredictiveApriori μας δίνει τους 3 ακόλουθους διαφορετικούς κανόνες:

1. CRIME=SUIC AGE=AGE2 252 ==> RANK=NOTH 252 acc:(0.99498)
2. CRIME=FORG SEX=WOMAN 243 ==> SOLVING=SOL 243 acc:(0.99497)
3. CRIME=HOM 228 ==> SOLVING=SOL CONDITION=ENDED 228 acc:(0.99496)

Ο πρώτος κανόνας ως προς το δεύτερο σκέλος του είναι αυτονόητος και αδιάφορος αφού λέει το προφανές, ότι όταν γίνεται αυτοκτονία ηλικίας από 18 έως 34 τότε είναι μη ποινικά τιμωρητέα. Ο δεύτερος όμως κανόνας προβλέπει ότι όταν η πλαστογραφία γίνεται από γυναίκα τότε το έγκλημα εξιχνιάζεται, ενώ ο τρίτος ότι όταν διαπράττεται ανθρωποκτονία με πρόθεση τότε είναι τετελεσμένη και εξιχνιάζεται. Ο συγκεκριμένος αλγόριθμος αναζητά τη βέλτιστη ανταλλαγή μεταξύ εμπιστοσύνης και υποστήριξης, επιστρέφοντας κανόνες που μεγιστοποιούν την αναμενόμενη ακρίβεια πρόβλεψης. Τα αποτελέσματα του δεύτερου σκέλους των μόλις 3 κανόνων συσχέτισης δεν βοηθούν ιδιαίτερα την Αστυνομία στις αναζητήσεις της, αφού δεν παρέχουν αξιοποιήσιμες για το έργο της πληροφορίες.

Σύμφωνα με την συσταδιοποίηση και τον αλγόριθμο k-means στα αποτελέσματα υπερτερεί με 91% σωστών κατηγοριοποιημένων περιπτώσεων, η συστάδα που είναι το Cluster 0: THEFT, MISD, UNSOL, ENDED, NAT, MAN, AGE2, NOT-KN. Οι τιμές αυτές επιβεβαιώνουν τα προηγούμενα αποτελέσματα των άλλων κατηγοριοποιητών (πλην του PredictiveApriori, όπου όπως εξηγήθηκε λειτουργεί με δυναμική τεχνική κλαδέματος και εξαιρούνται πολύ μεγάλα τμήματα του χώρου αναζήτησης). Από τα αποτελέσματα του k-means αλγορίθμου η Αστυνομία μπορεί να αντλήσει στοιχεία των πιο συχνών εγκλημάτων με συγκεκριμένα χαρακτηριστικά, ώστε οι υπεύθυνοι σχεδιασμού πρόληψης και καταστολής του εγκλήματος να πάρουν τα ανάλογα προληπτικά μέτρα. Οι ομάδες με σχετικά παρόμοιο επαναλαμβανόμενο μοτίβο μπορούν να αντιμετωπιστούν ενιαία. Για παράδειγμα στα αποτελέσματα παρατηρούνται ανεξιχνίαστες τελεσμένες κλοπές, πλημμελληματικού βαθμού, ημεδαπών ανδρών, ηλικίας 17-34 χρόνων, όπου ο λόγος τέλεσης είναι άγνωστος. Η Αστυνομία μπορεί να αυξήσει τις πεζές περιπολίες, να μεριμνήσει για τον κατάλληλο φωτισμό των περιοχών, να τοποθετήσει κάμερες και να εστιάσει σε αναζητήσεις με συγκεκριμένα χαρακτηριστικά.

Στην οπτικοποίηση των αποτελεσμάτων της συσταδιοποίησης σε σχέση με το έγκλημα και την ηλικία (σχήμα 4-20) παρατηρείται ότι τα περιστατικά των αυτοκτονιών μοιράζονται σε όλες τις κατηγορίες ηλικιών σχεδόν ισάριθμα με μικρή αυξητική

διαφορά στις ηλικίες 34-59. Ακόμη, διαπιστώνεται ότι στις ηλικίες 18-34 χρόνων υπάρχει αυξημένος αριθμός κλοπών, απάτης και ληστειών, ενώ στις ηλικίες 34-59 χρόνων υπάρχει έντονη παρουσία όλων των εγκλημάτων. Στην οπτικοποίηση σε σχέση με την εθνικότητα και τον αριθμό περιστατικών (σχήμα 4-21), φαίνεται η υπεροχή των ημεδαπών στις κλοπές και έπειτα στις πλαστογραφίες, ενώ των αλλοδαπών στις κλοπές και πλαστογραφίες με λιγότερες περιπτώσεις στις διακεκριμένες κλοπές και ληστείες. Ακόμη, στην οπτικοποίηση των εγκλημάτων (σχήμα 4-22) σε σχέση με τον αριθμό περιστατικών και την ηλικία, φαίνεται η υπεροχή της ηλικίας μεταξύ 18-34 στις κλοπές και κατόπιν στις απάτες, η ηλικία από 7-17 υπερτερεί σε κλοπές και ληστείες, ενώ στις ηλικίες 35 και έπειτα υπάρχει ποικιλία. Τέλος, στο σχήμα 4-23 φαίνεται η μεγάλη διαφορά στα εγκλήματα σε σχέση με τον αριθμό περιπτώσεων στις κλοπές και έπειτα στις ληστείες και απάτες. Τα αποτελέσματα αυτά μπορούν να αξιοποιηθούν από την Αστυνομία ώστε να αντιδράσει στοχευμένα και αποτελεσματικά. Για παράδειγμα να δοθεί προσοχή στις μικρές ηλικίες, σχετικά με την επιμόρφωσή τους στα σχολεία, μέσω σεμιναρίων για την πρόληψη αυτοκτονιών και παραβατικών πράξεων που αφορούν κλοπές και ληστείες και για τις ηλικίες από 18-34 που τελούν σε μεγάλο βαθμό οικονομικής φύσεως εγκλήματα (κλοπές, απάτες, ληστείες) να σχεδιαστούν τα ανάλογα μέτρα αντιμετώπισης και να ενημερωθούν οι αρμόδιες αρχές χάραξης πολιτικής της χώρας για αντίστοιχα μέτρα.

Από τις δοκιμές που έγιναν στην έρευνα με διάφορους αλγορίθμους κατηγοριοποίησης, κανόνων συσχέτισης και συσταδιοποίησης, παρατηρήθηκε ότι ο MultilayerPerceptron χρειάστηκε πολύ περισσότερο χρόνο έναντι των άλλων. Ένα από τα κριτήρια επιλογής σωστού αλγορίθμου είναι το υπολογιστικό κόστος, όπου σε πολύ μεγάλες βάσεις δεδομένων με ποσοτικά συνεχή αριθμητικά δεδομένα που χρειάζονται πολύπλοκους υπολογισμούς, είναι πιθανό η διαδικασία να διαρκέσει αρκετό χρόνο. Από τον συγκεντρωτικό πίνακα 5-1 φαίνεται ότι τα αποτελέσματα του κατηγοριοποιητή MultilayerPerceptron και με τις δύο τεχνικές (Cross-validation Folds=10 και Percentage split 66%) και του J48 με Cross-validation Folds=10 δεν διαφέρουν ιδιαίτερα, δίνοντας σχεδόν τα ίδια αποτελέσματα. Η μόνη μεγάλη διαφορά είναι ο χρόνος επεξεργασίας των δεδομένων που οφείλεται στον διαφορετικό τρόπο λειτουργίας τους.

6 Επίλογος

6.1 Σύνοψη και Συμπεράσματα

6.1.1 Σύνοψη

Η παρούσα διπλωματική εργασία έχει ως βασικό σκοπό την εξόρυξη άγνωστης, μέχρι πρότινος, γνώσης και πληροφορίας από μεγάλο όγκο δεδομένων που αφορούν εγκλήματα και τα χαρακτηριστικά τους. Η βάση δεδομένων των εγκλημάτων, τα οποία διαπράχθηκαν το 2016 στην Ελλάδα, αντλήθηκε κυρίως από πηγές της Αστυνομίας και περιλαμβάνει 123.631 περιστατικά εγκλημάτων (ανθρωποκτονία από αμέλεια, ανθρωποκτονία με πρόθεση, ληστεία, κλοπή, διακεκριμένη κλοπή, απάτη, παραχάραξη, πλαστογραφία και αυτοκτονία) από τα οποία το κάθε ένα έχει 8 χαρακτηριστικά (έγκλημα βαθμός εξιχνίαση κατάσταση εθνικότητα φύλο ηλικία και αιτία).

Επισημαίνεται η αξία απεικόνισης των εγκλημάτων σε γεωγραφικό σύστημα πληροφοριών και η άντληση γνώσης από την ταυτόχρονη χρήση τεχνικών εξόρυξης δεδομένων. Αφού γίνει εννοιολογική οριοθέτηση της εγκληματικότητας, διαχωρίζονται τα εγκλήματα σε κατηγορίες και επιλέγονται τα κατάλληλα προς έρευνα, καθώς και τα χαρακτηριστικά τους. Παρουσιάζονται οι εργασίες και τεχνικές της εξόρυξης δεδομένων και περιγράφεται η προετοιμασία και δημιουργία της βάσης δεδομένων στο πλαίσιο διεξαγωγής της έρευνας. Το λογισμικό εξόρυξης δεδομένων που χρησιμοποιείτε στη παρούσα εργασία είναι το ¨WEKA¨ και το γεωγραφικό σύστημα πληροφοριών το ¨CRIMEVIEW¨ (ArcGIS) της ESRI. Στην συνέχεια περιγράφεται ο τρόπος χρήσης τους, επιλέγονται οι κατάλληλες τεχνικές και μέθοδοι εξόρυξης δεδομένων και διεξάγεται εκπαίδευση μοντέλου πρόβλεψης εγκλήματος και των χαρακτηριστικών του. Τέλος, συγκρίνονται τα εξαγόμενα αποτελέσματα και αξιολογούνται για την χρησιμότητά τους στο έργο της αστυνομίας εναντίον της εγκληματικότητας.

6.1.2 Γενικά Συμπεράσματα

Η εγκληματικότητα αποτελεί κοινωνικό φαινόμενο που μαστίζει όλη την ανθρωπότητα και είναι αναπόσπαστο στοιχείο της εκάστοτε κοινωνίας. Η ανάπτυξη της

εγκληματολογικής επιστήμης που χωρίζεται σε υποκατηγορίες (Εγκληματολογία, Θυματολογία, Ανακριτική), μελετά την υπόσταση του εγκλήματος και στο έργο της συμβάλλουν επικουρικά και άλλες επιστήμες (Δικαστική Ψυχολογία, Κοινωνιολογία, Γραφολογία, Ιατροδικαστική). Οι αιτίες διάπραξης εγκλήματος ποικίλουν (οικονομικές, οικογενειακές, ηθικές, ψυχολογικές κτλ.) με βασικά κίνητρα τα οικονομικά, τα συναισθηματικά και τα ηθικά.

Οι κανόνες δικαίου καθορίζουν την συμπεριφορά των ανθρώπινων σχέσεων και χωρίζονται σε δημόσιους και ιδιωτικούς. Το Ουσιαστικό Ποινικό Δίκαιο καθορίζει ποια είναι τα εγκλήματα και αποτελεί κλάδο του Δημοσίου Δικαίου. Τα εγκλήματα κατηγοριοποιούνται σε ομάδες και ανάλογα με τη βαρύτητά τους, διαχωρίζονται σε κακουργήματα, πλημμελήματα και πταίσματα.

Η τεχνολογική εξέλιξη έχει συμβάλει στην ύπαρξη πληθώρας δεδομένων, όπου πλέον η συλλογή, αποθήκευση και ανάλυσή τους είναι εύκολη υπόθεση με την χρήση εξελιγμένων τεχνικών και εργαλείων. Σύγχρονοι αλγόριθμοι επικουρούν στην αποτελεσματικότερη επεξεργασία δεδομένων, παρέχοντας πολύτιμη γνώση. Η εξόρυξη δεδομένων βρίσκει εφαρμογή σε πολλούς τομείς (διαχείριση αγοράς, ρίσκου, απάτης, text mining κτλ.), σε εμπορικούς και επιστημονικούς χώρους, καθώς και σε διάφορους φορείς και οργανισμούς του κράτους (Αστυνομία, Λιμενικό, Περιφέρειες κτλ.). Οι βασικές εργασίες εξόρυξης δεδομένων είναι η κατηγοριοποίηση, η συσταδιοποίηση, οι κανόνες συσχέτισης, η παλινδρόμηση, η ανίχνευση ανωμαλιών, η ανάλυση χρονολογικών σειρών, τα πρότυπα ακολουθιών και η μείωση των διαστάσεων.

Τα δεδομένα τα οποία δίνουν πληροφορίες σχετικά με τη δική τους θέση (γεωγραφικό πλάτος και μήκος, διεύθυνση, διαμέριση κατά θέση) ονομάζονται χωρικά και οι βάσεις αυτών μπορούν να περιέχουν και μη χωρικές πληροφορίες. Εκτός των τεχνικών εξόρυξης δεδομένων, εκ των οποίων κάποια χρησιμοποιούνται και στην εξόρυξη χωρικών δεδομένων, δημιουργήθηκαν καινούργιες τεχνικές καθαρά για επεξεργασία χωρικών δεδομένων, καθώς και νέοι αλγόριθμοι. Τα γεωγραφικά συστήματα πληροφοριών χρησιμοποιούν τα χωρικά δεδομένα για παροχή πληροφοριών σχετικών με την γεωγραφική θέση των στοιχείων. Η εφαρμογή των γεωγραφικών συστημάτων πληροφοριών σε διάφορους κλάδους (π.χ. Πολεοδομία, περιβαλλοντικές και στατιστικές υπηρεσίες) έδωσε νέα θετική ώθηση στην αποτελεσματικότητα και αποδοτικότητα του έργου τους. Ο συνδυασμός των τεχνικών εξόρυξης δεδομένων με τη χρήση γεωγραφικών συστημάτων πληροφοριών παρέχει την ευκαιρία στους αναλυτές να

επιτύχουν εξόρυξη ακριβέστερης, πιο αξιόπιστης και εμπλουτισμένης γνώσης και πληροφορίας.

6.1.3 Συμπεράσματα έρευνας

Στην έρευνα της παρούσας διπλωματικής εργασίας έγινε προσπάθεια εξόρυξης γνώσης και πληροφορίας από μία μεγάλου όγκου βάση δεδομένων, αποτελούμενη από συγκεκριμένα εγκλήματα και τα χαρακτηριστικά τους. Χρησιμοποιήθηκε το λογισμικό ¨WEKA¨ στην εξόρυξη δεδομένων και δοκιμάστηκαν διάφοροι αλγόριθμοι, κάποιοι εκ των οποίων παρήγαγαν σημαντική γνώση για τα εγκλήματα και χρηστική για την διεξαγωγή του έργου της Αστυνομίας. Ακόμη, κάποιοι άλλοι αλγόριθμοι εξόρυξαν κανόνες συσχέτισης, οι οποίοι δεν παρείχαν επιπλέον πληροφορία ή πολύτιμη γνώση για να χρησιμοποιηθεί από την Αστυνομία.

Σύμφωνα με τα αποτελέσματα του κατηγοριοποιητή J48 παρατηρήθηκε υψηλό ποσοστό σωστών κατηγοριοποιημένων περιπτώσεων (95,395% και 91,277%), όπου αναδεικνύει την αποτελεσματικότητα και ακρίβεια της κατηγοριοποίησης. Τα λάθη στο δέντρο απόφασης των κατηγοριοποιητών είναι ελάχιστα και προτιμάται η πρώτη δοκιμή (Kappa statistic = 0.7651), όπου έχει καλύτερα αποτελέσματα δεικτών απ’ ότι η δεύτερη, με υψηλές τιμές των δεικτών TP Rate και ROC Area. Υπάρχει μεγάλη συμφωνία των κατηγοριοποιήσεων με τις πραγματικές κλάσεις και ξεχωρίζουν για το ποσοστό πραγματικών θετικών περιπτώσεων που έχουν ταξινομηθεί σωστά ως δεδομένη κλάση η Αυτοκτονία (SUIC=1,000), η Κλοπή (THEFT=0,997), η Διακεκριμένη κλοπή (DIS-TH=0,817), η Ληστεία (ROB=0,899) και η Πλαστογραφία (FORG=0,671).

Το ποσοστό σωστών κατηγοριοποιημένων περιπτώσεων με όλους τους κατηγοριοποιητές που χρησιμοποιήθηκαν είναι πάνω από 91%, δηλαδή πολύ ικανοποιητικός όπως και η τιμή των δεικτών ROC Area, ¨Precision¨ και ¨Recall¨ για τις τιμές ¨SUIC¨ και ¨THEFT¨ της κλάσης ¨Crime. Ακόμη διαπιστώθηκε ότι ο αλγόριθμος MultilayerPerceptron χρειάστηκε πολύ περισσότερο χρόνο έναντι των άλλων και πως τα αποτελέσματά του είναι σχεδόν ίδια με τον J48 στην πρώτη δοκιμή, όπου δεν αλλάχθηκαν οι ρυθμίσεις του αλγορίθμου. Επομένως, ο MultilayerPerceptron δεν ενδείκνυται για την μεγάλη βάση δεδομένων που χρησιμοποιήθηκε, καθαρά για λόγους υπολογιστικού κόστους, αφού προτιμάται ο J48 με τον οποίο η επεξεργασία διήρκησε πολύ λίγο χρόνο και πάρθηκαν σχεδόν τα ίδια αποτελέσματα. Τόσο από τις τιμές του

MultilayerPerceptron (και στις δύο δοκιμές) όσο και από του J48 (στην πρώτη δοκιμή) διαπιστώθηκε ότι το ποσοστό των θετικών προβλέψεων που είναι πραγματικά θετικές (Precision) και το ποσοστό των περιπτώσεων που πραγματικά είναι θετικές και είχαν προβλεφθεί θετικά (Recall), αλλά και ο δείκτης ROC Area είναι οι πιο αποτελεσματικοί κατηγοριοποιητές σε σχέση με τους υπολοίπους. Θα πρέπει να σημειωθεί πως για τις τιμές των ανθρωποκτονιών, των απατών και της παραχάραξης η κατηγοριοποίηση δεν ήταν αποτελεσματική, πιθανόν λόγω της ανισοκατανομής των κλάσεων και χρήζει περαιτέρω έρευνας, αφού δοκιμές με τεχνικές ¨ClassBalancer¨, ¨SMOTE¨ και ¨Resample¨ δεν άλλαξαν τα αποτελέσματα.

Παρότι τα αποτελέσματα των αλγορίθμων της κατηγοριοποίησης μπορεί να μην ήταν ακριβή για συγκεκριμένες τιμές, μπορούν να εξαχθούν χρήσιμα συμπεράσματα για την επίλυση του προβλήματος εντοπισμού εγκλημάτων που χρήζουν ιδιαίτερης προσοχής και ταλανίζουν την εύρυθμη λειτουργία της κοινωνίας. Η Αστυνομία μπορεί να χρησιμοποιήσει την εξαγόμενη γνώση για την αποτελεσματικότερη αντιμετώπιση των συγκεκριμένων εγκλημάτων που στην προκειμένη περίπτωση, βάσει αποτελεσμάτων είναι οι κλοπές, οι διακεκριμένες κλοπές, οι ληστείες, οι πλαστογραφίες και οι αυτοκτονίες.

Κατά την επεξεργασία των δεδομένων από τον κατηγοριοποιητή NaïveBayes δημιουργείται συγκεντρωτικός πίνακας με στατιστικά στοιχεία των εγκλημάτων και των χαρακτηριστικών τους, όπου συμπεραίνεται ότι οι ανεξιχνίαστες τελεσμένες πλημμελληματικές κλοπές, πλημμελληματικές ληστείες και κακουργηματικές απάτες ημεδαπών ανδρών δραστών με άγνωστους λόγους δράσης και ηλικίας από 18 έως 34 χρόνων, υπερτερούν έναντι των άλλων εγκλημάτων για το έτος 2016 στην Ελλάδα. Ακολουθούν οι εξιχνιασμένες τελεσμένες κακουργηματικές διακεκριμένες κλοπές αλλοδαπών και ημεδαπών (σχεδόν στο ίδιο ποσοστό) ανδρών δραστών με άγνωστους λόγους δράσης και ηλικίας από 35 έως 59 χρόνων. Κατόπιν, είναι οι εξιχνιασμένες τελεσμένες πλημμελληματικές πλαστογραφίες αλλοδαπών ανδρών δραστών με άγνωστους λόγους δράσης και ηλικίας από 35 έως 59 χρόνων και μετά οι απόπειρες αυτοκτονιών ανδρών ηλικίας από 35 έως 59 χρόνων. Επίσης, αξιοσημείωτο είναι ότι μετά από τους άγνωστους λόγους αυτοκτονιών (απόπειρες και τελεσμένες), αυτές τελούνται με κίνητρα πρωτίστως ασθένειας, αισθηματικά και οικογενειακά στο ίδιο ποσοστό και τελευταία οικονομικά. Οι συγκεκριμένες παρατηρήσεις μπορούν να ωφελήσουν την Αστυνομία για περαιτέρω προβληματισμό και έρευνα κατανόησης των πτυχών της εγκληματικότητας.

Στην χρήση των αλγορίθμων Apriori και PredictiveApriori για εξαγωγή κανόνων συσχέτισης επιδιώκεται η εξαγωγή γνώσης και χρήσιμων συμπερασμάτων και αναζητούνται οι όποιοι πιθανοί συνδυασμοί χαρακτηριστικών. Οι κανόνες έχουν πάρα πολύ υψηλό Confidence, ίσο ή μεγαλύτερο του 0,95. Οι πληροφορίες που αντλούνται από τον αλγόριθμο Apriori δίνουν ότι όταν το έγκλημα είναι κλοπή τότε είναι πλημμέλημα, όταν τα ανεξιχνίαστα εγκλήματα συνδυάζονται με ημεδαπούς δράστες τότε είναι άνδρες, το πλήθος των ημεδαπών δραστών διαπράττει πλημμελήματα. Παρατηρείται ότι υπάρχει επανάληψη των τιμών ¨Πλημμέλημα¨, ¨Ημεδαπός¨, ¨Άνδρας¨ και ¨Ανεξιχνίαστο έγκλημα¨. Εξάγονται πληροφορίες συσχετίσεων μεταξύ των εγκληματικών χαρακτηριστικών, όπου υπάρχουν και δεν είναι εμφανείς εξαρχής για την Αστυνομία και είναι χρήσιμες για τις έρευνες, ώστε να τις κατευθύνουν σε συγκεκριμένους στόχους, επιτυγχάνοντας γρηγορότερη και ευκολότερη εξιχνίαση του εγκλήματος και εντοπισμού και σύλληψης του δράστη. Επομένως, σύμφωνα με τα αποτελέσματα των κανόνων συσχέτισης του αλγορίθμου Apriori, αναζητείται για την διάπραξη ανεξιχνίαστου πλημμελήματος πιθανόν άνδρας δράστης ημεδαπός. Ανάλογα με τις ρυθμίσεις του αλγορίθμου (¨MetricType¨ το ¨Lift¨) επιτυγχάνονται πιο χαλαροί κανόνες, όπου προστίθενται τιμές χαρακτηριστικών (αν είναι ανεξιχνίαστη κλοπή τότε τείνει, αυξάνοντας τις πιθανότητες εμφανίσεις προς διάπραξη πλημμελήματος από ημεδαπό άνδρα ηλικίας 18 έως 34 χρόνων). Στην περίπτωση αυτή, επιδιώκεται η επιστροφή περισσότερων τιμών χαρακτηριστικών στα αποτελέσματα, ώστε να βοηθηθούν και επεκταθούν οι έρευνες της Αστυνομίας με περισσότερα πιθανά στοιχεία, όμως μειώνεται η ακρίβεια των αποτελεσμάτων. Αντιθέτως, μπορούν να περιοριστούν τα αποτελέσματα κανόνων, δίνοντας λιγότερα στοιχεία για αναζήτηση στην Αστυνομία με την πιθανότητα αυτά να μην φέρουν τα επιθυμητά αποτελέσματα ή να είναι γνωστά και να μην βοηθούν στις έρευνες της Αστυνομίας.

Τα αποτελέσματα του αλγόριθμου PredictiveApriori, ο οποίος ρυθμίστηκε να δώσει μόλις 3 κανόνες, δεν παρέχουν χρήσιμες και αξιοποιήσιμες πληροφορίες για τις έρευνες της Αστυνομίας, αφού προβλέπεται ότι οι αυτοκτονίες δεν τιμωρούνται, ότι όταν η πλαστογραφία γίνεται από γυναίκα τότε το έγκλημα εξιχνιάζεται και όταν διαπράττεται ανθρωποκτονία με πρόθεση τότε είναι τετελεσμένη και εξιχνιάζεται.

Κατά την συσταδιοποίηση και την χρήση του αλγορίθμου k-means εξάγονται συμπεράσματα με 91% σωστών κατηγοριοποιημένων περιπτώσεων με ανεξιχνίαστες τελεσμένες κλοπές, πλημμελληματικού βαθμού, ημεδαπών ανδρών, ηλικίας 17-34 χρόνων, όπου ο λόγος τέλεσης είναι άγνωστος. Οι συστάδες με σχετικά παρόμοιο

επαναλαμβανόμενο μοτίβο μπορούν να αντιμετωπιστούν ενιαία, αφού παρουσιάζουν συγκεκριμένα ίδια χαρακτηριστικά και η Αστυνομία να πάρει ανάλογα μέτρα καταστολής και πρόληψης (αύξηση πεζών περιπολιών, φωτισμός περιοχών, τοποθέτηση καμερών κτλ.). Ακόμη, από την οπτικοποίηση των αποτελεσμάτων του αλγορίθμου k-means με τις επιθυμητές ρυθμίσεις τιμών των αξόνων του διαγράμματος, βγαίνουν σημαντικά συμπεράσματα, όπως προαναφέρθηκαν στο 5^{O} κεφάλαιο και μπορούν να αξιοποιηθούν από την Αστυνομία, ώστε να αντιδράσει στοχευμένα και αποτελεσματικά στην πάταξη του εγκλήματος. Για παράδειγμα στις μικρές ηλικίες που παρατηρήθηκαν παραβατικές πράξεις που αφορούν κλοπές και ληστείες, αλλά και αυτοκτονίες, μπορούν να γίνουν επιμορφωτικά σεμινάρια στα σχολεία και να ενημερωθούν οι αρμόδιες πολιτικές αρχές για ανάλογα μέτρα αντιμετώπισης.

Τα αποτελέσματα από την χρήση των τεχνικών εξόρυξης δεδομένων εγκλημάτων αποτελούν πεδίο προβληματισμού και έρευνας για τις αρμόδιες αρχές. Οι αλγόριθμοι κατηγοριοποίησης χρησιμοποιήθηκαν για δημιουργία μοντέλου προβλέψεων, ώστε να βοηθηθεί η Αστυνομία στην επίλυση συγκεκριμένου προβλήματος και να λάβει μέτρα δράσης αντιμετώπισής του. Η συσταδιοποίηση και οι κανόνες συσχέτισης παρείχαν κρυμμένη γνώση για τα εγκλήματα και τα χαρακτηριστικά τους στην Αστυνομία, ώστε να εξαχθούν χρήσιμα συμπεράσματα που θα συμβάλουν επικουρικά στο έργο της για τον στρατηγικό σχεδιασμό αντιμετώπισης της εγκληματικότητας, την πρόληψη, την καταστολή και εξιχνίαση των εγκλημάτων.

6.2 Όρια και περιορισμοί της έρευνας

Σημαντικός παράγοντας που επηρέασε την υλοποίηση της παρούσας διπλωματικής εργασίας είναι η ύπαρξη ορίων και περιορισμών. Ο μεγάλος όγκος, όχι όμως και εύρος, δεδομένων εγκλημάτων και χαρακτηριστικών τους, καθιστά δύσκολη την συλλογή, αξιολόγηση και χρήση των πιο κατάλληλων για την δημιουργία της βάσης δεδομένων. Η συνεχής βελτίωση και εμφάνιση νέων αλγορίθμων στις εφαρμογές τεχνολογίας εξόρυξης δεδομένων, καθώς και καινούργιων τεχνικών, αλλά και η πληθώρα ρυθμίσεων και εμφάνιση νέων, βάζει όρια και καθιστά δύσκολο το έργο των σωστών επιλογών για δημιουργία μοντέλου πρόβλεψης και εξόρυξης γνώσης. Επίσης, η εφαρμογή ¨WEKA¨ που χρησιμοποιήθηκε για την εξόρυξη δεδομένων δεν υποστηρίζει την ελληνική γλώσσα, με συνέπεια να υπάρχουν προβλήματα κωδικοποίησης των

δεδομένων και κατά συνέπεια δυνατότητας ανάγνωσής τους. Περιοριστικός ερευνητικός παράγοντας, αποτέλεσε και η ανεύρεση εύχρηστου και αξιόπιστου, άνευ πληρωμής, γεωγραφικού συστήματος πληροφοριών για περαιτέρω χωρική ανάλυση των εγκλημάτων.

6.3 Μελλοντικές επεκτάσεις

Μελλοντική έρευνα της παρούσας διπλωματικής εργασίας μπορεί να αποτελέσει η χρήση νέων αλγορίθμων ή οι βελτιώσεις ήδη υπαρχόντων, καθώς και τεχνικών για την δημιουργία του μοντέλου πρόβλεψης και της εξόρυξης γνώσης. Ακόμη, η επιλογή διαφορετικών ή περισσοτέρων χαρακτηριστικών στην βάση δεδομένων μπορεί να επιφέρει διαφορετικά αποτελέσματα για σύγκριση με τα υπάρχοντα και εκτίμηση της χρησιμότητας τους και της επίδρασής τους στην πρόβλεψη και την εξόρυξη γνώσης. Τέλος, η ολοκληρωμένη και χαρτογραφημένη απεικόνιση του εγκλήματος από την Αστυνομία σε συνδυασμό με τα αποτελέσματα της εξόρυξης δεδομένων εγκλημάτων θα επιφέρει ακριβέστερη και πλουσιότερη γνώση.

Βιβλιογραφία

Βαζιργιάννης, Μ., Χαλκίδη, Μ., (2005). *Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό.* Αθήνα: Τυποθήτω-ΓΙΩΡΓΟΣ ΔΑΡΔΑΝΟΣ.

Κύρκος, Ε., (2015). *Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων*. [e-book]. Εκδόσεις Κάλλιπος. Διαθέσιμο: Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα και Βοηθήματα. [online] Available at: <https://repository.kallipos.gr/handle/11419/1226> [Προσπελάσιμο 08 Μαΐου 2017].

Κωστάρας, Α., (2001). *Έννοιες και Θεσμοί του Ποινικού Κώδικα.* Αθήνα: Εκδόσεις Αντ. Ν. Σάκκουλα.

Ματσατσίνης, Ν., (2010). *Συστήματα Υποστήριξης Αποφάσεων*. Αθήνα: Εκδόσεις Νέων Τεχνολογιών.

Νανόπουλος, Α., Μανωλόπουλος, Ι., (2008). *Εισαγωγή στην Εξόρυξη και τις Αποθήκες Δεδομένων.* Αθήνα: Εκδόσεις Νέων Τεχνολογιών.

Ραφτόπουλος, Π., (1996). *Ποινικό Δίκαιο.* Αθήνα: Εκδόσεις Παναγιώτης Δ. Ραφτόπουλος.

Ansari, S., Kale, K., (2014). *Mapping and Analysis of Crime in Aurangabad City using GIS.* [pdf] Journal of Computer Engineering. [online] Available at: <http://www.iosrjournals.org/iosr-jce/papers/Vol16-issue4/Version-7/L016476776.pdf> [Accessed 10 May 2017].

Becker, G., (1968). *Crime and Punishment: An Economic Approach.* [pdf] The Journal of Political Economy (Mar. - Apr. 1968). [online] Available at: <https://olis.leg.state.or.us/liz/2017R1/Downloads/CommitteeMeetingDocument/125036> [Accessed 20 April 2017].

Bhargava, N., Sharma, G., Bhargava, R., Mathuria, M., (2013). *Decision Tree Analysis on J48 Algorithm for Data Mining.* [pdf] International Journal of Advanced Research in Computer Science and Software Engineering, 2013. [online] Available at: <https://scholar.google.gr/scholar?q=Decision+Tree+Analysis+on+J48+Algorithm+for+Data+Mining&hl=el&as_sdt=0&as_vis=1&oi=scholart&sa=X&ved=

0ahUKEwi6rMXAxq7UAhWGOxQKHXUWDRMQgQMIIjAA> [Accessed 06 April 2017].

Chen, H., Chung, W., Qin, Y., Chau, M., Xu, J. J., Wang, G., Zheng, R., Atabakhsh, H., (2003). *Crime Data Mining, An Overview and Case Studies.* [pdf] annual national conference on Digital government research.2003, North America [online] Available at: <https://pdfs.semanticscholar.org/51c8/5213831025a6b3cb4a9122e35ade1cf5945a.pdf> [Accessed 22 May 2017].

Cs.usfca.edu, (2015). *Weka Data Analysis*. [pdf] University of San Francisco [online] Available at: <http://www.cs.usfca.edu/~pfrancislyon/courses/640fall2015/WekaDataAnalysis.pdf> [Accessed 04 April 2017].

Dunham, M., (2004). *Data Mining-Εισαγωγικά και Προηγμένα Θέματα Εξόρυξης Γνώσης από Δεδομένα.* Επιμέλεια: Βερύκιος, Β., Θεοδωρίδης, Γ., Αθήνα: Εκδόσεις Νέων Τεχνολογιών.

Data.gov.gr, (2016). *Στατιστική Επετηρίδα*. [pdf] Στατιστική επετηρίδα Ελληνικής Αστυνομίας, 2016. [online] Available at: <http://data.gov.gr/dataset/statistikh-epethrida> [Accessed 02 April 2017].

Goadrich, M., Davis, J., (2006). *The Relationship Between Precision-Recall and ROC Curves*. [pdf] 23rd international conference on Machine learning, June 2006, Pittsburgh, Pennsylvania, USA. [online] Available at: <https://www.biostat.wisc.edu/~page/rocpr.pdf> [Accessed 07 April 2017].

Johnson, C.P., (2000). *Crime Mapping and Analysis Using GIS.* [pdf] Conference on Geomatics in Electronic Governance, January 2000, Pune [online] Available at: <http://fac.ksu.edu.sa/sites/default/files/crim_mapping.pdf> [Accessed 10 May 2017].

Kumar, M., (2012). *Evaluating the performance of apriori and predictive apriori algorithm to find new association rules based on the statistical measures of datasets.* [pdf] International Journal of Engineering Research & Technology, Aug 2012. [online] Available at: <https://www.researchgate.net/publication/262979804_Evaluating_the_performance_of_apriori_and_predictive_apriori_algorithm_to_find_new_association_rules_based_on_the_statistical_measures_of_datasets> [Accessed 10 April 2017].

Leu, F.-Y., Wang, T.-H., (2006). *Data analysis using GIS and data mining.* [pdf] International Conference of Territorial Intelligence, Sep 2006, Alba Iulia, Romania. [online] Available at: <https://hal.archives-ouvertes.fr/halshs-00516476/> [Accessed 09 May 2017].

Ojiako, J.C., Okafor, C. K., Igbokwe, E.C., Enedah, I.C., (2016). *Modeling of Crime Pattern in Anaocha L.G.A, Anambra State, Nigeria Using GIS Approach.* [pdf] International Journal of Innovative Research in Engineering & Management, Sep 2016, [online] Available at: <http://www.ijirem.org/DOC/4_%20IREM388e82939b1-6973-4d27-8fcb-ff76ea9c85c9.pdf> [Accessed 28 April 2017].

Opengov.gr, (2017). *Έξυπνη Αστυνόμευση – Smart Policing.* [pdf] Υπουργείο Διοικητικής Ανασυγκρότησης-Εθνικό Κέντρο Δημόσιας Διοίκησης και Αυτοδιοίκησης-Μονάδα Τεκμηρίωσης και Καινοτομιών. [online] Available at: <http://www.opengov.gr/ypes/wp-content/uploads/downloads/2017/02/smart-policing.pdf> [Accessed 18 April 2017].

Patil, T.R., Sherekar, S.S., (2013). *Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification*. [pdf] International Journal Of Computer Science And Applications, Apr 2013. [online] Available at: <http://keddiyan.com/files/AHCI/week2/9.pdf> [Accessed 03 April 2017].

Phillips, P., Lee, I., (2012). *Mining co-distribution patterns for large crime datasets.* [pdf] Journal Expert Systems with Applications: An International Journal Oct 2012. [online] Available at: <http://www.sciencedirect.com/science/article/pii/S0957417412005945> [Accessed 26 April 2017].

Saputra, B.T., Yang, F.Y., (2015). *Improving ids alerts using predictive Apriori algorithm.* [doc] International Multi-Conference on Engineering and Technology Innovation 2015, Oct-Nov 2015, Kaohsiung, Taiwan. [online] Available at: <https://www.google.gr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0ahUKEwj2mP-i7rbUAhWCvxQKHfbECqYQFgguMAI&url=http%3A%2F%2Fir.lib.cyut.edu.tw%3A8080%2Fbitstream%2F310901800%2F26053%2F2%2F2015-11-09-Manuscript%2520No.F5022_Improving%2520IDS%2520Alert%2520using%

2520Predictive%2520Apriori.docx&usg=AFQjCNHCyjgd7Oazo8ev_GSOa6O006fiVA> [Accessed 09 April 2017].

Scheffer, T., (2001). *Finding Association Rules that Trade Support Optimally Against Confidence.* [pdf] In: 5th European Conference on Principles of Data Mining and Knowledge Discovery, 424-435, 2001. [online] Available at: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.25.4880> [Accessed 15 June 2017].

Tan, P., Steinbach, M., Kumar, V., (2015). *Εισαγωγή στην Εξόρυξη Δεδομένων.* Επιμέλεια: Βερύκιος, Β., Μετάφραση: Σουραβλάς, Σ., Αθήνα: Εκδόσεις ΤΖΙΟΛΑ.

Witten, I., Frank, E., (2005). *Data Mining - Practical Machine Learning Tools and Techniques.* USA: Elsevier.

ΙΣΤΟΣΕΛΙΔΕΣ

Σπυρόπουλος, Π., (2017). *Crime View, ένα χρήσιμο GPS για την πάταξη της εγκληματικότητας.* Thetoc [online] Available at: <http://www.thetoc.gr/koinwnia/article/crime-view-ena-xrisimo-gps-gia-tin-pataksi-tis-egklimatikotitas> [Accessed 19 April 2017].

Astynomia.gr, (2016). *Στατιστικά Στοιχεία.* [online] Available at: <http://www.astynomia.gr/index.php?option=ozo_content&perform=view&id=81&Itemid=73> [Accessed 05 April 2017].

Astynomia.gr, (2017). *Στατιστικά στοιχεία - απολογισμός συνολικής δραστηριότητας της Ελληνικής Αστυνομίας για το έτος 2016.* [online] Available at: <http://www.astynomia.gr/index.php?option=ozo_content&lang=%27..%27&perform=view&id=70674&Itemid=1866&lang=> [Accessed 15 April 2017].

Crimeview.cityofnorthlasvegas.com, (2017). *CrimeView Community.* [online] Available at: <http://crimeview.cityofnorthlasvegas.com/> [Accessed 20 April 2017].

Enallaxnews.gr, (2017). *Μεγάλη αύξηση της «εγκληματικότητας του δρόμου» στην Ελλάδα το 2016.* [online] Available at: <https://www.enallaxnews.gr/2017/04/25/megalh-afkshsh-ths-egklhmatikothtas-tou-dromou-sthn-ellada-to-2016/> [Accessed 20 April 2017].

List.waikato.ac.nz, (2012). *[Wekalist] J48: What is ROC Area?*. [online] Available at: <https://list.waikato.ac.nz/pipermail/wekalist/2012-May/055512.html> [Accessed 09 April 2017].

Marathondata.gr, (2017). *CrimeView Server*. [online] Available at: <http://www.marathondata.gr/products/omega.htm> [Accessed 19 April 2017].

Peoria.il.crimeviewcommunity.com, (2017). *CrimeView Community*. [online] Available at: <http://peoria.il.crimeviewcommunity.com/default.aspx> [Accessed 20 April 2017].

Policenet.gr, (2012). *Η εγκληματικότητα στον χάρτη. Crime View, ένα χρήσιμο εργαλείο για την πάταξη της εγκληματικότητας*. [online] Available at: <policenet.gr/article/crime-view-ένα-χρήσιμο-gps-για-την-πάταξη-της-εγκληματικότητας> [Accessed 15 April 2017].

Statistics.gr, (2016). *Ελληνική Στατιστική Αρχή*. [online] Available at: <http://www.statistics.gr/> [Accessed 03 April 2017].

Theomegagroup.com, (2012). *CrimeView Dashboard*. [online] Available at: <http://www.theomegagroup.com/police/omega_dashboard_police.html> [Accessed 15 April 2017].

Wikipedia.org, (2017). *Weka (μηχανική μάθηση)*. [online] Available at: <https://el.wikipedia.org/wiki/Weka_(μηχανική_μάθηση)> [Accessed 27 April 2017].

Weka.wikispaces.com, (2017). *I have unbalanced data - now what?*. [online] Available at: <http://weka.wikispaces.com/I+have+unbalanced+data+-+now+what%3F> [Accessed 27 April 2017].

Weka.wikispaces.com, (2017). *Can I process UTF-8 datasets or files?*. [online] Available at: <http://weka.wikispaces.com/Can+I+process+UTF-8+datasets+or+files%3 F> [Accessed 27 April 2017].

Weka.wikispaces.com, (2017). *Can I use CSV files?*. [online] Available at: <http://weka.wikispaces.com/Can+I+use+CSV+files%3F> [Accessed 27 April 2017].

Weka.wikispaces.com, (2017). *Can I use WEKA in commercial applications?*. [online] Available at: <http://weka.wikispaces.com/Can+I+use+WEKA+in+commercial+applications%3F> [Accessed 27 April 2017].

Weka.wikispaces.com, (2017). *Frequently Asked Questions?*. [online] Available at:

<http://weka.wikispaces.com/Frequently+Asked+Questions> [Accessed 27 April 2017].

Financesonline.com, (2017). *RapidMiner*. [online] Available at: <https://reviews.financesonline.com/p/rapidminer/> [Accessed 05 June 2017].

Wikipedia.org, (2017). *RapidMiner*. [online] Available at: <https://en.wikipedia.org/wiki/RapidMiner > [Accessed 04 June 2017].

Rapidminer.com, (2017). *RapidMiner*. [online] Available at: <https://rapidminer.com/> [Accessed 04 June 2017].

Παράρτημα 1

WEKA

Το ακρωνύμιο WEKA προέρχεται από το τις λέξεις Waikato Environment for Knowledge Analysis και επίσης, με το ίδιο όνομα υπάρχει στα νησιά της Νέας Ζηλανδίας ένα πουλί που δεν πετάει και έχει περίεργο χαρακτήρα. Το WEKA είναι ένα δημοφιλές λογισμικό μηχανικής μάθησης γραμμένο σε γλώσσα Java για εργασίες εξόρυξης δεδομένων. Ακόμη, είναι κατάλληλο εργαλείο για ανάπτυξη νέων συστημάτων μηχανικής μάθησης. Δημιουργήθηκε στο Πανεπιστήμιο του Waikato της Νέας Ζηλανδίας και είναι ελεύθερο λογισμικό υπό την άδεια GNU General Public License. Το λογισμικό αυτό διαθέτει στον χρήστη μια πληθώρα εργαλείων οπτικοποίησης, γραφικών διεπαφών και διάφορους αλγορίθμους για ανάλυση δεδομένων και προγνωστική μοντελοποίηση. Αρχικά η πρώτη έκδοση του WEKA ήταν ¨Tcl/Tk front-end¨ (γλώσσα προγραμματισμού σεναρίων που δημιουργήθηκε από τον John Ousterhout) και το 1997 κυκλοφόρησε η πρώτη πλήρης έκδοση σε γλώσσα Java με πολλούς τομείς εφαρμογής και όχι μόνο για επεξεργασία δεδομένων σε γεωργικούς τομείς, όπως αρχικά είχε σχεδιαστεί. Πλεονεκτήματα του WEKA μπορούν να θεωρηθούν η δωρεάν διαθεσιμότητα του, η φορητότητα του (γλώσσα προγραμματισμού Java) όπου τρέχει σχεδόν χωρίς πρόβλημα σε όλες τις σύγχρονες υπολογιστικές πλατφόρμες, η ευκολία χρήσης (εύκολα κατανοητά γραφικά διεπαφών χρήστη) και η διαθεσιμότητα ολοκληρωμένης συλλογής δεδομένων προεπεξεργασίας και τεχνικών μοντελοποίησης (wikipedia.org, 2017).

Οι βασικές διεργασίες του WEKA όπου έχεις την δυνατότητα επιλογής είναι η προεπεξεργασία δεδομένων, η κατηγοριοποίηση, η ομαδοποίηση, η παλινδρόμηση, οι κανόνες συσχέτισης και η απεικόνιση. Περιγράφονται κανονικά, αριθμητικά, ονομαστικά χαρακτηριστικά κατά βάση, καθώς και κάποιοι άλλοι τύποι και τα δεδομένα ότι είναι απλό αρχείο ή συσχέτιση. Το WEKA υποστηρίζει SQL βάσεις δεδομένων με Java Database Connectivity (wikipedia.org, 2017).

Το Weka είναι λογισμικό ανοικτού κώδικα με άδεια GNU General Public License, όπου τα πνευματικά δικαιώματα του περισσότερου κώδικα ανήκουν στο Πανεπιστήμιο του Waikato, περιέχει πολλά δωρεάν online μαθήματα που διδάσκουν μηχανική μάθηση και εξόρυξη δεδομένων και τα βίντεο για τα μαθήματα είναι διαθέσιμα

στο ¨YouTube¨. Ακόμη, μπορεί να εφαρμοστεί σε μεγάλες βάσεις δεδομένων και να χρησιμοποιηθούν αρχεία CSV αντί της προεπιλεγμένης μορφής αρχείου WEKA που είναι αρχεία arff. Όμως ο χρήστης θα πρέπει να γνψρίζει ότι τα αρχεία csv δεν μπορούν να διαβαστούν σταδιακά όπου για να καθοριστεί αν μια στήλη αν είναι αριθμητική ή ονομαστική, όλες οι γραμμές πρέπει να επιθεωρούνται πρωτίστως και η εκπαίδευση και δοκιμή αρχείων μπορεί να μην φέρει τα αναμενόμενα αποτελέσματα, αφού δεν περιέχουν καμία πληροφορία σχετικά με τα χαρακτηριστικά και το WEKA πρέπει να καθορίσει τις ετικέτες για τα ονομαστικά χαρακτηριστικά (weka.wikispaces.com, 2017).

Το λογισμικό WEKA μπορεί να χρησιμοποιηθεί για εμπορικές εφαρμογές, αρκεί να διανέμεται υπό την άδεια GNU General Public License αν το παράγωγο έργο κατανέμεται σε τρίτους, ειδάλλως αν δεν διανέμεται από την GPL, να αγοραστεί η κατάλληλη άδεια από τους κατόχους των πνευματικών δικαιωμάτων. Η εταιρεία Pentaho παρέχει εμπορικές άδειες για το WEKA σε εφαρμογές που εμπίπτουν στην περιοχή της επιχειρηματικής ευφυΐας (weka.wikispaces.com, 2017).

RapidMinner

Η εταιρεία RapidMiner κατασκευάζει λογισμικό για χρήση στον τομέα της επιστήμης των δεδομένων, όπου οι χρήστες μπορούν να γίνουν πιο παραγωγικοί μέσω της πλατφόρμας RapidMiner, η οποία παρέχει μεγάλη ταχύτητα στην προετοιμασία και απεικόνιση των δεδομένων, την εκμάθηση μηχανών, την βαθιά μάθηση, την εξόρυξη κειμένου, την στατιστική μοντελοποίηση, την παροχή προγνωστικών analytics και την ανάπτυξη μοντέλων. Έχει χρήση σε εμπορικές και επιχειρηματικές εφαρμογές αλλά και στην εκπαίδευση και έρευνα. Επίσης, χρησιμοποιείται για τη κατάρτιση, προτυποποίηση και ανάπτυξη εφαρμογών, υποστηρίζοντας όλα τα στάδια της μηχανικής μάθησης με την προετοιμασία δεδομένων, οπτικοποίηση αποτελεσμάτων, επικύρωση μοντέλων και βελτιστοποίηση (Wikipedia.org, 2017).

Το RapidMiner είναι γραμμένο στη γλώσσα προγραμματισμού Java και αναπτύσσεται σε ένα ανοικτού πυρήνα μοντέλο, χρησιμοποιώντας ένα client / server μοντέλο με διακομιστή είτε υπό προϋποθέσεις είτε σε δημόσιες ή ιδιωτικές υποδομές ¨cloud¨, παρέχοντας ένα εύκολο και ισχυρό γραφικό περιβάλλον για να σχεδιάσουν και να εκτελέσουν αναλυτικές ροές εργασίας. Ορισμένες λειτουργίες μπορούν να εκτελεσθούν από τη γραμμή εντολών, παρέχονται σχεδιαγράμματα μάθησης, μοντέλα

και αλγόριθμοι και μπορεί να επεκταθεί με τη χρήση ¨R¨ και ¨Python¨ scripts. Επιπλέον ¨plugins¨ διατίθενται μέσω RapidMiner Marketplace (Wikipedia.org, 2017).

Το RapidMiner παρέχει φιλικό ¨interface¨, εύκολο στη χρήση οπτικό περιβάλλον, ολοκληρωμένα εργαλεία και χαρακτηριστικά, παρέχοντας γραφικές παραστάσεις, απεικονίσεις και περιγραφική στατιστική για άντληση πληροφορίες και γνώσης, ώστε να ενισχυθεί η παραγωγικότητα και αποτελεσματικότητα του χρήστη με επιλογή σωστών αποφάσεων και στρατηγικών. Τα χαρακτηριστικά του RapidMiner είναι τα εξής:

- Σχεδίαση Ροής Εργασίας
- Πρόσβαση και Διαχείριση Δεδομένων
- Εξερεύνηση δεδομένων
- Περιγραφική στατιστική
- Γραφήματα και Οπτικοποίηση
- Προετοιμασία δεδομένων
- Δειγματοληψία δεδομένων
- Διαχωρισμός δεδομένων
- Αντικατάσταση δεδομένων
- Στάθμιση και Επιλογής
- Υπολογισμός Ομοιότητας
- Ομαδοποίηση
- Ανάλυση Καλαθιού Αγοράς
- Μπαϋεσιανό Μοντέλο
- Αξιολόγηση μοντέλων
- Σκορ
- Αυτοματοποίηση και Έλεγχος Διαδικασιών

(Financesonline.com, 2017)

Παράρτημα 2

❖ Αποτέλεσμα αλγόριθμου J48 με κλάση ¨Crime¨

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     RapidMinerData
Instances:    123631
Attributes:   8
              CRIMES
              RANK
              SOLVING
              CONDITION
              NATIONALITY
              SEX
              AGE
              CAUSE
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

RANK = MISD
|  SOLVING = SOL
|  |  AGE = AGE1
|  |  |  SEX = MAN
|  |  |  |  NATIONALITY = NAT
|  |  |  |  |  CONDITION = ENDED: THEFT (504.0/8.0)
|  |  |  |  |  CONDITION = ATT: FORG (4.0)
|  |  |  |  NATIONALITY = FOREI: THEFT (1004.0/6.0)
|  |  |  |  NATIONALITY = UNKN: THEFT (0.0)
|  |  |  SEX = WOMAN
|  |  |  |  NATIONALITY = NAT: DIS-TH (81.0/28.0)
|  |  |  |  NATIONALITY = FOREI: THEFT (259.0/24.0)
|  |  |  |  NATIONALITY = UNKN: THEFT (0.0)
|  |  AGE = AGE2
|  |  |  SEX = MAN
|  |  |  |  NATIONALITY = NAT
|  |  |  |  |  CONDITION = ENDED: THEFT (479.0/206.0)
|  |  |  |  |  CONDITION = ATT: FRAUD (319.0/70.0)
|  |  |  |  NATIONALITY = FOREI: THEFT (1425.0/423.0)
|  |  |  |  NATIONALITY = UNKN: THEFT (0.0)
|  |  |  SEX = WOMAN
|  |  |  |  CONDITION = ENDED: THEFT (3562.0/190.0)
|  |  |  |  CONDITION = ATT: DIS-TH (13.0)
|  |  AGE = AGE3
|  |  |  SEX = MAN: FORG (733.0/85.0)
|  |  |  SEX = WOMAN
|  |  |  |  NATIONALITY = NAT: DIS-TH (49.0/9.0)
|  |  |  |  NATIONALITY = FOREI: FORG (149.0/47.0)
|  |  |  |  NATIONALITY = UNKN: FORG (0.0)
|  |  AGE = AGE4
|  |  |  NATIONALITY = NAT
|  |  |  |  SEX = MAN: THEFT (93.0/23.0)
|  |  |  |  SEX = WOMAN: ROB (2.0)
|  |  |  NATIONALITY = FOREI
|  |  |  |  SEX = MAN: FORG (113.0/1.0)
|  |  |  |  SEX = WOMAN: FRAUD (10.0)
|  |  |  NATIONALITY = UNKN: FORG (0.0)
|  SOLVING = UNSOL
|  |  AGE = AGE1
|  |  |  NATIONALITY = NAT
|  |  |  |  SEX = MAN
|  |  |  |  |  CONDITION = ENDED: THEFT (1259.0/135.0)
|  |  |  |  |  CONDITION = ATT: ROB (36.0/10.0)
|  |  |  |  SEX = WOMAN: ROB (63.0)
|  |  |  NATIONALITY = FOREI
|  |  |  |  CONDITION = ENDED: DIS-TH (287.0/55.0)
|  |  |  |  CONDITION = ATT
|  |  |  |  |  SEX = MAN: FRAUD (16.0/6.0)
|  |  |  |  |  SEX = WOMAN: DIS-TH (7.0)
|  |  |  NATIONALITY = UNKN: THEFT (0.0)
|  |  AGE = AGE2
|  |  |  SEX = MAN: THEFT (95187.0/2376.0)
|  |  |  SEX = WOMAN
|  |  |  |  CONDITION = ENDED
|  |  |  |  |  NATIONALITY = NAT: THEFT (1098.0/98.0)
|  |  |  |  |  NATIONALITY = FOREI: DIS-TH (55.0/4.0)
|  |  |  |  |  NATIONALITY = UNKN: THEFT (0.0)
|  |  |  |  CONDITION = ATT: FRAUD (196.0)
|  |  AGE = AGE3
|  |  |  CONDITION = ENDED
|  |  |  |  SEX = MAN: THEFT (7233.0/685.0)
|  |  |  |  SEX = WOMAN
|  |  |  |  |  NATIONALITY = NAT: ROB (40.0)
|  |  |  |  |  NATIONALITY = FOREI: DIS-TH (7.0)
|  |  |  |  |  NATIONALITY = UNKN: ROB (0.0)
|  |  |  CONDITION = ATT
|  |  |  |  SEX = MAN: FRAUD (199.0)
|  |  |  |  SEX = WOMAN
|  |  |  |  |  NATIONALITY = NAT: FRAUD (16.0/1.0)
|  |  |  |  |  NATIONALITY = FOREI: DIS-TH (66.0/23.0)
|  |  |  |  |  NATIONALITY = UNKN: DIS-TH (0.0)
|  |  AGE = AGE4
|  |  |  SEX = MAN: THEFT (523.0)
|  |  |  SEX = WOMAN
|  |  |  |  NATIONALITY = NAT: ROB (3.0)
|  |  |  |  NATIONALITY = FOREI: DIS-TH (4.0/1.0)
|  |  |  |  NATIONALITY = UNKN: ROB (0.0)
RANK = FEL
|  SOLVING = SOL
|  |  AGE = AGE1
|  |  |  NATIONALITY = NAT: ROB (178.0)
|  |  |  NATIONALITY = FOREI
|  |  |  |  CONDITION = ENDED: DIS-TH (170.0/3.0)
|  |  |  |  CONDITION = ATT
|  |  |  |  |  SEX = MAN: ROB (50.0)
|  |  |  |  |  SEX = WOMAN: DIS-TH (21.0)
|  |  |  NATIONALITY = UNKN: ROB (0.0)
|  |  AGE = AGE2
|  |  |  SEX = MAN
|  |  |  |  NATIONALITY = NAT: DIS-TH (113.0/56.0)
|  |  |  |  NATIONALITY = FOREI: ROB (159.0/82.0)
|  |  |  |  NATIONALITY = UNKN: HOM (0.0)
|  |  |  SEX = WOMAN
|  |  |  |  NATIONALITY = NAT: ROB (394.0/7.0)
|  |  |  |  NATIONALITY = FOREI
|  |  |  |  |  CONDITION = ENDED: HOM (49.0/24.0)
|  |  |  |  |  CONDITION = ATT: ROB (50.0)
|  |  |  |  NATIONALITY = UNKN: ROB (0.0)
|  |  AGE = AGE3
|  |  |  NATIONALITY = NAT
|  |  |  |  SEX = MAN: DIS-TH (801.0/98.0)
|  |  |  |  SEX = WOMAN
|  |  |  |  |  CONDITION = ENDED: ROB (18.0/4.0)
|  |  |  |  |  CONDITION = ATT: DIS-TH (26.0/1.0)
|  |  |  NATIONALITY = FOREI
|  |  |  |  SEX = MAN: DIS-TH (362.0/134.0)
|  |  |  |  SEX = WOMAN
|  |  |  |  |  CONDITION = ENDED: FORG (17.0/9.0)
|  |  |  |  |  CONDITION = ATT: DIS-TH (48.0/4.0)
|  |  |  NATIONALITY = UNKN: DIS-TH (0.0)
|  |  AGE = AGE4
|  |  |  NATIONALITY = NAT
|  |  |  |  CONDITION = ENDED
|  |  |  |  |  SEX = MAN: DIS-TH (55.0/30.0)
|  |  |  |  |  SEX = WOMAN: FORG (6.0/1.0)
```

```
| | | | CONDITION = ATT: DIS-TH (10.0)
| | | NATIONALITY = FOREI: FORG (21.0/7.0)
| | | NATIONALITY = UNKN: DIS-TH (0.0)
| SOLVING = UNSOL
| | CONDITION = ENDED
| | | NATIONALITY = NAT: ROB (3855.0/454.0)
| | | NATIONALITY = FOREI
| | | | AGE = AGE1: ROB (20.0)
| | | | AGE = AGE2
| | | | | SEX = MAN: DIS-TH (457.0/157.0)
| | | | | SEX = WOMAN: THEFT (5.0)
| | | | AGE = AGE3: ROB (219.0/13.0)
| | | | AGE = AGE4: ROB (37.0)
| | | NATIONALITY = UNKN: ROB (0.0)
| | CONDITION = ATT
| | | AGE = AGE1: ROB (50.0/5.0)
| | | AGE = AGE2
| | | | SEX = MAN: FRAUD (82.0/32.0)
| | | | SEX = WOMAN: DIS-TH (40.0)
| | | AGE = AGE3
| | | | NATIONALITY = NAT
| | | | | SEX = MAN: FRAUD (34.0/15.0)
| | | | | SEX = WOMAN: ROB (26.0/1.0)
| | | | NATIONALITY = FOREI: FRAUD (57.0/17.0)
| | | | NATIONALITY = UNKN: FRAUD (0.0)
| | | AGE = AGE4: FRAUD (0.0)
RANK = NOTH: SUIC (1107.0)

Number of Leaves  :  86

Size of the tree :  143


Time taken to build model: 2.79 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      117938               95.3952 %
Incorrectly Classified Instances      5693                4.6048 %
Kappa statistic                          0.7651
Mean absolute error                      0.0181
Root mean squared error                  0.0953
Relative absolute error                 36.5984 %
Root relative squared error             60.5729 %
Total Number of Instances           123631

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure
MCC      ROC Area  PRC Area  Class
                 0,000    0,000    0,000      0,000    0,000      0,000
0,913    0,016     MANSL
                 0,096    0,000    0,379      0,096    0,154      0,191
0,997    0,313     HOM
                 0,174    0,001    0,845      0,174    0,289      0,376
0,729    0,306     FRAUD
                 1,000    0,000    1,000      1,000    1,000      1,000
1,000    1,000     SUIC
                 0,997    0,280    0,963      0,997    0,980      0,820
0,903    0,975     THEFT
                 0,817    0,005    0,775      0,817    0,795      0,791
0,994    0,808     DIS-TH
                 0,899    0,005    0,887      0,899    0,893      0,889
0,996    0,910     ROB
                 0,000    0,000    0,000      0,000    0,000      0,000
0,945    0,023     COUN
                 0,671    0,001    0,854      0,671    0,751      0,755
0,995    0,753     FORG
Weighted Avg.    0,954    0,247    0,949      0,954    0,943
0,805    0,904    0,941

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i   <-- classified as
     0     1     0     0    10     1     1     0    19 |     a =
MANSL
     0    22     4     0     2   143    52     0     5 |     b =
HOM
     0    22   785     0  3443   123    41     0    87 |     c =
FRAUD
     0     0     0  1107     0     0     0     0     0 |     d = SUIC
     0     1    80     0 108457    65   138     0     0 |     e =
THEFT
     0     6    26     0    69  2063   344     0    17 |     f =
DIS-TH
     0     0    34     0   245   222  4618     0    15 |     g =
ROB
     0     0     0     0    25     6     3     0     8 |     h = COUN
     0     6     0     0   382    40     7     0   886 |     i =
FORG
```

❖ Αποτέλεσμα αλγόριθμου J48 με μείωση χαρακτηριστικών με κλάση ¨Crime¨

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -U -M 10
Relation:     RapidMinerData-
weka.filters.unsupervised.attribute.Remove-R8-
weka.filters.unsupervised.attribute.Remove-R2
Instances:    122524
Attributes:   6
              CRIMES
              SOLVING
              CONDITION
              NATIONALITY
              SEX
              AGE
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 unpruned tree
------------------

SOLVING = SOL
| AGE = AGE1
| | CONDITION = ENDED
| | | NATIONALITY = NAT
| | | | SEX = MAN: THEFT (584.0/88.0)
| | | | SEX = WOMAN: DIS-TH (91.0/41.0)
| | | NATIONALITY = FOREI: THEFT (1420.0/200.0)
| | | NATIONALITY = UNKN: THEFT (0.0)
| | CONDITION = ATT
| | | SEX = MAN: ROB (130.0/17.0)
| | | SEX = WOMAN
| | | | NATIONALITY = NAT: ROB (25.0/5.0)
| | | | NATIONALITY = FOREI: DIS-TH (21.0)
| | | | NATIONALITY = UNKN: DIS-TH (0.0)
| AGE = AGE2
| | SEX = MAN
| | | NATIONALITY = NAT
| | | | CONDITION = ENDED: THEFT (592.0/305.0)
| | | | CONDITION = ATT: FRAUD (319.0/70.0)
| | | NATIONALITY = FOREI: THEFT (1584.0/573.0)
| | | NATIONALITY = UNKN: THEFT (0.0)
| | SEX = WOMAN
| | | CONDITION = ENDED: THEFT (3885.0/511.0)
| | | CONDITION = ATT: ROB (183.0/16.0)
| AGE = AGE3
| | NATIONALITY = NAT: DIS-TH (1267.0/499.0)
| | NATIONALITY = FOREI
| | | CONDITION = ENDED: FORG (888.0/453.0)
| | | CONDITION = ATT: DIS-TH (48.0/4.0)
| | NATIONALITY = UNKN: DIS-TH (0.0)
```

```
|  AGE = AGE4
|  |  NATIONALITY = NAT
|  |  |  SEX = MAN: THEFT (148.0/78.0)
|  |  |  SEX = WOMAN: DIS-TH (18.0/8.0)
|  |  NATIONALITY = FOREI
|  |  |  SEX = MAN: FORG (131.0/8.0)
|  |  |  SEX = WOMAN: FRAUD (13.0/3.0)
|  |  NATIONALITY = UNKN: FORG (0.0)
SOLVING = UNSOL
|  NATIONALITY = NAT
|  |  AGE = AGE1
|  |  |  CONDITION = ENDED
|  |  |  |  SEX = MAN: THEFT (1553.0/429.0)
|  |  |  |  SEX = WOMAN: ROB (61.0)
|  |  |  CONDITION = ATT: ROB (79.0/13.0)
|  |  AGE = AGE2
|  |  |  CONDITION = ENDED: THEFT (88006.0/5187.0)
|  |  |  CONDITION = ATT
|  |  |  |  SEX = MAN: THEFT (8506.0/500.0)
|  |  |  |  SEX = WOMAN: FRAUD (190.0/40.0)
|  |  AGE = AGE3
|  |  |  CONDITION = ENDED
|  |  |  |  SEX = MAN: THEFT (7647.0/999.0)
|  |  |  |  SEX = WOMAN: ROB (40.0)
|  |  |  CONDITION = ATT
|  |  |  |  SEX = MAN: FRAUD (119.0/15.0)
|  |  |  |  SEX = WOMAN: ROB (42.0/16.0)
|  |  AGE = AGE4: THEFT (528.0/3.0)
|  NATIONALITY = FOREI
|  |  AGE = AGE1
|  |  |  CONDITION = ENDED: DIS-TH (307.0/75.0)
|  |  |  CONDITION = ATT
|  |  |  |  SEX = MAN: FRAUD (18.0/8.0)
|  |  |  |  SEX = WOMAN: DIS-TH (14.0/7.0)
|  |  AGE = AGE2
|  |  |  SEX = MAN: THEFT (3457.0/406.0)
|  |  |  SEX = WOMAN
|  |  |  |  CONDITION = ENDED: DIS-TH (60.0/9.0)
|  |  |  |  CONDITION = ATT: FRAUD (46.0)
|  |  AGE = AGE3
|  |  |  CONDITION = ENDED: ROB (226.0/20.0)
|  |  |  CONDITION = ATT: FRAUD (237.0/60.0)
|  |  AGE = AGE4: ROB (41.0/3.0)
|  NATIONALITY = UNKN: THEFT (0.0)

Number of Leaves  :  44

Size of the tree :  75


Time taken to build model: 0.47 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      111837               91.2776 %
Incorrectly Classified Instances    10687                8.7224 %
Kappa statistic                          0.3983
Mean absolute error                      0.0383
Root mean squared error                  0.1385
Relative absolute error                 73.4279 %
Root relative squared error             85.7432 %
Total Number of Instances           122524

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure
MCC      ROC Area  PRC Area  Class
                 0,000    0,000    0,000      0,000    0,000      0,000
0,956    0,013     MANSL
                 0,000    0,000    0,000      0,000    0,000      0,000
0,981    0,063     HOM
                 0,160    0,001    0,802      0,160    0,266      0,349
0,735    0,276     FRAUD
                 0,999    0,673    0,921      0,999    0,959      0,541
0,742    0,940     THEFT
                 0,472    0,006    0,638      0,472    0,543      0,541
0,914    0,493     DIS-TH
                 0,144    0,001    0,889      0,144    0,247      0,349
0,682    0,265     ROB
                 0,000    0,000    0,000      0,000    0,000      0,000
0,970    0,014     COUN
                 0,422    0,004    0,548      0,422    0,477      0,476
0,989    0,482     FORG
Weighted Avg.    0,913    0,598    0,903      0,913    0,887
0,524    0,746    0,871

=== Confusion Matrix ===

     a     b     c     d     e     f     g     h   <-- classified as
     0     0     0    12    19     0     0     1 |     a = MANSL
     0     0     0   142    79     0     0     7 |     b = HOM
     0     0   718  3539    47    29     0   168 |     c =
FRAUD
     0     0    76 108631     5    29     0     0 |     d = THEFT
     0     0    71   988  1193    28     0   245 |     e = DIS-
TH
     0     0    27  4180   157   737     0    33 |     f = ROB
     0     0     0    27     8     0     0     7 |     g = COUN
     0     0     3   391   363     6     0   558 |     h = FORG
```

❖ Αποτέλεσμα αλγόριθμου NaïveBayes

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     RapidMinerData
Instances:    123631
Attributes:   8
              CRIME
              RANK
              SOLVING
              CONDITION
              NATIONALITY
              SEX
              AGE
              CAUSE
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

                Class
Attribute             MANSL       HOM      FRAUD       SUIC
THEFT   DIS-TH      ROB      COUN      FORG
                        (0)         (0)      (0.04)      (0.01)    (0.88)
(0.02)     (0.04)      (0)       (0.01)
=========================================
=========================================
==========
RANK
  MISD              31.0      3.0    4243.0      1.0  108523.0    531.0
475.0     32.0    1264.0
  FEL                3.0    227.0     260.0      1.0     220.0   1996.0
4661.0     12.0     59.0
  NOTH               1.0      1.0       1.0   1108.0       1.0      1.0
1.0       1.0      1.0
  [total]           35.0    231.0    4504.0   1110.0  108744.0
2528.0    5137.0     45.0    1324.0

SOLVING
  SOL               33.0    229.0     878.0   1108.0    6542.0
1473.0    835.0     43.0    1322.0
  UNSOL              1.0      1.0    3625.0      1.0  102201.0
1054.0    4301.0      1.0       1.0
  [total]           34.0    230.0    4503.0   1109.0  108743.0
2527.0    5136.0     44.0    1323.0
```

CONDITION
ENDED 33.0 229.0 3258.0 454.0 100587.0 2294.0 4700.0 43.0 1308.0
ATT 1.0 1.0 1245.0 655.0 8156.0 233.0 436.0 1.0 15.0
[total] 34.0 230.0 4503.0 1109.0 108743.0 2527.0 5136.0 44.0 1323.0

NATIONALITY
NAT 31.0 144.0 3854.0 1.0 102657.0 1376.0 4460.0 31.0 403.0
FOREI 3.0 86.0 649.0 1.0 6086.0 1151.0 676.0 13.0 920.0
UNKN 1.0 1.0 1.0 1108.0 1.0 1.0 1.0 1.0 1.0
[total] 35.0 231.0 4504.0 1110.0 108744.0 2528.0 5137.0 45.0 1324.0

SEX
MAN 30.0 194.0 4161.0 734.0 104127.0 2048.0 4348.0 30.0 1079.0
WOMAN 4.0 36.0 342.0 375.0 4616.0 479.0 788.0 14.0 244.0
[total] 34.0 230.0 4503.0 1109.0 108743.0 2527.0 5136.0 44.0 1323.0

AGE
AGE1 1.0 6.0 27.0 89.0 2859.0 495.0 903.0 1.0 19.0
AGE2 9.0 118.0 3287.0 253.0 98624.0 887.0 3516.0 28.0 367.0
AGE3 21.0 87.0 1170.0 633.0 6666.0 1103.0 674.0 13.0 788.0
AGE4 5.0 21.0 21.0 136.0 596.0 44.0 45.0 4.0 151.0
[total] 36.0 232.0 4505.0 1111.0 108745.0 2529.0 5138.0 46.0 1325.0

CAUSE
NOT-KN 33.0 229.0 4502.0 656.0 108742.0 2526.0 5135.0 43.0 1322.0
SENS 1.0 1.0 1.0 91.0 1.0 1.0 1.0 1.0 1.0
DISE 1.0 1.0 1.0 204.0 1.0 1.0 1.0 1.0 1.0
FAM 1.0 1.0 1.0 97.0 1.0 1.0 1.0 1.0 1.0
ECON 1.0 1.0 1.0 64.0 1.0 1.0 1.0 1.0 1.0
[total] 37.0 233.0 4506.0 1112.0 108746.0 2530.0 5139.0 47.0 1326.0

Time taken to build model: 0.14 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	113588	91.8766 %
Incorrectly Classified Instances	10043	8.1234 %
Kappa statistic	0.5956	
Mean absolute error	0.0265	
Root mean squared error	0.1177	
Relative absolute error	53.63 %	
Root relative squared error	74.8071 %	
Total Number of Instances	123631	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,000	0,000	0,000	0,000	0,000	0,000	0,979	0,025	MANSL
0,088	0,000	0,323	0,088	0,138	0,167	0,994	0,189	HOM
0,030	0,000	0,733	0,030	0,058	0,145	0,661	0,176	FRAUD
1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	SUIC
0,979	0,370	0,951	0,979	0,965	0,678	0,896	0,974	THEFT
0,482	0,019	0,343	0,482	0,401	0,392	0,971	0,424	DIS-TH
0,782	0,009	0,788	0,782	0,785	0,776	0,989	0,823	ROB
0,000	0,000	0,000	0,000	0,000	0,000	0,944	0,003	COUN
0,484	0,008	0,382	0,484	0,427	0,423	0,988	0,513	FORG
Weighted Avg. 0,919	0,326	0,916	0,919	0,905	0,656	0,895	0,921	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	<-- classified as
0	1	0	0	25	0	1	0	5	a = MANSL
0	20	0	0	2	164	42	0	0	b = HOM
0	0	137	0	3986	192	73	0	113	c = FRAUD
0	0	0	1107	0	0	0	0	0	d = SUIC
0	0	0	0	106454	1260	192	0	835	e = THEFT
0	28	45	0	402	1217	767	0	66	f = DIS-TH
0	0	0	0	452	657	4013	0	12	g = ROB
0	0	0	0	25	9	2	0	6	h = COUN
0	13	5	0	616	47	0	0	640	i = FORG

❖ Αποτέλεσμα αλγόριθμου Multilayer Perceptron (Percentage split 66%)

=== Run information ===

Scheme: weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation: RapidMinerData
Instances: 123631
Attributes: 8
CRIME
RANK
SOLVING
CONDITION
NATIONALITY
SEX
AGE
CAUSE
Test mode: split 66.0% train, remainder test

=== Classifier model (full training set) ===

Sigmoid Node 0
Inputs Weights
Threshold -3.620879078881608
Node 9 -0.053953199647587664
Node 10 -1.4761758027601266
Node 11 0.5137516004354755
Node 12 -0.8924657424986593
Node 13 0.055414451888778124
Node 14 -1.9432911141325744
Node 15 1.0927454220389743
Node 16 -3.116558654960106
Node 17 0.0915103461739577
Node 18 0.34343527566269905
Node 19 -0.572581345240217

Node 20 -1.967097727429514
Node 21 -0.9990952206612371
Sigmoid Node 1
Inputs Weights
Threshold -7.777393866297411
Node 9 -4.987711170438945
Node 10 -1.5397446792697547
Node 11 3.36549746951744
Node 12 -4.679324886097471
Node 13 -1.0520996286984448
Node 14 -6.959729557632323
Node 15 -3.7398686614282908
Node 16 -0.8689362805781682
Node 17 7.7112597379646
Node 18 -2.638017322458315
Node 19 4.881473778467356
Node 20 -11.221652794395526
Node 21 2.5860638261156517
Sigmoid Node 2
Inputs Weights
Threshold -11.633294483898556
Node 9 -3.529652915531489
Node 10 1.8372714573488504
Node 11 -5.918890502415622
Node 12 -1.7340032249795045
Node 13 7.505359472108959
Node 14 -5.22901291325648
Node 15 3.0683890417137616
Node 16 3.4808605371776804
Node 17 0.2883731938607695
Node 18 2.7459955433419534
Node 19 -5.373363505830023
Node 20 -1.685142340249153
Node 21 7.066376772384383
Sigmoid Node 3
Inputs Weights
Threshold -2.8579208743845457
Node 9 5.398141413153223
Node 10 -3.4295850525978624
Node 11 -1.0065680699475175
Node 12 7.795646341458244
Node 13 -2.171893702041364
Node 14 -0.5633225012698332
Node 15 -4.323060743458968
Node 16 -1.9918122241028255
Node 17 -0.04912318313412623
Node 18 -4.144525720772248
Node 19 -1.6767236464317614
Node 20 -1.08796146293532
Node 21 -5.117318047704845
Sigmoid Node 4
Inputs Weights
Threshold -11.668557873247286
Node 9 0.7131720379458378
Node 10 -0.9051409554051952
Node 11 7.741832490337367
Node 12 -3.0338717016798293
Node 13 -0.608672407192012
Node 14 0.85569689241911
Node 15 4.063893681892805
Node 16 -2.5289741272237207
Node 17 4.491688389065962
Node 18 8.334623618064851
Node 19 2.7486224871523466
Node 20 3.681810539111086
Node 21 -9.398785252571159
Sigmoid Node 5
Inputs Weights
Threshold -2.54490586453744
Node 9 -20.569434513255562
Node 10 -11.912190725322489
Node 11 1.7505975751174458
Node 12 2.2836842184582213
Node 13 -14.794846996879166
Node 14 4.260040880716976
Node 15 -3.735666523359472
Node 16 10.426556326472697
Node 17 7.541946327532242
Node 18 4.30811130383773
Node 19 0.5900510933692023
Node 20 6.710991560052112
Node 21 -6.609381879730758
Sigmoid Node 6
Inputs Weights
Threshold -2.6888403304076074
Node 9 4.708057645278279
Node 10 16.13276567568878
Node 11 1.971294020440417
Node 12 -5.322851140468292
Node 13 -7.772981069107973
Node 14 5.111102472037452
Node 15 -7.337825979783278
Node 16 -9.683407532313595
Node 17 -5.528922330513769
Node 18 -2.2591436620167507
Node 19 10.248649056603476
Node 20 -7.4233984923510565
Node 21 -4.933912153428831
Sigmoid Node 7
Inputs Weights
Threshold -3.927687527876921
Node 9 0.36002379053526595
Node 10 0.3544202888778869
Node 11 -0.34146538485876055
Node 12 -0.8542767414647714
Node 13 0.6293730748521332
Node 14 -1.7501128730244306
Node 15 -0.2940511236315807
Node 16 -2.079269267070483
Node 17 0.47682665984131484
Node 18 0.38843480122800134
Node 19 -0.896899754057045
Node 20 -2.342651333817252
Node 21 -0.10739961956975463
Sigmoid Node 8
Inputs Weights
Threshold -5.808368108108834
Node 9 -0.44604662319429816
Node 10 -5.6446628768496625
Node 11 0.7970200113859404
Node 12 -7.135965246999882
Node 13 -0.3838024550791895
Node 14 -3.0192341405833285
Node 15 6.937101943229818
Node 16 -0.3662848526448916
Node 17 6.691178971972694
Node 18 -4.966696688496146
Node 19 -3.5115057285760667
Node 20 -3.892523618332459
Node 21 0.4755908713199672
Sigmoid Node 9
Inputs Weights
Threshold -0.8851704347360124
Attrib RANK=MISD 3.0985415370943334
Attrib RANK=FEL -6.696595587215197
Attrib RANK=NOTH 4.504925029377472
Attrib SOLVING=UNSOL -10.445497327834904
Attrib CONDITION=ATT -5.044378264601183
Attrib NATIONALITY=NAT -4.859646266641237
Attrib NATIONALITY=FOREI 1.2474739444472465
Attrib NATIONALITY=UNKN 4.47197745370124
Attrib SEX=WOMAN -4.852294224844906
Attrib AGE=AGE1 -1.9352086955564645
Attrib AGE=AGE2 3.2178396121974364
Attrib AGE=AGE3 -3.745806216000776
Attrib AGE=AGE4 4.1723123330939105
Attrib CAUSE=NOT-KN -1.264161251606354
Attrib CAUSE=SENS 0.9834102239021679
Attrib CAUSE=DISE 0.9690824663001287
Attrib CAUSE=FAM 0.9583635529625409
Attrib CAUSE=ECON 0.96049440328871
Sigmoid Node 10

Inputs Weights
Threshold 0.5082756726793676
Attrib RANK=MISD -0.36989213848783664
Attrib RANK=FEL 3.309658383377947
Attrib RANK=NOTH -3.4858190504650106
Attrib SOLVING=UNSOL 14.87190245869057
Attrib CONDITION=ATT -8.985715796856816
Attrib NATIONALITY=NAT 2.139495332221033
Attrib NATIONALITY=FOREI 0.7788980922276825
Attrib NATIONALITY=UNKN -3.4822515161619303
Attrib SEX=WOMAN 4.842887331602476
Attrib AGE=AGE1 9.864340944180485
Attrib AGE=AGE2 3.9820605715583928
Attrib AGE=AGE3 -6.486922359878658
Attrib AGE=AGE4 -8.58108835849976
Attrib CAUSE=NOT-KN 0.863958614999178
Attrib CAUSE=SENS -0.7745103348492474
Attrib CAUSE=DISE -0.6608655433742888
Attrib CAUSE=FAM -0.406921776094124
Attrib CAUSE=ECON -0.5712142526889034
Sigmoid Node 11
Inputs Weights
Threshold -2.9053719336118915
Attrib RANK=MISD 6.36624656718079
Attrib RANK=FEL -2.9515374598765813
Attrib RANK=NOTH -0.5926537975351429
Attrib SOLVING=UNSOL -5.181270687175923
Attrib CONDITION=ATT -7.066786712608285
Attrib NATIONALITY=NAT 4.046837637191295
Attrib NATIONALITY=FOREI -0.5775949008183534
Attrib NATIONALITY=UNKN -0.5273772763264029
Attrib SEX=WOMAN -8.789137070681958
Attrib AGE=AGE1 -2.3088261412634914
Attrib AGE=AGE2 8.36368227143106
Attrib AGE=AGE3 1.7301644920284769
Attrib AGE=AGE4 -2.05829671894001
Attrib CAUSE=NOT-KN -0.18113784438875552
Attrib CAUSE=SENS 1.4994237195444453
Attrib CAUSE=DISE 2.5827779590159423
Attrib CAUSE=FAM 2.0357795822089018
Attrib CAUSE=ECON 2.6968356738039314
Sigmoid Node 12
Inputs Weights
Threshold 0.2848346056532555
Attrib RANK=MISD -0.003591724020697792
Attrib RANK=FEL -5.361174565653983
Attrib RANK=NOTH 5.064687721767851
Attrib SOLVING=UNSOL 1.845499432376252
Attrib CONDITION=ATT 4.053742147243348
Attrib NATIONALITY=NAT 2.436703001595406
Attrib NATIONALITY=FOREI -7.765163192465653
Attrib NATIONALITY=UNKN 5.052511754138258
Attrib SEX=WOMAN -1.5875663342002497
Attrib AGE=AGE1 -1.3591764701124207
Attrib AGE=AGE2 6.355155388286872
Attrib AGE=AGE3 -7.298717203387253
Attrib AGE=AGE4 1.7069424219344882
Attrib CAUSE=NOT-KN -0.35289441276406447
Attrib CAUSE=SENS -0.17675061234003006
Attrib CAUSE=DISE 0.036856217712320866
Attrib CAUSE=FAM -0.19456627760024048
Attrib CAUSE=ECON -0.25119112065737154
Sigmoid Node 13
Inputs Weights
Threshold -2.285527587216706
Attrib RANK=MISD 7.789033557879921
Attrib RANK=FEL -7.576881803620768
Attrib RANK=NOTH 1.9695255647387924
Attrib SOLVING=UNSOL 9.564228155369975
Attrib CONDITION=ATT 5.16963725911111
Attrib NATIONALITY=NAT 2.3771448507849327
Attrib NATIONALITY=FOREI -2.1111648540063817
Attrib NATIONALITY=UNKN 1.9291479263512046
Attrib SEX=WOMAN -9.836879149681199
Attrib AGE=AGE1 -3.934264631860191
Attrib AGE=AGE2 6.157841592704416
Attrib AGE=AGE3 -2.8108695317539647
Attrib AGE=AGE4 5.143088946744794
Attrib CAUSE=NOT-KN -1.9086853497527814
Attrib CAUSE=SENS 2.1356056990118018
Attrib CAUSE=DISE 2.304377102945799
Attrib CAUSE=FAM 2.057716420031137
Attrib CAUSE=ECON 2.267777308131391
Sigmoid Node 14
Inputs Weights
Threshold -1.5184909598358305
Attrib RANK=MISD 5.982520037424821
Attrib RANK=FEL -5.592936916614986
Attrib RANK=NOTH 1.16296837669959
Attrib SOLVING=UNSOL 16.44391191330868
Attrib CONDITION=ATT -7.426447316659258
Attrib NATIONALITY=NAT -2.309171649745005
Attrib NATIONALITY=FOREI 2.712584483480503
Attrib NATIONALITY=UNKN 1.1423124008593781
Attrib SEX=WOMAN -2.2820924910599176
Attrib AGE=AGE1 4.682768217056231
Attrib AGE=AGE2 -0.9435322544658548
Attrib AGE=AGE3 -2.0363069226449664
Attrib AGE=AGE4 1.5213656810156202
Attrib CAUSE=NOT-KN -1.163015650480024
Attrib CAUSE=SENS 1.2778773748805279
Attrib CAUSE=DISE 1.520620844181473
Attrib CAUSE=FAM 1.5235301589691712
Attrib CAUSE=ECON 1.5663946252778373
Sigmoid Node 15
Inputs Weights
Threshold -2.0647907886173416
Attrib RANK=MISD 10.237278202750868
Attrib RANK=FEL -5.985585744594955
Attrib RANK=NOTH -2.290246201467976
Attrib SOLVING=UNSOL -9.714623357391043
Attrib CONDITION=ATT 3.4321874376420176
Attrib NATIONALITY=NAT 2.3621757443993983
Attrib NATIONALITY=FOREI 1.933672812448058
Attrib NATIONALITY=UNKN -2.2628197472288023
Attrib SEX=WOMAN -4.841939176076014
Attrib AGE=AGE1 -5.406405491661814
Attrib AGE=AGE2 3.431166849401625
Attrib AGE=AGE3 4.413273564474558
Attrib AGE=AGE4 1.5076818598238588
Attrib CAUSE=NOT-KN -0.9657270273146619
Attrib CAUSE=SENS 1.930116105320853
Attrib CAUSE=DISE 1.8449681012090529
Attrib CAUSE=FAM 1.3335059544629146
Attrib CAUSE=ECON 1.9772547246377659
Sigmoid Node 16
Inputs Weights
Threshold 2.2084191754613665
Attrib RANK=MISD -4.803035601078484
Attrib RANK=FEL 4.433115392688242
Attrib RANK=NOTH -1.8329867430801075
Attrib SOLVING=UNSOL -12.011862506833852
Attrib CONDITION=ATT 7.750448371446895
Attrib NATIONALITY=NAT -5.389733695098063
Attrib NATIONALITY=FOREI 5.022553866673459
Attrib NATIONALITY=UNKN -1.7768209041934344
Attrib SEX=WOMAN 12.419820766804245
Attrib AGE=AGE1 -4.756414703833749
Attrib AGE=AGE2 1.779149666632129
Attrib AGE=AGE3 1.3519565180215358
Attrib AGE=AGE4 -2.7995300067490994
Attrib CAUSE=NOT-KN 0.7650166123511654
Attrib CAUSE=SENS -2.005309610884686
Attrib CAUSE=DISE -1.6003045118669004
Attrib CAUSE=FAM -1.9632285178282356
Attrib CAUSE=ECON -1.914545511167027
Sigmoid Node 17
Inputs Weights
Threshold -2.4418532955665864
Attrib RANK=MISD 4.795561500140707
Attrib RANK=FEL -3.699407522893962
Attrib RANK=NOTH 1.4361448244379118

Attrib SOLVING=UNSOL -4.147674851176912
Attrib CONDITION=ATT -12.33011618626401
Attrib NATIONALITY=NAT -1.6374310161847379
Attrib NATIONALITY=FOREI 2.6202554834522176
Attrib NATIONALITY=UNKN 1.3539671864691543
Attrib SEX=WOMAN -10.805447182331214
Attrib AGE=AGE1 3.0375165318552395
Attrib AGE=AGE2 12.40018792197772
Attrib AGE=AGE3 -11.275670111742004
Attrib AGE=AGE4 0.641153671199793
Attrib CAUSE=NOT-KN -1.9185301737802913
Attrib CAUSE=SENS 2.333659621513516
Attrib CAUSE=DISE 2.2855264147858296
Attrib CAUSE=FAM 2.2744730746177924
Attrib CAUSE=ECON 2.3324937047102208

Sigmoid Node 18
Inputs Weights
Threshold -1.181372616306832
Attrib RANK=MISD 9.97124089450444
Attrib RANK=FEL -8.316145045425607
Attrib RANK=NOTH -0.4953567438340525 4
Attrib SOLVING=UNSOL 7.901923578319927
Attrib CONDITION=ATT 2.049240671204496
Attrib NATIONALITY=NAT 2.727582945605351
Attrib NATIONALITY=FOREI -1.0450575255840309
Attrib NATIONALITY=UNKN -0.5072514129200171
Attrib SEX=WOMAN 3.4520946649104585
Attrib AGE=AGE1 2.352368875442363
Attrib AGE=AGE2 5.133594535080086
Attrib AGE=AGE3 -6.6084540438228085
Attrib AGE=AGE4 1.5945788149930398
Attrib CAUSE=NOT-KN 0.13063326640221384
Attrib CAUSE=SENS 0.6741614315448179
Attrib CAUSE=DISE 0.871742634247936
Attrib CAUSE=FAM 0.7871032844445977
Attrib CAUSE=ECON 1.0173945422219859

Sigmoid Node 19
Inputs Weights
Threshold -2.177793647723339
Attrib RANK=MISD -1.722889149528577
Attrib RANK=FEL 4.925431471277383
Attrib RANK=NOTH -1.0526653146201936
Attrib SOLVING=UNSOL 5.796071791461737
Attrib CONDITION=ATT -0.4341478879383603
Attrib NATIONALITY=NAT 9.302832283736718
Attrib NATIONALITY=FOREI -6.0994652972144845
Attrib NATIONALITY=UNKN -1.0087934662134146
Attrib SEX=WOMAN 10.530682674467306
Attrib AGE=AGE1 5.3635422124598255
Attrib AGE=AGE2 -0.3484458924469601
Attrib AGE=AGE3 -7.612501738234174
Attrib AGE=AGE4 6.823481151369662
Attrib CAUSE=NOT-KN -0.15191789810603742
Attrib CAUSE=SENS 1.4412590089971757
Attrib CAUSE=DISE 1.6149950878806163
Attrib CAUSE=FAM 1.8201755721787163
Attrib CAUSE=ECON 1.7506260513028784

Sigmoid Node 20
Inputs Weights
Threshold -0.2577220262808919
Attrib RANK=MISD 3.5673858222860195
Attrib RANK=FEL -2.486562235509235
Attrib RANK=NOTH -0.8827480522877809
Attrib SOLVING=UNSOL 12.930973500671744
Attrib CONDITION=ATT -8.212283582223051
Attrib NATIONALITY=NAT 3.5918789564526086
Attrib NATIONALITY=FOREI -2.5119963611819216
Attrib NATIONALITY=UNKN -0.8492726916959299
Attrib SEX=WOMAN 8.156450623328592
Attrib AGE=AGE1 13.85842872729407
Attrib AGE=AGE2 -4.080309369438002
Attrib AGE=AGE3 -4.797385179645655
Attrib AGE=AGE4 -4.55256967355791
Attrib CAUSE=NOT-KN 0.6148000703173232
Attrib CAUSE=SENS -0.37432107651063656
Attrib CAUSE=DISE 0.09712128604777534
Attrib CAUSE=FAM 0.08374744083606721
Attrib CAUSE=ECON 0.11591724414271047

Sigmoid Node 21
Inputs Weights
Threshold 1.5779819363818388
Attrib RANK=MISD 1.0797097321359577
Attrib RANK=FEL 0.7127366077589701
Attrib RANK=NOTH -3.469566752677553
Attrib SOLVING=UNSOL 9.146525844617878
Attrib CONDITION=ATT -5.089652939358787
Attrib NATIONALITY=NAT 5.45514544344419
Attrib NATIONALITY=FOREI -3.601566284906817
Attrib NATIONALITY=UNKN -3.374835806551886
Attrib SEX=WOMAN -2.7141774345758893
Attrib AGE=AGE1 -4.498616994868366
Attrib AGE=AGE2 -0.03154210707399752
Attrib AGE=AGE3 -0.6882269871900213
Attrib AGE=AGE4 2.116003914547982
Attrib CAUSE=NOT-KN 1.4189134625163078
Attrib CAUSE=SENS -1.648261192429706
Attrib CAUSE=DISE -1.6075426032112201
Attrib CAUSE=FAM -1.3992045655291616
Attrib CAUSE=ECON -1.5233010157655984

Class MANSL
Input
Node 0

Class HOM
Input
Node 1

Class FRAUD
Input
Node 2

Class SUIC
Input
Node 3

Class THEFT
Input
Node 4

Class DIS-TH
Input
Node 5

Class ROB
Input
Node 6

Class COUN
Input
Node 7

Class FORG
Input
Node 8

Time taken to build model: 696.88 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.81 seconds

=== Summary ===

Correctly Classified Instances	40071	95.3277 %
Incorrectly Classified Instances	1964	4.6723 %
Kappa statistic	0.7599	
Mean absolute error	0.0172	
Root mean squared error	0.0959	
Relative absolute error	34.7545 %	
Root relative squared error	61.163 %	
Total Number of Instances	42035	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,000	0,000	0,000	0,000	0,000	0,000	0,841	0,044	MANSL

	0,282	0,001	0,272	0,282	0,277	0,275	0,994	0,228	HOM
	0,187	0,001	0,834	0,187	0,305	0,386	0,767	0,313	FRAUD
	1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	SUIC
	0,998	0,282	0,963	0,998	0,980	0,821	0,914	0,977	THEFT
	0,803	0,005	0,759	0,803	0,780	0,776	0,991	0,758	DIS-TH
	0,881	0,004	0,902	0,881	0,891	0,887	0,993	0,907	ROB
	0,000	0,000	0,000	0,000	0,000	0,000	0,512	0,006	COUN
	0,642	0,001	0,848	0,642	0,731	0,735	0,981	0,704	FORG
Weighted Avg.	0,953	0,248	0,949	0,953	0,943	0,805	0,915	0,941	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	<-- classified as
0	0	0	0	5	0	0	0	8	a = MANSL
0	22	2	0	1	53	0	0	0	b = HOM
0	17	291	0	1161	51	3	0	35	c = FRAUD
0	0	0	366	0	0	0	0	0	d = SUIC
0	4	27	0	36921	17	35	0	0	e = THEFT
0	0	13	0	25	670	124	0	2	f = DIS-TH
0	37	9	0	78	75	1512	0	5	g = ROB
0	0	1	0	9	3	1	0	2	h = COUN
0	1	6	0	138	14	2	0	289	i = FORG

❖ Αποτέλεσμα αλγόριθμου Multilayer Perceptron (Cross-validation & Folds=10)

=== Run information ===

Scheme: weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation: RapidMinerData
Instances: 123631
Attributes: 8
CRIME
RANK
SOLVING
CONDITION
NATIONALITY
SEX
AGE
CAUSE
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Sigmoid Node 0
Inputs Weights
Threshold -3.620879078881608
Node 9 -0.053953199647587664
Node 10 -1.4761758027601266
Node 11 0.5137516004354755
Node 12 -0.8924657424986593
Node 13 0.055414451888778124
Node 14 -1.9432911141325744
Node 15 1.0927454220389743
Node 16 -3.116558654960106
Node 17 0.0915103461739577
Node 18 0.34343527566269905
Node 19 -0.572581345240217
Node 20 -1.967097727429514
Node 21 -0.9990952206612371
Sigmoid Node 1
Inputs Weights
Threshold -7.777393866297411
Node 9 -4.987711170438945
Node 10 -1.5397446792697547
Node 11 3.36549746951744
Node 12 -4.679324886097471
Node 13 -1.0520996286984448
Node 14 -6.959729557632323
Node 15 -3.7398686614282908
Node 16 -0.8689362805781682
Node 17 7.7112597379646
Node 18 -2.638017322458315
Node 19 4.881473778467356
Node 20 -11.221652794395526
Node 21 2.5860638261156517
Sigmoid Node 2
Inputs Weights
Threshold -11.633294483898556
Node 9 -3.529652915531489
Node 10 1.8372714573488504
Node 11 -5.918890502415622
Node 12 -1.7340032249795045
Node 13 7.505359472108959
Node 14 -5.22901291325648
Node 15 3.0683890417137616
Node 16 3.4808605371776804
Node 17 0.2883731938607695
Node 18 2.7459955433419534
Node 19 -5.373363505830023
Node 20 -1.685142340249153
Node 21 7.066376772384383
Sigmoid Node 3
Inputs Weights
Threshold -2.8579208743845457
Node 9 5.398141413153223
Node 10 -3.4295850525978624
Node 11 -1.0065680699475175
Node 12 7.795646341458244
Node 13 -2.171893702041364
Node 14 -0.5633225012698332
Node 15 -4.323060743458968
Node 16 -1.9918122241028255
Node 17 -0.04912318313412623
Node 18 -4.144525720772248
Node 19 -1.6767236464317614
Node 20 -1.08796146293532
Node 21 -5.117318047704845
Sigmoid Node 4
Inputs Weights
Threshold -11.668557873247286
Node 9 0.7131720379458378
Node 10 -0.9051409554051952
Node 11 7.741832490337367
Node 12 -3.0338717016798293
Node 13 -0.608672407192012
Node 14 0.85569689241911
Node 15 4.063893681892805
Node 16 -2.5289741272237207
Node 17 4.491688389065962
Node 18 8.334623618064851
Node 19 2.7486224871523466
Node 20 3.681810539111086
Node 21 -9.398785252571159
Sigmoid Node 5
Inputs Weights
Threshold -2.54490586453744
Node 9 -20.569434513255562
Node 10 -11.912190725322489
Node 11 1.7505975751174458
Node 12 2.2836842184582213
Node 13 -14.794846996879166
Node 14 4.260040880716976
Node 15 -3.735666523359472
Node 16 10.426556326472697
Node 17 7.541946327532242
Node 18 4.30811130383773

Node 19 0.5900510933692023
Node 20 6.710991560052112
Node 21 -6.609381879730758
Sigmoid Node 6
Inputs Weights
Threshold -2.6888403304076074
Node 9 4.708057645278279
Node 10 16.13276567568878
Node 11 1.971294020440417
Node 12 -5.322851140468292
Node 13 -7.772981069107973
Node 14 5.111102472037452
Node 15 -7.337825979783278
Node 16 -9.683407532313595
Node 17 -5.528922330513769
Node 18 -2.2591436620167507
Node 19 10.248649056603476
Node 20 -7.4233984923510565
Node 21 -4.933912153428831
Sigmoid Node 7
Inputs Weights
Threshold -3.927687527876921
Node 9 0.36002379053526595
Node 10 0.3544202888778869
Node 11 -0.34146538485876055
Node 12 -0.8542767414647714
Node 13 0.6293730748521332
Node 14 -1.7501128730244306
Node 15 -0.2940511236315807
Node 16 -2.079269267070483
Node 17 0.47682665984131484
Node 18 0.38843480122800134
Node 19 -0.896899754057045
Node 20 -2.342651333817252
Node 21 -0.10739961956975463
Sigmoid Node 8
Inputs Weights
Threshold -5.808368108108834
Node 9 -0.44604662319429816
Node 10 -5.6446628768496625
Node 11 0.7970200113859404
Node 12 -7.135965246999882
Node 13 -0.3838024550791895
Node 14 -3.0192341405833285
Node 15 6.937101943229818
Node 16 -0.3662848526448916
Node 17 6.691178971972694
Node 18 -4.966696688496146
Node 19 -3.5115057285760667
Node 20 -3.892523618332459
Node 21 0.4755908713199672
Sigmoid Node 9
Inputs Weights
Threshold -0.8851704347360124
Attrib RANK=MISD 3.0985415370943334
Attrib RANK=FEL -6.696595587215197
Attrib RANK=NOTH 4.504925029377472
Attrib SOLVING=UNSOL -10.445497327834904
Attrib CONDITION=ATT -5.044378264601183
Attrib NATIONALITY=NAT -4.859646266641237
Attrib NATIONALITY=FOREI 1.2474739444472465
Attrib NATIONALITY=UNKN 4.47197745370124
Attrib SEX=WOMAN -4.852294224844906
Attrib AGE=AGE1 -1.9352086955564645
Attrib AGE=AGE2 3.2178396121974364
Attrib AGE=AGE3 -3.745806216000776
Attrib AGE=AGE4 4.1723123330939105
Attrib CAUSE=NOT-KN -1.264161251606354
Attrib CAUSE=SENS 0.9834102239021679
Attrib CAUSE=DISE 0.9690824663001287
Attrib CAUSE=FAM 0.9583635529625409
Attrib CAUSE=ECON 0.96049440328871
Sigmoid Node 10
Inputs Weights
Threshold 0.5082756726793676
Attrib RANK=MISD -0.36989213848783664
Attrib RANK=FEL 3.309658383377947
Attrib RANK=NOTH -3.4858190504650106
Attrib SOLVING=UNSOL 14.87190245869057
Attrib CONDITION=ATT -8.985715796856816
Attrib NATIONALITY=NAT 2.139495332221033
Attrib NATIONALITY=FOREI 0.7788980922276825
Attrib NATIONALITY=UNKN -3.4822515161619303
Attrib SEX=WOMAN 4.842887331602476
Attrib AGE=AGE1 9.864340944180485
Attrib AGE=AGE2 3.9820605715583928
Attrib AGE=AGE3 -6.486922359878658
Attrib AGE=AGE4 -8.58108835849976
Attrib CAUSE=NOT-KN 0.863958614999178
Attrib CAUSE=SENS -0.7745103348492474
Attrib CAUSE=DISE -0.6608655433742888
Attrib CAUSE=FAM -0.406921776094124
Attrib CAUSE=ECON -0.5712142526889034
Sigmoid Node 11
Inputs Weights
Threshold -2.9053719336118915
Attrib RANK=MISD 6.36624656718079
Attrib RANK=FEL -2.9515374598765813
Attrib RANK=NOTH -0.5926537975351429
Attrib SOLVING=UNSOL -5.181270687175923
Attrib CONDITION=ATT -7.066786712608285
Attrib NATIONALITY=NAT 4.046837637191295
Attrib NATIONALITY=FOREI -0.5775949008183534
Attrib NATIONALITY=UNKN -0.5273772763264029
Attrib SEX=WOMAN -8.789137070681958
Attrib AGE=AGE1 -2.3088261412634914
Attrib AGE=AGE2 8.36368227143106
Attrib AGE=AGE3 1.7301644920284769
Attrib AGE=AGE4 -2.05829671894001
Attrib CAUSE=NOT-KN -0.18113784438875552
Attrib CAUSE=SENS 1.4994237195444453
Attrib CAUSE=DISE 2.5827779590159423
Attrib CAUSE=FAM 2.0357795822089018
Attrib CAUSE=ECON 2.6968356738039314
Sigmoid Node 12
Inputs Weights
Threshold 0.2848346056532555
Attrib RANK=MISD -0.003591724020697792
Attrib RANK=FEL -5.361174565653983
Attrib RANK=NOTH 5.064687721767851
Attrib SOLVING=UNSOL 1.845499432376252
Attrib CONDITION=ATT 4.053742147243348
Attrib NATIONALITY=NAT 2.436703001595406
Attrib NATIONALITY=FOREI -7.765163192465653
Attrib NATIONALITY=UNKN 5.052511754138258
Attrib SEX=WOMAN -1.5875663342002497
Attrib AGE=AGE1 -1.3591764701124207
Attrib AGE=AGE2 6.355155388286872
Attrib AGE=AGE3 -7.298717203387253
Attrib AGE=AGE4 1.7069424219344882
Attrib CAUSE=NOT-KN -0.35289441276406447
Attrib CAUSE=SENS -0.17675061234003006
Attrib CAUSE=DISE 0.036856217712320866
Attrib CAUSE=FAM -0.19456627760024048
Attrib CAUSE=ECON -0.25119112065737154
Sigmoid Node 13
Inputs Weights
Threshold -2.285527587216706
Attrib RANK=MISD 7.789033557879921
Attrib RANK=FEL -7.576881803620768
Attrib RANK=NOTH 1.9695255647387924
Attrib SOLVING=UNSOL 9.564228155369975
Attrib CONDITION=ATT 5.16963725911111
Attrib NATIONALITY=NAT 2.3771448507849327
Attrib NATIONALITY=FOREI -2.1111648540063817
Attrib NATIONALITY=UNKN 1.9291479263512046
Attrib SEX=WOMAN -9.836879149681199
Attrib AGE=AGE1 -3.934264631860191
Attrib AGE=AGE2 6.157841592704416
Attrib AGE=AGE3 -2.8108695317539647
Attrib AGE=AGE4 5.143088946744794
Attrib CAUSE=NOT-KN -1.9086853497527814

Attrib CAUSE=SENS 2.1356056990118018
Attrib CAUSE=DISE 2.304377102945799
Attrib CAUSE=FAM 2.057716420031137
Attrib CAUSE=ECON 2.267777308131391
Sigmoid Node 14
Inputs Weights
Threshold -1.5184909598358305
Attrib RANK=MISD 5.982520037424821
Attrib RANK=FEL -5.592936916614986
Attrib RANK=NOTH 1.16296837669959
Attrib SOLVING=UNSOL 16.44391191330868
Attrib CONDITION=ATT -7.426447316659258
Attrib NATIONALITY=NAT -2.309171649745005
Attrib NATIONALITY=FOREI 2.712584483480503
Attrib NATIONALITY=UNKN 1.1423124008593781
Attrib SEX=WOMAN -2.2820924910599176
Attrib AGE=AGE1 4.682768217056231
Attrib AGE=AGE2 -0.9435322544658548
Attrib AGE=AGE3 -2.0363069226449664
Attrib AGE=AGE4 1.5213656810156202
Attrib CAUSE=NOT-KN -1.163015650480024
Attrib CAUSE=SENS 1.2778773748805279
Attrib CAUSE=DISE 1.520620844181473
Attrib CAUSE=FAM 1.5235301589691712
Attrib CAUSE=ECON 1.5663946252778373
Sigmoid Node 15
Inputs Weights
Threshold -2.0647907886173416
Attrib RANK=MISD 10.237278202750868
Attrib RANK=FEL -5.9855855744594955
Attrib RANK=NOTH -2.290246201467976
Attrib SOLVING=UNSOL -9.714623357391043
Attrib CONDITION=ATT 3.4321874376420176
Attrib NATIONALITY=NAT 2.3621757443993983
Attrib NATIONALITY=FOREI 1.933672812448058
Attrib NATIONALITY=UNKN -2.2628197472288023
Attrib SEX=WOMAN -4.841939176076014
Attrib AGE=AGE1 -5.406405491661814
Attrib AGE=AGE2 3.431166849401625
Attrib AGE=AGE3 4.413273564474558
Attrib AGE=AGE4 1.5076818598238588
Attrib CAUSE=NOT-KN -0.9657270273146619
Attrib CAUSE=SENS 1.930116105320853
Attrib CAUSE=DISE 1.8449681012090529
Attrib CAUSE=FAM 1.3335059544629146
Attrib CAUSE=ECON 1.9772547246377659
Sigmoid Node 16
Inputs Weights
Threshold 2.2084191754613665
Attrib RANK=MISD -4.803035601078484
Attrib RANK=FEL 4.433115392688242
Attrib RANK=NOTH -1.8329867430801075
Attrib SOLVING=UNSOL -12.011862506833852
Attrib CONDITION=ATT 7.750448371446895
Attrib NATIONALITY=NAT -5.389733695098063
Attrib NATIONALITY=FOREI 5.022553866673459
Attrib NATIONALITY=UNKN -1.7768209041934344
Attrib SEX=WOMAN 12.419820766804245
Attrib AGE=AGE1 -4.756414703833749
Attrib AGE=AGE2 1.779149666632129
Attrib AGE=AGE3 1.3519565180215358
Attrib AGE=AGE4 -2.7995300067490994
Attrib CAUSE=NOT-KN 0.7650166123511654
Attrib CAUSE=SENS -2.005309610884686
Attrib CAUSE=DISE -1.6003045118669004
Attrib CAUSE=FAM -1.9632285178282356
Attrib CAUSE=ECON -1.914545511167027
Sigmoid Node 17
Inputs Weights
Threshold -2.4418532955665864
Attrib RANK=MISD 4.795561500140707
Attrib RANK=FEL -3.699407522893962
Attrib RANK=NOTH 1.4361448244379118
Attrib SOLVING=UNSOL -4.147674851176912
Attrib CONDITION=ATT -12.33011618626401
Attrib NATIONALITY=NAT -1.6374310161847379
Attrib NATIONALITY=FOREI 2.6202554834522176
Attrib NATIONALITY=UNKN 1.3539671864691543
Attrib SEX=WOMAN -10.805447182331214
Attrib AGE=AGE1 3.0375165318552395
Attrib AGE=AGE2 12.40018792197772
Attrib AGE=AGE3 -11.275670111742004
Attrib AGE=AGE4 0.641153671199793
Attrib CAUSE=NOT-KN -1.9185301737802913
Attrib CAUSE=SENS 2.333659621513516
Attrib CAUSE=DISE 2.2855264147858296
Attrib CAUSE=FAM 2.2744730746177924
Attrib CAUSE=ECON 2.3324937047102208
Sigmoid Node 18
Inputs Weights
Threshold -1.181372616306832
Attrib RANK=MISD 9.97124089450444
Attrib RANK=FEL -8.316145045425607
Attrib RANK=NOTH -0.49535674383405254
Attrib SOLVING=UNSOL 7.901923578319927
Attrib CONDITION=ATT 2.049240671204496
Attrib NATIONALITY=NAT 2.727582945605351
Attrib NATIONALITY=FOREI -1.0450575255840309
Attrib NATIONALITY=UNKN -0.5072514129200171
Attrib SEX=WOMAN 3.4520946649104585
Attrib AGE=AGE1 2.352368875442363
Attrib AGE=AGE2 5.133594535080086
Attrib AGE=AGE3 -6.6084540438228085
Attrib AGE=AGE4 1.5945788149930398
Attrib CAUSE=NOT-KN 0.13063326640221384
Attrib CAUSE=SENS 0.6741614315448179
Attrib CAUSE=DISE 0.871742634247936
Attrib CAUSE=FAM 0.7871032844445977
Attrib CAUSE=ECON 1.0173945422219859
Sigmoid Node 19
Inputs Weights
Threshold -2.177793647723339
Attrib RANK=MISD -1.722889149528577
Attrib RANK=FEL 4.925431471277383
Attrib RANK=NOTH -1.0526653146201936
Attrib SOLVING=UNSOL 5.796071791461737
Attrib CONDITION=ATT -0.4341478879383603
Attrib NATIONALITY=NAT 9.302832283736718
Attrib NATIONALITY=FOREI -6.0994652972144845
Attrib NATIONALITY=UNKN -1.0087934662134146
Attrib SEX=WOMAN 10.530682674467306
Attrib AGE=AGE1 5.3635422124598255
Attrib AGE=AGE2 -0.3484458924469601
Attrib AGE=AGE3 -7.612501738234174
Attrib AGE=AGE4 6.823481151369662
Attrib CAUSE=NOT-KN -0.15191789810603742
Attrib CAUSE=SENS 1.4412590089971757
Attrib CAUSE=DISE 1.6149950878806163
Attrib CAUSE=FAM 1.8201755721787163
Attrib CAUSE=ECON 1.7506260513028784
Sigmoid Node 20
Inputs Weights
Threshold -0.2577220262808919
Attrib RANK=MISD 3.5673858222860195
Attrib RANK=FEL -2.486562235509235
Attrib RANK=NOTH -0.8827480522877809
Attrib SOLVING=UNSOL 12.930973500671744
Attrib CONDITION=ATT -8.212283582223051
Attrib NATIONALITY=NAT 3.5918789564526086
Attrib NATIONALITY=FOREI -2.5119963611819216
Attrib NATIONALITY=UNKN -0.8492726916959299
Attrib SEX=WOMAN 8.156450623328592
Attrib AGE=AGE1 13.85842872729407
Attrib AGE=AGE2 -4.080309369438002
Attrib AGE=AGE3 -4.797385179645655
Attrib AGE=AGE4 -4.55256967355791
Attrib CAUSE=NOT-KN 0.6148000703173232
Attrib CAUSE=SENS -0.37432107651063656
Attrib CAUSE=DISE 0.09712128604777534
Attrib CAUSE=FAM 0.08374744083606721
Attrib CAUSE=ECON 0.11591724414271047
Sigmoid Node 21

Inputs Weights
Threshold 1.5779819363818388
Attrib RANK=MISD 1.0797097321359577
Attrib RANK=FEL 0.7127366077589701
Attrib RANK=NOTH -3.469566752677553
Attrib SOLVING=UNSOL 9.146525844617878
Attrib CONDITION=ATT -5.089652939358787
Attrib NATIONALITY=NAT 5.45514544344419
Attrib NATIONALITY=FOREI -3.601566284906817
Attrib NATIONALITY=UNKN -3.374835806551886
Attrib SEX=WOMAN -2.7141774345758893
Attrib AGE=AGE1 -4.498616994868366
Attrib AGE=AGE2 -0.03154210707399752
Attrib AGE=AGE3 -0.6882269871900213
Attrib AGE=AGE4 2.116003914547982
Attrib CAUSE=NOT-KN 1.4189134625163078
Attrib CAUSE=SENS -1.648261192429706
Attrib CAUSE=DISE -1.6075426032112201
Attrib CAUSE=FAM -1.3992045655291616
Attrib CAUSE=ECON -1.5233010157655984
Class MANSL
Input
Node 0
Class HOM
Input
Node 1
Class FRAUD
Input
Node 2
Class SUIC
Input
Node 3
Class THEFT
Input
Node 4
Class DIS-TH
Input
Node 5
Class ROB
Input
Node 6
Class COUN
Input
Node 7
Class FORG
Input
Node 8

Time taken to build model: 701.44 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	117862	95.3337 %
Incorrectly Classified Instances	5769	4.6663 %
Kappa statistic	0.762	
Mean absolute error	0.0161	
Root mean squared error	0.0964	
Relative absolute error	32.5836 %	
Root relative squared error	61.2651 %	
Total Number of Instances	123631	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,000	0,000	0,000	0,000	0,000	0,000	0,930	0,026	MANSL
0,250	0,001	0,368	0,250	0,298	0,302	0,990	0,305	HOM
0,172	0,001	0,836	0,172	0,285	0,370	0,739	0,300	FRAUD
0,999	0,000	1,000	0,999	1,000	1,000	1,000	1,000	SUIC
0,997	0,281	0,963	0,997	0,980	0,819	0,911	0,978	THEFT
0,790	0,005	0,770	0,790	0,780	0,776	0,979	0,791	DIS-TH
0,896	0,005	0,887	0,896	0,891	0,886	0,993	0,900	ROB
0,000	0,000	0,000	0,000	0,000	0,000	0,813	0,005	COUN
0,666	0,001	0,852	0,666	0,748	0,751	0,992	0,759	FORG
Weighted Avg. 0,953	0,247	0,949	0,953	0,943		0,804	0,911	0,943

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	<-- classified as
0	1	0	0	10	1	1	0	19	a = MANSL
0	57	3	0	2	131	35	0	0	b = HOM
0	26	772	0	3444	126	40	0	93	c = FRAUD
0	0	0	1106	0	0	1	0	0	d = SUIC
0	8	80	0	108452	63	138	0	0	e = THEFT
0	37	33	0	74	1996	363	0	22	f = DIS-TH
0	21	33	0	243	227	4599	0	11	g = ROB
0	0	0	0	25	7	2	0	8	h = COUN
0	5	2	0	385	41	8	0	880	i = FORG

❖ Αποτέλεσμα αλγόριθμου Apriori (Cross-validation & Folds=10)

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: RapidMinerData-weka.filters.unsupervised.attribute.Remove-R8
Instances: 123631
Attributes: 7
CRIME
RANK
SOLVING
CONDITION
NATIONALITY
SEX
AGE
=== Associator model (full training set) ===

Apriori
=======

Minimum support: 0.85 (105086 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 3

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 9

Size of set of large itemsets L(3): 1

Best rules found:

1. CRIME=THEFT 108741 ==> RANK=MISD 108522 <conf:(1)> lift:(1.07) lev:(0.06) [7289] conv:(34.13)
2. SOLVING=UNSOL NATIONALITY=NAT 106771 ==> SEX=MAN 105331 <conf:(0.99)> lift:(1.04) lev:(0.04) [4509] conv:(4.13)

3. SOLVING=UNSOL 111177 ==> SEX=MAN 109487 <conf:(0.98)> lift:(1.04) lev:(0.04) [4505] conv:(3.66)
4. NATIONALITY=NAT SEX=MAN 108215 ==> SOLVING=UNSOL 105331 <conf:(0.97)> lift:(1.08) lev:(0.06) [8017] conv:(3.78)
5. SOLVING=UNSOL SEX=MAN 109487 ==> NATIONALITY=NAT 105331 <conf:(0.96)> lift:(1.05) lev:(0.04) [5304] conv:(2.28)
6. SOLVING=UNSOL 111177 ==> NATIONALITY=NAT 106771 <conf:(0.96)> lift:(1.05) lev:(0.04) [5200] conv:(2.18)
7. NATIONALITY=NAT 112948 ==> SEX=MAN 108215 <conf:(0.96)> lift:(1.01) lev:(0.01) [1560] conv:(1.33)
8. SOLVING=UNSOL 111177 ==> RANK=MISD 106295 <conf:(0.96)> lift:(1.03) lev:(0.02) [2795] conv:(1.57)
9. RANK=MISD 115094 ==> SEX=MAN 109394 <conf:(0.95)> lift:(1.01) lev:(0.01) [713] conv:(1.12)
10. NATIONALITY=NAT 112948 ==> RANK=MISD 107269 <conf:(0.95)> lift:(1.02) lev:(0.02) [2120] conv:(1.37)

❖ Αποτέλεσμα αλγόριθμου Apriori (MetricType=Lift, NumRules=10)

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 1 -C 1.1 -D 0.1 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: RapidMinerData-weka.filters.unsupervised.attribute.Remove-R8
Instances: 123631
Attributes: 7
CRIME
RANK
SOLVING
CONDITION
NATIONALITY
SEX
AGE
=== Associator model (full training set) ===

Apriori
=======

Minimum support: 0.7 (86542 instances)
Minimum metric <lift>: 1.1
Number of cycles performed: 3

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 21

Size of set of large itemsets L(3): 35

Size of set of large itemsets L(4): 32

Size of set of large itemsets L(5): 13

Size of set of large itemsets L(6): 2

Best rules found:

1. CRIME=THEFT SOLVING=UNSOL 102200 ==> RANK=MISD NATIONALITY=NAT SEX=MAN AGE=AGE2 89811 conf:(0.88) < lift:(1.17)> lev:(0.1) [12944] conv:(2.04)
2. RANK=MISD NATIONALITY=NAT SEX=MAN AGE=AGE2 92985 ==> CRIME=THEFT SOLVING=UNSOL 89811 conf:(0.97) < lift:(1.17)> lev:(0.1) [12944] conv:(5.08)
3. CRIME=THEFT NATIONALITY=NAT SEX=MAN 99047 ==> RANK=MISD SOLVING=UNSOL AGE=AGE2 89811 conf:(0.91) < lift:(1.16)> lev:(0.1) [12471] conv:(2.35)
4. RANK=MISD SOLVING=UNSOL AGE=AGE2 96536 ==> CRIME=THEFT NATIONALITY=NAT SEX=MAN 89811 conf:(0.93) < lift:(1.16)> lev:(0.1) [12471] conv:(2.85)
5. RANK=MISD SOLVING=UNSOL 106295 ==> CRIME=THEFT NATIONALITY=NAT SEX=MAN AGE=AGE2 89811 conf:(0.84) < lift:(1.16)> lev:(0.1) [12274] conv:(1.74)
6. CRIME=THEFT NATIONALITY=NAT SEX=MAN AGE=AGE2 90182 ==> RANK=MISD SOLVING=UNSOL 89811 conf:(1) < lift:(1.16)> lev:(0.1) [12274] conv:(33.99)
7. CRIME=THEFT SOLVING=UNSOL 102200 ==> RANK=MISD CONDITION=ENDED NATIONALITY=NAT SEX=MAN 90006 conf:(0.88) < lift:(1.16)> lev:(0.1) [12117] conv:(1.99)
8. RANK=MISD CONDITION=ENDED NATIONALITY=NAT SEX=MAN 94222 ==> CRIME=THEFT SOLVING=UNSOL 90006 conf:(0.96) < lift:(1.16)> lev:(0.1) [12117] conv:(3.87)
9. RANK=MISD SOLVING=UNSOL CONDITION=ENDED 97331 ==> CRIME=THEFT NATIONALITY=NAT SEX=MAN 90006 conf:(0.92) < lift:(1.15)> lev:(0.1) [12029] conv:(2.64)
10. CRIME=THEFT NATIONALITY=NAT SEX=MAN 99047 ==> RANK=MISD SOLVING=UNSOL CONDITION=ENDED 90006 conf:(0.91) < lift:(1.15)> lev:(0.1) [12029] conv:(2.33)

❖ Αποτέλεσμα αλγόριθμου Apriori (MetricType=Lift, NumRules=3)

=== Run information ===

Scheme: weka.associations.Apriori -N 3 -T 1 -C 1.1 -D 0.1 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: RapidMinerData-weka.filters.unsupervised.attribute.Remove-R8
Instances: 123631
Attributes: 7
CRIME
RANK
SOLVING
CONDITION
NATIONALITY
SEX
AGE
=== Associator model (full training set) ===

Apriori
=======

Minimum support: 0.8 (98905 instances)
Minimum metric <lift>: 1.1
Number of cycles performed: 2

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 19

Size of set of large itemsets L(3): 14

Size of set of large itemsets L(4): 4

Best rules found:

1. RANK=MISD SOLVING=UNSOL 106295 ==> CRIME=THEFT SEX=MAN 101006 conf:(0.95) < lift:(1.13)> lev:(0.09) [11480] conv:(3.17)
2. CRIME=THEFT SEX=MAN 104126 ==> RANK=MISD SOLVING=UNSOL 101006 conf:(0.97) < lift:(1.13)> lev:(0.09) [11480] conv:(4.68)
3. RANK=MISD SOLVING=UNSOL 106295 ==> CRIME=THEFT NATIONALITY=NAT 99006 conf:(0.93) < lift:(1.12)> lev:(0.09) [10744] conv:(2.47)

❖ Αποτέλεσμα αλγόριθμου Predictive Apriori (MetricType=Lift, NumRules=3)

=== Run information ===

Scheme: weka.associations.PredictiveApriori -N 3 -c -1
Relation: RapidMinerData-weka.filters.unsupervised.attribute.Remove-R8
Instances: 123631
Attributes: 7
CRIME
RANK
SOLVING
CONDITION
NATIONALITY
SEX
AGE
=== Associator model (full training set) ===

PredictiveApriori
==========

Best rules found:

1. CRIME=SUIC AGE=AGE2 252 ==> RANK=NOTH 252 acc:(0.99498)
2. CRIME=FORG SEX=WOMAN 243 ==> SOLVING=SOL 243 acc:(0.99497)
3. CRIME=HOM 228 ==> SOLVING=SOL CONDITION=ENDED 228 acc:(0.99496)

❖ Αποτέλεσμα αλγόριθμου SimpleKMeans

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation: RapidMinerData
Instances: 123631
Attributes: 8
CRIME
RANK
SOLVING
CONDITION
NATIONALITY
SEX
AGE
CAUSE
Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 70456.0

Initial starting points (random):

Cluster 0: THEFT,MISD,UNSOL,ENDED,NAT,MAN,AGE2,NOT-KN
Cluster 1: THEFT,MISD,UNSOL,ATT,NAT,MAN,AGE2,NOT-KN

Missing values globally replaced with mean/mode

Final cluster centroids:

	Cluster#		
Attribute	Full Data	0	1
	(123631.0)	(112897.0)	(10734.0)
CRIME	THEFT	THEFT	THEFT
RANK	MISD	MISD	MISD
SOLVING	UNSOL	UNSOL	UNSOL
CONDITION	ENDED	ENDED	ATT
NATIONALITY	NAT	NAT	NAT
SEX	MAN	MAN	MAN
AGE	AGE2	AGE2	AGE2
CAUSE	NOT-KN	NOT-KN	NOT-KN

Time taken to build model (full training data) : 0.28 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 112897 (91%)
1 10734 (9%)