Μελέτη της δυνατότητας, αλλά και της αξίας της εξόρυξης δεδομένων, σε βάσεις δεδομένων που χειρίζονται οι Υγειονομικές Υπηρεσίες

Γιώργος Σάββα

ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΚΑΙ ΕΦΑΡΜΟΣΜΕΝΩΝ ΕΠΙΣΤΗΜΩΝ

Μάιος 2011

Ημερομηνία 29.05.2011 Έκδοση 3.6

Filename : George Savva Thesis 29.05.2011 Edition 3.6.docx

Περιεχόμενα

Περίληψη

Η παρούσα διπλωματική εργασία αναπτύχθηκε στα πλαίσια του μεταπτυχιακού προγράμματος σπουδών της ειδίκευσης “Πληροφοριακά Συστήματα” του Ανοικτού Πανεπιστημίου Κύπρου. Το θέμα της εργασίας είναι «Μελέτη της δυνατότητας, αλλά και της αξίας της εξόρυξης δεδομένων, σε βάσεις δεδομένων που χειρίζονται οι Υγειονομικές Υπηρεσίες». Η παρούσα εργασία παρουσιάζει συνοπτικά την τεχνική που ονομάζεται Εξόρυξη Δεδομένων και εξετάζει την εφαρμογή της μεθόδου αυτής στο χώρο της Δημόσιας Υγείας στην Κύπρο. Η καινοτομία της παρούσας εργασίας έγκειται στη μελέτη για πρώτη φορά τυχών χρήσιμων εφαρμογών της μεθόδου Εξόρυξης Δεδομένων στο χώρο της Δημόσιας Υγείας και συγκεκριμένα στο χώρο της Περιβαλλοντικής Υγιεινής στην Κύπρο.

Η υπό αναφορά εργασία περιλαμβάνει πραγματικά περιστατικά εφαρμογής της εξόρυξης δεδομένων σε δεδομένα της Κύπρου. Έγινε αξιολόγηση των βάσεων δεδομένων που υπάρχουν και μελετήθηκε η δυνατότητα ανακάλυψης γνώσης μέσα από τα δεδομένα που υπάρχουν αποθηκευμένα σε αυτές και έγινε εξόρυξη δεδομένων και ανακάλυψη γνώσης σε δεδομένα του προγράμματος παρακολούθησης του διασωληνωμένου πόσιμου νερού με τη χρήση του λογισμικού WEKA.

Από τα συμπεράσματα φαίνεται ότι πρόκειται για μια ανεκμετάλλευτη πηγή από την οποία, οι Αρχές που ασχολούνται με την ασφάλεια των τροφίμων και του νερού, μπορούν να εξάγουν χρήσιμη πληροφόρηση. Η υπό αναφορά πληροφόρηση βρέθηκε ότι μπορεί να αποκαλύψει τάσεις και μοτίβα στην εμφάνιση περιστατικών όπου απειλείται η δημόσια υγεία και έτσι να επιτρέψει στις Αρχές να αντεπεξέλθουν αποτελεσματικά σε περιστατικά που αποτελούν κίνδυνο για τη δημόσια υγεία, αναδιοργανώνοντας και επικεντρώνοντας τους πόρους τους ή ακόμη επιτρέποντάς τους να δράσουν προληπτικά/προδραστικά.

Abstract

The current Master thesis was developed in the context of the MSc program "Information Systems" run by the Cyprus Open University. The subject is «Study of the applicability and value of using Data Mining in databases that are kept by the Public Health Services in Cyprus".

The work deals with implementing data mining into real data from Cyprus. An evaluation of the databases, currently held was done and the potential of knowledge discovery from the data held there was done. Moreover data mining and knowledge discovery techniques were applied to a set of data concerning the monitoring of the quality of the drinking water through the software WEKA.

From the conclusions it seems that this is an unexploited source from where authorities that deal with food and water safety can yield useful information. The information in question was found that it can reveal tendencies and motifs in the appearance of incidences where the public health is under threat and thus allow the authorities to handle imminent incidents by reorganizing and focusing their resources or by even acting proactively.

Πρόλογος

Η πρόταση για επιλογή ενός θέματος που θα προερχόταν από τον τομέα της Εξόρυξης Δεδομένων έγινε στα τέλη του 2008 σε συνεργασία με τον επιβλέποντα καθηγητή μου Δρ Θανάση Χατζηλάκο, λαμβάνοντας υπόψη τις ακαδημαϊκές μου σπουδές και τα ενδιαφέροντά μου τα οποία εμπίπτουν σε τρεις κατευθύνσεις: (α) τη Δημόσια και Περιβαλλοντική Υγιεινή, (β) τους Υπολογιστές και (γ) τη Διαχείριση της Πληροφορίας.

Αφορμή αποτέλεσε μια ενότητα στο βιβλίο των Navathe και Elmasri, (2006) σχετικά με την εξόρυξη δεδομένων και μια άσκηση σε ένα λογισμικό εξόρυξης δεδομένων. Αρχικά, αντιμετώπισα με κάποια επιφυλακτικότητα την προοπτική ενασχόλησής μου με ένα χώρο που συναντούσα για πρώτη φορά. Παρ' όλο που η αναφορά του βιβλίου στην Εξόρυξη Δεδομένων είναι πολύ σύντομη, ο συγγραφέας φροντίζει να τονίσει ότι πρόκειται για ένα πρόσφατο κλάδο της επιστήμης που βρίσκεται υπό εξέλιξη.

1. Εισαγωγή

Σχέση πληροφορίας και γνώσης

Το Oxford Dictionary, για να δώσει τη σχέση μεταξύ πληροφορίας και γνώσης, αναφέρει ότι, η γνώση προέρχεται από πληροφορίες οι οποίες εντοπίζονται, επιλέγονται, επεξεργάζονται και μεταποιούνται σε γνώση. Με αυτή τη λογική, η πληροφορία αποτελεί την πρώτη ύλη για τη γνώση (Oxford Dictionary).

Τεκμηριωμένη λήψη και εφαρμογή αποφάσεων στη φροντίδα της υγείας

Η αξία της γνώσης και πληροφορίας είναι αδιαμφισβήτητη και έχει λάβει τέτοια αναγνώριση που έχει δημιουργηθεί επιστήμη της βιβλιοθηκονομίας/διαχείρισης της γνώσης που ασχολείται αποκλειστικά με την εύρεση, οργάνωση διανομή και αξιοποίηση τους. Η σημασία αυτών των δύο στοιχείων (πληροφορίας και γνώσης) και η αξιοποίηση της σε όλους τους τομείς της ζωής έχει πρόσφατα αναγνωριστεί σε μεγάλο βαθμό. Στους χώρους εργασίας, η πληροφορία και η γνώση αποτελούν βασικά στοιχεία για διασφάλιση της αποδοτικότητας και αποτελεσματικότητας.

Οι παραδοσιακοί τρόποι άσκησης των καθηκόντων βασίζονταν στις απηρχαιωμένες γνώσεις, τη ρουτίνα, την εμπειρία και την προαίσθηση. Αυτή η πρακτική αποδείχθηκε ανεπαρκής και επικίνδυνη για τη φροντίδα της υγείας και γι' αυτό αντικαταστάθηκε με το σύγχρονο σύστημα της τεκμηριωμένης λήψης και εφαρμογής αποφάσεων (Evidence Based Practice). Είναι μια κίνηση η οποία

άρχισε το 1992 και έθεσε τη φροντίδα υγείας στις ορθές της διαστάσεις. Στα πλαίσια αυτής της κίνησης, οι λειτουργοί, **είναι υποχρεωμένοι να** επιλέγουν ενσυνείδητα τις πλέον πρόσφατες, τεκμηριωμένες (έγκυρες) πληροφορίες και γνώσεις και με βάση αυτές να παίρνουν και να εφαρμόζουν τις αποφάσεις τους κατά την άσκηση των καθηκόντων τους (Scott, Heyworth, & Fairweather, 2000). Με το ίδιο σκεπτικό, οι Castillo και Abraham (2008) επισημαίνουν ότι, χωρίς τη συστηματική εξασφάλιση και κριτική αξιολόγηση τεκμηριωμένων πληροφοριών/γνώσεων που πηγάζουν μέσα από κλινικές έρευνες, οι επαγγελματίες υγείας θα είναι υποχρεωμένοι να βασίζουν τα καθήκοντά τους σε ακατάλληλες, απηρχαιωμένες πληροφορίες/γνώσεις (Castillo & Abraham, 2008).

Ο Verhoeven, (1996) υποστηρίζει ότι οι έμπειροι ιατρικοί λειτουργοί πρέπει να χρησιμοποιούν γύρω στα δύο εκατομμύρια πληροφορίες/γνώσεις για να μπορέσουν να διαχειριστούν/φροντίσουν σωστά τους ασθενείς τους. Υπό αυτή την έννοια, η κλινική πληροφορία/γνώση θα μπορούσε να οριστεί ως “τα αγαθά που είναι απαραίτητα στοιχεία που βοηθούν στη λήψη τεκμηριωμένων αποφάσεων για τη φροντίδα των ασθενών” (Wyatt, 1996). Συνεπώς, οι επαγγελματίες υγείας δεν μπορούν να εφαρμόσουν ποιοτική φροντίδα χωρίς να αναβαθμίζουν συνεχώς τις πληροφορίες/γνώσεις που βρίσκονται στο μυαλό τους (González, 2007). Δεν μπορούν να βασίζουν τη φροντίδα της υγείας σε ξεπερασμένες, επικίνδυνες πληροφορίες/γνώσεις. Σύμφωνα με τον καθηγητή του Harvard University Dr Sydney Burwell, το ήμισυ των γνώσεων που αποκτούν οι μαθητές κατά την εκπαίδευση, σε 10 χρόνια θα είναι ξεπερασμένες και επικίνδυνες. Το πρόβλημα είναι ότι κανένας καθηγητής δεν είναι σε θέση να καθορίσει ποιες είναι αυτές που είναι ξεπερασμένες και επικίνδυνες. Άρα πρέπει να ανανεώνονται όλες συστηματικά.

Ένα άλλο πρόβλημα που αντιμετωπίζουν οι λειτουργοί είναι η επιλογή των πιο κατάλληλων πληροφοριών/γνώσεων για κάθε δεδομένη ανάγκη. Αυτό οφείλεται στον τεράστιο όγκο των πληροφοριών/γνώσεων που υπάρχει στις μέρες μας και που συνεχώς αυξάνεται. Σύμφωνα με την Breivik (1993), ο συνολικός όγκος των

πληροφοριών/γνώσεων διπλασιάστηκε από το 1750 μέχρι το 1900. Σήμερα, υπολογίζεται ότι διπλασιάζονται κάθε 2-3 χρόνια. Με το ρυθμό που αυξάνονται, προβλέπεται ότι, μέχρι το 2020 θα διπλασιάζονται κάθε 73 μέρες (Breivik, 1993).
Η Breivik (1993), τονίζει ότι, η δημιουργία ηλεκτρονικών συστημάτων διαχείρισης της πληροφορίας/γνώσης δεν αποτελεί επιθυμητό στόχο αλλά πρακτική επιβίωσης και ανάπτυξης.

Χαρακτηριστικά του προσωπικού των Υγειονομικών Υπηρεσιών

Λόγω της φύσης του επαγγέλματός τους, οι Υγειονομικοί λειτουργοί ασχολούνται με το πιο σημαντικό αγαθό των πολιτών, την υγεία, γι αυτό και χρειάζονται τις πληροφορίες και τις γνώσεις που μπορούν να τους βοηθήσουν ώστε να διεκπεραιώσουν την εργασία τους πιο αποτελεσματικά και αποδοτικά.Είναι σε αυτά τα πλαίσια που αναζητούνται τεχνικές εύρεσης χρήσιμων πληροφοριών και αξιοποίηση τους.

Βάσεις Δεδομένων και Πληροφορίες

Η ανάπτυξη των ηλεκτρονικών υπολογιστών από τη δεκαετία του ‘60 και έπειτα έδωσε ώθηση στην ευρεία χρήση τους σε όλο το φάσμα της ζωής μας ιδιαίτερα στους τομείς της οικονομίας. Η συλλογή στοιχείων μέσω των μηχανών έχει μετατραπεί σε μια εύκολη υπόθεση και η αποθήκευσή τους, που γίνεται με ψηφιακό τρόπο, έχει μειώσει το φυσικό όγκο που αυτά καταλαμβάνουν. Υπάρχει λοιπόν η ευχέρεια να δημιουργούμε τεράστιες και συνεχώς αυξανόμενες βάσεις δεδομένων. Η δημιουργία βάσεων δεδομένων, τις τρεις τελευταίες δεκαετίες, έχει δημιουργήσει τεράστια αποθέματα δεδομένων, που ωστόσο μεγάλο μέρος τους δεν επεξεργάζονται και ερμηνεύονται, για να αντληθούν και αξιοποιηθούν όλες οι δυνατές πληροφορίες (Navathe & Elmasri, 2006).

Στον επιχειρησιακό κόσμο υπάρχει η γενική παραδοχή ότι πέραν από το ανθρώπινο δυναμικό το σημαντικότερο κτήμα μιας επιχείρησης είναι οι πληροφορίες που διαθέτει, έτσι πολλές επιχειρήσεις δαπανούν μεγάλο μέρος του προϋπολογισμού τους στην εξεύρεση πληροφοριών. Συγκεκριμένα,

δαπανούνται μεγάλα ποσά για την εξεύρεση πληροφοριών μέσω ερευνών αγοράς, αγοράζοντας δεδομένα και με πολλούς άλλους ευφάνταστους τρόπους. Τότε ηγέρθηκε το ερώτημα, μήπως μπορεί να αντλήσουμε χρήσιμες πληροφορίες από στοιχεία που διατηρούμε στην κατοχή μας μετά από κατάλληλη επεξεργασία;

Ανάγκη για ανακάλυψη γνώσης μέσα από τις πληροφορίες που βρίσκονται στις βάσεις δεδομένων

Από τα πιο πάνω προέκυψε η ανάγκη για δημιουργία τεχνικών που να ανιχνεύουν χρήσιμη πληροφόρηση μέσα από μεγάλο όγκο δεδομένων. Για τον σκοπό αυτό έχουν γίνει διάφορες προσεγγίσεις στο θέμα μέσα από τον κλάδο της στατιστικής, των βάσεων δεδομένων και της τεχνητής νοημοσύνης. Οι διάφορες προσεγγίσεις συνοψίζονται κάτω από τη γενική και ευρεία έννοια που αποκαλείται «Εξόρυξη Δεδομένων»(Data Mining). Το "Data Mining" είναι ένα εργαλείο που συνδυάζει στατιστική, εκμάθηση μηχανής, αλγόριθμους ομαδοποίησης (clustering), μεθόδους οπτικοποίησης (visualization) και βάσεις δεδομένων.

Η πληροφορία που βρίσκεται κρυμμένη σε ένα σύνολο δεδομένων αναγνωρίστηκε πλέον ως σημαντική και καταβλήθηκε κάθε δυνατή προσπάθεια εξόρυξης και αξιοποίησής της. Για παράδειγμα, η πληροφορία ότι η πλειοψηφία των πελατών ενός καταστήματος που αγόρασαν το προϊόν Α αγόρασαν και το προϊόν Β, αν αξιοποιηθεί από την διεύθυνση του καταστήματος μπορεί να αυξήσει τις πωλήσεις. Έτσι η εξόρυξη δεδομένων μετατράπηκε σε ένα αναντικατάστατο εργαλείο κάθε επιτυχημένης εφαρμογής, σύμφωνα με το Gartner Report όπως αναφέρεται από τους Navathe και Elmasri (2006).

Μετά από σύντομη μελέτη του θέματος διαφάνηκε ότι η εξόρυξη δεδομένων απασχολούσε πολλούς επιστήμονες, αλλά και ταυτόχρονα είχε και άμεσες βιομηχανικές εφαρμογές και είναι ένας από τους πιο ραγδαία αναπτυσσόμενους τομείς έρευνας. Η εξόρυξη δεδομένων παρόλο που έχει ευρύ φάσμα και

πολλαπλές εφαρμογές κυρίως στα οικονομικά (ανάλυση δεδομένων εταιρειών για θέματα πωλήσεων, διαχείριση ρίσκου, τάσεις καταναλωτών, ανίχνευση απάτης κτλ) τυγχάνει πολύ μικρής χρήσης στη Δημόσια Υγεία.

Η εξόρυξη δεδομένων στη Δημόσια Υγιεινή στην Κύπρο

Στην Κύπρο όπως και σε πολλές άλλες χώρες διατηρούμε σε ηλεκτρονική μορφή διάφορα δεδομένα σχετικά με τη Δημόσια Υγιεινή, τα οποία λόγω της συσσώρευσης αλλά και των αυξημένων ελέγχων έχουν αποκτήσει τεράστιο όγκο. Τα δεδομένα αυτά, αν και μέσω διάφορων στατιστικών προσεγγίσεων γίνεται προσπάθεια αξιοποίησης τους προς όφελος της δημόσιας υγείας, δεν έχουν αναλυθεί μέσω της μεθόδου «Εξόρυξη Δεδομένων» παρόλο που σε άλλες χώρες υπήρξαν περιπτώσεις που έγιναν τέτοιες προσεγγίσεις με πολύ ενδιαφέροντα αποτελέσματα όπως περιγράφονται στο ¨Possibilities for Applying Data Mining for Early Warning in Food Supply Networks¨ των Beulens, Kramer και Vorst(2006).

Το ερώτημα που προκύπτει είναι αν υπάρχουν κάποιες κρυμμένες πληροφορίες στα πιο πάνω δεδομένα που με την αποκάλυψη τους μπορούν να ωφελέσουν την δημόσια υγεία. Συγκεκριμένα, οι κρυμμένες πληροφορίες πιθανό να είναι χρήσιμες στην αναγνώριση περιοχών όπου είτε χρονικά ή θεματικά αναμένεται να υπάρξει πρόβλημα, επιτρέποντας στις αρμόδιες αρχές να αξιοποιούν καλύτερα τους πόρους τους εστιάζοντας την προσοχή τους στις περιπτώσεις των προβλέψιμων αναδυόμενων κινδύνων.

2. Η υποκείμενη τεχνολογία: Εισαγωγή για τον Υγειονομικό Λειτουργό

Οι βάσεις δεδομένων χρησιμοποιούνται εδώ και δεκαετίες με αποτέλεσμα να συσσωρεύουμε ψηφιακά αποτελέσματα τεραστίων μεγεθών. Μέσα από αυτά τα δεδομένα υπάρχουν κάποια κρυφά νοήματα, επεξεργασμένα δεδομένα με

εννοιολογική σημασία τα οποία είναι χρήσιμα. Ωστόσο η εξόρυξη των εννοιών αυτών είναι πολύ δύσκολη λόγω των μεγεθών και της πολυπλοκότητας της δόμησης των δεδομένων αυτών. Γι αυτό χρησιμοποιώντας αρχές από τα πεδία της στατιστικής, της μηχανικής εκμάθησης, της θεωρίας της πληροφορίας και των υπολογιστικών διαδικασιών, έχει δημιουργηθεί μια νέα επιστήμη με δυναμικά εργαλεία η οποία καλείται «Εξόρυξη Δεδομένων (ΕΔ)» (Data Mining) και είναι μέρος της διαδικασίας «Ανακάλυψης Γνώσης από Βάσεις Δεδομένων» (Knowledge Discovery in Databases - KDD). Τα εργαλεία της ΕΔ είναι οι αλγόριθμοί της, οι οποίοι επιχειρούν να βρουν χρήσιμα και κατανοητά πρότυπα στα δεδομένα.

2.1. Αποθήκες Δεδομένων

Σύμφωνα με τον Inmon, (1992) μια «Αποθήκη Δεδομένων» είναι μια συλλογή δεδομένων ενός οργανισμού που χρησιμοποιείται κυρίως για τη λήψη αποφάσεων, έχει θεματικό προσανατολισμό και ολοκληρωμένα δεδομένα, τα οποία διατηρούνται χρονικά.

Όλες οι αποθήκες δεδομένων έχουν ένα σχήμα το οποίο είναι το σύνολο των εννοιών που περιγράφουν τη δομή τους και είναι σχεδόν πάντα σταθερό. Τα δεδομένα είναι αποθηκευμένα στα προκαθορισμένα πεδία της βάσης και αποτελούν τις πηγές από τις οποίες η Αποθήκη Δεδομένων αντλεί δεδομένα.

Οι διαφορές της βάσης δεδομένων με την αποθήκη δεδομένων εντοπίζονται σε δέκα στο τεχνικό/κατασκευαστικό επίπεδο από τους Velicanu και Matei(2007), στο άρθρο τους «Database versus Data Warehouse». Κατά τη γνώμη μου οι διαφορές μεταξύ των δύο, μπορούν να αναχθούν σε διαφορές που σχετίζονται με το στόχο χρήσης του κάθε λογισμικού, η βάση δεδομένων έχει ως σκοπό να καταγράψει ενώ η αποθήκη δεδομένων έχει ως σκοπό να αποκριθεί στις ερωτήσεις ανάλυσης, που είναι κρίσιμες για έναν οργανισμό.

2.1.1 Βάσεις Δεδομένων

Οι Navathe και Elmasri, (2006) δίνουν ίσως τον πιο απλό ορισμό της βάσης δεδομένων, ορίζοντας την ως μια συλλογή δεδομένων που συσχετίζονται μεταξύ τους. Σαν δεδομένα ορίζονται γνωστά στοιχεία, τα οποία μπορούν να καταγραφούν και να έχουν αναμφισβήτητο νόημα.

Όσον αφορά τη σχεδίαση και απεικόνιση βάσεων δεδομένων, υπάρχουν διάφορες προσεγγίσεις που κύριος στόχος τους είναι να παρέχουν μια αφηρημένη όψη των δεδομένων, αποκρύπτοντας από το χρήστη λεπτομέρειες σχετικά με την αναπαράσταση και την αποθήκευσή τους. Συνήθως τα δεδομένα αναπαριστώνται σε 3 επίπεδα αφαίρεσης (abstraction levels):

- Το εσωτερικό επίπεδο (internal level), που είναι το χαμηλότερο επίπεδο αφαίρεσης, όπου περιγράφεται με λεπτομέρεια η αποθήκευση των δεδομένων και οι τρόποι προσπέλασης σε αυτά (Navathe & Elmasri, 2006).
- Το εννοιολογικό επίπεδο (conceptual level), το οποίο περιγράφει τη δομή ολόκληρης της Β.Δ. και αναπαριστά τα δεδομένα και τις μεταξύ τους σχέσεις (Navathe & Elmasri, 2006).
- Το Εξωτερικό επίπεδο ή επίπεδο όψεων (external ή view level), που είναι το υψηλότερο επίπεδο αφαίρεσης σύμφωνα με το άρθρο "The Challenge of Knowledge Soup" του John F. Sowa (2006), και προσφέρει την όψη που παρουσιάζεται στα λογισμικά εφαρμογής για τη βάση δεδομένων.

Οι βάσεις δεδομένων μεταβάλλονται με την πάροδο του χρόνου, καθώς προστίθενται νέες πληροφορίες και αφαιρείται ή τροποποιείται το στιγμιότυπο (database or snapshot) της βάσης, το σύνολο δηλαδή της πληροφορίας το οποίο βρίσκεται αποθηκευμένο σε μια βάση δεδομένων σε μια συγκεκριμένη χρονική στιγμή (Navathe & Elmasri, 2006).

Τα σημερινά συστήματα διαχείρισης βάσεων δεδομένων, όπως για παράδειγμα η γλώσσα σχεσιακών βάσεων δεδομένων Stuctured Querry Language (SQL) παρέχουν μια ολοκληρωμένη γλώσσα η οποία περιλαμβάνει δομικά στοιχεία για τη δημιουργία του σχήματος και των όψεων, καθώς και για τον χειρισμό δεδομένων. Επίσης, τα σημερινά συστήματα παρέχουν τα εργαλεία, για να γίνουν καταχωρήσεις, αναζητήσεις, ενώ μπορούν να αποτελέσουν και δίαυλο επικοινωνίας με το χρήστη ή να του επιτρέψουν να ενωθεί μέσω άλλου λογισμικού, που να εξυπηρετεί σαν διαδραστικό μέσο αλληλεπίδρασης μεταξύ του χρήστη και της βάσης.

2.2 Ανακάλυψη γνώσης σε βάσεις δεδομένων

Το Γενικότερο Πλαίσιο

Η «ανακάλυψη γνώσης» σε βάσεις δεδομένων (*Knowledge Discovery in Databases - KDD*) είναι η διαδικασία για τον προσδιορισμό έγκυρων, νέων, χρήσιμων και κατανοητών σχέσεων-προτύπων σε δεδομένα. Πρόκειται για μια μεγάλη διαδικασία που αποτελεί μια σημαντική εφαρμογή σε πραγματικές συνθήκες και σε μεγάλη κλίμακα των ερευνητικών αποτελεσμάτων της Στατιστικής, των Βάσεων Δεδομένων και της Τεχνητής Νοημοσύνης. Σύμφωνα με τους *Fayyad, Piatetsky-Shapiro, και Smyth (1996)*, η ανακάλυψη γνώσης σε βάσεις δεδομένων σε ένα αφηρημένο επίπεδο ασχολείται με την ανάπτυξη μεθόδων και τεχνικών που σκοπό έχουν την εξαγωγή νοήματος από δεδομένα. Η ανακάλυψη γνώσης είναι μια ολοκληρωμένη διαδικασία που περιλαμβάνει αρκετά στάδια που αφορούν την επεξεργασία των δεδομένων, την εφαρμογή των αλγορίθμων ανακάλυψης γνώσης και την ερμηνεία των αποτελεσμάτων μεταξύ των οποίων είναι και η εξόρυξη δεδομένων.

Από την πιο πάνω προσέγγιση διαφαίνεται να υπάρχει μια διφορούμενη άποψη για το τι είναι «Ανακάλυψη γνώσης» σε βάσεις δεδομένων και τι «εξόρυξη δεδομένων». Σε μεγάλο μέρος της βιβλιογραφίας ο ορισμός της εξόρυξη

δεδομένων περιορίζεται στο υπολογιστικό μέρος της όλης διαδικασίας, ωστόσο πλέον φαίνεται να έχουν συνυφαστεί οι δύο όροι και σήμερα ο όρος εξόρυξη δεδομένων έχει επικρατήσει να χρησιμοποιείται, για να περιγράψει ολόκληρη τη διαδικασία ανακάλυψης γνώσης.

Οι συμβατικές προσεγγίσεις στατιστικών αναλύσεων που αφορούν ανάλυση ενός περιορισμένου δείγματος δεν μπορούν να εφαρμοστούν αποτελεσματικά σε βάσεις δεδομένων με πολύ μεγάλο όγκο, που παράγονται στις μέρες μας (Karimipouri, Delavari, & Kinaie, 2005). Έτσι προκύπτει αναντίρρητα η ανάγκη χρήσης άλλων μεθόδων ανάλυση που μπορούν να χειριστούν μεγάλο όγκο δεδομένων και να αποκαλύψουν τις κρυμμένες σχέσεις και μελλοντικές τάσεις που κρύβονται σε αυτά. Η μέθοδος της Εξόρυξης Δεδομένων έρχεται να λύσει αυτό το πρόβλημα και να προσφέρει λύσεις στο χειρισμό πολύπλοκων και μεγάλων βάσεων δεδομένων.

Μια άλλη τεχνική για την οποία υπάρχει σύγχυση με την ανακάλυψη γνώσης είναι η «Μηχανική Μάθησης». Στην περίπτωση αυτή είναι δυο διακριτές τεχνικές που μπορεί να προσομοιάζουν αλλά διαφέρουν τελείως στον αντικειμενικό τους στόχο αφού η Μηχανική Μάθησης προσπαθεί να ανακαλύψει τεχνικές μάθησης ως μίμηση της ανθρώπινης συμπεριφοράς δηλαδή την παραγωγή προγραμμάτων που να μαθαίνουν, ενώ η ανακάλυψη γνώσης στοχεύει στην ανακάλυψη πληροφορίας που θα είναι χρήσιμη στον άνθρωπο. (Dunham, 2003).

2.3 Εισαγωγή στις βασικές έννοιες της εξόρυξης δεδομένων

Σύμφωνα με τον Ackoff (1989) οι συνοπτικοί ορισμοί των δεδομένων, της πληροφορίας, της γνώσης και της κατανόησης είναι οι ακόλουθοι:

1. **Δεδομένα** : σύμβολα
2. **Πληροφορίες** : δεδομένα τα οποία επεξεργάζονται, για να γίνουν χρήσιμα. Δίνονται απαντήσεις στις ερωτήσεις: «ποιος;», «πού;» και «πότε;»

3. **Γνώση**: Εφαρμογή των δεδομένων και πληροφοριών. Δίνει απάντηση στην ερώτηση «Πώς;»
4. **Κατανόηση**: εκτίμηση των «Γιατί;»
5. **Σοφία**: αποτιμημένη κατανόηση.

Ορισμός Εξόρυξης Δεδομένων

Όσον αφορά την εξόρυξη δεδομένων (Data mining) μιας και είναι σχετικά σύγχρονη έννοια σαν πιο δόκιμος ορισμός προτείνεται αυτός των Navathe και Elmasri (2006) οι οποίοι αναφέρουν ότι εξόρυξη γνώσης αποκαλείται η εξεύρεση νέων *πληροφοριών από επαναλαμβανόμενα πρότυπα* (patterns) ή κανόνες (rules) σε μεγάλους όγκους δεδομένων.

Με άλλα λόγια, η εξόρυξη δεδομένων μπορεί να θεωρηθεί ως μια προσέγγιση για να καθορίσει *έγκυρα, νέα, χρήσιμα* και *κατανοητά* μοτίβα δεδομένων που προκύπτουν από μεγάλου όγκου βάσεις δεδομένων (Miller and Han, 2001). Ο όρος *έγκυρα* αναφέρεται στην ικανότητα των εξαγόμενων μοτίβων να εφαρμοστούν σε νέες βάσεις δεδομένων, πέραν των δεδομένων από τα οποία έχουν εξαχθεί. Ο όρος *νέα* αναφέρεται στο γεγονός ότι τα μοτίβα που προκύπτουν είναι απρόβλεπτα. Ο όρος *χρήσιμα* αναφέρεται στη δυνατότητα των εξαγόμενων μοτίβων να χρησιμοποιηθούν σε μελλοντικές δραστηριότητες, καθώς η μέθοδος Εξόρυξης Δεδομένων χρησιμοποιείται ως μέσο για να στηρίξει συστήματα λήψης αποφάσεων. Τέλος, ο *όρος* κατανοητά αναφέρεται στο γεγονός ότι οι σχέσεις που προκύπτουν από την ανάλυση πρέπει να είναι απλές και ερμηνεύσιμες (Karimipouri, Delavari, & Kinaie, 2005).

2.4. Παρουσίαση των βασικών βημάτων της διαδικασίας ανακάλυψης γνώσης από βάσεις δεδομένων

Οι Han και Kamber (2006) έχουν κωδικοποιήσει την διαδικασίας ανακάλυψης γνώσης από βάσεις δεδομένων σε επτά στάδια ως ακολούθως :

I. Καθαρισμός δεδομένων

Στο στάδιο αυτό απομακρύνονται ασυνεπή, αντιφατικά στοιχεία και στοιχεία τα οποία γενικά δυσχεραίνουν την εξόρυξη δεδομένων χωρίς να προσφέρουν στην διαδικασία εξόρυξης δεδομένων.

II. Ενοποίηση των δεδομένων

Δεδομένα από διάφορες πηγές ενσωματώνονται σε ένα ενιαίο σύνολο για να μπορέσουν να εξαχθούν συμπεράσματα από το συνδυασμό δεδομένων.

III. Επιλογή Δεδομένων

Δημιουργείται το σύνολο δεδομένων στο οποίο θα εφαρμοστεί η αναζήτηση (*training dataset selection*) με επιλογή στοιχείων (πινάκων, πεδίων) από σχεσιακές βάσεις δεδομένων .

IV. Αλλαγή μορφής δεδομένων

Αλλαγή της μορφής των δεδομένων, για να μπορούν να επεξεργαστούν ή αλλαγή σε μορφή, που να υποβοηθά μια συγκεκριμένη τεχνική.

V. Επιλογή αλγορίθμου εξόρυξης δεδομένων και εφαρμογή του

Καθορίζεται τι είδους γνώση θα αναζητηθεί, κάτι που έμμεσα προσδιορίζει και την κατηγορία αλγορίθμου που θα χρησιμοποιηθεί.

Τα παράγωγα της διαδικασίας ανακάλυψης γνώσης μπορεί να είναι:

- πρότυπα πληροφόρησης - informative patterns *(μάθηση χωρίς επίβλεψη)*
- μοντέλα πρόβλεψης - predictive models *(μάθηση με επίβλεψη)*.

Στο στάδιο αυτό εφαρμόζονται έξυπνες τεχνικές για ανίχνευση μοτίβων.

Είναι ένα καθαρά υπολογιστικό στάδιο, στο οποίο γίνεται η ουσιαστική αναζήτηση της γνώσης στα δεδομένα (**εξόρυξη σε δεδομένα).**

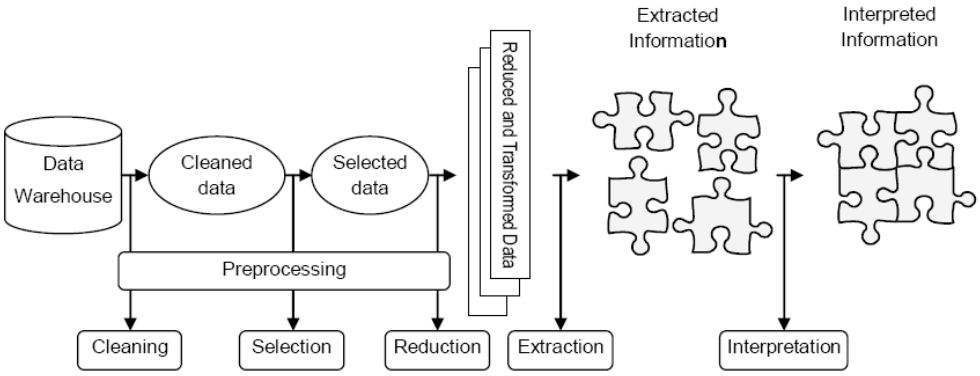
VI. Αξιολόγηση των μοτίβων

Γίνεται *ερμηνεία* και *αξιολόγηση* των ευρεθέντων προτύπων. Αναγνωρίζονται τα πραγματικά ενδιαφέροντα μοτίβα τα οποία με την ερμηνεία τους θα αποδώσουν γνώση.

VII. Παρουσίαση της γνώσης

Στάδιο όπου χρησιμοποιούνται τεχνικές απεικόνισης και παρουσίασης της γνώσης από τις εξορυγμένες γνώσεις με υποβοήθηση γραφικών απεικονίσεων των προτύπων ή/και των δεδομένων που περιγράφονται από το πρότυπο (*pattern/data visualization*).

Υπάρχουν διάφορες προσεγγίσεις για το θέμα και μια άλλη προσέγγιση από τους Karimipouri, Delavari, & Kinaie (2005) παρουσιάζεται στο Σχήμα 1 έτσι όπως δημοσιεύτηκε στο άρθρο του.



Σχήμα 1. Στάδια εξόρυξης Δεδομένων (Karimipouri, Delavari, & Kinaie, 2005)

3. Τύποι γνώσης που μπορούν να ανακαλυφθούν με την εξόρυξη δεδομένων και τεχνικές εξόρυξης δεδομένων

Υπάρχουν αρκετές προσεγγίσεις/αλγόριθμοι στην εξόρυξη δεδομένων οι οποίες εμπίπτουν στις πιο κάτω κατηγορίες (Navathe & Elmasri, 2006):

1. Συσταδοποίηση - Clustering: εύρεση ενός συνόλου από ομάδες με όμοια στοιχεία.
2. Ταξινόμηση - Classification: εκμάθηση μιας συνάρτησης – κατασκευή ενός μοντέλου που απεικονίζει τη σχέση των στοιχείων σε ένα σύνολο από προκαθορισμένες κλάσεις.
3. Εύρεση Συχνών Προτύπων, Εξαρτήσεων και Συσχετίσεων – Dependencies and associations: εύρεση σημαντικών/συχνών εξαρτήσεων μεταξύ γνωρισμάτων.
4. Συνοψίσεις - Summarization: εύρεση μιας συνοπτικής περιγραφής του συνόλου δεδομένων ή ενός υποσυνόλου του.

Σύμφωνα με τους Kamber και Han (2006) υπάρχουν δύο λειτουργίες που μπορούν να εκτελεστούν με την εξόρυξη δεδομένων, η μια είναι η περιγραφική και η άλλη η προφητική. Η περιγραφική χαρακτηρίζει τις γενικές ιδιότητες των δεδομένων της βάσης ενώ η προφητική αποστολή παρεμβαίνει στα δεδομένα και εξάγει συμπεράσματα λογικής ανάλυσης στα παρών δεδομένα, για να μπορεί να διεξάγει προβλέψεις.

Σύμφωνα με τα πιο πάνω κάποιες μέθοδοι εξόρυξης δεδομένων μπορούν να διακριθούν ως περιγραφικές. Πιο κάτω παρουσιάζεται μια συνοπτική περιγραφή τους :

3.1 Περιγραφή Εννοιών/κατηγοριών: χαρακτηρισμός και διάκριση (Concept class description: Characterization and discrimination)

Η Περιγραφή Εννοιών/κατηγοριών: χαρακτηρισμός και διάκριση (Concept class description: Characterization and discrimination) είναι μια περιγραφική διαδικασία στην οποία δεδομένα μπορούν να συσχετιστούν με έννοιες και κατηγορίες και να δώσουν στοιχεία γι αυτές. Μέσω των «Συνοψίσεων» έχουμε μια συνοπτική περιγραφή του συνόλου δεδομένων ή ενός υποσυνόλου του το οποίο μπορεί και να μας φανεί χρήσιμο.

3.2 Εύρεση κανόνων συσχέτισης προτύπων και σχέσεων

Η εύρεση κανόνων συσχέτισης προτύπων και σχέσεων είναι επίσης μια περιγραφική διαδικασία ενός συνόλου από εγγραφές που η κάθε μία έχει ένα αριθμό από στοιχεία από κάποιο δοσμένο σύνολο στο οποίο θα πρέπει να βρούμε κανόνες εξάρτησης που προβλέπουν την παρουσία ενός στοιχείου με βάση την παρουσία άλλων στοιχείων.

Ένα άλλο είδος κανόνων που μπορούμε να βρούμε είναι τα ακολουθιακά πρότυπα συσχέτισης (Sequential Pattern Discovery), που είναι επίσης μια περιγραφική διαδικασία την οποία βρίσκουμε σε ένα σύνολο από εγγραφές, πρότυπα διακύμανσης/εμφάνισης κάποιων παραμέτρων, δεδομένης της διακύμανσης/εμφάνισης κάποιων παραμέτρων σε ακολουθία. Πρόκειται επίσης για περιγραφική διαδικασία που μας επιτρέπει να προχωρήσουμε σε πρόβλεψη, με την ανακάλυψη της ακολουθίας.

Όπως είναι αντιληπτό αυτά τα πρότυπα/σχέσεις δεν ακολουθούνται σε ολόκληρο το σύνολο των δεδομένων με αποτέλεσμα μια πρόβλεψη με άλλη να μπορεί να διακριθεί και στο βαθμό εμπιστοσύνης που έχουμε για την πρόβλεψη μας. Αυτός ο βαθμός σιγουριάς στην πρόβλεψη μας καλείται «βαθμός εμπιστοσύνης (Confidence)» Kamber and Han (2006).

Στην εξόρυξη δεδομένων, η εύρεση κανόνων συσχέτισης για εκμάθηση ενός κανόνα ένωσης είναι μια δημοφιλής και καλά εδραιωμένη μέθοδος για τις ενδιαφέρουσες σχέσεις μεταξύ των μεταβλητών στις μεγάλες βάσεις δεδομένων.

3.2.1 Διαστάσεις στις σχέσεις - Dimensions in association

Καθώς στην εξόρυξη δεδομένων, αναζητούμε σχέσεις (associations) και συσχετίσεις (correlations) μεταξύ δεδομένων, σε ένα συγκεκριμένο δείγμα θα

πρέπει πρώτα από όλα να διακρίνουμε τη διαφορά μεταξύ των όρων «σχέσεων» και «συσχετίσεων».

Σύμφωνα με το Oxford Dictionary ο όρος «συσχέτιση» αναφέρεται σε μια αμοιβαία συγγένεια ή ένωση μεταξύ δύο ή περισσοτέρων πραγμάτων, ενώ ο όρος «σχέση» αναφέρεται σε μια σύνδεση ή συνεργατικό δεσμό. Οι δύο έννοιες είναι τόσο συνυφασμένες που η Wikipedia αφιέρωσε συγκεκριμένο τμήμα της, για να επεξηγήσει τη διαφορά μεταξύ των δύο. Σύμφωνα με την Wikipedia η ειδοποιός διαφορά μεταξύ «συσχέτισης» και «σχέσης» έγκειται στον τρόπο που σχετίζονται δύο αντικείμενα. Ο όρος «συσχέτιση» δεικνύει γραμμική σχέση μεταξύ δύο αντικειμένων, ενώ ο όρος «σχέση» δεν ορίζει τη σχέση μεταξύ δύο αντικειμένων.

Οι σχέσεις διακρίνονται επιμέρους σε δύο διαστάσεις: (α) στις μονοδιάστατες σχέσεις, όπου μια παράμετρος έχει γραμμική σχέση με μια άλλη παράμετρο και (β) στις πολυδιάστατες σχέσεις, όπου μια παράμετρος εξαρτάται από δύο ή περισσότερες παραμέτρους (Kamber & Han, 2006).

Παρατίθενται δύο (πλασματικά) παραδείγματα πιο κάτω για να επεξηγηθούν οι όροι «μονοδιάστατες σχέσεις» και «πολυδιάστατες σχέσεις».

Μια μονοδιάστατη σχέση μπορεί να είναι η αύξηση των παρατηρήσεων ύπαρξης ψηλού μικροβιακού φορτίου στον έλεγχο του πόσιμου νερού, όταν η αύξηση είναι ανάλογη με τη θερμοκρασία της περιόδου κατά την οποία λήφθηκε το δείγμα. Μια πολυδιάστατη σχέση στο πιο πάνω παράδειγμα μπορεί να είναι η αύξηση των παρατηρήσεων ύπαρξης ψηλού μικροβιακού φορτίου στον έλεγχο του πόσιμου νερού, όταν η αύξηση είναι ανάλογη με τη θερμοκρασία της περιόδου σε συνδυασμό με την ύπαρξη χαμηλής αγωγιμότητας στο δείγμα.

3.2.2 Κύριες μέθοδοι εύρεσης κανόνων συσχέτισης προτύπων και σχέσεων

Στην εύρεση κανόνων συσχέτισης ο αλγόριθμος «Apriori» είναι ο πιο διαδεδομένος και χρησιμοποιείται στις βάσεις δεδομένων που περιέχουν συναλλαγές.

Ο αλγόριθμος «Apriori» χρησιμοποιεί πρώτη σε εύρος αναζήτηση μια δομή δέντρων,για να μετρήσει τα σύνολα των υποψηφίων στοιχείων αποτελεσματικά. Ο αλγόριθμος αρχικά παράγει σύνολα υποψηφίων στοιχείων μήκους K από σύνολα στοιχείων μήκους K − 1. Κατόπιν διαχωρίζει τους υποψηφίους που έχουν ένα σπάνιο υπο- σχέδιο. Το σύνολο υποψηφίων υποσυνόλων περιέχει όλα τα συχνά σύνολα στοιχείων K-μήκους. Μετά από αυτόν, ανιχνεύει τη βάση δεδομένων συναλλαγής, για να καθορίσει τα συχνά σύνολα στοιχείων μεταξύ των υποψηφίων. Σύμφωνα με την Wikipedia (2010) ο «Apriori» αλγόριθμος, ενώ έχει ιστορική σημασία , πάσχει από διάφορες ανεπάρκειες ή ανταλλαγές, οι οποίες έχουν ωοτοκήσει άλλους αλγόριθμους. Άλλοι αλγόριθμοι κανόνων συσχέτισης προτύπων και σχέσεων είναι ο «Eclat» και ο «FP-Growth».

3.3 Κατηγοριοποίηση/Ταξινόμηση (Classification) και πρόβλεψη

Η ταξινόμηση είναι μια προβλεπτική (Predictive) διαδικασία όπου δοθέντος ενός συνόλου εγγραφών (σύνολο εκπαίδευσης -training set), κάθε εγγραφή έχει ένα σύνολο από γνωρίσματα ένα εκ των οποίων είναι η κλάση (ή κατηγορία), βρίσκει ένα μοντέλο για το γνώρισμα της κλάσης ως συνάρτηση της τιμής των άλλων γνωρισμάτων.

Στόχος της ταξινόμησης είναι να αναθέτει σε εγγραφές που δεν έχουμε δει μια κλάση με τη μεγαλύτερη δυνατή ακρίβεια. Για να χαρακτηρίσουμε την ακρίβεια του μοντέλου χρησιμοποιούμε ένα σύνολο ελέγχου (test set). Συνήθως, το δοθέν

σύνολο δεδομένων χωρίζεται σε ένα σύνολο εκπαίδευσης και σε ένα σύνολο ελέγχου – το πρώτο χρησιμοποιείται για την κατασκευή του μοντέλου και το δεύτερο για τον έλεγχό του.

Στην αγγλική είναι καθιερωμένος ο όρος Classification, όμως στην Ελληνική υπάρχουν αναφορές στον όρο «Κατηγοριοποίηση» και «Ταξινόμηση» που όμως και στις δύο περιπτώσεις αναφέρονται στο ίδιο θέμα.

3.3.1 Κύριες μέθοδοι Ταξινόμησης (Classification)

Decision tree learning

Σύμφωνα με την Wikipedia(2010), η εκμάθηση δέντρων απόφασης, που χρησιμοποιείται στην εξόρυξη δεδομένων και στην εκμάθηση μηχανών, χρησιμοποιεί ένα δέντρο απόφασης ως προβλεπτικό μοντέλο που χαρτογραφεί τις παρατηρήσεις για ένα στοιχείο στα συμπεράσματα για την τιμή του στόχου του στοιχείου.

Σε αυτές τις δομές δέντρων, τα φύλλα αντιπροσωπεύουν τις ταξινομήσεις και οι κλάδοι αντιπροσωπεύουν τις κλίσεις των χαρακτηριστικών γνωρισμάτων που οδηγούν σε εκείνες τις ταξινομήσεις. Στην ανάλυση απόφασης, ένα δέντρο απόφασης μπορεί να χρησιμοποιηθεί οπτικά και ρητά να αντιπροσωπεύσει τις αποφάσεις και τη λήψη απόφασης. Στην εξόρυξη δεδομένων, ένα δέντρο απόφασης περιγράφει τα στοιχεία, αλλά όχι τις αποφάσεις, μάλλον το προκύπτον δέντρο ταξινόμησης μπορεί να είναι μια διαδικασία λήψης απόφασης.

Ο στόχος είναι να δημιουργηθεί ένα πρότυπο, που προβλέπει την αξία μιας μεταβλητής στόχων βασισμένης σε διάφορες μεταβλητές εισαγωγής. Κάθε εσωτερικός κόμβος αντιστοιχεί σε μια από τις μεταβλητές εισαγωγής, όπου υπάρχουν παιδιά για κάθε μια από τις πιθανές τιμές εκείνης της μεταβλητής εισαγωγής. Κάθε φύλλο αντιπροσωπεύει μια αξία τη μεταβλητή στόχων δεδομένων των τιμών των μεταβλητών εισαγωγής, που αντιπροσωπεύονται από

την πορεία από τη ρίζα στο φύλλο. Ένα δέντρο μπορεί να είναι " learned" με το διαχωρισμό η πηγή έθεσε στα υποσύνολα βασισμένα σε μια δοκιμή αξίας ιδιοτήτων.

Αυτή η διαδικασία επαναλαμβάνεται σε κάθε παραγόμενο υποσύνολο κατά τρόπο επαναλαμβανόμενο αποκαλούμενο επαναλαμβανόμενος χωρισμός. Αναδρομικά ολοκληρώνεται, όταν το υποσύνολο έχει σε έναν κόμβο την ίδια αξία της μεταβλητής στόχων ή όταν ο διαχωρισμός δεν προσθέτει πλέον την αξία στις προβλέψεις.

Bayesian Classification

Ένας ταξινομητής Bayes είναι ένας απλός πιθανολογικός ταξινομητής που υποθέτει ότι η παρουσία (ή απουσία) ενός ιδιαίτερου χαρακτηριστικού γνωρίσματος μιας κατηγορίας είναι ανεξάρτητη από την παρουσία (ή την απουσία) οποιουδήποτε χαρακτηριστικού. Ακόμη και αν τα χαρακτηριστικά αυτά εξαρτώνται το ένα από το άλλο ή στην ύπαρξη άλλου χαρακτηριστικού, αυτά ανεξάρτητα το κάθε ένα συμβάλλουν στην πρόβλεψη. Ανάλογα με την ακριβή φύση του προτύπου πιθανότητας, οι αφελείς ταξινομητές Bayes μπορούν να εκπαιδευθούν πολύ αποτελεσματικά σε μια εποπτευμένη ρύθμιση εκμάθηση (Wikipedia, 2010).

Σύμφωνα με τους Pedro D και Pedro M *(1997) ο Bayesian classifier* έχει μια πολύ μεγαλύτερη δυνατότητα εφαρμογής από ότι αναμενόταν. Στο άρθρο τους ¨On the Optimality of the Simple Bayesian Classifier under Zero-One Loss¨, αποδεικνύεται ότι για να είναι βέλτιστο για την εκμάθηση των κλίσεων και των αποσυνδέσεων, ακόμα κι αν παραβιάζουν την υπόθεση ανεξαρτησίας. Επίσης αναφέρουν ότι οι μελέτες τεχνικής νοημοσύνης δείχνουν ότι η αποδοτικότητά τους ξεπερνά συχνά τους ισχυρότερους ταξινομητές για τα κοινά μεγέθη εκπαιδευόμενων συνόλων και αριθμών χαρακτηριστικών ιδιοτήτων.

k-nearest neighbor algorithm

Σύμφωνα με την Wikipedia (2010) ο K-κοντινότερος αλγόριθμος γειτόνων είναι μεταξύ του απλούστερου όλων των αλγορίθμων εκμάθησης μηχανών: ένα αντικείμενο ταξινομείται από έναν ψήφο πλειοψηφίας των γειτόνων του, με το αντικείμενο που κατατάσσεται στην κατηγορία την πιο κοντινή μεταξύ των κοντινότερων γειτόνων του K (το K είναι ένας θετικός ακέραιος αριθμός, χαρακτηριστικά μικρός). Εάν K = 1, έπειτα το αντικείμενο κατατάσσεται απλά στην κατηγορία κοντινότερου γείτονά του. Η ίδια μέθοδος μπορεί να χρησιμοποιηθεί για την οπισθοδρόμηση, με απλά λόγια η αξία ενός στοιχείου για το αντικείμενο ορίζεται ως ο μέσος όρος των τιμών των πιο k-κοντινών γειτόνων της. Μπορεί να είναι χρήσιμο να σταθμιστούν οι συνεισφορές των γειτόνων, έτσι ώστε οι κοντινότεροι γείτονες να συμβάλλουν περισσότερο στο μέσο όρο από τους πιο απόμακρους. Οι γείτονες λαμβάνονται από ένα σύνολο αντικειμένων για το οποίο η σωστή ταξινόμηση είναι γνωστή. Αυτό μπορεί να θεωρηθεί ως σύνολο κατάρτισης για τον αλγόριθμο, αν και κανένα ρητό βήμα κατάρτισης δεν απαιτείται.

Τεχνητό νευρονικό δίκτυο (Artificial neural network)

Ένα τεχνητό νευρικό δίκτυο (Artificial neural network), είναι ένα μαθηματικό πρότυπο ή υπολογιστικό πρότυπο το οποίο είναι ένα προσαρμοστικό σύστημα, που αλλάζει τη δομή του βασισμένο στις εξωτερικές ή εσωτερικές πληροφορίες, που διατρέχουν του δικτύου κατά τη διάρκεια της φάσης εκμάθησης. Τα νευρονικά δίκτυα χρησιμοποιούνται συνήθως, για να διαμορφώσουν τις σύνθετες σχέσεις μεταξύ των εισαγωγών και των αποτελεσμάτων ή για να βρούν τα σχέδια στα στοιχεία. Αφού δικτυωθούν τα στοιχεία στη συνέχεια χαρτογραφούνται μετρούνται τα διανύσματα και αρχειοθετούνται σε γραμμικό ταξινομητή που καθορίζει αυτό που είναι γνωστό ως μέγιστος ταξινομητής περιθωρίου (Wikipedia, 2010)

3.4. Συσταδοποίηση (Clustering)

Η συσταδοποίηση (clustering) είναι μια περιγραφική διαδικασία, όπου δοθέντος ενός συνόλου από σημεία, που το καθένα έχει κάποια γνωρίσματα διεξάγουμε μια μέτρηση ομοιότητας μεταξύ τους με σκοπό την εύρεση **συστάδων** (**clusters)** τέτοιων, ώστε τα σημεία σε μία συστάδα να είναι πιο όμοια μεταξύ τους και τα σημεία σε διαφορετικές συστάδες να είναι λιγότερα όμοια μεταξύ τους.

Με άλλα λόγια η συσταδοποίηση είναι η τμηματοποίηση (partitioning) ενός συνόλου δεδομένων σε συστάδες. Έτσι ώστε τα στοιχεία του συνόλου δεδομένων που ανήκουν σε μία συστάδα να έχουν περισσότερες ομοιότητες μεταξύ τους παρά με στοιχεία των άλλων συστάδων ή με στοιχεία, που δεν ανήκουν σε καμία συστάδα. Με σκοπό να αποκαλύψει την οργάνωση προτύπων σε «λογικές» συστάδες, οι οποίες θα μας επιτρέψουν να ανακαλύψουμε ομοιότητες και διαφορές, καθώς επίσης και να εξάγουμε συμπεράσματα. Σε αντίθεση με την ταξινόμηση, οι συστάδες δεν είναι γνωστές από πριν.

3.4.1 Κύριες μέθοδοι συσταδοποίησης (Clustering)

Οι κύριες μέθοδοι συσταδοποίησης στις οποίες βασίζονται οι πιο κύριοι αλγόριθμοι είναι οι εξής :

1. Διαιρετική
2. Ιεραρχική
3. Πυκνότητας

Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)

Στην ιεραρχική μέθοδο κάνουμε τμηματοποίηση στη βάση δεδομένων και οργανώνουμε τα τμήματα έτσι, ώστε το κάθε τμήμα να περιέχει τουλάχιστο ένα τμήμα και ένα αντικείμενο να ανήκει μόνο σε ένα τμήμα. Οι συστάδες έχουν υποσυστάδες, που είναι ένα σύνολο από ένθετες συστάδες που είναι οργανωμένες σαν δέντρο (Wikipedia, 2010).

Ιεραρχικοί Αλγόριθμοι Συσταδοποίησης

BIRCH

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

Ο αλγόριθμος BIRCH σαρώνει τη βάση δεδομένων για να δημιουργήσει ένα αρχικό. Το CF-δέντρο είναι ένα ισοζυγισμένο δέντρο με δυο παραμέτρους: τον παράγοντα διακλάδωσης B που καθορίζεται από το μέγεθος του block και το κατώφλι T που καθορίζει την ποιότητα της συσταδοποίησης και στη συνέχεια χρησιμοποιεί ένα αλγόριθμο συσταδοποίησης για να οργανώσει σε συστάδες τους κόμβους φύλλα του CF-tree (Wikipedia, 2010).

CURE

Ο αλγόριθμος CURE χρησιμοποιώντας συνδυασμό τυχαίας δειγματοποίησης και τμηματοποίησης, αναγνωρίζει clusters με πιο περίεργες γεωμετρίες (που έχουν μη σφαιρικά σχήματα) χρησιμοποιεί πολλαπλά αντιπροσωπευτικά σημεία αξιολογεί τις αποστάσεις ανάμεσα στις συστάδες και προσαρμόζεται καλά σε αφηρημένα σχήματα συστάδων. Ένα άλλο χαρακτηριστικό του είναι ότι μπορεί και χειρίζεται αποτελεσματικά δεδομένα με outliers, ενώ ένα μειονέκτημα του είναι ότι χρειάζεται εκ των προτέρων γνώση του αριθμού συστάδων (Wikipedia, 2010).

ROCK

Ο αλγόριθμος ROCK εισάγει δύο νέες έννοιες: γείτονες σημείου και σύνδεσμοι, ενώ χρησιμοποίει links για να μετρήσουμε ομοιότητα/εγγύτητα και δεν βασίζεται σε αποστάσεις. Χρησιμοποιείται για Boolean και κατηγορικά δεδομένα.

Αλγόριθμοι Πυκνότητας (Density-Based Clustering Algorithms)

Τα κύρια χαρακτηριστικά συσταδοποίησης με βάση την πυκνότητα (local cluster criterion), όπως πυκνά συνδεδεμένα σημεία, είναι να ανακαλύψουμε συστάδες τυχαίων σχημάτων και ακολούθως να αναγνωρίσουμε τις πραγματικές συστάδες από τους τυχαίους συνδυασμούς (θόρυβο).

DBSCAN

Ο αλγόριθμος DBSCAN βασίζεται στην έννοια της πυκνότητας της συστάδας cluster και ορίζει συστάδες που περιέχουν τα σημεία που είναι πυκνά συνδεδεμένα. Χαρακτηριστικό του είναι η ικανότητά του να ανακαλύπτει συστάδες περίεργης γεωμετρίας σε βάσεις δεδομένων με υψηλό θόρυβο.

DENCLUE

Ο αλγόριθμος DENCLUE χρησιμοποιεί πλέγμα (grid) σε στατιστικές συναρτήσεις πυκνότητας και είναι κατάλληλος για δεδομένα με μεγάλο ποσοστό θορύβου, ενώ επιτρέπει περιγραφή συστάδων με περίεργα σχήματα σε υψηλές διαστάσεις. Πλεονέκτημα του είναι ότι είναι γρηγορότερος από τους άλλους αλγορίθμους πυκνότητας αλλά απαιτεί μεγάλο αριθμό από παραμέτρους.

3.5 Ανάλυση Περιθωριακών τιμών (Outlier analysis)

Η ανάλυση περιθωριακών τιμών (Outlier analysis) είναι μια προφητική διαδικασία, όπου σε ένα σύνολο δεδομένων μιας βάσης μπορεί κάποια δεδομένα να μην συνάδουν με τη γενική συμπεριφορά των δεδομένων ή του μοντέλου των δεδομένων τα οποία και αποκαλούνται περιθωριακές τιμές. Οι περισσότερες μέθοδοι εξόρυξης δεδομένων απορρίπτουν τις περιθωριακές τιμές, γιατί θεωρούν ότι αλλοιώνουν τα δεδομένα. Ωστόσο κάποιες μέθοδοι, όπως αυτή της ανίχνευσης απάτης ασχολούνται με τον εντοπισμό τέτοιων τιμών.

2.1.2 Κύριες μέθοδοι Εύρεσης Εκτόπων

Οι αλγόριθμοι ανίχνευσης εκτόπων συνήθως βασίζονται στην απόσταση ή την πυκνότητα ή τη διανομή των δεδομένων. Επίσης βασίζονται κυρίως στην οργάνωση των στοιχείων και μέτρηση της απόκλισης των στοιχείων χρησιμοποιώντας τη στατιστική.

Ένας από του πρώτους που ασχολήθηκε με τις έκτοπες τιμές είναι ο Hawkins (1980) ενώ αναφορές για αλγόριθους και διαθεματικές εφαρμογές έχει ο Edwin

M. Knorr, που έχει αρκετές δημοσιεύσεις για το θέμα. Ιδιαίτερα στο άρθρο του 'Algorithms for mining distance-based outliers in large datasets' (Knorr & Raymond, 1998) περιγράφονται ιδιαίτερα ενδιαφέρον αλγόριθμοι και εφαρμογές. Ενδιαφέρον εισήγηση του είναι ο αλγόριθμος που προσδιορίζει ακραίες τιμές με τον υπολογισμό του αριθμού γειτόνων μέσα σε μια διευκρινισμένη ακτίνα ενός σημείου στοιχείων. Η ακτίνα και ο αριθμός κατώτατων ορίων σημείων είναι οι μόνες δύο παράμετροι της προσέγγισης. Η προσέγγιση είναι απλή αλλά είναι ανεπαρκής για το στοιχείο που διανέμεται με την ανώμαλη πυκνότητα όπου πρέπει να ποικίλει για να αντιμετωπίσει τις αλλαγές.

4. Εξόρυξη δεδομένων σε δεδομένα που αφορούν την ασφάλεια τροφίμων

4.1 Ο έλεγχος των τροφίμων στην Κύπρο: εισαγωγή για τον επιστήμονα πληροφορικής

Επειδή το θέμα της ασφάλειας των τροφίμων είναι άγνωστο στους επιστήμονες πληροφορικής και η γνώση σχετικά με κάποιες από τις πτυχές της είναι απαραίτητη για την κατανόηση της προσπάθειας και προσέγγισης που καταβλήθηκε, ακολουθεί σύντομη ενημέρωση αναφορικά με την οργάνωση και διενέργεια των ελέγχων όσον αφορά την ασφάλεια των τροφίμων στην Κύπρο.

4.1.1 Δομή των Υπηρεσιών που είναι επιφορτισμένες με την ασφάλεια των τροφίμων στην Κύπρο

Η ασφάλεια των τροφίμων στην Κύπρο είναι ένα μεγάλο κεφάλαιο το οποίο λόγω μεγέθους χωρίζεται στις αρμοδιότητες τριών Υπηρεσιών:

Υγειονομικές Υπηρεσίες

Οι Υγειονομικές Υπηρεσίες του Υπουργείου Υγείας έχουν σαν αποστολή τους τη διασφάλιση της Δημόσιας Υγιεινής όσον αφορά διάφορους παράγοντες που δυνατό να την επηρεάσουν.

Συγκεκριμένα, οι Υγειονομικές Υπηρεσίες του Υπουργείου Υγείας είναι η Αρμόδια Αρχή για την εφαρμογή και τον έλεγχο της νέας εναρμονισμένης Κοινοτικής νομοθεσίας για τα τρόφιμα. Σε συνεργασία με όλα τα ενδιαφερόμενα μέρη(καταναλωτές/υπεύθυνους επιχειρήσεων τροφίμων/άλλες ελέγχουσες αρχές) αναπτύσσουν μια ενιαία και ολοκληρωμένη πολιτική με σκοπό την διασφάλιση της ασφάλειας των τροφίμων. Για να επιτύχουν τους στόχους αυτούς, σε συνεργασία με τις τοπικές αρχές ή και άλλες εμπλεκόμενες Υπηρεσίες, εφαρμόζουν διάφορα προγράμματα τα οποία και επιθεωρούνται από την Ευρωπαϊκή Επιτροπή και οργανώνουν τους ελέγχους στα τρόφιμα μη ζωικής προέλευσης ενώ έχουν αρμοδιότητα για άσκηση επίσημων ελέγχων σε όλα τα στάδια της τροφικής αλυσίδας από την παραγωγή μέχρι και τη λιανική πώληση για τα τρόφιμα αυτά, καθώς και για την εφαρμογή μέτρων για την προστασία της υγείας των καταναλωτών.

Όσον αφορά τα τρόφιμα ζωικής προέλευσης, έχουν την αρμοδιότητα ελέγχου στο στάδιο της λιανικής πώλησης. Επίσης έχουν αρμοδιότητα σε όλα τα στάδια της τροφικής αλυσίδας για τις επιχειρήσεις μελιού και παγωτού. Πέραν των πιο πάνω έχουν ευθύνη για την κατάρτιση σχεδίων για την αντιμετώπιση εκτάκτων καταστάσεων καθώς και κρίσεων που σχετίζονται με κινδύνους που αφορούν την ασφάλεια των τροφίμων, ενώ είναι και το Εθνικό Σημείο Επαφής της Κύπρου όσον αφορά το Σύστημα Έγκαιρης Προειδοποίησης για τα Τρόφιμα και τις Ζωοτροφές.

Εκτός από τον έλεγχο των τροφίμων, στις Υγειονομικές Υπηρεσίες λειτουργεί και ο τομέας της Περιβαλλοντικής Υγιεινής που καλύπτει ένα ευρύ φάσμα δραστηριοτήτων όπου οι Υγειονομικές Υπηρεσίες σε συνεργασία με άλλες εμπλεκόμενες Υπηρεσίες έχουν αρμοδιότητα για την εφαρμογή και

παρακολούθηση της νομοθεσίας και λήψη μέτρων για πληθώρα θεμάτων Δημόσιας Υγείας. Στον τομέα αυτό μεταξύ άλλων διεξάγουν δραστηριότητες όπως η παρακολούθηση και έλεγχος της ποιότητας του νερού ανθρώπινης κατανάλωσης, η καταπολέμηση εντόμων ιατρικής σπουδαιότητας, η επιθεώρηση /έλεγχος δημοσίων και ιδιωτικών χώρων, η υγειονομική διαφώτιση, ο έλεγχος δημοσίων κολυμβητικών δεξαμενών και των θαλάσσιων περιοχών λουομένων, ο έλεγχος της παρασκευής και διάθεσης απορρυπαντικών, η διερεύνηση παραπόνων για οχληρίες, ο έλεγχος του καπνίσματος και των καπνικών προϊόντων και εφαρμογή της σχετικής με το κάπνισμα νομοθεσίας και η πρόληψη και διερεύνηση κρουσμάτων μολυσματικών ασθενειών συμπεριλαμβανομένων και των τροφικών δηλητηριάσεων (*Υγειονομικές Υπηρεσίες*, 2010).

Τμήμα Κτηνιατρικών Υπηρεσιών

Οι Κτηνιατρικές Υπηρεσίες του Υπουργείου Γεωργίας Φυσικών Πόρων και Περιβάλλοντος έχουν σαν κύριους τομείς ενασχόλησης τον Τομέα Υγείας και Ευημερίας των Ζώων, τον Τομέα Κτηνιατρικής Δημόσιας Υγείας, τα Εργαστήρια Δημόσιας Υγείας και Υγείας των Ζώων και τον Έλεγχο και κυκλοφορία των Κτηνιατρικών Φαρμακευτικών Προϊόντων.

Ο Τομέας Κτηνιατρικής Δημόσιας Υγείας έχει άμεση σχέση με τον έλεγχο των τροφίμων, καθώς οι Κτηνιατρικές Υπηρεσίες έχουν την αποκλειστική αρμοδιότητα για τον έλεγχο της παραγωγής, χειρισμού, μεταφοράς, αποθήκευσης και διάθεσης στην αγορά των προϊόντων ζωικής προέλευσης με εξαίρεση το παγωτό και το μέλι καθώς και τον έλεγχο και την έγκριση των εγκαταστάσεων όπου αυτά παράγονται. Πέραν των πιο πάνω, ο εν λόγω τομέας ασχολείται επίσης με τον έλεγχο του ενδοκοινοτικού εμπορίου, την έκδοση πιστοποιητικών για εξαγωγή προϊόντων ζωικής προέλευσης σε τρίτες χώρες και την εφαρμογή του προγράμματος παρακολούθησης των καταλοίπων (Κτηνιατρικές Υπηρεσίες, 2010).

Τμήμα Γεωργίας

Το Τμήμα Γεωργίας του Υπουργείου Γεωργίας Φυσικών Πόρων και Περιβάλλοντος έχει σαν κύριο στόχο την ανάπτυξη του γεωργοκτηνοτροφικού τομέα μέσω της επιμόρφωσης και της καθοδήγησης των αγροτών του σχεδιασμού και της εφαρμογής αναπτυξιακών προγραμμάτων.

Το Τμήμα Γεωργίας έχει διάφορα τμήματα τα οποία και ασχολούνται με πληθώρα θεμάτων, όσον αφορά όμως την τροφική αλυσίδα έχει τον έλεγχο της πρωτογενής παραγωγής φρούτων και λαχανικών ενώ είναι επίσης υπεύθυνο για τους ελέγχους των ζωοτροφών (Τμήμα Γεωργίας, 2010).

4.1.2 Τι είναι το Σύστημα Έγκαιρης Προειδοποίησης για τα Τρόφιμα και τις Ζωοτροφές (RASFF)

Το Σύστημα Έγκαιρης Προειδοποίησης για τα Τρόφιμα και τις Ζωοτροφές (Rapid Alert System for Food and Feed) είναι ένα δίκτυο για την έγκαιρη και αποτελεσματική ανταλλαγή πληροφοριών μεταξύ των κρατών-μελών της Ευρωπαϊκής Ένωσης και τη λήψη μέτρων, όταν διαπιστώνονται κίνδυνοι για την ανθρώπινη υγεία σε τρόφιμα και ζωοτροφές (Κανονισμός (ΕΚ) Αριθ. 178/2002).

Δημιουργία και θεσμοθέτηση

Μετά από ένα περιστατικό το 1979 που αφορούσε υδράργυρο σε πορτοκάλια από το Ισραήλ, στο οποίο εμπλέκονταν διάφορα Ευρωπαϊκά Κράτη διαφάνηκε η ανάγκη για ένα μέσο ανταλλαγής πληροφοριών αναφορικά με την ασφάλεια των τροφίμων μεταξύ των αρμοδίων Υπηρεσιών των Ευρωπαϊκών Κρατών (*European Commission, 2009).* Το δίκτυο RASFF συστάθηκε το 1979 και σταδιακά θεσμοθετήθηκε με τη σημερινή του μορφή να αποδίδεται στον Κανονισμό (ΕΚ) αριθ. 178/2002 του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου της 28ης Ιανουαρίου 2002. Ο κανονισμός αυτός καθορίζει τις γενικές αρχές και απαιτήσεις της νομοθεσίας για τα τρόφιμα, προνοεί την ίδρυση της Ευρωπαϊκής Αρχής για την Ασφάλεια των Τροφίμων και καθορίζει

διαδικασίες σε θέματα ασφάλειας των τροφίμων. Σε μια προσπάθεια της η Επιτροπή να θεσμοθετήσει την ενιαία και ομοιογενή εφαρμογή του RASFF σε όλη την Ευρωπαϊκή Ένωση, στις 10ης Ιανουαρίου 2011 ψηφίστηκε ο Κανονισμός (ΕΕ) αριθ. 16/2011 της Επιτροπής για τον καθορισμό μέτρων εφαρμογής του συστήματος έγκαιρης προειδοποίησης για τρόφιμα και ζωοτροφές, κανονισμός που παρουσιάζει ενδιαφέρον για τον ΕΟΧ(Κανονισμός (ΕΕ) αριθ. 16/2011).

Μέλη και διάρθρωση του συστήματος

Στο δίκτυο του RASSF συμμετέχουν η Ευρωπαϊκή Επιτροπή (Γενική Διεύθυνση Υγείας και προστασίας των καταναλωτών), οι αρμόδιοι φορείς για την ασφάλεια τροφίμων στα Κράτη-Μέλη της ΕΕ και η Ευρωπαϊκή Αρχή για την Ασφάλεια των Τροφίμων (ΕΑΑΤ) καθώς και η Νορβηγία, το Λιχτενστάιν και η Ισλανδία. Επειδή στα Κράτη Μέλη υπάρχει ανομοιομορφία στη διάρθρωση των Υπηρεσιών που ασχολούνται με την ασφάλεια των τροφίμων και των ζωοτροφών ορίζεται ένα Σημείο Επαφής για κάθε μέλος του Συστήματος, το οποίο και είναι υπεύθυνο για τη διαχείριση των εισερχόμενων και εξερχόμενων πληροφοριών. Η υπηρεσία αυτή λειτουργεί σε εικοσιτετράωρη βάση και εξασφαλίζει την αποστολή, παραλαβή και διεκπεραίωση των επειγουσών κοινοποιήσεων σε όσο το δυνατόν πιο σύντομο χρονικό διάστημα (Σάββα, 2009).

Η Κύπρος το 2004 με την ένταξη της στην Ευρωπαϊκή Ένωση έγινε και μέλος του RASFF. Την Κυπριακή συμμετοχή απαρτίζουν οι Κτηνιατρικές Υπηρεσίες, το Τμήμα Γεωργίας και οι Υγειονομικές Υπηρεσίες, ο Προϊστάμενος των οποίων έχει καθοριστεί ως το εθνικό σημείο επαφής.

Ροή πληροφοριών στο σύστημα

Το RASFF δίνει τη δυνατότητα ταχείας και αποτελεσματικής ανταλλαγής πληροφοριών μεταξύ των Κρατών Μελών και της Επιτροπής. Όταν ένα μέλος του RASFF λάβει οποιαδήποτε πληροφορία για την ύπαρξη άμεσου ή έμμεσου κινδύνου για την ανθρώπινη υγεία που προέρχεται από τρόφιμα ή ζωοτροφές

που διατίθενται στην αγορά ή στα σύνορά του, οφείλει να ενημερώσει την Επιτροπή μέσω του RASFF. Η Επιτροπή διαβιβάζει αμέσως την πληροφορία στα μέλη του δικτύου.

Πέραν των πιο πάνω, σε περιπτώσεις όπου χρειάζεται συνεργασία είτε με κράτη μη μέλη του συστήματος, είτε με οργανισμούς ή με άλλα δίκτυα όπως το δίκτυο των διεθνών αρχών για την ασφάλεια των τροφίμων της Παγκόσμιας Οργάνωσης Υγείας (INFOSAN), το RASFF διασφαλίζει αυτή τη συνεργασία και αποτελεί το μακρύ βραχίονα των Κρατών Μελών, της Ευρωπαϊκής Επιτροπής και της EFSA.

RASFF και ασφαλέστερες εισαγωγές

Πέραν των κινδύνων που εντοπίζονται στην Κοινή αγορά, οι περιπτώσεις απόρριψης εισαγωγής φορτίων από τρίτες χώρες επίσης δημοσιοποιούνται στο σύστημα (border rejection notification) και όταν εντοπίζεται ένα τέτοιο προϊόν, το RASFF ενημερώνει την εμπλεκόμενη τρίτη χώρα προκειμένου να αποφευχθεί η επανεμφάνιση του προβλήματος ή προσπάθεια εισαγωγής του ίδιου φορτίου από άλλο Σημείο Εισόδου της Ε.Ε.

Σε περίπτωση που διαπιστώνεται η ύπαρξη ενός σοβαρού και επίμονου προβλήματος, η Επιτροπή είτε απευθύνει γραπτή επιστολή στις αρμόδιες αρχές της χώρας προέλευσης του προϊόντος ζητώντας την εφαρμογή μέτρων προς διόρθωση του προβλήματος ή μπορεί να προχωρήσει σε σύσταση προς τα Κράτη Μέλη ή απόφαση για την επιβολή άμεσων μέτρων, όπως η απαγόρευση εξαγωγής ή η διεξαγωγή εντατικότερων ελέγχων.

Πέραν των πιο πάνω τα ίδια τα Κράτη Μέλη έχουν την δυνατότητα να εφαρμόζουν ενισχυμένους και στοχευόμενους ελέγχους σύμφωνα με τα ευρήματα τους για τρόφιμα ζωικής προέλευσης σύμφωνα με τις πρόνοιες του άρθρου 24 της Οδηγίας 97/78/ΕΚ για καθορισμό των αρχών οργανώσεως των κτηνιατρικών ελέγχων των προϊόντων, που εισάγονται στην Κοινότητα από τρίτες

χώρες(Οδηγία 97/78/ΕΚ) και για τρόφιμα μη ζωικής προέλευσης σύμφωνα με το άρθρο 15 του Κανονισμού (ΕΚ) αριθ. 882/2004 για τη διενέργεια επίσημων ελέγχων της συμμόρφωσης προς τη νομοθεσία περί ζωοτροφών και τροφίμων και προς τους κανόνες για την υγεία και την καλή διαβίωση των ζώων (Κανονισμός 882/2004).

Δικτύωση του RASFF στην Κύπρο

Οι Υπηρεσίες μας ενεργούν ως Εθνικό Σημείο Επαφής, λαμβάνουν τις σχετικές πληροφορίες από την Ευρωπαϊκή Επιτροπή, τις αξιολογούν και τις αποστέλλουν για σχετική ενημέρωση και τυχόν ενέργεια ως ακολούθως: Για τρόφιμα ζωικής προέλευσης οι πληροφορίες διαβιβάζονται τόσο στις Κτηνιατρικές Υπηρεσίες όσο και στις Υγειονομικές Υπηρεσίες για έλεγχο στην αγορά. Οι περιπτώσεις τροφίμων μη ζωικής προέλευσης διαβιβάζονται στις Υγειονομικές Υπηρεσίες για ανάλογη ενέργεια, οι δε περιπτώσεις ζωοτροφών και τροφίμων φυτικής παραγωγής (πρωτογενής παραγωγή) προς το Τμήμα Γεωργίας (Σάββα, 2009).

Στο Γραφείο του Π.Υ.Υ. υπάρχει ομάδα η οποία ασχολείται με τον διαχωρισμό των γνωστοποιήσεων που αποστέλλονται από την Ε.Ε. και η ομάδα αυτή είναι υπεύθυνη για την διαβίβαση των γνωστοποιήσεων που συμπληρώνονται από τα Επαρχιακά Γραφεία και τις Συνεργαζόμενες Υπηρεσίες στην Ε.Ε.. Η ομάδα εποπτεύεται άμεσα από το Εθνικό Σημείο Επαφής.

Οι Υγειονομικές Υπηρεσίες των Επαρχιών έχουν καταρτισμένες ομάδες που ασχολούνται με το RASFF. Σε κάθε Επαρχία υπάρχουν λειτουργοί οι οποίοι εκτελούν χρέη Υπεύθυνου Λειτουργίας (ΥΛ) για την επαρχία τους, δηλαδή είναι επιφορτισμένοι με το έργο της επικοινωνίας με το Γραφείο του ΠΥΥ, με την παραλαβή γνωστοποιήσεων για διερεύνηση και την παροχή πληροφοριών στο Γραφείο του ΠΥΥ.

Η συνεισφορά του RASFF στη διασφάλιση της υγείας των Ευρωπαίων καταναλωτών

Σε περίπτωση που βρεθεί στην Ευρωπαϊκή αγορά τρόφιμο ή ζωοτροφή η κατανάλωση του οποίου δυνατό να αποτελεί άμεσο ή έμμεσο κίνδυνο για την ανθρώπινη υγεία τότε το μέλος τους συστήματος, στην επικράτεια του οποίου έχει βρεθεί το υπό αναφορά προϊόν συμπληρώνει και αποστέλλει τυποποιημένη κοινοποίηση στο RASFF, η οποία και διαβαθμίζεται σε δύο επίπεδα σημαντικότητας ανάλογα με το κατά πόσο ο κίνδυνος, που έχει διαπιστωθεί αφορά προϊόν που βρίσκεται ήδη στην αγορά και είναι απαραίτητη η άμεση ενέργεια (κοινοποιήσεις προειδοποίησης και κοινοποιήσεις πληροφοριακού περιεχομένου) και ενημερώνει τα μέλη του συστήματος (Κανονισμός (ΕΕ) αριθ. 16/2011).

Αξιολογώντας τις κοινοποιήσεις που λαμβάνουν τα Κράτη Μέλη μπορούν να εκτιμήσουν αν επηρεάζονται και να αντιδράσουν αναλόγως, διασφαλίζοντας τη συνεπή και ταυτόχρονη δράση σε ολόκληρη την ΕΕ και προστατεύοντας την υγεία των καταναλωτών. Σύμφωνα με την ομιλία της Επιτρόπου Υγείας της Ευρωπαϊκής Ένωσης κα Ανδρούλλα Βασιλείου, κατά τους εορτασμούς για τα 30 χρόνια λειτουργίας του RASFF, που έγιναν στις Βρυξέλλες τον Ιούνιο του 2009, χάρη στην αποτελεσματικότητα και αποδοτικότητα του RASFF στην ανταλλαγή πληροφοριών αποφεύχθηκαν πολλοί κίνδυνοι για τους καταναλωτές πριν καταστούν ζημιογόνοι για την υγεία τους (Σάββα, 2009).

Πέραν των συνηθισμένων περιστατικών το RASFF δοκιμάστηκε δεκάδες φορές σε διατροφικές κρίσεις που συγκλόνισαν την Ευρώπη με πιο πρόσφατες την περίπτωση στις αρχές του 2011, που αφορούσε διοξίνες σε ζωοτροφές στη Γερμανία για το δυνητικό κίνδυνο, που προκύπτει με την πιθανή εισαγωγή τροφίμων ή ζωοτροφών μολυσμένων με ραδιονουκλίδια μετά το ατύχημα στο πυρηνικό εργοστάσιο παραγωγής ηλεκτρικής ενέργειας της Fukushima. Για όλες τις πιο πάνω υποθέσεις, το RASFF συνέβαλε στο συντονισμό των ενεργειών των κρατών μελών, ελαχιστοποιώντας με τον τρόπο αυτό τις συνέπειες από τα συμβάντα επιμόλυνσης και βοηθώντας στο συντονισμό αλλά και στη διαχείριση του κινδύνου.

Από τα πιο πάνω μπορούμε να καταλήξουμε ότι το RASFF αποτελεί τον ακρογωνιαίο λίθο της Δημόσιας Υγείας της Ευρώπης όσον αφορά την ασφάλεια τροφίμων και ζωοτροφών.

4.2. Η παρούσα Έρευνα σε σχέση με την ασφάλεια τροφίμων

Η παρούσα έρευνα εστιάζεται στο να ανακαλύψει κατά πόσο υπάρχουν επαναλαμβανόμενες ακολουθίες και τάσεις στην εμφάνιση των αποτελεσμάτων του ελέγχου των τροφίμων, που μπορεί να μας δώσουν τη δυνατότητα πρόβλεψης μελλοντικών προβλημάτων που μπορεί να παρατηρηθούν. Ένα σημαντικό πρόβλημα που είχαμε να αντιμετωπίσουμε για τη διεξαγωγή της έρευνας ήταν η πρόσβαση στα δεδομένα.

Παραχώρηση άδειας πρόσβασης στα δεδομένα

Για να εξασφαλισθεί η εν λόγω άδεια έπρεπε να πειστεί η διεύθυνση των Υγειονομικών Υπηρεσιών για να δώσει άδεια πρόσβασης σε αυτά και σε μορφή που μπορεί να επεξεργαστεί. Για το θέμα είχα αποταθεί στον Προϊστάμενο Υγειονομικών Υπηρεσιών με τον οποίο και είχα μακροσκελή συνάντηση όπου συζητήσαμε εκτενώς το θέμα. Κατά τη διάρκεια της συνάντησής μας του ανέλυσα του σκοπούς της εργασίας και τον τρόπο που σκόπευα να εργαστώ, καθώς και τα πιθανά ωφέλη για την Υπηρεσία. Επειδή τα δεδομένα δεν ανήκουν στην Υπηρεσία μας αλλά στην Ευρωπαϊκή Επιτροπή, ο Προϊστάμενος Υγειονομικών Υπηρεσιών που είναι και το Εθνικό Σημείο Επαφής για το RASFF, είχε αποταθεί στην Ευρωπαϊκή Επιτροπή, για να ζητήσει άδεια πρόσβασης στα δεδομένα. Τελικά η εν λόγω άδεια μου δόθηκε. Στο Παράρτημα Ι επισυνάπτεται ολόκληρη η αλληλογραφία με τον κ Jose De Filippe που είναι ο υπεύθυνος για τη λειτουργία, συντονισμό και ανάπτυξη του RASFF στην Ευρωπαϊκή Επιτροπή.

4.3 Εφαρμογή

4.3.1 Οργάνωση Πληροφοριών του RASFF

Για τη λειτουργία του RASFF, έχει δημιουργηθεί από την Ε.Ε. μια βάση δεδομένων η οποία ονομάζεται “CIRCA” και είναι το αρκτικόλεξο της φράσης “Communication & Information Resource Centre Administrator”. Οι πληροφορίες του RASFF είναι αναρτημένες στη διαδικτυακά προσβάσιμη ιστοσελίδα/βάση δεδομένων όπου και μπορούν να ανακληθούν. Η βάση αυτή απευθύνεται στα Σημεία Επαφής στα Κράτη Μέλη και μόνο άτομα με κωδικό σύνδεσης έχουν πρόσβαση στο περιεχόμενο της.

Τα τμήματα τα οποία περιέχει είναι:

- Information (πληροφορίες): Γενικές Πληροφορίες και Σύνδεσμοι
- Library (βιβλιοθήκη): Το τμήμα όπου φορτώνονται όλες οι ανακοινώσεις
- Directory (αρχείο): Στοιχεία των χρηστών
- Meetings (συνεδριάσεις): Δυνατότητα εικονικών συνεδριάσεων
- Newsgroup (ομάδα συζήτησης): Χώρος για συζητήσεις
- Administration (διοίκηση): Προσβάσιμο για τη διοίκηση
- Email (ηλεκτρονικό ταχυδρομείο): Επιτρέπει την αποστολή ηλεκτρονικών μηνυμάτων σε άλλους χρήστες
- Search (αναζήτηση): Αναζήτηση πλήρους κειμένου
- Help (βοήθεια): Αρχείο βοήθειας

Το ενεργό τμήμα το οποίο περιέχει όλες τις πληροφορίες για το RASFF είναι το Library (βιβλιοθήκη): Το τμήμα όπου φορτώνονται όλες οι πληροφορίες του RASFF, ενώ τα άλλα μέρη εξυπηρετούν για τη διαχείριση των διαδικαστικών /οργανωτικών θεμάτων του RASFF.

R_A_S (RASFF Search tool)

Επίσης αναρτημένη είναι και η βάση δεδομένων στην Microsoft Access 2002 όπου είναι καταγραμμένα όλα τα κύρια στοιχεία της γνωστοποίησης καθώς και Μέτα Δεδομένα (Meta Data) για κάθε γνωστοποίηση. Η βάση δεδομένων αυτή ονομάζεται R_A_S και έχει μέγεθος 41MB με 28017 καταχωρημένες γνωστοποιήσεις σε 57 πίνακες και δεκάδες πεδία που αντιστοιχούν στην κάθε γνωστοποίηση.

Η Ε.Ε. δημιούργησε το R_A_S ένα εργαλείο το οποίο είναι φτιαγμένο, για να διευκολύνει την έρευνα στα στοιχεία των γνωστοποιήσεων, επειδή ο όγκος των στοιχείων που διατηρούνται για το RASFF είναι τεράστιος και είναι αδύνατη η αναζήτηση στοιχείων με την απλή ανάγνωση γνωστοποιήσεων. Πρόκειται για μια βάση δεδομένων δημιουργημένη στην Microsoft Access 2002.

Ένα συμπιεσμένο αντίγραφο της βάσης δεδομένων τοποθετείται κάθε βράδυ στο CIRCA-RASFF, στον φάκελο 4-GENERAL INFORMATION, κάτω από το τμήμα Library, το οποίο ενημερώνεται καθημερινά με τα νέα στοιχεία των γνωστοποιήσεων που καταγράφονται στην εν λόγω βάση δεδομένων. Για να αποκτήσουμε το εργαλείο φορτώνουμε το αρχείο με την ονομασία “RASFF database“ από το φάκελο 4-GENERAL INFORMATION στον υπολογιστή μας όπου και δημιουργεί το συμπιεσμένο αρχείο R_A_S και αν πατήσουμε διπλό click η βάση αποσυμπιέζεται αυτόματα και δημιουργείται ένα αρχείο της Microsoft Access το οποίο και είναι το εργαλείο. Για την χρήση και λειτουργίες του εργαλείου έχει συνταχθεί οδηγός από την Ε.Ε.

4.3.2 Προσέγγιση

Στόχος μου ήταν η λήψη αυτής της βάσης δεδομένων, η μετατροπή των στοιχείων σε μορφή που να είναι επεξεργάσιμη από το πρόγραμμα εξόρυξης δεδομένων και η αποτύπωση τυχόν προτύπων, ακολουθιών και τάσεων που είναι επαναλαμβανόμενες και μπορούν να οδηγήσουν σε δυνατότητα πρόβλεψης σε μελλοντικά προβλήματα που μπορεί να παρατηρηθούν.

4.4 Αποτέλεσμα

Δυστυχώς μετά από μελέτη των στοιχείων φάνηκε το εγχείρημα αυτό να μην είναι εφικτό, γιατί τα στοιχεία που διατηρούνται στο R_A_S αναφέρονται μόνο σε αρνητικά αποτελέσματα, δηλαδή τρόφιμα που δεν είναι ασφαλή. Στο R_A_S δεν υπάρχουν αναφορές στο πλήθος των ελέγχων που διενεργήθηκαν, έτσι ώστε να μπορούν να εξαχθούν συμπεράσματα. Έτρεξα δοκιμαστικά διάφορους αλγόριθμους στο R_A_S, αλλά τα αποτελέσματα που πήρα ήταν χωρίς οποιαδήποτε εγκυρότητα και σοβαρότητα αφού δεν απεικόνιζαν πραγματικές προβλέψεις.

Ένα παράδειγμα μπορεί να αποτυπώσει πιο εμπεριστατωμένα τους ενδοιασμούς μου. Σίγουρα μπορούμε να δούμε ότι η χώρα Χ μπορεί να έχει περισσότερα μη ασφαλή τρόφιμα τύπου Υ από οποιαδήποτε άλλη χώρα, αλλά ποια η χρησιμότητα της πληροφορίας αυτής αν δεν γνωρίζουμε πόσοι έλεγχοι διεξήχθηκαν. Αν έχουμε την πλήρη εικόνα μπορεί να παρατηρήσουμε ότι η Χώρα Χ είναι ο μεγαλύτερος παραγωγός του προϊόντος Υ και για αυτό γίνονται εντατικοί έλεγχοι από την ίδια, αλλά και από άλλες χώρες στα προϊόντα της που στην πραγματικότητα μπορεί τα μη ασφαλή προϊόντα Υ να αποτελούν ένα πολύ μικρό ποσοστό της όλης παραγωγής. Όμως η χώρα Ζ που επίσης παράγει το τρόφιμο Υ μπορεί να έχει ποσοστιαία πολύ χειρότερα αποτελέσματα από την χώρα Υ, που όμως αριθμητικά λόγω της περιορισμένης παραγωγής της να εμφανίζεται να έχει μικρό ολικό αριθμό περιστατικών μη ασφαλή τροφίμων Υ. Αποτέλεσμα είναι να προχωρούμε σε αυθαίρετα συμπεράσματα με τα οποία να στοχοποιούμε χώρες, παρασκευαστές και είδη τροφίμων χωρίς να είναι κατ΄ ανάγκη δυνητικά λιγότερο ασφαλή από άλλα και να εφαρμόζουμε επισταμένους ελέγχους σε αυτά, με αποτέλεσμα να μην αξιοποιούμε τους περιορισμένους μας πόρους επαρκώς

Για το λόγο αυτό τερματίστηκε η προσπάθεια εξόρυξης δεδομένων από τη βάση R_A_S. Για να συνεχιστεί η εργασία ζητήθηκε από το Εθνικό Σημείο του RASFF

πρόσβαση στα δεδομένα του αριθμού των ελέγχων που έγιναν, αλλά αυτά τα δεδομένα δεν ήταν διαθέσιμα, γιατί δε ζητούνται από την Ευρωπαϊκή Επιτροπή με αποτέλεσμα να χρειαστεί να ζητήσω από όλα τα Κράτη Μέλη να μου αποστείλουν όλα τα στοιχεία που διέθεταν σχετικά με τα δεδομένα αυτά.

Μια άλλη προσπάθεια που έγινε ήταν να συγκριθούν οι αριθμοί των ελέγχων που έγιναν από τις Υγειονομικές Υπηρεσίες με τις γνωστοποιήσεις που υπέβαλαν οι Υγειονομικές Υπηρεσίες στο RASFF, όμως αυτό δεν έχει νόημα, αφού για το έτος 2009 δηλώθηκαν μόλις 53 γνωστοποιήσεις από την Κύπρο και επιπλέον οι Υγειονομικές Υπηρεσίες δεν είναι μηχανογραφημένες με αποτέλεσμα να μην διατηρούν τα δεδομένα των ελέγχων των τροφίμων σε ηλεκτρονική μορφή. Με βάση τα πιο πάνω κρίθηκε ότι δεν υπήρχε νόημα να εφαρμόσει κάποιος εξόρυξη δεδομένων σε αυτό το δείγμα αφού τα αποτελέσματα σε τόσο μικρό σύνολο δεδομένων μπορεί να εξαχθούν μετά από σύντομη παρατήρηση.

4.5 Μελλοντικές αναζητήσεις στην εξόρυξη όσον αφορά την ασφάλεια τροφίμων

Η Υγειονομική Υπηρεσία άρχισε να εφαρμόζει σχέδια για την μηχανογράφησή της και αναμένεται να μηχανογραφηθεί εντός του 2012. Πλέον όλα τα αποτελέσματα του ελέγχου των τροφίμων θα είναι διαθέσιμα σε ηλεκτρονική μορφή, οπότε μπορεί να γίνει αντιπαραβολή των στοιχείων των Υγειονομικών Υπηρεσιών με αυτά της Επιτροπής και στο μέλλον να γίνει μελέτη για τυχόν διαφορές, που μπορεί να αποκαλύψουν χρήσιμα στοιχεία όπως το ότι στην Κύπρο έχουμε υπερβολική εμφάνιση αρνητικών αποτελεσμάτων σε συγκεκριμένα τρόφιμα. Αυτό δυνατό να έχει δύο ερμηνείες: είτε εφαρμόζουμε δυσανάλογα περισσότερους ελέγχους σε σύγκριση με άλλα Κράτη Μέλη ή κάτι υπάρχει στην παραγωγή, επεξεργασία ή διάθεση του εμπλεκόμενου είδους τροφίμου στην Κύπρο που το καθιστά ευάλωτο.

Άλλο θέμα που μπορεί να ερευνηθεί είναι η εξόρυξη δεδομένων μέσα από όλα τα δεδομένα του ελέγχου των τροφίμων που διεξήχθηκαν στην Κύπρο χωρίς να περιοριστούμε στις υποθέσεις που δηλώθηκαν στο RASFF. Αυτή η προσέγγιση μπορεί επίσης να γίνει ακόμη και πάλι μετά τη μηχανογράφηση των Υγειονομικών Υπηρεσιών, όπου τα πιο πάνω δεδομένα θα είναι διαθέσιμα σε ηλεκτρονική μορφή.

5. Εξόρυξη δεδομένων στα αποτελέσματα εργαστηριακών εξετάσεων πόσιμων νερών

Στην Κύπρο η αρμόδια Υπηρεσία για την παρακολούθηση και τον έλεγχο του πόσιμου νερού είναι οι Υγειονομικές Υπηρεσίες του Υπουργείου Υγείας, οι οποίες σε συνεργασία με άλλες Υπηρεσίες, εφαρμόζουν την προληπτική πολιτική του ελέγχου της ποιότητας του πόσιμου νερού στο επίπεδο της διανομής του νερού. Ακολουθεί σύντομη ανασκόπηση του νομικού πλαισίου και της εμπλοκής των διάφορων Υπηρεσιών στον έλεγχο του πόσιμου νερού.

5.1 Ο έλεγχος της ποιότητας του πόσιμου διασωληνωμένου νερού στην Κύπρο: Εισαγωγή για τον επιστήμονα πληροφορικής

5.1.1 Νομοθετικό Πλαίσιο

Μέχρι πρόσφατα, πριν από την ένταξη της Κυπριακής Δημοκρατίας στην ΕΕ, ο έλεγχος που γινόταν στο νερό, που προοριζόταν για ανθρώπινη κατανάλωση ήταν σύμφωνα με τις προδιαγραφές της Παγκόσμιας Οργάνωσης Υγείας.

Με την ένταξη της Κύπρου στην ΕΕ έχει υιοθετηθεί σαν ισχύουσα Νομοθεσία, η οδηγία για την ποιότητα του πόσιμου νερού 80/778/ΕΟΚ και η αναθεωρημένη Οδηγία 98/83/ΕΚ(ΟΔΗΓΙΑ *98/83/ΕΚ*) η οποία:

- Υποχρεώνει τα κράτη μέλη να παρακολουθούν την ποιότητα του νερού, που χρησιμοποιείται για ανθρώπινη κατανάλωση.
- Καθιερώνει αυστηρά ποιοτικά πρότυπα για το νερό, που προορίζεται για ανθρώπινη κατανάλωση.
- Καθορίζει μέγιστες επιτρεπτές και ενδεικτικές τιμές για τα επιμέρους συστατικά στοιχεία του ύδατος.
- Προβλέπει διαφάνεια και ενημέρωση του κοινού για τα αποτελέσματα του ελέγχου.

Ο έλεγχος πλέον γίνεται βάση της οδηγίας 98/83/ΕΚ του Συμβουλίου, της 3ης Νοεμβρίου 1998, σχετικά με την ποιότητα των νερών που προορίζονται για την ανθρώπινη κατανάλωση, που εναρμονίζεται στην Κυπριακή Νομοθεσία με τον περί της Ποιότητας του Νερού Ανθρώπινης Κατανάλωσης Παρακολούθηση και Έλεγχος Νόμος Ν87(Ι)2001.

5.1.2 Αρμοδιότητες και εμπλεκόμενες Υπηρεσίες

Το διασωληνωμένο πόσιμο νερό στην Κύπρο

Το διασωληνωμένο νερό που χρησιμοποιείται για πόση στην Κύπρο προέρχεται κυρίως από:

• Επιφανειακά νερά, που συγκεντρώνονται σε υδατοφράκτες
• Γεωτρήσεις
• Φυσικές πηγές
• Τη θάλασσα (μονάδες αφαλάτωσης)

Τα νερά αυτά τυγχάνουν αναλόγως επεξεργασίας σε διυλιστήρια του Τμήματος Αναπτύξεως Υδάτων (ΤΑΥ).

Στη συνέχεια το νερό διατίθεται στον καταναλωτή από τα Συμβούλια Υδατοπρομήθειας και από τις Τοπικές Αρχές και ο έλεγχος της ποιότητας του διανεμημένου νερού γίνεται από τις Υγειονομικές Υπηρεσίες.

Τμήμα Αναπτύξεως Υδάτων

Το Τμήμα Αναπτύξεως Υδάτων είναι υπεύθυνο για την υλοποίηση της υδατικής πολιτικής του Υπουργείου Γεωργίας, Φυσικών Πόρων και Περιβάλλοντος με σκοπό την ορθολογική ανάπτυξη και διαχείριση των υδάτινων πόρων της Κύπρου.

Οι βασικές αρμοδιότητες του Τμήματος Αναπτύξεως Υδάτων είναι:

- Η συλλογή και επεξεργασία υδρολογικών, υδρογεωλογικών και γεωτεχνικών στοιχείων για τη μελέτη και την ασφάλεια των αναπτυξιακών έργων.
- Η προστασία των υδάτινων πόρων από τις μολύνσεις και τις ρυπάνσεις του περιβάλλοντος.
- Η μελέτη, σχεδίαση, εκτέλεση, λειτουργία και συντήρηση έργων υποδομής, όπως φράγματα, λιμνοδεξαμενές, αρδευτικά, υδρευτικά και αποχετευτικά δίκτυα, διυλιστήρια νερού, μονάδες επεξεργασίας και επαναχρησιμοποίησης λυμάτων και μονάδες αφαλάτωσης νερού (Τμήμα Αναπτύξεως Υδάτων, 2010).

Συμβούλια Υδατοπρομήθειας και Τοπικές Αρχές

Τα Συμβούλια Υδατοπρομήθειας και οι Τοπικές Αρχές ασχολούνται με τη διανομή του νερού στους πολίτες. Τα Συμβούλια Υδατοπρομήθειας σύμφωνα με τον περί της Ποιότητας του Νερού Ανθρώπινης Κατανάλωσης Παρακολούθηση και Έλεγχο Νόμο Ν87(Ι)2001, ορίζονται σαν οι Φορείς Ύδρευσης.

Σύμφωνα με το άρθρο 5 του περί της Ποιότητας του Νερού Ανθρώπινης Κατανάλωσης Παρακολούθηση και Έλεγχος Νόμος Ν87(Ι)2001, κάθε Φορέας ύδρευσης οφείλει να λαμβάνει τα κατάλληλα μέτρα, ώστε το νερό ανθρώπινης κατανάλωσης που προμηθεύει στο κοινό, να διατηρείται πάντοτε υγιεινό και καθαρό και για το σκοπό αυτό οφείλει να συμμορφώνεται με τις σχετικές υποδείξεις του Αρχιεπιθεωρητή ή των Επιθεωρητών.

Γενικό Χημείο του Κράτους

Το Γενικό Χημείο του Κράτους (ΓΧΚ) είναι Τμήμα του Υπουργείου Υγείας που ιδρύθηκε το 1932 αρχικά ως το "Κρατικό Χημείο" εντός του Τμήματος Ιατρικών Υπηρεσιών και το 1981, αναβαθμίστηκε σε ανεξάρτητο Τμήμα του Υπουργείου. Από το 2001 δεκαέξι από τα εργαστήρια του ΓΧΚ έχουν διαπιστευθεί σύμφωνα με το Ευρωπαϊκό/Διεθνές πρότυπο ISO/IEC/EN 17025, ενώ τα υπόλοιπα τρία εργαστήρια του ΓΧΚ, εφαρμόζουν το ίδιο Σύστημα Διασφάλισης Ποιότητας με τα διαπιστευμένα εργαστήρια (Γενικό Χημείο του Κράτους, 2010).

Τα εργαστήρια του Γενικού Χημείου του Κράτους :

1. ΣΥΣΤΑΣΗΣ, ΠΟΙΟΤΗΤΑΣ ΚΑΙ ΘΡΕΠΤΙΚΗΣ ΑΞΙΑΣ ΤΡΟΦΙΜΩΝ	12. ΒΙΟΜΗΧΑΝΙΚΩΝ ΕΙΔΩΝ ΚΑΙ ΚΑΠΝΙΚΩΝ ΠΡΟΙΟΝΤΩΝ
2. ΓΕΝΙΚΩΝ ΑΝΑΛΥΣΕΩΝ ΝΕΡΩΝ	13. ΕΛΕΓΧΟΥ ΥΛΙΚΩΝ ΣΕ ΕΠΑΦΗ ΜΕ ΤΡΟΦΙΜΑ ΚΑΙ ΠΑΙΔΙΚΩΝ ΠΑΙΧΝΙΔΙΩΝ
3. ΔΙΚΑΝΙΚΗΣ ΧΗΜΕΙΑΣ ΚΑΙ ΤΟΞΙΚΟΛΟΓΙΑΣ	14. ΠΡΟΣΘΕΤΩΝ ΟΥΣΙΩΝ ΚΑΙ ΕΙΔΙΚΩΝ ΑΝΑΛΥΣΕΩΝ ΤΡΟΦΙΜΩΝ
4. ΕΛΕΓΧΟΥ ΦΑΡΜΑΚΩΝ, ΚΑΛΛΥΝΤΙΚΩΝ ΚΑΙ ΣΥΜΠΛΗΡΩΜΑΤΩΝ ΔΙΑΤΡΟΦΗΣ	15. ΠΕΡΙΒΑΛΛΟΝΤΙΚΗΣ κ.α. ΕΠΙΒΑΡΥΝΣΗΣ ΤΡΟΦΙΜΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΤΟΞΙΝΩΝ
5. ΥΠΟΛΕΙΜΜΑΤΩΝ ΚΤΗΝΙΑΤΡΙΚΩΝ ΦΑΡΜΑΚΩΝ	16. ΜΙΚΡΟΒΙΟΛΟΓΙΚΟΥ ΕΛΕΓΧΟΥ ΝΕΡΩΝ, ΦΑΡΜΑΚΩΝ ΚΑΙ ΠΕΡΙΒΑΛΛΟΝΤΟΣ
6. ΠΕΡΙΒΑΛΛΟΝΤΙΚΗΣ ΧΗΜΕΙΑΣ Ι	17. ΜΙΚΡΟΒΙΟΛΟΓΙΚΟΥ ΕΛΕΓΧΟΥ ΤΡΟΦΙΜΩΝ
7. ΟΙΚΟΤΟΞΙΚΟΛΟΓΙΑΣ	18. ΠΕΡΙΒΑΛΛΟΝΤΙΚΗΣ ΙΟΛΟΓΙΑΣ
8. ΥΠΟΛΕΙΜΜΑΤΩΝ ΦΥΤΟΦΑΡΜΑΚΩΝ	19. ΑΝΙΧΝΕΥΣΗΣ ΓΕΝΕΤΙΚΑ ΤΡΟΠΟΠΟΙΗΜΕΝΩΝ ΟΡΓΑΝΙΣΜΩΝ
9. ΡΑΔΙΕΝΕΡΓΕΙΑΣ ΣΤΑ ΤΡΟΦΙΜΑ ΚΑΙ ΠΕΡΙΒΑΛΛΟΝΤΙΚΗΣ ΡΑΔΙΕΝΕΡΓΕΙΑΣ	
10. ΠΕΡΙΒΑΛΛΟΝΤΙΚΗΣ ΧΗΜΕΙΑΣ ΙΙ	
11. SNIF-NMR	

Τα πιο πάνω εργαστήρια εφαρμόζουν σε συνεργασία με διάφορες Υπηρεσίες 49 Εθνικά Προγράμματα Ελέγχου εκ των οποίων τα δύο προγράμματα, τα οποία αφορούν τον ποιοτικό έλεγχο του νερού ανθρώπινης κατανάλωσης γίνονται σε συνεργασία με τις Υγειονομικές Υπηρεσίες. Τα εργαστήρια του Γενικού Χημείου, που εμπλέκονται στον έλεγχο της χημικής ποιότητας του νερού είναι το Εργαστήριο Αρ.2, 'Γενικών Αναλύσεων Νερών', που ασχολείται με τη χημική ποιότητα των νερών και το Εργαστήριο Αρ.6 'Περιβαλλοντικής Χημείας', το οποίο ασχολείται με τον προσδιορισμό των οργανικών ρυπαντών γεωργικών και βιομηχανικών στα Νερά (επιφανειακά, υπόγεια και πόσιμα). Όσον αφορά την μικροβιολογική ποιότητα του νερού ασχολείται το εργαστήριο 15 'Εργαστηριακός

έλεγχος Νερών και Φαρμάκων, που είναι και το Επίσημο Εργαστήριο για τον μικροβιολογικό έλεγχο των νερών. Και τα τρία πιο πάνω εργαστήρια είναι διαπιστευμένα σύμφωνα με το πρότυπο ISO/IEC/EN 17025 (Γενικό Χημείο του Κράτους, 2010).

Υγειονομική Υπηρεσία

Στην Κύπρο η αρμόδια Υπηρεσία για την παρακολούθηση και τον έλεγχο του πόσιμου νερού είναι οι Υγειονομικές Υπηρεσίες του Υπουργείου Υγείας, οι οποίες σε συνεργασία με άλλες Υπηρεσίες, εφαρμόζουν την προληπτική πολιτική του ελέγχου της ποιότητας του πόσιμου νερού στο επίπεδο του δικτύου. Οι αρμοδιότητες τους εκτείνονται από την εκροή της δημόσιας υδατοδεξαμενής μέχρι την εισροή σε ιδιωτική περιουσία, δηλαδή για όλο το δίκτυο. Σε συνεργασία με το Γενικό Χημείο του Κράτους καταρτίζονται δύο εθνικά προγράμματα ελέγχου για τον έλεγχο των νερών ανθρώπινης κατανάλωσης:

1. Μικροβιολογικός έλεγχος πόσιμου νερού για εφαρμογή της Νομοθεσίας Ν87(Ι)/2001 (Οδηγία 98/83/ΕΕ), συμπεριλαμβανομένου δικτύων υδατοπρομήθειας, βυτιοφόρων και νερού κερματοδεκτών.
2. Χημικός έλεγχος πόσιμου νερού για εφαρμογή της Νομοθεσίας Ν87(Ι)/2001 (Οδηγία 98/83/ΕΕ), συμπεριλαμβανομένου δικτύων υδατοπρομήθειας, βυτιοφόρων και νερού κερματοδεκτών κ.α.

Ο Διευθυντής Ιατρικών Υπηρεσιών και Υπηρεσιών Δημόσιας Υγείας, σύμφωνα με το άρθρο 6 του περί της Ποιότητας του Νερού Ανθρώπινης Κατανάλωσης Παρακολούθηση και Έλεγχος Νόμος Ν87(Ι)2001, ορίζεται ως η αρμόδια αρχή για τους σκοπούς του παρόντος Νόμου και έχει καθήκον να μεριμνά για την παρακολούθηση και έλεγχο του νερού ανθρώπινης κατανάλωσης και να συμβουλεύει τον Υπουργό. Οι Υγειονομικές Υπηρεσίες είναι η αρμόδια Υπηρεσία του Τμήματος Ιατρικών Υπηρεσιών και Υπηρεσιών Δημόσιας Υγείας που χειρίζεται το εν λόγω θέμα.

Ο Υπουργός Υγείας, σύμφωνα με το άρθρο 7 του περί της Ποιότητας του Νερού Ανθρώπινης Κατανάλωσης Παρακολούθηση και Έλεγχος Νόμος Ν87(Ι)2001, μπορεί με γνωστοποίηση, που δημοσιεύεται στην Επίσημη Εφημερίδα της Δημοκρατίας να διορίσει Αρχιεπιθεωρητή και Επιθεωρητές, οι οποίοι ενεργούν υπό τις οδηγίες του Διευθυντή για την εφαρμογή των διατάξεων του παρόντος Νόμου και των Κανονισμών που εκδίδονται με βάση το Νόμο. Οι λειτουργοί των Υγειονομικών Υπηρεσιών των Ιατρικών Υπηρεσιών και Υπηρεσιών Δημόσιας Υγείας, του Υπουργείου Υγείας και οι Υγειονομικοί Επιθεωρητές των Δήμων σύμφωνα με τις πρόνοιες του άρθρου αυτού, έχουν διοριστεί σαν οι αρμόδιοι λειτουργοί για την εφαρμογή των διατάξεων του Νόμου.

5.1.3 Έλεγχος του πόσιμου νερού

Οι Υγειονομικές Υπηρεσίες και οι τοπικές Αρχές παρακολουθούν τη χημική και μικροβιολογική ποιότητα του παρεχομένου στο κοινό νερού και το σύστημα παρακολούθησης περιλαμβάνει:

- Επιθεώρηση των πηγών και των περιβαλλοντικών συνθηκών κάθε πηγής
- Την καταγραφή όλων των πηγών υδατοπρομήθειας
- Δειγματοληψία για χημική και μικροβιολογική εξέταση
- Αξιολόγηση αποτελεσμάτων
- Παρακολούθηση του βαθμού χλωρίωσης και της υπολειμματικότητας του χλωρίου
- Διερεύνηση αιτιών πιθανής μόλυνσης
- Ενημέρωση αρμοδίων φορέων και λήψη μέτρων όπου χρειάζεται.

Σύμφωνα με το άρθρο 7 του περί της Ποιότητας του Νερού Ανθρώπινης Κατανάλωσης Παρακολούθηση και Έλεγχος Νόμος Ν87(Ι)2001, αποτελεί καθήκον του Αρχιεπιθεωρητή να προβαίνει σε **δοκιμαστική** και **ελεγκτική** παρακολούθηση του νερού σε όλες τις περιοχές τις Δημοκρατίας σύμφωνα με τις διατάξεις του Παραρτήματος ΙΙ της σχετικής νομοθεσίας. Η δοκιμαστική παρακολούθηση πραγματοποιείται με αναλύσεις των παραμέτρων που αναφέρονται στον Πίνακα Α του παραρτήματος ΙΙ και η ελεγκτική

παρακολούθηση με αναλύσεις όλων των παραμέτρων που αναφέρονται στα Μέρη Α, Β, και Γ του Παραρτήματος Ι. Οι Επιθεωρητές λαμβάνουν τα κατάλληλα μέτρα, ώστε τα δείγματα που λαμβάνουν να μην υφίστανται οποιεσδήποτε αλλοιώσεις μέχρι να διαβιβαστούν στο Γενικό Χημείο του Κράτους. Η δειγματοληψία, διενεργείται σύμφωνα με το πρωτόκολλο δειγματοληψίας και μεταφοράς δειγμάτων πόσιμου νερού που έχει καταρτισθεί από τις Υγειονομικές Υπηρεσίες.

Η νομοθεσία προβλέπει συχνότητα ελέγχου που καθορίζεται ανάλογα με το μέγεθος του πληθυσμού που υδρεύεται από το κάθε δίκτυο ύδρευσης. Οι Υγειονομικές Υπηρεσίες έχουν καταρτίσει τακτικά προγράμματα παρακολούθησης της ποιότητας του πόσιμου νερού, όπου γίνονται δειγματοληψίες νερού και εργαστηριακή εξέταση του στα διαπιστευμένα εργαστήρια του Γενικού Χημείου του Κράτους. Σκοπός η διασφάλιση της ποιότητας του μέσω της παρακολούθησης των χημικών και μικροβιολογικών παραμέτρων του νερού ανθρώπινης κατανάλωσης.

Αξιολόγηση των Αποτελεσμάτων

Σε περίπτωση που ο Διευθυντής, διαπιστώνει ή έχει λόγους να πιστεύει ότι στο νερό που προμηθεύει οποιοσδήποτε Φορέας, περιέχονται ουσίες ή μικροοργανισμοί που ενώ δεν αποτελούν παραμέτρους που αναφέρονται στο Παράρτημα 1, εντούτοις περιέχονται σε τέτοιες ποσότητες ή αριθμούς, ώστε να δημιουργείται κίνδυνος στη δημόσια υγεία από τη χρήση του νερού αυτού, σύμφωνα με τις πρόνοιες του άρθρου 15 του περί της Ποιότητας του Νερού Ανθρώπινης Κατανάλωσης Παρακολούθηση και Έλεγχος Νόμος Ν87(Ι)2001, οφείλει να μεριμνά ώστε να διενεργείται συμπληρωματική παρακολούθηση του νερού. Να πληροφορεί κατάλληλα το Φορέα ή τους Φορείς ύδρευσης, που παρέχουν το νερό και να εισηγείται τη λήψη κατάλληλων μέτρων για εξάλειψη του

κινδύνου και να παρακολουθεί κατά πόσο λαμβάνονται μέτρα που επιτυγχάνουν το ίδιο αποτέλεσμα.

Σε περίπτωση που η υπέρβαση αφορά τη δοκιμαστική παρακολούθηση, τότε ο Διευθυντής εξετάζει κατά πόσο το γεγονός αυτό δημιουργεί οποιοδήποτε κίνδυνο για τη δημόσια υγεία, λαμβάνοντας υπόψη το βαθμό απόκλισης από την καθορισμένη παραμετρική τιμή και τη χρονική περίοδο που διαρκεί ή δυνατό να διαρκέσει η απόκλιση. Πληροφορεί κατόπιν το Φορέα ή τους Φορείς ύδρευσης που προμηθεύουν το νερό αναφορικά με τις διαπιστώσεις του και εισηγείται τη λήψη μέτρων που κρίνει αναγκαία προς αποκατάσταση της ποιότητας του νερού.

Διορθωτικές Ενέργειες

Σύμφωνα με το ¨Εγχειρίδιο για την παρακολούθηση και έλεγχο της ποιότητας του νερού ανθρώπινης κατανάλωσης¨(2009) σε περίπτωση απόκλισης σε δείγμα που λήφθηκε από το δίκτυο θα πρέπει να γίνουν οι ακόλουθες ενέργειες:

Μόλις γίνει προφορική ενημέρωση από το Γενικό Χημείο του Κράτους, ενημερώνεται τηλεφωνικά ο Φορέας ΄Υδρευσης για τα αποτελέσματα των εργαστηριακών εξετάσεων και ο Προϊστάμενος Υγειονομικών Υπηρεσιών, ενώ προγραμματίζεται παράλληλα επαναληπτική δειγματοληψία. Ακολουθεί γραπτή ενημέρωση του Φορέα μετά τη λήψη της εργαστηριακής έκθεσης με εισήγηση για άμεση διερεύνηση με σκοπό τον εντοπισμό της αιτίας που προκάλεσε την απόκλιση, καθώς και λήψη όλων των αναγκαίων μέτρων για προστασία της δημόσιας υγείας ενώ μπορεί να γίνουν και εισηγήσεις. Στη συνέχεια γίνεται παρακολούθηση των διορθωτικών ενεργειών, που λαμβάνονται από το Φορέα ΄Υδρευσης για αποκατάσταση της ποιότητας του νερού. Μετά τη λήψη των διορθωτικών μέτρων από το Φορέα ΄Υδρευσης, ακολουθεί επαναληπτική δειγματοληψία από τον Υγειονομικό Επιθεωρητή για επαλήθευση της αποτελεσματικότητας των διορθωτικών μέτρων που λήφθηκαν.

Ο Φορέας οφείλει να προβεί χωρίς καθυστέρηση στις κατάλληλες ενέργειες για προστασία της δημόσιας υγείας, περιλαμβανομένης της διακοπής της προμήθειας του νερού και της λήψης μέτρων αποκατάστασης της ποιότητας του και ο Φορέας πρέπει να πληροφορήσει το κοινό για τα μέτρα που λαμβάνονται, εκτός εάν ο Διευθυντής και Φορέας αποφασίσουν από κοινού ότι η απόκλιση είναι άνευ σημασίας.

5.1.4 Ενημέρωση του κοινού

Εάν από τις αναλύσεις οποιουδήποτε δείγματος που λαμβάνεται σύμφωνα με τις διατάξεις του Νόμου, διαπιστωθεί ότι η τιμή οποιασδήποτε από τις παραμέτρους υπερβαίνει την τιμή που καθορίζεται για την παράμετρο αυτή στη νομοθεσία, τότε ο Διευθυντής σύμφωνα με το άρθρο 14 του περί της Ποιότητας του Νερού Ανθρώπινης Κατανάλωσης Παρακολούθηση και Έλεγχος Νόμος Ν87(Ι)2001, χωρίς καθυστέρηση οφείλει να προβαίνει στις κατάλληλες ενέργειες, για να πληροφορηθεί το κοινό, που χρησιμοποιεί το νερό από το οποίο λήφθηκε το δείγμα για τον ενδεχόμενο κίνδυνο στην υγεία. Παράλληλα πληροφορεί το Φορέα ή τους Φορείς ύδρευσης, που προμηθεύουν το νερό αναφορικά με τις διαπιστώσεις του και εισηγείται τη λήψη μέτρων για προστασία της δημόσιας υγείας, περιλαμβανομένης της διακοπής της παροχής ή περιορισμού χρήσης του νερού.

Πέραν των πιο πάνω ο Διευθυντής Ιατρικών Υπηρεσιών και Υπηρεσιών Δημόσιας Υγείας σύμφωνα με το άρθρο 16 του περί της Ποιότητας του Νερού Ανθρώπινης Κατανάλωσης Παρακολούθηση και Έλεγχος Νόμος Ν87(Ι)2001, όρισε τον Προϊστάμενο Υγειονομικών Υπηρεσιών να μεριμνά όπως σε καθορισμένα χρονικά διαστήματα δημοσιεύονται πληροφορίες αναφορικά με τη γενική ποιοτική κατάσταση του πόσιμου νερού. Η ενημέρωση για την ποιοτική κατάσταση του πόσιμου νερού γίνεται ανά τετραμηνία και περιλαμβάνει τα αποτελέσματα εργαστηριακών εξετάσεων δειγμάτων νερού, τα οποία λαμβάνονται από τους Υγειονομικούς Λειτουργούς. Σε αυτή καταγράφονται

επίσης τα μέτρα που λαμβάνονται από τους Υγειονομικούς Λειτουργούς και τους φορείς ύδρευσης ξεχωριστά για κάθε σημείο δειγματοληψίας, στην περίπτωση μη συμμόρφωσης με τις παραμετρικές τιμές που περιέχονται στη Νομοθεσία.

Κάθε Κράτος Μέλος σύμφωνα με το άρθρο 13 της οδηγίας 98/83/ΕΚ του Συμβουλίου, της 3ης Νοεμβρίου 1998, σχετικά με την ποιότητα των νερών που προορίζονται για την ανθρώπινη κατανάλωση, δημοσιεύει ανά τριετία έκθεση για την ποιότητα του νερού ανθρώπινης κατανάλωσης, με σκοπό την ενημέρωση των καταναλωτών. Η πρώτη από τις εκθέσεις αυτές καλύπτει τα έτη 2002, 2003 και 2004. Κάθε έκθεση αφορά τουλάχιστον τις ατομικές παροχές νερού που υπερβαίνουν τα 1 000 m3 ημερησίως κατά μέσο όρο, ή εξυπηρετούν ανά των 5 000 ατόμων. Η έκθεση καλύπτει τρία ημερολογιακά έτη και δημοσιεύεται πριν από το τέλος του ημερολογιακού έτους που έπεται της περιόδου στην οποία αναφέρεται.

5.2 Η παρούσα έρευνα σε σχέση με την ποιότητα του πόσιμου νερού-ποιότητα πόσιμου νερού και εξόρυξη δεδομένων

Αν και στην Κύπρο δεν είχαμε πρόσφατα φαινόμενα μεγάλων επιδημιών, που οφείλονταν στο μολυσμένο νερό με τους τεράστιους αριθμούς των θυμάτων, όπως συμβαίνει με αναπτυσσόμενες χώρες, σύμφωνα με έρευνα του Pacific Institute for Studies in Development, Environment, and Security (Gleick, 2002) από το 2002 μέχρι το 2020 θα πεθάνουν 135 εκατομμύρια άνθρωποι στον κόσμο από λοιμώδη ασθένειες, που προέρχονται από την πόση μη ασφαλούς νερού, αν η κατάσταση παραμείνει η ίδια όσον αφορά την ασφάλεια του πόσιμου νερού στον κόσμο.

Κατά την περίοδο της αποικιοκρατίας, αν ανατρέξουμε στα αρχεία θα διαπιστώσουμε ότι και στην Κύπρο είχαμε αρκετές περιπτώσεις κρουσμάτων ακόμη και θανάτους τυφοειδούς πυρετού, σε κοινότητες που παραδοσιακά

υδρεύονταν με λάκκους και το νερό τους εμολύνετο υπόγεια από αποχετεύσεις παρακειμένων αποχωρητηρίων ή απορροφητικών λάκκων (Υγειονομικές Υπηρεσίες, 2010). Το νερό της βρύσης που οι περισσότεροι θεωρούμε ότι είναι ασφαλές και ούτε που περνά από το μυαλό μας ότι μπορεί να απειλήσει την Υγεία μας, στην πραγματικότητα μπορεί να αποτελέσει ένα σιωπηλό όπλο μαζικής καταστροφής αν δεν λαμβάνονται όλα τα μέτρα για την προστασία της Δημόσιας Υγείας.

Σε διεθνές επίπεδο, έχει αναδειχθεί πάμπολλες φορές η σημασία που δίνει η διεθνής κοινότητα στην ασφάλεια του νερού, ενώ υπάρχουν διεθνής που το θέμα είναι στις βασικές τους αρμοδιότητες. Παράδειγμα αποτελεί η συμπερίληψη της παροχής νερού και της προστασίας της ποιότητας στην Ατζέντα 21 της διάσκεψης του Ρίο το 1992 (www.un.org).

Η σημασία της διαχείρισης της ποιότητας του νερού ανάγκασε αρκετές χώρες να διερευνήσουν τρόπους ελέγχου της μόλυνσης του νερού. Στα πλαίσια αυτής της προσπάθειας, εφαρμόστηκαν τεχνικές όπως η Γεωχωρική εξόρυξη δεδομένων (Geospatial Data Mining). Η Γεωχωρική εξόρυξη δεδομένων αναφέρεται στο χειρισμό γεωχωρικών δεδομένων, δηλαδή δεδομένων που σχετίζονται με μια συγκεκριμένη γεωγραφική περιοχή. Λόγω της εξάρτησης που υπάρχει μεταξύ των δεδομένων και μιας συγκεκριμένης περιοχής, οι πληροφορίες που εξάγονται από ένα δείγμα που αφορά μια συγκεκριμένη περιοχή πιθανό να μην είναι γενικεύσιμες, να μην ισχύουν για άλλες περιοχές. Παράδειγμα τέτοιων αναλύσεων αποτελεί το έργο “The Alabama Watershed Demonstration” , το οποίο συσχετίζει μοτίβα καλλιέργειας της Γης με την ποιότητα του νερού, μέσω της Γεωχωρικής Εξόρυξης Δεδομένων (Flynn, 1999). Επίσης, οι Karimipouri, Delavari και Kinaie (2005) σε μια μελέτη περίπτωσης, διερεύνησαν τη συσχέτιση μεταξύ βιομηχανικής μόλυνσης και ποιότητας του νερού. Τα αποτελέσματα της έρευνάς τους αποκάλυψαν τη σχέση μεταξύ του αριθμού και της τοποθεσίας βιομηχανικών αποβλήτων και της ποιότητας του νερού.

Για να μπορέσουν να ανταποκριθούν στις πιο πάνω υποχρεώσεις οι Υγειονομικές Υπηρεσίες διατηρούν στοιχεία για την ποιότητα των νερών σε ηλεκτρονική μορφή και αυτά τα στοιχεία δυνατό να είναι αρκετά ενδιαφέρον ερευνητικά, αφού λόγω του πλήθους τους κανείς δεν μπορεί να τα επεξεργαστεί αποτελεσματικά και μεμονωμένες παρατηρήσεις εξάγονται μόνο από απλή παρατήρηση.

Παραχώρηση άδειας πρόσβασης στα δεδομένα των ελέγχων του πόσιμου νερού

Για να εξασφαλισθεί η εν λόγω άδεια έπρεπε να πειστεί η διεύθυνση των Υγειονομικών Υπηρεσιών, για να δώσει άδεια πρόσβασης σε αυτά και σε μορφή που μπορεί να επεξεργαστεί. Για το θέμα είχα αποταθεί στον Προϊστάμενο Υγειονομικών Υπηρεσιών με τον οποίο και είχα μακροσκελή συνάντηση όπου συζητήσαμε εκτενώς το θέμα. Κατά τη διάρκεια της συνάντησης μας του ανάλυσα του σκοπούς της εργασίας και τον τρόπο που σκόπευα να εργαστώ, καθώς και τα πιθανά ωφέλη για την Υπηρεσία και η εν λόγω άδεια μου δόθηκε. Στο Παράρτημα Ι επισυνάπτεται ολόκληρη η αλληλογραφία.

5.3 Εφαρμογή

5.3.1 Προσέγγιση

Η έρευνα, για να ανακαλυφθεί κατά πόσο υπάρχουν επαναλαμβανόμενες ακολουθίες και τάσεις στην εμφάνιση των αποτελεσμάτων του ελέγχου του πόσιμου νερού, που μπορεί να μας δώσουν τη δυνατότητα πρόβλεψης σε μελλοντικά προβλήματα, που μπορεί να παρατηρηθούν δεν ήταν επίσης απλή αλλά απαιτούσε χρόνο και πόρους.

Πρώτα από όλα, όπως και την περίπτωση της Εξόρυξης δεδομένων στα αποτελέσματα εργαστηριακών εξετάσεων πόσιμων νερών ήταν πρόβλημα η πρόσβαση στα εν λόγω δεδομένα. Για να εξασφαλιστεί η εν λόγω άδεια έπρεπε

να πειστεί η διεύθυνση των Υγειονομικών Υπηρεσιών, για να δώσει άδεια πρόσβασης σε αυτά και σε μορφή που μπορεί να επεξεργαστεί.

Για το θέμα είχα αποταθεί γραπτώς (δες Παράρτημα Ι) στον Προϊστάμενο Υγειονομικών Υπηρεσιών με τον οποίο και είχα μακροσκελή συνάντηση με όπου συζητήσαμε εκτενώς το θέμα, του ανάλυσα του σκοπούς μου και τον τρόπο που σκόπευα να εργαστώ καθώς και τα πιθανά ωφέλη για την Υπηρεσία. Τελικά η εν λόγω άδεια μου δόθηκε και επισυνάπτεται στο Παράρτημα Ι.

Η αξιολόγηση των αποτελεσμάτων που μπορεί να γίνει είναι ποσοτική και ποιοτική. Δηλαδή αν ασχοληθούμε με τις μετρήσεις που υπάρχουν για τις διάφορες παραμέτρους μπορούμε να εντοπίσουμε μοτίβα στις τιμές ή ακραίες τιμές, που να χρήζουν περαιτέρω διερεύνησης. Για αυτό τον τύπο ανάλυσης θα χρειαστούμε όλες τις παραμέτρους του κάθε δείγματος. Το αποτέλεσμα της εργαστηριακής ανάλυσης θεωρείται η εξαρτημένη ποσοτική μεταβλητή. Η εξαρτημένη μεταβλητή, θα μπορούσε να είναι και η αξιολόγηση του αποτελέσματος που είναι ποιοτική μεταβλητή, η οποία είναι εναλλακτικός τρόπος παρουσίασης του αποτελέσματος στηριζόμενη στο αποτέλεσμα και το όριο που καθορίζει η νομοθεσία. Δεν συστήνεται να εξετάζουμε δείγματα με 2 εξαρτώμενες μεταβλητές πόσο μάλλον μια ποιοτική και μια ποσοτική. Ουσιαστικά είναι η ίδια μεταβλητή αλλά με διαφορετική παρουσίαση.

Έτσι πέραν από την ποσοτική αξιολόγηση των αποτελεσμάτων μπορεί να γίνει και ποιοτική αξιολόγηση τους εφόσον για το κάθε δείγμα υπάρχει αξιολόγηση κατά πόσο το δείγμα συνάδει ή όχι με τη νομοθεσία.

5.3.2 Περιγραφή δεδομένων ελέγχου πόσιμου νερού

Τα δεδομένα για τον έλεγχο του πόσιμου νερού δημιουργούνται με τον εξής τρόπο. Ο έλεγχος γίνεται με συχνότητα που καθορίζεται από τα εθνικά προγράμματα ελέγχου :

1. Μικροβιολογικός έλεγχος πόσιμου νερού
2. Χημικός έλεγχος πόσιμου νερού

Η διαδικασία προγραμματισμού και διενέργειας του δειγματοληπτικού ελέγχου καλύπτεται από εγχειρίδια και οδηγούς των Υγειονομικών Υπηρεσιών που αναφέρονται στο σημείο 4.3. Τα σημεία δειγματοληψίας είναι καταγραμμένα και έχουν κωδικό αναγνώρισης. Ο Λειτουργός των Υγειονομικών Υπηρεσιών που προβαίνει σε δειγματοληψία συμπληρώνει σε προκαθορισμένο έντυπο τα στοιχεία για τη δειγματοληψία. Στο εν λόγω έντυπο καταγράφονται στοιχεία του δειγματολήπτη (ονοματεπώνυμο, τίτλος, Υπηρεσία), ή η ημερομηνία της δειγματοληψίας, ο κωδικός του σημείου δειγματοληψίας και η ονομασία του δείγματος (χώρος που λήφθηκε, χρονική στιγμή, παρατηρήσεις) και ο σκοπός της δειγματοληψίας (π.χ. μικροβιολογική/ χημική ανάλυση). Στη συνέχεια ο Υγειονομικός Λειτουργός μεταφέρει το δείγμα στο Γενικό Χημείο του Κράτους κάτω από κατάλληλες συνθήκες φύλαξης (π.χ θερμοκρασία) και το παραδίδει στο Γενικό Χημείο του Κράτους.

Στη συνέχεια το Γενικό Χημείο του Κράτους καταχωρεί τα δείγματα στο αρχείο του και τους αποδίδει ένα μοναδικό κωδικό αναγνώρισης τον Αρ. Γενικού Χημείου του δείγματος, ο οποίος αποτελεί και το κλειδί αναγνώρισης του δείγματος. Το δείγμα παραδίνεται στο αρμόδιο εργαστήριο (Εργαστήριο Μικροβιολογικής Εξέτασης Νερών/Εργαστήριο Χημικής Εξέτασης Νερών) όπου θα πραγματοποιήσει την εργαστηριακή ανάλυση. Στη συνέχεια πραγματοποιούνται οι εργαστηριακές εξετάσεις, καταγράφονται τα αποτελέσματα και εκδίδονται τα πιστοποιητικά εργαστηριακής εξέτασης. Στην περίπτωση που τα αποτελέσματα δε συνάδουν με τη νομοθεσία, τότε άμεσα ειδοποιείται τηλεφωνικά η Υπηρεσία που διενέργησε τη δειγματοληψία για άμεση λήψη διορθωτικών μέτρων (Η διαδικασία επεξηγείται αναλυτικά στο 4.3.3).

Το Γενικό Χημείο του Κράτους έχει δημιουργήσει το λογισμικό LIMS (**Laboratory Information Management System**) στην Microsoft Visual Foxpro 9.0 όπου και αποθηκεύει όλα τα πιο πάνω δεδομένα. Στις Υγειονομικές Υπηρεσίες τόσο στην

Κεντρική Υπηρεσία όσο και σε κάθε Επαρχία έχουν ορισθεί σημεία επαφής στους Ηλεκτρονικούς Υπολογιστές των οποίων έχει εγκατασταθεί το εν λόγω λογισμικό και σε εβδομαδιαία βάση αποστέλλεται αρχείο το οποίο και περιέχει δεδομένα του τρέχοντος έτους, το οποίο και προστίθεται στη βάση δεδομένων αντικαθιστώντας το προηγούμενο.

Τα δεδομένα είναι αρχειοθετημένα ανά εργαστήριο, όμως τα εργαστήρια που ασχολούνται με τον έλεγχο του πόσιμου διασωληνωμένου νερού ασχολούνται και με άλλα θέματα και για αυτό αναμένεται τα δεδομένα να είναι αναμεμειγμένα με άλλα δεδομένα που αφορούν τον έλεγχο εμφιαλωμένων νερών, κερματοδεκτών πόσιμου νερού, βυτιοφόρων νοσοκομείων, πηγών/γεωτρήσεων και άλλα πολλά δεδομένα, τα οποία δεν αφορούσαν τον έλεγχο του πόσιμου διασωληνωμένου νερού.

Τα πιο πάνω δεδομένα είναι αυτά που χρησιμοποιούνται από τις Υγειονομικές Υπηρεσίες για δημιουργία εκθέσεων, ερευνών καθώς και για τη δημιουργία των ετήσιων εκθέσεων για το Υπουργείου Υγείας και των τριετών εκθέσεων για την Ευρωπαϊκή Επιτροπή.

Μετά τη συγκατάθεση του Προϊστάμενου Υγειονομικών Υπηρεσιών, μου έχει δοθεί άδεια στα δεδομένα των εργαστηριακών εξετάσεων του πόσιμου νερού για τα έτη 2008, 2009 και 2010.

5.4 Επιλογή Λογισμικού

Για την εξόρυξη δεδομένων επιλέχθηκε το Weka, που αναπτύσσεται στο πανεπιστήμιο Waikato, στη Νέα Ζηλανδία και είναι ένα δημοφιλής λογισμικό που εργάζεται σε περιβάλλον Java. Το Weka είναι ένα ελεύθερα διαθέσιμο λογισμικό που διαθέτει μια συλλογή εργαλείων και αλγορίθμων απεικόνισης για την ανάλυση στοιχείων και το προβλεπτικό μοντέλο, μαζί με τα γραφικά ενδιάμεσα με το χρήστη για την εύκολη πρόσβαση σε αυτήν τη λειτουργία. Η αρχική έκδοση

μη-Java, Weka ήταν αλγόριθμοι μιας προηγούμενης διαμόρφωσης που εφαρμόστηκε σε άλλες γλώσσες προγραμματισμού, για τα πειράματα εκμάθησης μηχανών. Αυτή η αρχική έκδοση σχεδιάστηκε πρώτιστα ως εργαλείο για επεξεργασία στοιχείων από γεωργικές περιοχές, (Holmes, Donkin και Witten, 1994) αλλά η πιο πρόσφατη έκδοση (Weka 3), βασισμένη στη Java, χρησιμοποιείται τώρα σε πολλούς διαφορετικούς τομείς εφαρμογής, και ιδιαίτερα για εκπαιδευτικούς λόγους και την έρευνα.

Το Weka υποστηρίζει διάφορους τυποποιημένους στόχους εξόρυξης δεδομένων, πιο συγκεκριμένα, τα στοιχεία που προεπεξεργάζονται, που συγκεντρώνονται, ταξινόμηση, οπισθοδρόμηση, απεικόνιση, και επιλογή χαρακτηριστικών γνωρισμάτων. Το εν λόγω λογισμικό, παρέχει την πρόσβαση στις βάσεις δεδομένων SQL χρησιμοποιώντας τη συνδετικότητα βάσεων δεδομένων της Java και μπορεί να επεξεργαστεί το αποτέλεσμα, που επιστρέφεται από μια ερώτηση βάσεων δεδομένων.

Σύμφωνα με τους Reutemann, Pfahringer και Frank (2004) το WEKA δεν είναι ικανό της πολυ-συγγενικής ανάσυρσης δεδομένων, αλλά υπάρχει χωριστό λογισμικό για τη μετατροπή μιας συλλογής των συνδεμένων πινάκων βάσεων δεδομένων σε έναν ενιαίο πίνακα, που είναι κατάλληλος για την επεξεργασία που χρησιμοποιεί Weka.

Σύμφωνα με τη Wikipedia, (2010) τα πλεονεκτήματα του Weka περιλαμβάνουν:

1. Ελεύθερα διαθέσιμο
2. Επειδή εφαρμόζεται πλήρως στη γλώσσα προγραμματισμού της Java, μπορεί να τρέξει σχεδόν σε οποιαδήποτε σύγχρονη πλατφόρμα υπολογιστή
3. Πολύ περιεκτική συλλογή εργαλείων/μοντέλων προεπεξεργασίας στοιχείων και των τεχνικών εξόρυξης
4. Ευκολία στη χρήση λόγω των γραφικών που περιέχει

Πέραν των πιο πάνω υπάρχει στο WEKA επίσης η δυνατότατα για συστηματική σύγκριση της προβλεπτικής απόδοσης των αλγορίθμων εκμάθησης.

5.5 Προετοιμασία δεδομένων

Τα δεδομένα όπως περιγράφεται στο 6.3 είναι αναρτημένα στο λογισμικό LIMS (**Laboratory Information Management System**), που είναι μια βάση δεδομένων δημιουργημένη στην Microsoft Visual Foxpro 9.0. Ο μοναδικός τρόπος εξόδου των δεδομένων είναι σε μορφή πίνακα στην Microsoft Excel, έτσι έγινε εξαγωγή των αποτελεσμάτων των εργαστηριακών αποτελεσμάτων, που έγιναν στα εργαστήρια Μικροβιολογικού ελέγχου του πόσιμου νερού, σε φύλλα (spreadsheets) της Excel

Τα δεδομένα εξήχθηκαν σε διαφορετικά αρχεία, ξεχωριστά τα αποτελέσματα των δύο εργαστηρίων. Για να μπορέσει να πραγματοποιηθεί η εξαγωγή χρειάστηκε να γίνουν ξεχωριστά αρχεία για κάθε έτος γιατί η excel 2003 δεν μπορούσε να χειριστεί τόσο μεγάλο όγκο δεδομένων.

5.5.1 Καθαρισμός των δεδομένων

Τα αποτελέσματα, εκτός από τις εργαστηριακές εξετάσεις πόσιμου νερού, περιείχαν αποτελέσματα εξέτασης θαλάσσιου νερού και κολυμβητικών δεξαμενών. Τα δεδομένα που υπήρχαν στα αρχεία χρειαζόταν να καθαριστούν και σαν πρώτη κίνηση αυτά τα δεδομένα θα έπρεπε να αποκλειστούν από το σύνολο των δεδομένων. Οι κίνδυνοι να χειριστούμε αποτελέσματα που να αφορούν θέματα πέραν του πόσιμου νερού είναι πολλαπλοί. Εκτός από το να έχουμε τιμές, που μπορεί να μας αλλοιώσουν τα αποτελέσματα μπορεί να έχουμε και διαφορετικά όρια με αποτέλεσμα μια τιμή για την ίδια παράμετρο σε μια χρήση να θεωρείται επιθυμητή, ενώ σε άλλη ανεπιθύμητη (π.χ. υπολειματικότητα χλωρίου σε νερό κολυμβητικών δεξαμενών και σε πόσιμο νερό).

Εξετάζοντας τα δεδομένα που αφορούσαν αποτελέσματα ελέγχου του πόσιμου νερού παρατηρήθηκε ότι τα δεδομένα περιλάμβαναν διάφορους ελέγχους, πέραν από αυτόν του δικτύου υδατοπρομήθειας, όπως κερματοδέκτες, νερό για ειδικούς σκοπούς (νοσοκομειακή χρήση, κτλ), νερό πλοίων, πηγών/γεωτρήσεων, ψυκτών, βυτιοφόρων κ.α. Έτσι χρειάστηκε να απομονωθούν τα αποτελέσματα που μας ενδιαφέρουν, απομονώνοντας από το κάθε αχρείο τα αποτελέσματα που αφορούσαν μόνο πόσιμο νερό.

Στη συνέχεια, έγινε αξιολόγηση των πεδίων που διατηρούνται στη βάση δεδομένων. Στα δεδομένα που εξήχθηκαν από το LIMS υπάρχει μια γραμμή, που αντιστοιχεί σε κάθε παράμετρο που εξετάστηκε ένα δείγμα, έτσι ένα δείγμα απαντάται σε αριθμό γραμμών, ισάριθμο των παραμέτρων που εξετάστηκαν. Με την μορφή που πήρε η βάση δεδομένων μας, δεν υπάρχει μοναδικό πεδίο αναγνώρισης για κάθε εγγραφή, αλλά σαν «primary key» μπορεί να χρησιμοποιηθεί ο συνδυασμός που προκύπτει από την παράμετροεξέτασης, την ημερομηνία δειγματοληψίας και τον Αριθμό Γενικού Χημείου του Κράτους που λαμβάνει κάθε δείγμα με την παραλαβή του από το Γενικό Χημείο του Κράτους. Στο LIMS για την κάθε ανάλυση διατηρούνται 49 πεδία για τον έλεγχο των νερών από τα οποία 23 δεν συμπληρώνονται, γιατί το LIMS έγινε για να καλύψει όλα τα προϊόντα που γίνονται εργαστηριακή εξέταση στο Γενικό Χημείο του Κράτους και προφανώς όλες οι παράμετροι δεν εφαρμόζονται στον έλεγχο των νερών. Τα κύρια στοιχεία που αποθηκεύονται στη βάση δεδομένων είναι ο Αρ. Γενικού Χημείου του δείγματος, ο κωδικός του σημείου δειγματοληψίας και η ονομασία του, η ημερομηνία της δειγματοληψίας, ο δειγματολήπτης, η παράμετρος για την οποία έγινε εργαστηριακή εξέταση, το αποτέλεσμα και η αξιολόγηση του αποτελέσματος (δηλαδή αν συνάδει ή όχι με τη νομοθεσία).

Στη συνέχεια έγινε απομόνωση των δεδομένων, διαγράφοντας τις κενές στήλες, και περεταίρω επεξεργασίατους . Συγκεκριμένα, παρατηρήθηκε ότι υπήρχαν δεδομένα για σημεία δειγματοληψίας για τα οποία δεν υπήρχε κωδικός και αφού ο κωδικός θα χρησιμοποιόταν ωςτο μοναδικό στοιχείο για τον καθορισμό του

σημείου δειγματοληψίας χρειάστηκε να απαλειφθούν τέτοια δεδομένα όπως επίσης και ελλειπείς καταχωρήσεις, για παράδειγμα εργαστηριακές εξετάσεις χωρίς καταχώρηση της παραμέτρου ή της αξιολόγησης, για να καταλήξουμε στη τελική μορφή των δεδομένων που θα χρησιμοποιούνταν για ανάλυση.

Η ανάλυση των αποτελεσμάτων που μπορεί να γίνει είναι ποσοτική και ποιοτική. Δηλαδή αν ασχοληθούμε με τις μετρήσεις που υπάρχουν για τις διάφορες παραμέτρους μπορούμε να εντοπίσουμε μοτίβα στις τιμές ή ακραίες τιμές που να χρήζουν περαιτέρω διερεύνησης. Για αυτό τον τύπο ανάλυσης θα χρειαστούμε όλες τις παραμέτρους του κάθε δείγματος.

Πέραν από την ποσοτική ανάλυση των αποτελεσμάτων μπορεί να γίνει και ποιοτική ανάλυσή τους. Για το κάθε δείγμα υπάρχει αξιολόγηση για το κατά πόσο το δείγμα συνάδει ή όχι με τη νομοθεσία. Ένα σημείο που πρέπει να διευκρινισθεί είναι ότι στην περίπτωση που μια παράμετρος δεν συνάδει με τη νομοθεσία τότε στην αξιολόγηση των αποτελεσμάτων σε όλες τις παραμέτρους της συγκεκριμένης δειγματοληψίας, η αξιολόγηση αναφέρει ότι δεν συνάδει για τη νομοθεσία, γιατί η αξιολόγηση αναφέρεται στο δείγμα και όχι στις επί μέρους παραμέτρους.

Για την ανάλυση ένα πρόβλημα που αντιμετωπίστηκε ήταν η πολλαπλή εγγραφή των αποτελεσμάτων, αφού υπάρχει μια γραμμή που αντιστοιχεί σε κάθε παράμετρο που εξετάστηκε ένα δείγμα. Έτσι ένα δείγμα απαντάται σε αριθμό γραμμών, ισάριθμο των παραμέτρων που εξετάστηκε. Ο μοναδικός κωδικός αναγνώρισης, που υπάρχει για κάθε δείγμα και χρησιμοποιείται σαν «primary key» σε συνδυασμό με την ημερομηνία και την παράμετρο για τη βάση δεδομένων, είναι ο Αριθμός Γενικού Χημείου που επαναλαμβάνεται και με τον τρόπο αυτό μπορούμε να καταλάβουμε ότι όσες αναλύσεις έχουν τον ίδιο Αριθμό Γενικού Χημείου αφορούν εξέταση διαφορετικών παραμέτρων του ιδίου δείγματος. Λύση για το θέμα βρέθηκε με τη δημιουργία μιας βάσης δεδομένων στη Microsoft Access όπου τέθηκε σαν «primary key» ο συνδυασμός της

ημερομηνίας δειγματοληψίας με τον Αριθμό Γενικού Χημείου, παραλείποντας δηλαδή από το συνδυασμό την παράμετρο εξέτασης. Με αυτό τον τρόπο από κάθε εξέταση θα έχουμε απάλειψη όλων των πολλαπλών εγγραφών που αναφέρονται στο ίδιο δείγμα και την ίδια ημερομηνία δειγματοληψίας. Η απάλειψη αυτή μπορεί να λειτουργήσει για τον λόγο ότι αν μια παράμετρος δε συνάδει με τη νομοθεσία τότε στην αξιολόγηση των αποτελεσμάτων σε όλες τις παραμέτρους η αξιολόγηση αναφέρει ότι δεν συνάδει με τη νομοθεσία, έτσι δεν έχει σημασία ποια παράμετρο θα απαλείψουμε, φτάνει να μένει μια παράμετρος από κάθε δείγμα με την αξιολόγηση που αφορά το σύνολο των παραμέτρων.

5.5.2 Ενοποίηση των δεδομένων

Τα δεδομένα στη βάση δεδομένων LIMS εξήχθηκαν σε μορφή EXCEL σε διάφορα αρχεία. Επειδή, είχαμε τα δεδομένα διασπασμένα σε διάφορα αρχεία, για να μπορέσουμε να τα επεξεργαστούμε χρειάστηκε να γίνει ενοποίηση τους σε ένα αρχείο. Προσπάθησα να πετύχω το πιο πάνω με απλή αντιγραφή των στοιχείων αλλά ο περιορισμό των γραμμών της excel 2003 σε 64k (2^16) δεν μου το επέτρεψε να ενοποιήσω τα αποτελέσματα λόγω μεγέθους των δεδομένων. Τα μικροβιολογικά αποτελέσματα για τα έτη 2008, 2009, και 2010 απαριθμούνταν σε σύνολο 71538 γραμμές όποτε υπερβήκαμε κατά πολύ το όριο της excel 2003. Για να μπορέσω να αντιμετωπίσουμε το θέμα, χρειάστηκε να προχωρήσω με τη χρήση της excel 2007 που επιτρέπει τη χρήση 1,048,576 γραμμών. Έτσι τα δεδομένα που εξήχθηκαν από το LIMS σε μορφή excel worksheet, μετατράπηκαν σε excel 2007.

5.5.3 Επιλογή Δεδομένων

Η αξιολόγηση των αποτελεσμάτων που μπορεί να γίνει είναι ποσοτική και ποιοτική. Δηλαδή στην ποσοτική προσέγγιση θα ασχοληθούμε με τις μετρήσεις που υπάρχουν για τις διάφορες παραμέτρους μπορούμε να εντοπίσουμε μοτίβα στις τιμές ή ακραίες τιμές, που να χρήζουν περαιτέρω διερεύνησης. Για αυτό τον

τύπο ανάλυσης θα χρειαστούμε όλες τις παραμέτρους του κάθε δείγματος. Πέραν από την ποσοτική αξιολόγηση των αποτελεσμάτων μπορεί να γίνει και ποιοτική αξιολόγηση τους. Για το κάθε δείγμα υπάρχει αξιολόγηση κατά πόσο το δείγμα συνάδει ή όχι με τη νομοθεσία. Σημείο που πρέπει να αναφερθεί είναι ότι αν μια παράμετρος δεν συνάδει με τη νομοθεσία τότε στην αξιολόγηση των αποτελεσμάτων σε όλες τις παραμέτρους η αξιολόγηση αναφέρει ότι δεν συνάδει για τη νομοθεσία γιατί η αξιολόγηση αναφέρεται στο δείγμα και όχι στις επί μέρους παραμέτρους.

Συγκεκριμένα στο εργαστήριο που γίνεται η μικροβιολογική εξέταση των δειγμάτων υπάρχουν τρεις τιμές που μπορεί να λάβει ένα δείγμα : συνάδει με τη νομοθεσία (s) δεν συνάδει με τη νομοθεσία (u) ή είναι ύποπτο (y). Το αν συνάδει ή όχι με τη νομοθεσία εννοούμε αν προς τη συγκεκριμένη παράμετρο που αναλύθηκε η τιμή της ανάλυσης είναι εντός του πεδίου των τιμών που καθορίζει ως αποδεκτές η νομοθεσία (Κεφάλαιο 4.3), ενώ ύποπτες καθορίζονται οι τιμές οι οποίες υποδηλούν αυξημένη μικροβιολογική παρουσία σε παράμετρο που δεν προβλέπεται η αξιολόγηση της με βάση τη νομοθεσία. Για το λόγο αυτό είναι αναγκαία η κωδικοποίηση των αποτελεσμάτων για τη σύγκρισή τους.

Για την ποιοτική προσέγγιση χρησιμοποιήθηκαν τα ακόλουθα στοιχεία :
Η αξιολόγηση του αποτελέσματος θεωρείται η εξαρτημένη ποιοτική μεταβλητή, ενώ τα ακόλουθα στοιχεία αποτελούν τις ανεξάρτητες μεταβλητές:

1. χώρος δειγματοληψίας,
2. ημερομηνία δειγματοληψίας
3. δειγματολήπτης

Μετά που έγινε εξόρυξη στα δεδομένα με διάφορους αλγόριθμους βρέθηκε να μην έχει τη σημασία που ήθελα να έχω όσον αφορά την ημερομηνία αφού οι αλγόριθμοι προσπαθούσαν να βρουν συνεχόμενες περιόδους με σημασία ενώ αυτό που πραγματικά αποζητούσα ήταν ο πιθανός σχηματισμός εναλλασσόμενων μοτίβων ανά χρονικές περιόδους. Οι πιθανές ομοιότητες που

δυνατό να προέκυπταν και να είχαν ενδιαφέρον, όπως προκύπτει μέσα από τη μελέτη που διεξήγαγα, είναι οι μεταβολές στην ποιότητα ανάλογα με τις κλιματολογικές συνθήκες. Οι μετεωρολογικές παρατηρήσεις είναι εκτός του φάσματος των στοιχείων που διατηρούνται στη βάση δεδομένων μας, ωστόσο ειδικά στην Κύπρο, οι κλιματολογικές συνθήκες είναι συνυφασμένες με την χρονική περίοδο. Για να γίνει ανάλυση με βάση τα πιο πάνω ερωτήματα αποφασίστηκε να επαναληφθεί η διαδικασία επιλογής των δεδομένων και αλλαγή της μορφής των δεδομένων στην μορφή ARFF σημείο 5.5.4.

Έτσι αυτή την φορά επιλέγηκε να χρησιμοποιηθεί στην θέση της ημερομηνίας ο μήνας μόνο, έτσι ώστε να είναι δυνατός ο εντοπισμός τυχών μοτίβων που προκύπτουν ανάλογα με το μήνα, καθώς υπάρχει η γενική παραδοχή ότι οι μετεωρολογικές συνθήκες στην Κύπρο είναι παρόμοιες τον μήνα κάθε χρόνου. Τα δεδομένα μεταφέρθηκαν και πάλι στην Microsoft Excel και μέσω επεξεργασίας προστέθηκε επιπλέον στήλη και με εξίσωση της Excel από την ημερομηνία εξήχθηκε ο μήνας σε διψήφιο αριθμό από 1-12 (π.χ. Ιανουάριος =01, Φεβρουάριος=02, κτλ). Στη συνέχεια επαναλήφθηκε η διαδικασία που περιγράφεται στο 5.5.4 για να έρθουν τα δεδομένα σε μορφή που να επιτρέπει την περαιτέρω επεξεργασία τους.

5.5.4 Αλλαγή μορφής δεδομένων

Το λογισμικό WEKA μπορεί να εισάγει δεδομένα από αρχεία μόνο σε μορφή ARFF, CSV, C4.5 ή σε διατεταγμένη σειριακά μορφή την οποία και θα πρέπει να καθορίσουμε. Στο σημείο αυτό υπήρξε μεγάλη δυσκολία αφού συνήθως τα αρχεία των δεδομένων επεξεργάζονται σε λογισμικά όπως το MATLAB που επιτρέπουν την αποθήκευση των δεδομένων σε συμβατή μορφή, όμως στην περίπτωση μας τα αρχεία από την Microsoft EXCEL 2007 δεν μπορούσαν να μετατραπούν σε κατάλληλη μορφή. Η Excel μπορεί να επιτρέπει την αποθήκευση των δεδομένων σε μορφή CSV αλλά δεν μπορούν να εισαχθούν στο WEKA γιατί

το WEKA απαιτεί οι τιμές να μπορούν να διαχωριστούν με χαρακτήρες που θα πρέπει να καθορίσουμε όπως και το τέλος της γραμμής.

Για να μπορέσουμε να καταχωρήσουμε τα δεδομένα στο WEKA επιλέχθηκε η μετατροπή τους σε αρχείο ARFF, όμως για να γίνει δυνατή η μετατροπή αυτή χρειάστηκε επισταμένη μελέτη της μορφής και της δομής των αρχείων ARFF ώστε να επιτευχθεί η μετανάστευση των δεδομένων στην Microsoft Access 2007. Στη συνέχεια εγκαταστάθηκε ένα module στην Access με ένα κώδικα στη Visual Basic το οποίο φροντίζει τα διαστήματα στα στοιχεία των τιμών και πεδίων, βρίσκει τις μοναδικές τιμές των ονομαστικών μεταβλητών, αναθέτει τον ανάλογο τύπο πληροφορίας ARFF (datatype) και αντικαθιστά τις ελλείπουσες τιμές με τα ερωτηματικά.

Πρόβλημα αντιμετωπίστηκε με τις ημερομηνίες τόσο στην μεταφορά τους από EXCEL σε ACCESS όσο και στην μεταφορά τους σε ARFF. Η ημερομηνία μεταφερόταν σαν τύπος κείμενο από το πρώτο στο δεύτερο λογισμικό και για να υπερπηδηθεί αυτή η δυσκολία χρειάστηκε να αλλάξουν οι τύποι δεδομένων τόσο στην Access όσο και στην Excel. Επίσης, το ARFF αρχείο χρειάστηκε να ανοιχθεί με επεξεργαστή κειμένου και να αλλάξει με το να προστεθεί ο τύπος της ημερομηνίας στον ορισμό της.

Πέραν των πιο πάνω δυσκολιών, ένα άλλο σημείο που κόστισε χρόνο στην εργασία μου ήταν η χρήση χαρακτήρων στο σύνολο των δεδομένων μου των οποίων η χρήση είναι δεσμευμένη για τη δομή των ARFF αρχείων όπως ο χαρακτήρας [¨](π.χ. Χ¨Σάββα). Έγινε κωδικοποίηση των δεσμευμένων χαρακτήρων και αλλαγή τους με μη δεσμευμένους.

Ένα άλλο σημείο ήταν και η αναγκαιότητα κωδικοποίησης των δεδομένων που αφορούσαν τις μικροβιολογικές εξετάσεις των δειγμάτων. Στις μικροβιολογικές εξετάσεις υπήρχαν τρεις τιμές που μπορούσε να λάβει ένα δείγμα: συνάδει με τη νομοθεσία (s), είναι ύποπτο (y), ή δεν συνάδει με τη νομοθεσία (u). Τα

αποτελέσματα κωδικοποιήθηκαν με τους κωδικούς 0, 1, και 2 αντίστοιχα. Ο κωδικός του σημείου δειγματοληψίας ήταν της μορφής ΠΣ047-002-Δ1, όπου τα πρώτα 2 ψηφία δείχνουν τον τύπο του δείγματος, τα 3 επόμενα τη γενική περιοχή δειγματοληψίας, τα επόμενα 3 την ειδική περιοχή δειγματοληψίας και τα τελευταία 2 τον τύπο υδατοπρομήθειας. Για να είναι πιο φιλικοί οι κωδικοί σημείου δειγματοληψίας προς το WEKA έγιναν αντικαταστάσεις, αφού το Weka δεν μπορούσε να χειριστεί τους ελληνικούς χαρακτήρες, ως ακολούθως: Α = 1, Β = 2, Γ = 3, Δ = 4, Ε = 5, Ζ =6, Η = 7, Θ = 8, Ι = 9, Κ = 10,Λ = 11, Μ = 12, Ν = 13, Ξ = 14, Ο = 15, Π = 16, Ρ = 17, Σ = 18, Τ = 19, Υ = 20, Φ = 21, Χ = 22, Ψ = 23, Ω = 24, -=99.

Στο σημείο αυτό τα δεδομένα σχημάτισαν ένα αρχείο ενοποιημένο με τα δεδομένα για το 2008 και τα δεδομένα του 2009 και 2010 μαζί, που έχει 9588 καταχωρήσεις από τις οποίες οι 6764 συνάδουν με τη νομοθεσία οι 2535 όχι ενώ 289 κατατάσσονται σαν ύποπτα.

5.5.5. Επιλογή αλγόριθμου εξόρυξης δεδομένων και εφαρμογή του

Στο σημείο αυτό καθορίζεται τι είδους γνώση θα αναζητηθεί, κάτι που έμμεσα προσδιορίζει και την κατηγορία αλγόριθμου που θα χρησιμοποιηθεί. Τα παράγωγα της διαδικασίας ανακάλυψης γνώσης μπορεί να είναι:

- πρότυπα πληροφόρησης - informative patterns *(μάθηση χωρίς επίβλεψη)*
- μοντέλα πρόβλεψης - predictive models *(μάθηση με επίβλεψη)*.

Στο στάδιο αυτό εφαρμόζονται έξυπνες τεχνικές για ανίχνευση μοτίβων.

Είναι ένα καθαρά υπολογιστικό στάδιο, στο οποίο γίνεται η ουσιαστική αναζήτηση της γνώσης στα δεδομένα (**εξόρυξη σε δεδομένα).**

Αποφασίστηκε να χρησιμοποιηθούν διάφορα προβλεπτικά μοντέλα και ειδικά τεχνικές όπως η ταξινόμηση (classification) που είναι μια προβλεπτική (Predictive) διαδικασία, όπου δοθέντος ενός συνόλου εγγραφών (σύνολο εκπαίδευσης -training set), όπου κάθε εγγραφή έχει ένα σύνολο από γνωρίσματα

ένα εκ των οποίων είναι η κλάση (ή κατηγορία), βρίσκει ένα μοντέλο για το γνώρισμα της κλάσης ως συνάρτηση της τιμής των άλλων γνωρισμάτων. Ο στόχος της ταξινόμησης είναι να αναθέτει σε εγγραφές που δεν έχουμε δει μια κλάση με την μεγαλύτερη δυνατή ακρίβεια και για να χαρακτηρίσουμε την ακρίβεια του μοντέλου χρησιμοποιούμε ένα σύνολο ελέγχου (test set). Συνήθως το δοθέν δεδομένο σύνολο χωρίζεται σε ένα σύνολο εκπαίδευσης και σε ένα σύνολο ελέγχου – το πρώτο χρησιμοποιείται για την κατασκευή του μοντέλου και το δεύτερο για τον έλεγχό του.

Χαρακτηριστική μέθοδος είναι η Decision tree learning, που χρησιμοποιεί ένα δέντρο απόφασης ως προβλεπτικό μοντέλο, το οποίο χαρτογραφεί τις παρατηρήσεις για ένα στοιχείο στα συμπεράσματα για την τιμή του στόχου του στοιχείου. Άλλοι αλγόριθμοι αναφέρονται και περιγράφονται στο σημείο 3.3.

Έγιναν διάφορες δοκιμές μέσω τησυλοποίηση διάφορων αλγόριθμων ταξινόμησης, ωστόσο υπήρξαν περιορισμοί που προέρχονται από τη δομή και φύση των δεδομένων, παραδείγματος χάρη οι περισσότεροι αλγόριθμοι δεν δέχονταν κατηγορικές τιμές (Nominal values). Έτσι έγινε ανάκληση και εγκατάσταση αλγόριθμων πέραν από αυτούς που είναι προεγκατεστημένοι στο λογισμικό WEKA(trees.RandomForest kai trees.RandomTree), από τα εργαλεία του WEKA μέσω του package manager..

5.6 Εξόρυξη δεδομένων

Αλγόριθμοι

Για να δοκιμασθούν διάφοροι αλγόριθμοι έγινε συγκριτική δοκιμή εφαρμογής από όλες τις ομάδες των αλγόριθμων της ομάδας ταξινόμησης. Επειδή τα δεδομένα δεν ήταν ισοβαρή κατανεμημένα, με την πληθώρα των τιμών να είναι κατανεμημένες στο s. Αυτό είχε ως αποτέλεσμα μεγάλος αριθμός των αλγόριθμων κατάτασσαν όλες τις τιμές στο s με αποτέλεσμα να παρουσιάζουν από τη μία σημαντική επιτυχία κατάταξης, αφού πέραν του 70% των τιμών ήταν

σωστά κατανεμημένες, αλλά από την άλλη να μην έχουν καμιά αξία αφού δεν γινόταν καμιά πρόβλεψη για περιπτώσεις που δεν συνάδουν με τη νομοθεσία. Όμως ο αλγόριθμος j48 για τον οποίο και έτρεφα πολλές ελπίδες ήταν ο πλέον ιδανικός. Ένα σημείο που με απασχόλησε ήταν η απρόσμενη παύση της εκπαίδευσης αλγόριθμου λόγω έλλειψης μνήμης, όπως έγινε στην περίπτωση αλγόριθμος REPTree.

Με βάση τα πιο πάνω παρόλο που δοκιμάστηκαν πέραν των 30 αλγορίθμων, κατέληξα στην αξιοποίηση των αποτελεσμάτων των πιο κάτω αλγόριθμων :

1. bayes.BayesNet
2. bayes.NaiveBayes
3. bayes.NaiveBayesUpdateable
4. lazy.LWL
5. rules.DecisionTable
6. trees.RandomForest
7. trees.RandomTree

Σύνολο δεδομένων εκπαίδευσης αλγόριθμων – Data Training set

Οι πιο πάνω αλγόριθμοι αρχικά δοκιμάστηκαν με την επιλογή «use training set» όπου ολόκληρο το σύνολο των δεδομένων χρησιμοποιείται σαν σύνολο εκπαίδευσης. Όμως η πιο πάνω πρακτική είχε παράξει πολύ παράξενα αποτελέσματα, αφού για τους αλγόριθμους Random Forest και Random Tree, τα επίπεδα σωστής πρόβλεψης ήταν στο 100%. Μετά από μελέτη των εν λόγω αλγόριθμων διαφάνηκε ότι με τον τρόπο λειτουργίας των πιο πάνω αλγόριθμων τα αποτελέσματα ήταν φυσιολογικά, αφού δημιουργούσαν ένα μονοπάτι με αποφάσεις στα δεδομένα εκπαίδευσης. Στη συνέχεια όταν δίνεται το πραγματικό σύνολο δεδομένων ακολουθείται το ίδιο το μονοπάτι. Όπως είναι φυσιολογικό όταν δόσεις το 100% των δεδομένων σαν δεδομένα εκπαίδευση θα καταλήξεις με 100% επιτυχή πρόβλεψη. Αυτή η δοκιμή με έπεισε για την ακατάλληλότητα των Random Forest και Random Tree.

Στη συνέχεια, για να γίνει σύγκριση των αποτελεσμάτων για σκοπούς αξιολόγησης της αξιοπιστίας των αλγόριθμων και των αποτελεσμάτων, αφού η

προκειμένη περίπτωση είναι άσκηση πρόβλεψης και όπως είναι φυσικό δεν θα έχουμε πρόσβαση σε όλα τα δεδομένα, έγινε εκπαίδευση των αλγορίθμων. Χρησιμοποιήσαμε το 66% των δεδομένων εκμεταλλευόμενοι την επιλογή «percentage split» της WEKA. Οι αλγόριθμοι έτρεξαν ξεχωριστά για τα ακόλουθα σύνολα δεδομένων, για να δούμε τη διακύμανση των αποτελεσμάτων ανά ζεύγη αρχικά αλλά και με όλους τους παράγοντες μαζί, στη συνέχεια:

1. Ποιοτικό αποτέλεσμα-Σημείο δειγματοληψίας
2. Ποιοτικό αποτέλεσμα- Δειγματολήπτης
3. Ποιοτικό αποτέλεσμα-Μήνας Δειγματοληψίας
4. Ποιοτικό αποτέλεσμα-Σημείο δειγματοληψίας- Δειγματολήπτης- Μήνας Δειγματοληψίας

Ειδικά για την τέταρτη δοκιμή που αφορά την εξόρυξη δεδομένων Ποιοτικό αποτέλεσμα-Σημείο δειγματοληψίας- Δειγματολήπτης- Μήνας Δειγματοληψίας για τον αλγόριθμο Random Forest παρουσιάστηκε πρόβλημα απρόσμενης παύσης της λειτουργίας του αλγόριθμου λόγω έλλειψης μνήμης, στο στάδιο εκπαίδευσης του αλγόριθμου. Γι αυτό και ειδικά για τη συγκεκριμένη περίπτωση έτρεξα τον αλγόριθμο με 52% σύνολο δεδομένων εκπαίδευσης.

Το WEKA προφανώς, για κάθε αλγόριθμο δίνει διαφορετικά αποτελέσματα τόσο για το σύνολο των δεδομένων αλλά και για συγκεκριμένες τιμές. Ωστόσο υπάρχει μια κοινή περιληπτική μορφή όπου για τον σύνολο των δεδομένων δίνονται μετρήσεις σε κοινές παραμέτρους. Αυτό επιτρέπει τη σύγκριση της αποδοτικότητας των αλγόριθμων αλλά και τη γρήγορη εξαγωγή συμπερασμάτων. Όσον αφορά την τρίτη δοκιμή «Ποιοτικό αποτέλεσμα-Μήνας Δειγματοληψίας» από την εξόρυξη κατατάγηκαν όλες οι τιμές στο s με όλους τους αλγόριθμους, γεγονός που φαίνεται να οφείλεται στην μη ομαλή(ισοβαρή) κατανομή των δεδομένων. Τα αποτελέσματα με τη χρησιμοποίηση του 66% σαν σύνολο εκπαίδευσης παρατίθενται σε συνοπτική μορφή στο παράρτημα 4.

5.7 Αποτελέσματα

Πιο κάτω αναφέρονται συνοπτικά τα αποτελέσματα που προέκυψαν μετά από δοκιμή με τις παραμέτρους: Ποιοτικό αποτέλεσμα, Σημείο δειγματοληψίας, Δειγματολήπτης και Μήνας Δειγματοληψίας :

Πίνακας: 1 - Τα συνοπτικά αποτελέσματα της ανάλυσης με εκπαίδευση με το 66% του συνόλου των δεδομένων

Αλγόριθμος	=== Evaluation on test split === === Summary ===
1. .bayes.BayesNet	Correctly Classified Instances 2309 70.8282 % Incorrectly Classified Instances 951 29.1718 % Kappa statistic 0.356 Mean absolute error 0.2287 Root mean squared error 0.3615 Relative absolute error 79.6295 % Root relative squared error 95.6863 % Coverage of cases (0.95 level) 98.0675 % Mean rel. region size (0.95 level) 62.5256 % Total Number of Instances 3260
2. bayes.NaiveBayes	Correctly Classified Instances 2326 71.3497 % Incorrectly Classified Instances 934 28.6503 % Kappa statistic 0.348 Mean absolute error 0.2315 Root mean squared error 0.3553 Relative absolute error 80.6127 % Root relative squared error 94.0451 % Coverage of cases (0.95 level) 98.6503 % Mean rel. region size (0.95 level) 64.5706 % Total Number of Instances 3260
3. bayes.NaiveBayesUpdateable	Correctly Classified Instances 2326 71.3497 % Incorrectly Classified Instances 934 28.6503 % Kappa statistic 0.348 Mean absolute error 0.2315 Root mean squared error 0.3553 Relative absolute error 80.6127 % Root relative squared error 94.0451 % Coverage of cases (0.95 level) 98.6503 % Mean rel. region size (0.95 level) 64.5706 % Total Number of Instances 3260

4. lazy.LWL	Correctly Classified Instances 2344 71.9018 % Incorrectly Classified Instances 916 28.0982 % Kappa statistic 0.1269 Mean absolute error 0.2588 Root mean squared error 0.3568 Relative absolute error 90.121 % Root relative squared error 94.4408 % Coverage of cases (0.95 level) 98.773 % Mean rel. region size (0.95 level) 67.9243 % Total Number of Instances 3260
5. rules.DecisionTable	Correctly Classified Instances 2344 71.9018 % Incorrectly Classified Instances 916 28.0982 % Kappa statistic 0.1536 Mean absolute error 0.2571 Root mean squared error 0.3543 Relative absolute error 89.5285 % Root relative squared error 93.7894 % Coverage of cases (0.95 level) 99.2025 % Mean rel. region size (0.95 level) 74.5194 % Total Number of Instances 3260
6. trees.RandomForest	Correctly Classified Instances 3195 69.4263 % Incorrectly Classified Instances 1407 30.5737 % Kappa statistic 0.2927 Mean absolute error 0.2288 Root mean squared error 0.3957 Relative absolute error 79.5278 % Root relative squared error 104.6955 % Coverage of cases (0.95 level) 91.2429 % Mean rel. region size (0.95 level) 54.0055 % Total Number of Instances 4602
7. trees.RandomTree	Correctly Classified Instances 2261 69.3558 % Incorrectly Classified Instances 999 30.6442 % Kappa statistic 0.29 Mean absolute error 0.2194 Root mean squared error 0.4375 Relative absolute error 76.4084 % Root relative squared error 115.8097 % Coverage of cases (0.95 level) 75.3681 % Mean rel. region size (0.95 level) 40.3476 % Total Number of Instances 3260

Πίνακας: 2 - Τα συνοπτικά αποτελέσματα της ανάλυσης με εκπαίδευση του 66% του συνόλου των δεδομένων (πλην του 6 όπου έγινε με 52%)

Αλγόριθμος	=== Detailed Accuracy By Class ===
1. bayes.BayesNet	TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.757 0.399 0.823 0.757 0.789 0.752 s 0.614 0.213 0.503 0.614 0.553 0.79 u 0.364 0.019 0.379 0.364 0.371 0.889 y Weighted Avg. 0.708 0.339 0.726 0.708 0.715 0.766
2. bayes.NaiveBayes	TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.78 0.435 0.814 0.78 0.796 0.751 s 0.581 0.2 0.505 0.581 0.541 0.789 u 0.303 0.012 0.435 0.303 0.357 0.874 y Weighted Avg. 0.713 0.361 0.722 0.713 0.716 0.765
3. bayes.NaiveBayes Updateable	TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.78 0.435 0.814 0.78 0.796 0.751 s 0.581 0.2 0.505 0.581 0.541 0.789 u 0.303 0.012 0.435 0.303 0.357 0.874 y Weighted Avg. 0.713 0.361 0.722 0.713 0.716 0.765
4. lazy.LWL	TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.961 0.868 0.73 0.961 0.83 0.703 s 0.123 0.036 0.545 0.123 0.2 0.731 u 0.172 0.002 0.708 0.172 0.276 0.843 y Weighted Avg. 0.719 0.625 0.681 0.719 0.649 0.714
5. rules.DecisionTable	TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.942 0.825 0.736 0.942 0.826 0.702 s 0.193 0.056 0.55 0.193 0.286 0.752 u 0.01 0 0.5 0.01 0.02 0.846 y Weighted Avg. 0.719 0.6 0.681 0.719 0.661 0.72
6. trees.RandomForest	TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.778 0.502 0.791 0.778 0.785 0.685 s 0.506 0.196 0.477 0.506 0.491 0.712 u 0.343 0.015 0.41 0.343 0.374 0.785 y Weighted Avg. 0.694 0.407 0.697 0.694 0.696 0.695
7. trees.RandomTree	TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.78 0.502 0.791 0.78 0.785 0.644 s 0.501 0.197 0.472 0.501 0.486 0.658 u 0.333 0.016 0.402 0.333 0.365 0.71 y Weighted Avg. 0.694 0.408 0.697 0.694 0.695 0.649

Πιο κάτω αναφέρονται τα Confusion Matrix στη δοκιμή με τις παραμέτρου Ποιοτικό αποτέλεσμα, Σημείο δειγματοληψίας, Δειγματολήπτης και Μήνας Δειγματοληψίας:

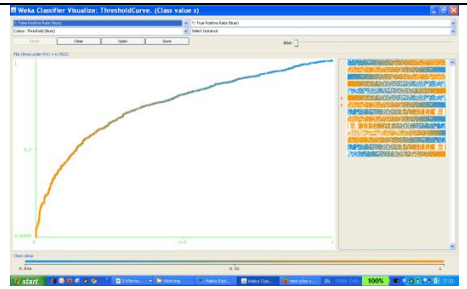
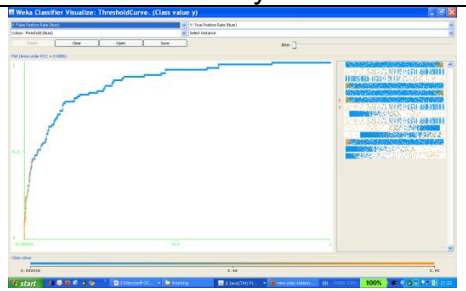
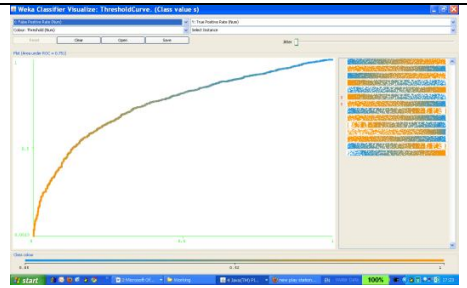
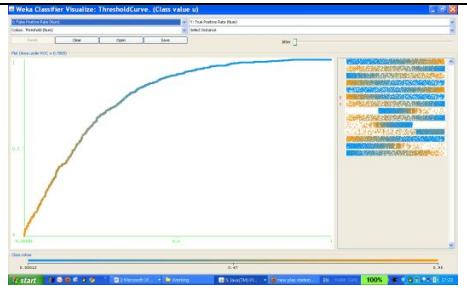
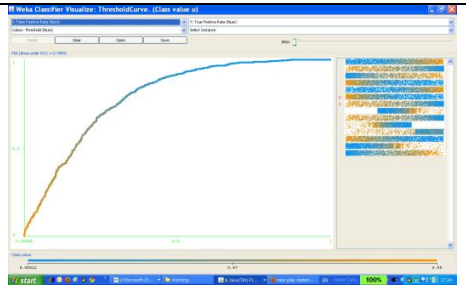
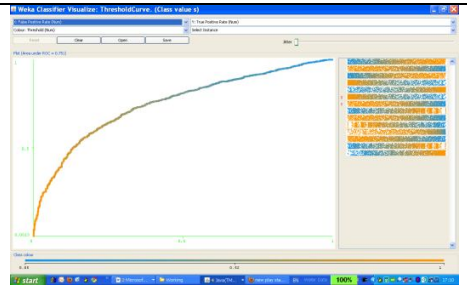
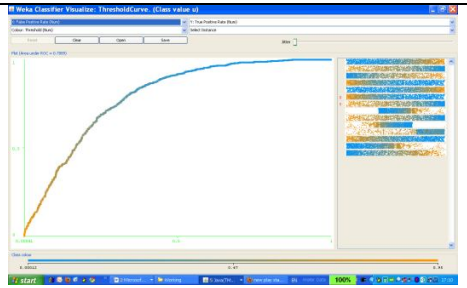
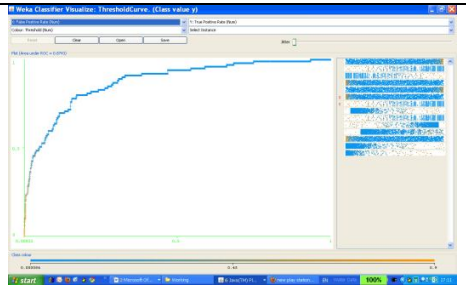
Πίνακας: 3 - Τα confusion matrix για όλους τους αλγόριθμους με εκπαίδευση με το 66% του συνόλου των δεδομένων(πλην του 6 όπου έγινε με 52%)

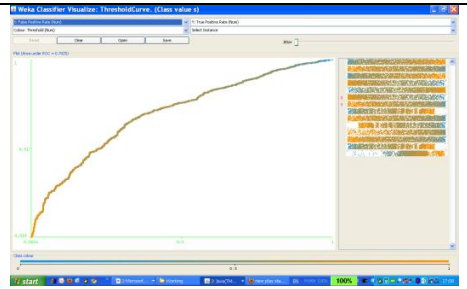
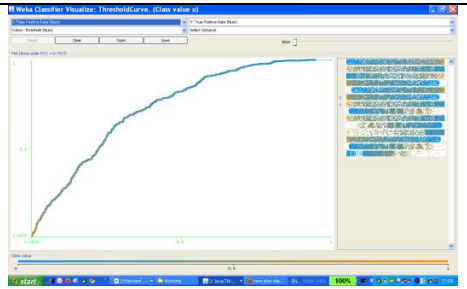
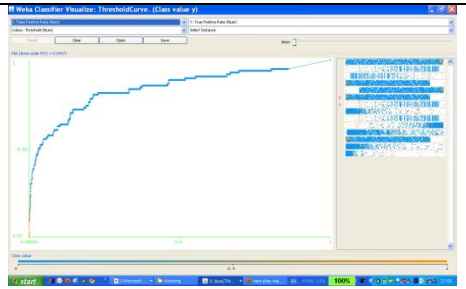
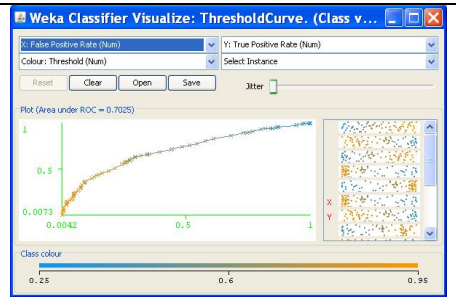
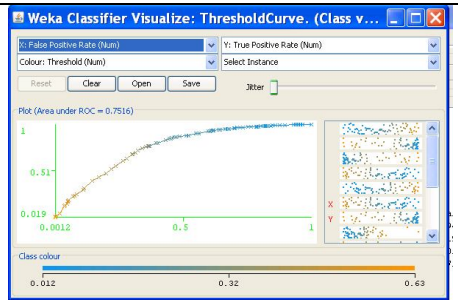
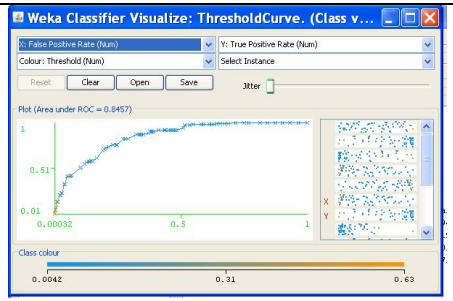
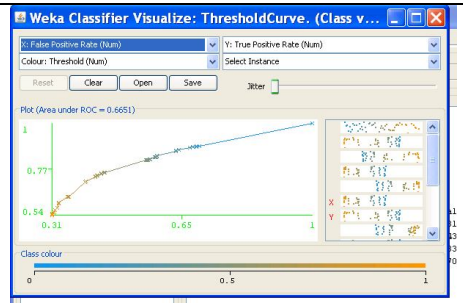
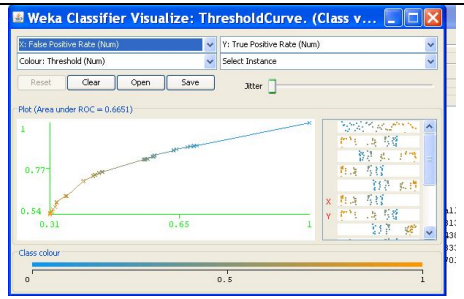
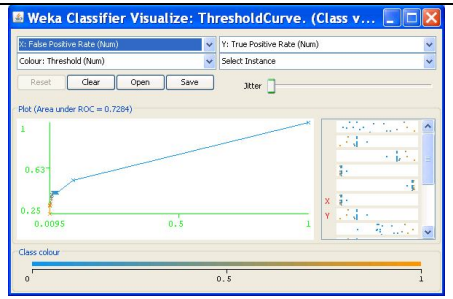
Αλγόριθμος	=== Confusion Matrix ===
1. bayes.BayesNet	a b c <-- classified as 1752 512 49 | a = s 317 521 10 | b = u 61 2 36 | c = y
2. bayes.NaiveBayes	a b c <-- classified as 1803 478 32 | a = s 348 493 7 | b = u 64 5 30 | c = y
3. bayes.NaiveBayesUpdateable	a b c <-- classified as 1803 478 32 | a = s 348 493 7 | b = u 64 5 30 | c = y
4. lazy.LWL	a b c <-- classified as 2223 86 4 | a = s 741 104 3 | b = u 81 1 17 | c = y
5. rules.DecisionTable	a b c <-- classified as 2179 133 1 | a = s 684 164 0 | b = u 97 1 1 | c = y
6. trees.RandomForest	a b c <-- classified as 2540 661 62 | a = s 585 607 7 | b = u 87 5 48 | c = y
7. trees.RandomTree	a b c <-- classified as 1803 470 40 | a = s 414 425 9 | b = u 61 5 33 | c = y

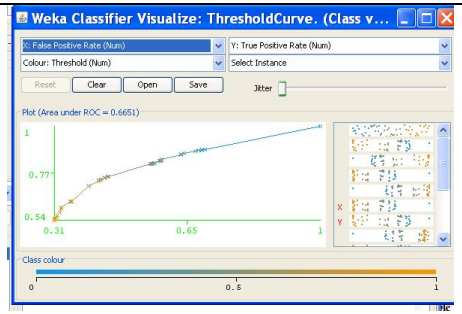
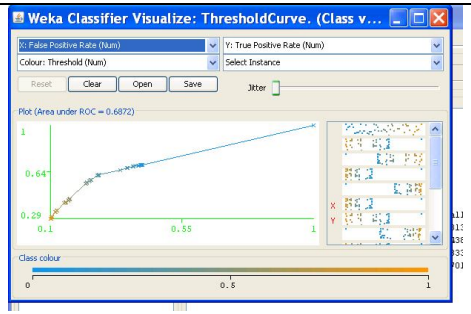
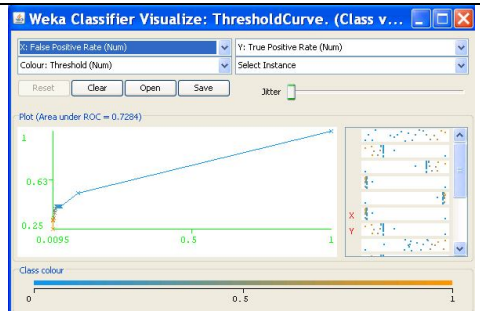
Στον πίνακα 4 παρατίθενται οι καμπύλες ROC για το s, u και y μετά από την εφαρμογή του κάθε αλγόριθμου για τις παραμέτρους Αποτέλεσμα-Δειγματολήπτης , Σημείο Δειγματοληψίας και Μήνας Δειγματοληψίας :

Πίνακας: 4 – Πίνακας με τις καμπύλες ROC για όλους τους αλγόριθμους με εκπαίδευση με το 66% του συνόλου των δεδομένων(πλην του 6 όπου έγινε με 52%)

ROC +confusion matrix 66%

Αλγόριθμος	S	U	y
1. weka.classifiers.bayes.BayesNet			
2. weka.classifiers.bayes.NaiveBayes			
3. weka.classifiers.bayes.NaiveBayesUpdateable			

4. weka.classifiers.lazy.LWL			
5. weka.classifiers.rules.DecisionTable			
6. weka.classifiers.trees.RandomForest			

7. weka.classifiers.trees.RandomTree			

5.7.1 Επαλήθευση των αποτελεσμάτων με στατιστική ανάλυση

Τα αποτελέσματα της εξόρυξης δεδομένων έδειξαν πολύ ενδιαφέροντα αποτελέσματα, αλλά η εγκυρότητα των αποτελεσμάτων είναι αμφισβητούμενη εφόσον είναι η πρώτη φορά που ο ερευνητής έχει επιχειρήσει εξόρυξη δεδομένων, είναι η πρώτη φορά που εφαρμόζεται εξόρυξη δεδομένων στο συγκεκριμένο σύνολο δεδομένων και είναι η πρώτη φορά που χρησιμοποιεί ο ερευνητής το συγκεκριμένο λογισμικό για εξόρυξη δεδομένων. Έτσι στη συνέχεια έγινε στατιστική ανάλυση των δεδομένων.

Για να γίνει σύγκριση των αποτελεσμάτων και της ορθότητας τους έγινε παράλληλη ανάλυση των αποτελεσμάτων με το στατιστικό πακέτο SPSS που είναι γενικά αποδεκτό εργαλείο για στατιστική ανάλυση το οποίο και χρησιμοποιείται με ευρεία εφαρμογή. Μετά από μελέτη έγινε επιλογή του Chi-Square Test, για να γίνει η στατιστική ανάλυση, καθώς το δείγμα περιλάμβανε ποιοτικές μεταβλητές. Έγιναν ξεχωριστά τεστ για την κάθε μια από τις ακόλουθες ανεξάρτητες μεταβλητές: Ημερομηνία (μήνας), Δειγματολήπτης και Σημείο Δειγματολειψίας. Η εξαρτημένη μεταβλητή σε όλα τα τεστ ήταν τα αποτελέσματα που αφορούσαν τη συμβατότητα του δείγματος με τη νομοθεσία (συνάδει, δεν συνάδει, ύποπτο). Τα αποτελέσματα έδειξαν ότι υπάρχει στατιστικά σημαντική συσχέτιση μεταξύ των ζευγών μήνα - αποτέλεσμα, του δειγματολήπτη αποτέλεσμα και του σημείου με το αποτέλεσμα.

Πιο κάτω παρατίθεται ο πίνακας με μέρος της ανάλυσης του μήνα συγκριτικά με τα αποτελέσματα:

Πίνακας 5. Αποτελέσματα δειγματοληψίας σε σχέση με το μήνα δειγματοληψίας του δείγματος

Date * Results Crosstabulation

			Results			
			.00=s	1.00=y	2.00=u	Total
Date	1.00	Count	515	4	131	650
		% within Date	79.2%	.6%	20.2%	100.0%
	2.00	Count	558	10	170	738
		% within Date	75.6%	1.4%	23.0%	100.0%
	3.00	Count	572	9	146	727
		% within Date	78.7%	1.2%	20.1%	100.0%
	4.00	Count	408	24	136	568
		% within Date	71.8%	4.2%	23.9%	100.0%
	5.00	Count	771	19	331	1121
		% within Date	68.8%	1.7%	29.5%	100.0%
	6.00	Count	743	15	455	1213
		% within Date	61.3%	1.2%	37.5%	100.0%
	7.00	Count	786	22	381	1189
		% within Date	66.1%	1.9%	32.0%	100.0%
	8.00	Count	674	57	289	1020
		% within Date	66.1%	5.6%	28.3%	100.0%
	9.00	Count	957	84	292	1333
		% within Date	71.8%	6.3%	21.9%	100.0%
	10.00	Count	292	21	89	402
		% within Date	72.6%	5.2%	22.1%	100.0%
	11.00	Count	300	13	90	403
		% within Date	74.4%	3.2%	22.3%	100.0%
	12.00	Count	188	11	25	224
		% within Date	83.9%	4.9%	11.2%	100.0%
Total		Count	6764	289	2535	9588
		% within Date	70.5%	3.0%	26.4%	100.0%

5.8 Ερμηνεία των αποτελεσμάτων

Τα αποτελέσματα του WEKA χρειάζονται ερμηνεία που για να δοθεί αυτή θα πρέπει να γίνουν κατανοητοί κάποιοι όροι.

Ο όρος true positive (TP) αναφέρεται στα σωστά ταξινομημένα θετικά ενώ το false positive (FP) αναφέρεται στις λάθος ταξινομημένες σαν σωστές τιμές. Ο όρος "Correctly Classified Instances" " σωστά ταξινομημένες περιπτώσεις" αναφέρεται στις σωστά ταξινομημένες τιμές σύμφωνα με το πρότυπο μετά την επιβεβαίωση και το "Incorrectly Classified Instances" "Ανακριβώς ταξινομημένες τιμές" που αναφέρεται στις λάθος περιπτώσεις.

Η ορθότητα (accuracy) ορίζεται σαν οι τιμές των ορθών ταξινομημένων περιπτώσεων (number of correctly classified instances) διαιρούμενος με το συνολικό αριθμό των περιπτώσεων (number of instances). Η ακρίβεια προσέγγισης (precision) ορίζεται σαν ο αριθμός ορθών ταξινομημένων περιπτώσεων μιας τάξης X “number of correctly classified instances of class X” διαιρούμενος με τον αριθμό των ταξινομημένων περιπτώσεων που ανήκουν στην τάξη X “number of instances classified as belonging to class X”. Η ανάκληση (recall) ισούται με τον αριθμός ορθών ταξινομημένων περιπτώσεων μιας τάξης X “number of correctly classified instances of class X” διαιρούμενος με τον αριθμό των ταξινομημένων περιπτώσεων που ανήκουν στην τάξη X “number of instances in class X”

$$\text{accuracy} = \frac{\text{number of correctly classified instances}}{\text{number of instances}}$$

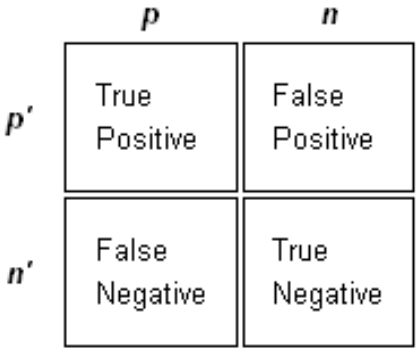
$$\text{precision}(X) = \frac{\text{number of correctly classified instances of class X}}{\text{number of instances classified as belonging to class X}}$$

$$\text{recall}(X) = \frac{\text{number of correctly classified instances of class X}}{\text{number of instances in class X}}$$

Πίνακας με τις επεξηγήσεις των όρων που υπάρχουν στην παρουσίαση των αποτελεσμάτων :

1.	true positive (TP)	Αληθινό θετικό
2.	true negative (TN)	αληθινό αρνητικό
3.	false positive (FP)	ψεύτικο θετικό (false alarm)
4.	false negative (FN)	δεν βρέθηκε η αρνητική τιμή
5.	sensitivity or true positive rate (TPR)	$TPR = TP / P = TP / (TP + FN)$
6.	false positive rate (FPR)	$FPR = FP / N = FP / (FP + TN)$
7.	accuracy (ACC)	$ACC = (TP + TN) / (P + N)$
8.	specificity (SPC) or True Negative Rate	$SPC = TN / N = TN / (FP + TN) = 1 - FPR$
9.	positive predictive value (PPV)= precision	$PPV = TP / (TP + FP)$
10.	negative predictive value (NPV)	$NPV = TN / (TN + FN)$
11.	false discovery rate (FDR)	$FDR = FP / (FP + TP)$

Άλλος πολύ σημαντικός όρος είναι το “confusion matrix” που παρουσιάζει τον αριθμό αληθινών/ψεύτικων θετικών και ψεύτικων αρνητικών τιμών.. Ο τρόπος που διαβάζεται το confusion matrix δύο τιμών είναι ο ακόλουθος :



Σχήμα 2 Επεξήγηση του confusion matrix (Wikipedia, 2010)

Οι σωστά ταξινομημένες τιμές βρίσκονται στα κουτιά true positive και true negative έτσι ώστε οι σωστές ταξινομήσεις να είναι στη διαγώνιο. Επειδή όμως έχουν τρεις δυνατές τιμές που μπορεί να λάβουν, το confusion matrix που προκύπτει είναι ένας πίνακας 3Χ3 που στην πρώτη γραμμή αναφέρεται η κατανομή των τιμών του s, στη δεύτερη η κατανομή του u και στην τρίτη η κατανομή του y. Οι

σωστές ταξινομήσεις εξακολουθούν να είναι ταξινομημένες στη διαγώνιο από πάνω αριστερά σε κάτω δεξιά.

Καμπύλη ROC

Σύμφωνα με τους Kamber και Han (2006), οι καμπύλες receiver operating characteristic (ROC) είναι απότοκο της θεωρία ανίχνευσης σημάτων, που αναπτύχθηκε κατά το Δεύτερο Παγκόσμιο Πόλεμο για την ανίχνευση των εχθρικών αντικειμένων στους τομείς μάχης.

Αποτελεί μια γραφική απεικόνιση του ρυθμού των θετικών έναντι του ρυθμού των αρνητικών και μπορεί να αντιπροσωπευθεί ισοδύναμα με τη χάραξη του μέρους των αληθινών θετικών από τα θετικά (TPR = αληθινό θετικό ποσοστό) εναντίον του μέρους των ψεύτικων θετικών από τα αρνητικά (FPR = ψευδοθετικό ποσοστό). Η καμπύλη ROC χρησιμοποιείται για τη σύγκριση μοντέλων κατηγοριοποίησης. Με βάση την καμπύλη ROC μπορεί κάποιος να επιλέξει τα ενδεχομένως βέλτιστα πρότυπα και για να απορρίψει άλλα πιο ανεξάρτητα. Όσο πιο κοντά είναι η καμπύλη ROC στη διαγώνιο μιας γραφική παράστασης δύο αξόνων, τόσο λιγότερη ακρίβεια το μοντέλο αυτό παρέχει. Για να γίνει αξιολόγηση της καμπύλης μετρούμε την περιοχή της καμπύλης και όσο πιο κοντά στο 0,5 βρίσκεται το αποτέλεσμα τότε τόσο πιο ανακριβές είναι το μοντέλο μας ενώ όσο πλησιάζει στο 1 αυξάνεται η ακρίβειά του.

5.9 Ερμηνεία των αποτελεσμάτων

Πρώτα από όλα, θα πρέπει να γίνει η παραδοχή ότι λόγω των περιορισμένων δεδομένων που έχουν αναλυθεί που καλύπτουν τη διάρκεια μόνο 3 ετών 2007-2009, τα αποτελέσματα μας αναμένεται να έχουν χαμηλά επίπεδα εμπιστοσύνης και η ανάλυση αυτή πρέπει να αντιμετωπιστεί σαν πιλοτική. Ωστόσο, λόγω του μεγέθους και του εύρους κάλυψης των δεδομένων, είναι περιορισμένη η εγκυρότητα της πρόβλεψης αλλά όσο περνά ο χρόνος, τόσο περισσότερα δεδομένα συσσωρεύονται στις βάσεις δεδομένων της παρακολούθησης της

ποιότητας των νερών με αποτέλεσμα ο βαθμός εμπιστοσύνης στην πρόβλεψη να μπορεί να αυξηθεί με την πρόσβαση σε περισσότερα δεδομένα.

Από τους αλγόριθμους που δοκιμάστηκαν στην εξόρυξη δεδομένων με το σύνολο των παραμέτρων όσον αφορά την ανίχνευση των τιμών που δεν συνάδουν με τη νομοθεσία, από το confusion matrix που φαίνεται στον Πίνακα 3 μπορούμε να συγκρίνουμε την απόδοση των αλγόριθμων όσον αφορά την πρόβλεψη του αποτελέσματος u και να παρατηρήσουμε ότι ο bayes(BayesNet) έχει τα καλύτερα αποτελέσματα. Πρέπει να σημειωθεί ότι όσον αφορά τον αλγόριθμο Random Forest στον Πίνακα 3, η εκπαίδευση έγινε με το 52% αντί το 66% του συνόλου των δεδομένων οπότε οι τιμές του u αυξάνονται από 848 στο 34% (υπολειπόμενο του 66%) σε 1119 στο 48%(υπολειπόμενο του 52%).

Όσον αφορά τον πίνακα 4 μπορούμε να δούμε από την στήλη που παρατίθενται οι καμπύλες ROC για την παράμετρο u ότι η περίπτωση του αλγόριθμου bayes(BayesNet) έχει την μεγαλύτερη περιοχή κάλυψης., γεγονός που επιβεβαιώνει την αποδοτικότητα του αλγορίθμου BayesNet. Στις 9588 καταχωρήσεις 2535 δεν συνάδουν με τη νομοθεσία που αποτελεί το 26,4%, από αυτό το 66% χρησιμοποιήθηκε σαν σύνολο εκπαίδευσης- test-set ενώ το 33% αποτελεί το σύνολο ελέγχου. Από τις 848 καταχωρήσεις που αντιστοιχούν στο 33%, φαίνεται ξεκάθαρα να υπάρχει σημαντική δυνατότητα πρόβλεψης, αφού έγινε σωστή πρόβλεψη για 521 περιπτώσεις γεγονός που αποτελεί το 61,4% που είναι θεαματικά αποτελέσματα.

Από την επί μέρους δοκιμή των ζευγών «Ποιοτικό αποτέλεσμα-Σημείο δειγματοληψίας» και «Ποιοτικό αποτέλεσμα- Δειγματολήπτης» προκύπτει επίσης ότι υπάρχει δυνατότητα πρόβλεψης, και στατιστική συσχέτιση μεταξύ τους όσον αφορά τα δείγματα που δεν συνάδουν με τη νομοθεσία. Τα πιο πάνω επιβεβαιώνονται με τα αποτελέσματα της στατιστικής ανάλυσης. Διαφαίνεται να υπάρχει συσχέτιση μεταξύ των σημείων δειγματοληψίας και των αποτελεσμάτων, κάτι που είναι αξιοσημείωτο, αφού αναγνωρίζοντας τα εν λόγω σημεία, μπορούμε

να λάβουμε μέτρα για την πιθανή λήψη προληπτικών μέτρων και την εστίαση των ελέγχων.

Όσον αφορά τη συσχέτιση που βρέθηκε μεταξύ του δειγματολήπτη και του αποτελέσματος η σχέση αυτή με την πρώτη ματιά φαίνεται να δίνει πολύ ενδιαφέροντες συνειρμούς. Όμως μετά από μια πιο προσεκτική μελέτη του θέματος, η σχέση αυτή δυνατό να είναι πλασματική και να αποδίδεται στο γεγονός ότι ο κάθε λειτουργός έχει την περιοχή ευθύνης του, στην οποία και είναι ο δειγματολήπτης. Εφόσον υπάρχει σημαντική στατιστική συσχέτιση με το σημείο δειγματοληψίας και το αποτέλεσμα, δυνατό να είναι αυτή που προκαλεί τη "θυματοποίηση" του δειγματολήπτη. Αφού εάν κάποιος έχει στην περιοχή ευθύνης του αρκετά σημεία που παρουσιάζουν συνέπεια ως προς ένα αποτέλεσμά, αυτό του προσδίδει και προσωπική σχέση όσον αφορά τον επηρεασμό του αποτελέσματος, κάτι που δεν ισχύει. Πρόκειται για προβολή της σχέσης του σημείου δειγματοληψίας με αυτή του αποτελέσματος.

Όσον αφορά τη διακύμανση των αποτελεσμάτων στους μήνες, από τις παρατηρήσεις φαίνονται τα ποσοστά των δειγμάτων που να μην συνάδουν με τη νομοθεσία να παρουσιάζουν παράξενη απόκλιση κατά τους μήνες Μάιο έως και Αύγουστο, όπου τα ποσοστά αποτελεσμάτων που δεν συνάδουν με τη νομοθεσία, όπως φαίνεται από τον Πίνακα 5 είναι τριπλάσια (28%) συγκριτικά με τους υπόλοιπους μήνες (20%) . Τα πιο πάνω μπορούν να αποδοθούν στα πιο κάτω:

1. Η πιο προφανής εξήγηση είναι η διακύμανση της θερμοκρασίας. Επειδή τα μικρόβια πολλαπλασιάζονται με βάση τη θερμοκρασία, τους μήνες Μάιο – Σεπτέμβριο όπου υπάρχει σημαντική αύξηση της θερμοκρασίας υποβοηθά την αύξηση του ρυθμού πολλαπλασιασμού των μικροβίων.
2. Κατά τους καλοκαιρινούς μήνες λόγω λειψυδρίας σαν μέτρο εξοικονόμησης νερού γίνεται διακοπή της παροχής του νερού. Η διακοπή και επανεκκίνηση της παροχής νερού αλλοιώνει τη

μικροβιολογική ποιότητα του νερού, αφού επηρεάζει την περιεκτικότητα χλωρίου στο νερό, που χρησιμοποιείται σαν απολυμαντικό.,

3. Κάτι που επίσης συνδέεται με τις διακοπές της παροχής του νερού είναι οι αυξανόμενες βλάβες που παρατηρούνται στα δίκτυα διανομής του νερού. Λόγω της διακοπής και επανεκκίνησης της παροχής νερού αυξομειώνεται η πίεση στους σωλήνες του δικτύου διανομής με αποτέλεσμα να παρατηρούνται βλάβες στο δίκτυο διανομής. Έτσι προκαλείται η επιμόλυνση του νερού με μικροοργανισμούς.
4. Πέραν των πιο πάνω, σύμφωνα με το άρθρο «Effects of temperature and biodegradable organic matter on control of biofilms by free chlorine in a model drinking water distribution system»των Ndiongue, Huck και Slawson (2005), η υποδειγματικότατα του χλωρίου που χρησιμοποιείται σαν απολυμαντικό στο νερό, επηρεάζεται από την αύξηση της θερμοκρασίας. Ενόψει των πιο πάνω διαφαίνεται ότι η θερμοκρασία επηρεάζει ποικιλοτρόπως τη μικροβιολογική ποιότητα του νερού.

6. Αξιολόγηση αποδοτικότητας της χρήσης της εξόρυξης δεδομένων στην ασφάλεια τροφίμων

Στη μελέτη αυτή έγινε αξιολόγηση των πληροφοριών, που διατηρούνται σε ηλεκτρονική μορφή στα αρχεία της Υγειονομική Υπηρεσίας και μελετήθηκε η δυνατότητα αλλά και αξία εξόρυξης δεδομένων μέσα από αυτά.

Όσον αφορά το κύριο εργαλείο που χρησιμοποιήθηκε, το Weka, αυτό είναι ένα ελεύθερα διαθέσιμο λογισμικό, που διαθέτει μια συλλογή εργαλείων και αλγορίθμων απεικόνισης για την ανάλυση στοιχείων και κατασκευή προβλεπτικών μοντέλων, με γραφικό λογισμικό διεπαφής για την εύκολη λειτουργία. Μεγάλα πλεονεκτήματα του είναι ότι είναι ελεύθερα διαθέσιμο και μπορεί να τρέξει σχεδόν σε οποιαδήποτε σύγχρονη πλατφόρμα υπολογιστή. Επειδή εφαρμόζεται πλήρως στη γλώσσα προγραμματισμού της Java, διαθέτει

πολύ περιεκτική συλλογή εργαλείων/μοντέλων προεπεξεργασίας στοιχείων και τεχνικών εξόρυξης και επιπλέον προσφέρει την ευκολία στη χρήση λόγω της χρησιμοποίησης γραφικών στο λογισμικό διεπαφής.

Από τα αποτελέσματα φαίνεται ότι το πρόγραμμα αυτό έχει πολύ καλές δυνατότητες και μπορεί να παρέχει πολύ χρήσιμη πληροφόρηση αν αξιοποιηθεί σε υπάρχουσες πηγές δεδομένων στην Κύπρο. Πρέπει να ληφθεί υπόψη ότι η Υγειονομική Υπηρεσία άρχισε να εφαρμόζει σχέδια για την μηχανογράφησή της και αναμένεται να μηχανογραφηθεί εντός του 2012. Επιπλέον όλα τα αποτελέσματα του ελέγχου των τροφίμων θα είναι διαθέσιμα σε ηλεκτρονική μορφή, οπότε μπορεί να γίνει μελλοντική εκμετάλλευση τους είτε με την αντιπαραβολή τους με στατιστικά στοιχεία της Ευρωπαϊκής Επιτροπής, για να βρεθούν τυχόν διαφορές. Οι διαφορές αυτές μπορεί να αποκαλύψουν χρήσιμα στοιχεία, όπως για παράδειγμα το ότι στην Κύπρο έχουμε υπερβολική εμφάνιση αρνητικών αποτελεσμάτων σε συγκεκριμένα τρόφιμα. Με την ανάλυση αυτή μπορούμε να αναπροσαρμόσουμε τους ελέγχους μας, μειώνοντάς τους σε περιοχές που Πανευρωπαϊκά παρατηρείται μειωμένος κίνδυνος και αυξάνοντάς τους σε περιοχές που έχουμε εθνικά παραμελήσει, ενώ στην υπόλοιπη Ευρώπη παρουσιάζουν αυξανόμενη επικινδυνότητα.

Άλλο θέμα που μπορεί να ερευνηθεί στο μέλλον είναι η εξόρυξη δεδομένων μέσα από όλα τα δεδομένα του ελέγχου των τροφίμων που διεξήχθηκαν στην Κύπρο με χρήση των δεδομένων. Αυτά θα προκύψουν από τη μηχανογράφηση των Υγειονομικών Υπηρεσιών και επιπλέον όλα τα πιο πάνω δεδομένα θα είναι διαθέσιμα σε ηλεκτρονική μορφή.

Από τα αποτελέσματα της πιλοτικής εφαρμογής στα δεδομένα μικροβιολογικού ελέγχου του πόσιμου νερού, φαίνεται να υπάρχει σαφώς έδαφος εφαρμογής της εξόρυξης δεδομένων στα δεδομένα Δημόσιας Υγιεινής στην Κύπρο. Παρόλο που λόγω των περιορισμένων δεδομένων που έχουν αναλυθεί που καλύπτουν τη διάρκεια μόνο 3 ετών, τα αποτελέσματα έχουν χαμηλά επίπεδα εμπιστοσύνης,

ωστόσο από τα συμπεράσματα φαίνεται ότι πρόκειται για μια ανεκμετάλλευτη πηγή, από την οποία οι αρχές που ασχολούνται με την ασφάλεια των τροφίμων και του νερού, μπορούν να εξάγουν χρήσιμη πληροφόρηση αναφορικά με τάσεις και μοτίβα στην εμφάνιση περιστατικών, όπου απειλείται η δημόσια υγεία. Λόγω του μεγέθους και του εύρους κάλυψης των δεδομένων, η εγκυρότητα της πρόβλεψης είναι περιορισμένη αλλά όσο περνά ο χρόνος, τόσο περισσότερα δεδομένα συσσωρεύονται στις βάσεις δεδομένων της παρακολούθησης της ποιότητας των νερών με αποτέλεσμα ο βαθμός εμπιστοσύνης στην πρόβλεψη να μπορεί να αυξηθεί με την πρόσβαση σε περισσότερα δεδομένα.

Η πληροφόρηση που μπορεί να εξαχθεί με την εξόρυξη δεδομένων, μπορεί να βοηθήσει τις αρχές ελέγχου να αντεπεξέλθουν πιο αποτελεσματικά και αποδοτικά στα καθήκοντα τους, αναδιοργανώνοντας και επικεντρώνοντας τους πόρους τους ή ακόμη επιτρέποντάς τους να δράσουν προληπτικά/προδραστικά.

7. Βιβλιογραφία

Abramowicz, Witold and Zurada, J., *Knowledge discovery for business information systems,* (Boston MA: Kluwer Academic Publishers, 2001).

Ackoff, R. L., 'From data to wisdom', *Journal of Applied Systems Analysis*, (1989): 16, 3-9.

Batyrshin, Ildar et al., *Perception-based Data Mining and Decision Making in Economics and Finance,* (Berlin: Springer, - Springerlink Engineering, 2007), http://ezproxy.uws.edu.au:2048/login?url=http://dx.doi.org/, accessed October 2010.

Beulens, A.J.M. et al., 'Possibilities for applying data mining for early warning in food supply networks,' in: Proceedings of the Workshop on Methodologies and Tools for Complex System Modeling and Integrated Policy Assessment. (Laxenburg: International Institute for Applied Systems Analysis, 2006).

Bramer, Max, *Principles of Data Mining*, (London : Springer-Verlag, 2007), http://ezproxy.uws.edu.au:2048/login?url=http://dx.doi.org/, accessed September 2010.

Breivik, P. S. "Information literacy: Liberal education for the Information Age". Liberal Education 79 no 1 (1993): pp 24-29.

Brusilovsky, P. Kobsa and Alfred, Nejdl, W., *The Adaptive Web:Methods and Strategies of Web Personalization.* (Berlin: Springer-Verlag, 2007), http://ezproxy.uws.edu.au:2048/login?url=http://dx.doi.org/, accessed September 2010.

Castillo, D. L. and Abraham, N. S., 'How to keep up with theiterature. Clinical Gastroenterology and Hepatology. 6 no 12 (2008): pp 1294-300.

Chang, N. B. et al. Comparative data mining analysis for information retrieval of MODIS images: Monitoring lake turbidity changes at lake Okeechobee, Florida. Journal of Applied Remote Sensing 3 no 1 (2009).

Chen, Guoqing et al., *Intelligent Data Mining:Techniques and Application,* (Berlin: Springer, 2005)

Dubitzky, W., Berrar, D. and Granzow, M., *Fundamentals of Data Mining in Genomics and Proteomics,* (Boston, MA : Springer Science, 2007), http://ezproxy.uws.edu.au:2048/login?url=http://dx.doi.org/, accessed October 2010.

Dunham, Margaret H., Data Mining: Εισαγωγή και Προηγμένα Θέματα Εξόρυξης Γνώσης από Δεδομένα, (Upper Saddle River, N.J. : Prentice Hall/Pearson Education, 2003).

Dz□eroski, Sas□o and Todorovski, Ljupc□o, *Computational Discovery of Scientific Knowledge: Techniques, and Applications in Environmental and Life Sciences,* (Berlin: Springer-Verlag, 2007), http:// www.springerlink.com/openurl.asp? accesed September 2010

Elmasri, R., and Navathe, B. S. (2007), *Fundamentals of Database Systems,* (California: Benjamin & Cummings Publishing Co, 2007).

European Commission, *The Rapid Alert System for Food and Feed of the European Union,* (Luxembourg: Office for Official Publications of the European Communities, 2009).

Ευρωπαϊκή Ένωση, 'Κανονισμός (ΕΕ) αριθ. 16/2011 της Επιτροπής, της 10ης Ιανουαρίου 2011 , για τον καθορισμό μέτρων εφαρμογής του

συστήματος έγκαιρης προειδοποίησης για τρόφιμα και ζωοτροφές Κείμενο που παρουσιάζει ενδιαφέρον για τον ΕΟΧ.' (Ευρωπαϊκή Ένωση, 2011).

Ghosh, Ashish and Jain, Lakhmi C., *Evolutionary Computation in Data Mining,* (Berlin: Springer-Verlag, 2005), http://ezproxy.uws.edu.au:2048/login?url=http://dx.doi.org/, accessed November 2010.

Gleick, Peter H, 'Dirty Water: Estimated Deaths from Water-Related Diseases 2000-2020' in Pacific Institute Research Report. August 15, 2002 (California: Pacific Institute for Studies in Development, Environment, and Security 2002).

Guillet, Fabrice J. and Hamilton, Howard J., *Quality Measures in Data Mining,* (Berlin: Springer-Verlag, 2007), *http://ezproxy.uws.edu.au:2048/login?url=http://dx.doi.org/*, accessed October 2010.

González, A. I. Et al., 'Information Needs and Information-Seeking Behavior of Primary Care Physicians.' *Annals of Family Medicine* 5 no 4 (2007): pp 345-52.

Gustafson, J. Perry, Shoemaker, Randy and Snape, John W., *Genome Exploitation: Data Mining the Genome*, (Norwell: Springer Science+Business Media, Inc., 2006), http://ezproxy.uws.edu.au:2048/login?, accessed September 2010.

Han, Jiawei and Kamber, Micheline, *Data mining: Concepts and techniques,* (San Francisco, Calif.: The Morgan Kaufmann series in data management systems, 2001).

Hawkins, D.M., Identification of outliers. (London: Chapman and Hall, 1980).

Holmes, G., Donkin, A. and Witten, I. H., "Weka: A machine learning workbench", in Proc Second Australia and New Zealand Conference on

Intelligent Information Systems, Brisbane, Australia, (New York: Institute of Electrical and Electronics Engineers, 1994).

Huisman, Leendert M., *Data mining and diagnosing IC fails: Frontiers in electronic testing,* (Boston, MA: Springer Science, 2005).

Inmon William H., Building the Data Warehouse, (New York: John Wiley & Sons, Inc., 1992).

Kao, Anne Poteet and Stephen. R., *Natural Language Processing and Text Mining*, (London: Springer-Verlag, 2007), http://ezproxy.uws.edu.au:2048/login?url=http://dx.doi.org/, accessed September 2010.

Karimipouri, F., Delavari, M. R. and Kinaie, M., 'Water quality management using GIS data mining.' Journal of Environmental Informatics, 5, no 2 (2005): pp61-72.

Knorr, E. M. and Ng, Raymond, 'Algorithms for mining distance-based outliers in large datasets', in Proceedings of the 24th VLDB Conference, (Pittsburgh: The Pennsylvania State University CiteSeer Archives, 1998).

Knorr, E. M., Ng, Raymond T. and Tucakov, Vladimir 'Distance-based outliers: algorithms and applications', *The VLDB Journal 8* (2000): nos 3-4. p.p. 237-253.

Kokotos, D., & Linardatos, D., 'An application of data mining tools for the study of shipping safety in restricted waters'. Safety Science, 49 (2011): pp192-197.

Lavrac Nada, Keravnou Elpida T. and Zupan, Blaz, *Intelligent data analysis in medicine and pharmacology*, (Boston: Kluwer Academic Publishers, 1997).

Miller, H.J. and Han, J., Geographic Data Mining and Knowledge Discovery, (London : Taylor & Francis, 2001).

Ogwueleka, T. D., and Ogwueleka, F. N. 'Data mining application in predicting cryptosporidium SPP, oocysts and Giardia SPP, cysts concentrations in rivers.' Journal of Engineering Science and Technology, 5, no 3 (2010), 342-349.

Oxford English Dictionary, 2nd ed. (Oxford: Oxford University Press, 2000).

Pedro D and Michael Pazzani. 'On the Optimality of the Simple Bayesian Classifier under Zero-One Loss', *Machine Learning* 29 (1997): pp 103–130.

Σάββα, Γεώργιος, 'Σύστημα έγκαιρης προειδοποίησης για τα τρόφιμα και τις ζωοτροφές', *Γεωργικά Νέα* 63 (2009): nos 9-10ος , pp 52, 53.

Scott, I. Heyworth, R. and Fairweather, P., 'The use of evidence-based medicine in the practice of consultant physicians. Results of a questionnaire survey.' *Australian and New Zealand Journal of Medicine* 30 no 3 (2000): 309-310

Sowa, John F., "The Challenge of Knowledge Soup" in: Research Trends in Science, Technology and Mathematics Education. Edited by J. Ramadas & S. Chunawala, Homi Bhabha (Mumbai: Homi Bhabha Centre, 2006).

Terzi, O., 'Data mining approach for estimation evaporation from free water surface', Journal of Applied Sciences, 7, no 4 (2007): pp 593-596.

Τμήμα Κτηνιατρικών Υπηρεσιών, Website of the Ministry of Agriculture, Republic of Cyprus, http://www.cyprus.gov.cy/moa/Agriculture.nsf, accessed August, 2010.

Τμήμα Αναπτύξεως Υδάτων, Website of the Ministry of Agriculture, Republic of Cyprus, http://www.moa.gov.cy/moa/wdd/Wdd.nsf, accessed August, 2010.

Usama Fayyad, Shapiro, G. P. and Padhraic Smyth "From Data Mining to Knowledge Discovery in Databases", *AI Journal*, (Fall 1996).

Vaidya, Jaideep, Clifton, Christopher W. and Zhu, Yu Michael, *Privacy Preserving Data Mining,* (Boston, MA: Springer Science+Business Media, Inc., 2006), http://dx.doi.org/, accessed September 2010

Velicanu, Manole and Matei, Gheorghe, Database versus Data Warehouse. (2007), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=994176, accessed November 2010.

Verhoeven, A., 'What clinical information do doctors need?' *BMJ* 313 (1996): 1062-1068

Wang, C., Liu, B., and Qiu, E., 'The prediction of river water pollution density based on data mining technology.' Advanced Materials Research, (2010): pp113-116, 1285-1288.

Wang, Wei and Yang, Jiong, *Mining Sequential Patterns from Large Data Sets,* (Boston, MA: Springer Science + Business Media, Inc., 2005). http://ezproxy.uws.edu.au:2048/login?url=http://dx.doi.org/, accessed November 2010.

Wikipedia, http://en.wikipedia.org, accessed August 2010.

Witten, I. H. and *Eibe*, *Frank*, *Data mining: practical machine learning tools and techniques with Java implementation,* (San Francisco, Calif.: Morgan Kaufmann, 2000).

Wyatt, J., 'Medical informatics, artefacts or science? *Methods of Information in Medicine.'* 1996 Sep;35(3):197-200.

Υγειονομικές Υπηρεσίες, Υπουργείο Υγείας Κυπριακής Δημοκρατίας, *Εγχειρίδιο για την παρακολούθηση και έλεγχο της ποιότητας του νερού ανθρώπινης κατανάλωσης,* (Λευκωσία: Υπουργείο Υγείας Κυπριακής Δημοκρατίας, Νοέμβριος 2009)

Υγειονομικές Υπηρεσίες, Website of the Ministry of Health, Republic of Cyprus, http://www.moh.gov.cy/Moh/mphs/phs.nsf,, accessed August 2010.

Παράρτημα 1. Διαδικασία έγκρισης πρόσβασης σε πληροφορίες

From: George Savva [gsavva@mphs.moh.gov.cy] Sent: Τετ 21/1/2009 3:38
To: 'George Georgallas'
Cc:
Subject: Request to process the data in the MS Access EC RASFF database

κ. Γιωργαλλά,

Σε συνέχεια της συνομιλίας μας, σας διαβιβάζω σχετικό DRAFT email που θα σας παρακαλούσα αν συμφωνείτε με το περιεχόμενο του να το διαβιβάσετε στην Ομάδα RASFF της ΕΕ.

Ευχαριστώ,
Γιώργος Σάββα

DRAFT

Dear Jan,

I would like to let you know that George Savva – a member of the Cyprus NCP team – asked me to pass on to you a request to process the data in the MS Access EC RASFF database (found on CIRCA), which he already has access to, with a technique called Data Mining. Data mining is a mean to extract useful information from large and complex business data; his findings – significant correlations and patterns – might be of particular importance to our sector as they can reveal trends towards potential dangers that are more likely to occur than others, allowing us to reallocate our attention and limited resources.

Mr Savva is going to pursue this analysis for his master thesis in the context of a postgraduate program in Computer Systems he is attending to, at the Cyprus Open University. He will be working under the guidance of Professor Dr Thanasi Hadjilako, also a Cyprus Government employee, who is not going to have any access to the data. All the outcomes and reports that might result from the proposed analysis of the data will be available to any party, including the Cyprus Open University, only after the data will be altered so as company names, brand names and countries will be altered to pseudo names. At the end of his study he will make all his work available to both our service and your Team, if it is of any interest to you. Furthermore, if he wishes to publish any reports with the full data to any other party, he would be obliged to get your prior permission.

Given that the proposed work might be beneficial to our services, by revealing significant correlations with several factors that we haven't thought of so far I would like to ask that Mr Savva is given the permission to carry out his proposal. If there are established procedures on the issue that he should follow, please do let me know about them.

Thanking you in advance,

For
George Georgallas
Head of Health Services
Ministry of Health
1, Prodromou Str
1449, Nicosia, Cyprus
Tel: ++357 22605554
Fax: ++357 22305345
e-mail: ggeorgallas@mphs.moh.gov.cy

From: George Georgallas [mailto:ggeorgallas@mphs.moh.gov.cy]
Sent: Friday, January 23, 2009 7:21 AM
To: DE FELIPE GARDON Jose (SANCO)
Cc: George Savva
Subject: Access to EC RASFF Database

Dear Jose,

I hope you are in good health.

I would like to let you know that George Savva – a member of the Cyprus NCP team – asked me to pass on to you a request to process the data in the MS Access EC RASFF database (found on CIRCA), which he already has access to, with a technique called Data Mining. Data mining is a mean to extract useful information from large and complex business data; his findings – significant correlations and patterns – might be of particular importance to our sector as they can reveal trends towards potential dangers that are more likely to occur than others, allowing us to reallocate our attention and limited resources.

Mr. Savva is going to pursue this analysis for his master thesis in the context of a postgraduate program in Computer Systems he is attending to, at the Cyprus Open University. He will be working under the guidance of Professor Dr Thanasi Hadjiliako, also a Cyprus Government employee, who is not going to have any access to the data. All the outcomes and reports that might result from the proposed analysis of the data will be available to any party, including the Cyprus Open University, only after the data will be altered so as company names, brand names and countries will be altered to pseudo names. At the end of his study he will make all his work available to both our service and your Team, if it is of any interest to you. Furthermore, if he wishes to publish any reports with the full data to any other party, he would be obliged to get your prior permission.

Given that the proposed work might be beneficial to our services, by revealing significant correlations with several factors that we haven't thought of so far, I would like to ask that Mr. Savva is given the permission to carry out his proposal. If there are established procedures on the issue that he should follow, please do let me know about them.

Thanking you in advance and looking forward to hearing from you soon,

George Georgallas
Head of Health Services
Medical and Public Health Services
Ministry of Health
1 Prodromos Street
1449 Nicosia Cyprus
Tel.: +357 22605554
Fax: +357 22305387

E-mail: ggeorgallas@mphs.moh.gov.cy

From: Sanco-Rasff@ec.europa.eu [mailto:Sanco-Rasff@ec.europa.eu]
Sent: Δευτέρα, 26 Ιανουαρίου 2009 6:43 μμ
To: ggeorgallas@mphs.moh.gov.cy
Cc: gsavva@mphs.moh.gov.cy
Subject: log100758 RE: Access to EC RASFF Database

Next Previous

Dear George,
I think that there no problem to use the RASFF information. George Savva is working for the RASFF in Cyprus and he will use this information for the benefic of the Cyprus consumers and who knows if also for other EU countries. As you mention all the sensible information should not be used (companies name, clients etc.)
I am sure that George Savva will do a good work with this information. Please keep us informed of the outcome.
Best regards
José Luis De Felipe

Head of Sector
RASFF
EUROPEAN COMMISSION
HEALTH AND CONSUMERS DIRECTORATE-GENERAL
Directorate E - Safety of the Food chain
E2 - Food Hygiene, Alert System and Training
Official address: Rue de la Loi 200, B-1049 Bruxelles/Wetstraat 200,
B-1049 Brussel - Belgium
Office:Rue Belliard 232, 4th floor, room 53.
Telephone: direct (+32-2) 299.3880.
Fax: (+32-2) 296.76.74.
PLEASE ADDRESS YOUR REPLIES TO THE FOLLOWING E-MAIL ADDRESS:
sanco-rasff@ec.europa.eu
Website: http://europa.eu.int/comm/dgs/health_consumer/index_en.htm

Γιώργος Σάββα
Αγίου Μεθοδίου 3,
Στρόβολος, 2055

20 Οκτωβρίου, 2010

Προϊστάμενο
Υγειονομικών Υπηρεσιών

Θέμα: Άδεια σε πρόσβαση στα αποτελέσματα των ποσίμων νερών

Επιθυμώ να αναφερθώ στο πιο πάνω θέμα και να σας πληροφορήσω τα ακόλουθα:

Από το 2007 φοιτώ με τη μέθοδο της εξ αποστάσεως μάθησης στο μεταπτυχιακό πρόγραμμα «Πληροφοριακά Συστήματα» στο Ανοικτό Πανεπιστήμιο Κύπρου το οποίο και αναμένω να ολοκληρώσω το 2011. Για σκοπούς ολοκλήρωσης του προγράμματος σπουδών μου θα πρέπει να ετοιμάσω μεταπτυχιακή εργασία.

Επέλεξα να ασχοληθώ με την εξόρυξη δεδομένων (Data Mining) που καλείται η εξεύρεση νέων *πληροφοριών από επαναλαμβανόμενα Προτύπα* (patterns) ή κανόνες (rules) σε μεγάλους όγκους δεδομένων.

Παρακαλώ όπως μου επιτραπεί να έχω πρόσβαση στα αποτελέσματα των εργαστηριακών εξετάσεων του πόσιμου νερού για να τα χρησιμοποιήσω στην εργασία μου. Αν μου επιτραπεί η πιο πάνω πρόσβαση τότε δυνατό από την επεξεργασία των στοιχείων να προκύψουν ευρήματα που να ωφελήσουν την δημόσια υγεία αναγνωρίζοντας περιοχές όπου είτε χρονικά ή θεματικά αναμένεται να υπάρξει πρόβλημα επιτρέποντας στις αρμόδιες αρχές την καλύτερη αξιοποίηση των πόρων τους με την επικέντρωση τους στις περιπτώσεις των προβλέψιμων αναδυόμενων κινδύνων.

Ευχαριστώ εκ των προτέρων για την υποστήριξη σας.

Γιώργος Σάββα

ΚΥΠΡΙΑΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ ΥΓΕΙΑΣ

ΤΜΗΜΑ
ΙΑΤΡΙΚΩΝ ΥΠΗΡΕΣΙΩΝ ΚΑΙ
ΥΠΗΡΕΣΙΩΝ ΔΗΜΟΣΙΑΣ ΥΓΕΙΑΣ
ΥΓΕΙΟΝΟΜΙΚΗ ΥΠΗΡΕΣΙΑ
1449 ΛΕΥΚΩΣΙΑ

Αρ. Φακ.: 5.10.009
Τηλ. 22605554
Email: healthservices@mphs.moh.gov.cy

01 Νοεμβρίου, 2010

κ. Γιώργο Σάββα
Αγίου Μεθοδίου 3,
Στρόβολος, 2055

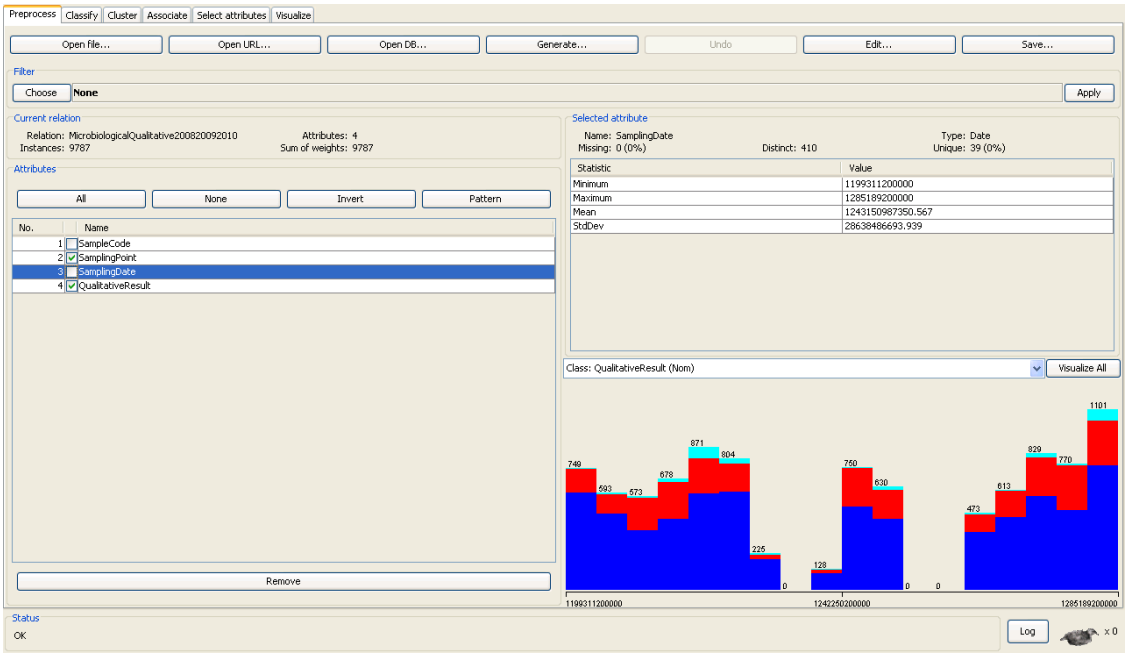
Θέμα : <u>Άδεια σε πρόσβαση στα αποτελέσματα των ποσίμων νερών</u>

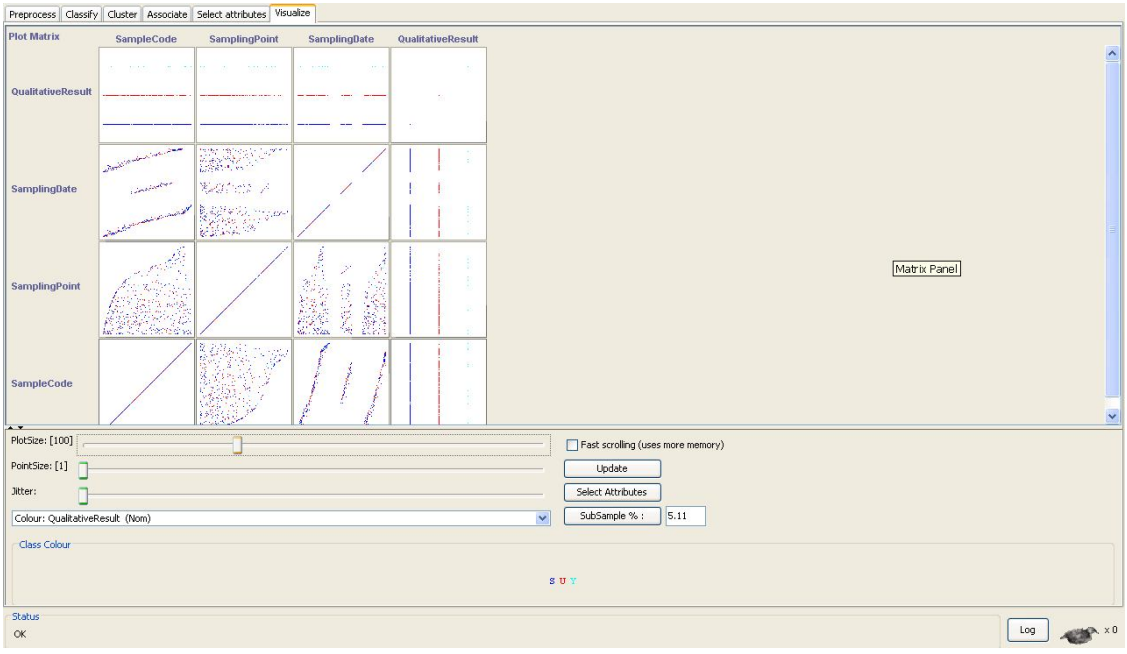
Επιθυμώ να αναφερθώ στην επιστολή σας ημερομηνίας 20.10.2010 αναφορικά με αίτημα για Άδεια σε πρόσβαση στα αποτελέσματα των ποσίμων νερών για σκοπούς της μεταπτυχιακής σας διατριβής και να σας πληροφορήσω ότι αυτό έχει εγκριθεί. Παρακαλώ όπως επικοινωνήσετε με την Λειτουργό του Γραφείου κα Χαραλάμπους για να παραλάβετε τα δεδομένα.

Παρακαλώ όπως ενημερωθώ για τα αποτελέσματα της έρευνας σας.

(Γιώργος Γεωργαλλάς)
Προϊστάμενος
Υγειονομικών Υπηρεσιών

Παράτημα 2. Στιγμιότυπα από την εξόρυξη δεδομένων





Crosstabs Month vs Results

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Date * Results	9588	100.0%	0	.0%	9588	100.0%

Date * Results Crosstabulation

			Results			
			.00	1.00	2.00	Total
Date	1.00	Count	515	4	131	650
		% within Date	79.2%	.6%	20.2%	100.0%
	2.00	Count	558	10	170	738
		% within Date	75.6%	1.4%	23.0%	100.0%
	3.00	Count	572	9	146	727
		% within Date	78.7%	1.2%	20.1%	100.0%
	4.00	Count	408	24	136	568
		% within Date	71.8%	4.2%	23.9%	100.0%
	5.00	Count	771	19	331	1121
		% within Date	68.8%	1.7%	29.5%	100.0%
	6.00	Count	743	15	455	1213
		% within Date	61.3%	1.2%	37.5%	100.0%
	7.00	Count	786	22	381	1189
		% within Date	66.1%	1.9%	32.0%	100.0%
	8.00	Count	674	57	289	1020
		% within Date	66.1%	5.6%	28.3%	100.0%
	9.00	Count	957	84	292	1333
		% within Date	71.8%	6.3%	21.9%	100.0%
	10.00	Count	292	21	89	402
		% within Date	72.6%	5.2%	22.1%	100.0%
	11.00	Count	300	13	90	403
		% within Date	74.4%	3.2%	22.3%	100.0%
	12.00	Count	188	11	25	224

	% within Date	83.9%	4.9%	11.2%	100.0%
Total	Count	6764	289	2535	9588
	% within Date	70.5%	3.0%	26.4%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	315.451[a]	22	.000
Likelihood Ratio	314.113	22	.000
Linear-by-Linear Association	2.722	1	.099
N of Valid Cases	9588		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.75.

Crosstabs Tester vs Results

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Tester * Results	9588	100.0%	0	.0%	9588	100.0%

Tester * Results Crosstabulation

		Results			
		.00	1.00	2.00	Total
Tester Σάββα Γ.	Count	1	0	0	1
	% within Tester	100.0%	.0%	.0%	100.0%
Total	Count	6764	289	2535	9588
	% within Tester	70.5%	3.0%	26.4%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2765.393[a]	238	.000
Likelihood Ratio	2571.899	238	.000
N of Valid Cases	9588		

a. 153 cells (42.5%) have expected count less than 5. The minimum expected count is .03.

Crosstabs Place vs Results (*Set used: first 999 samples)

Notes

Output Created		21-Mar-2011 19:01:27
Comments		
Input	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	999
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics for each table are based on all the cases with valid data in the specified range(s) for all variables in each table.
Syntax		CROSSTABS /TABLES=Place BY Results /FORMAT=AVALUE TABLES /STATISTICS=CHISQ /CELLS=COUNT ROW /COUNT ROUND CELL.
Resources	Processor Time	00 00:00:00.078
	Elapsed Time	00 00:00:00.080
	Dimensions Requested	2
	Cells Available	174762

Case Processing Summary

	Cases

	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Place * Results	999	100.0%	0	.0%	999	100.0%

Place * Results Crosstabulation

			Results			
			.00	1.00	2.00	Total
Place	ΠΧ001-001-Δ1	Count	12	0	2	14
		% within Place	85.7%	.0%	14.3%	100.0%
Total		Count	803	51	145	999
		% within Place	80.4%	5.1%	14.5%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	879.248[a]	406	.000
Likelihood Ratio	668.618	406	.000
N of Valid Cases	999		

a. 574 cells (93.8%) have expected count less than 5. The minimum expected count is .05.

Παράτημα 4. Αποτελέσματα Ανάλυσης με το WEKA

Sampling Point – Result Qualitative	Sampling Officer – Result Qualitative	Sampling Point –Officer –Month - Result
=== Run information === Scheme: weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5 Relation: RenameNominalQualitativeMbQualitativeNom-weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.attribute.Remove-R2-3 Instances: 9588 Attributes: 2 AreaCode Qualitative Test mode: split 66.0% train, remainder test === Classifier model (full training set) === Bayes Network Classifier not using ADTree #attributes=2 #classindex=1 Network structure (nodes followed by parents) AreaCode(1570): Qualitative Qualitative(3): LogScore Bayes: -73869.8441973634	=== Run information === Scheme: weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5 Relation: RenameNominalQualitativeMbQualitativeNom-weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.attribute.Remove-R1-2 Instances: 9588 Attributes: 2 Officer Qualitative Test mode: split 66.0% train, remainder test === Classifier model (full training set) === Bayes Network Classifier not using ADTree #attributes=2 #classindex=1 Network structure (nodes followed by parents) Officer(120): Qualitative Qualitative(3): LogScore Bayes: -46625.731047036665 LogScore BDeu: -49089.865794303594	=== Run information === Scheme: weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5 Relation: RenameNominalQualitativeMbQualitativeNom-weka.filters.unsupervised.attribute.Remove-R2 Instances: 9588 Attributes: 4 AreaCode SamMonth1 Officer Qualitative Test mode: split 66.0% train, remainder test === Classifier model (full training set) === Bayes Network Classifier not using ADTree #attributes=4 #classindex=3 Network structure (nodes followed by parents) AreaCode(1570): Qualitative SamMonth1(12): Qualitative

LogScore BDeu: -124341.59157275126
LogScore MDL: -110277.82740870766
LogScore ENTROPY: -88691.14135497512
LogScore AIC: -93400.14135497512

Time taken to build model: 0.08 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances 2340 71.7791 %
Incorrectly Classified Instances 920 28.2209 %
Kappa statistic 0.2643
Mean absolute error 0.2473
Root mean squared error 0.3556
Relative absolute error 86.1103 %
Root relative squared error 94.137 %
Coverage of cases (0.95 level) 98.8037 %
Mean rel. region size (0.95 level) 69.7342 %
Total Number of Instances 3260

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.863	0.633	0.769	0.863	0.813	0.718	s
0.388	0.131	0.51	0.388	0.441	0.754	u
0.152	0.002	0.75	0.152	0.252	0.839	y
Weighted Avg. 0.718	0.483	0.701	0.718			

LogScore MDL: -48664.24306992267
LogScore ENTROPY: -47018.539036520684
LogScore AIC: -47377.539036520684

Time taken to build model: 0.05 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances 2344 71.9018 %
Incorrectly Classified Instances 916 28.0982 %
Kappa statistic 0.1536
Mean absolute error 0.2478
Root mean squared error 0.3534
Relative absolute error 86.2699 %
Root relative squared error 93.5532 %
Coverage of cases (0.95 level) 98.7423 %
Mean rel. region size (0.95 level) 69.4785 %
Total Number of Instances 3260

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.942	0.825	0.736	0.942	0.826	0.708	s
0.193	0.056	0.55	0.193	0.286	0.752	u
0.01	0	0.5	0.01	0.02	0.855	y
Weighted Avg. 0.719	0.6	0.681	0.719	0.661	0.724	

=== Confusion Matrix ===

Officer(120): Qualitative
Qualitative(3):
LogScore Bayes: -136511.18113546423
LogScore BDeu: -189565.19795752075
LogScore MDL: -175056.21691288674
LogScore ENTROPY: -151681.7186780211
LogScore AIC: -156780.7186780211

Time taken to build model: 0.13 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances 2309 70.8282 %
Incorrectly Classified Instances 951 29.1718 %
Kappa statistic 0.356
Mean absolute error 0.2287
Root mean squared error 0.3615
Relative absolute error 79.6295 %
Root relative squared error 95.6863 %
Coverage of cases (0.95 level) 98.0675 %
Mean rel. region size (0.95 level) 62.5256 %
Total Number of Instances 3260

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.757	0.399	0.823	0.757	0.789	0.752	s
0.614	0.213	0.503	0.614	0.553	0.79	u

0.699 0.731 === Confusion Matrix === a b c <-- classified as 1996 314 3 | a = s 517 329 2 | b = u 82 2 15 | c = y	a b c <-- classified as 2179 133 1 | a = s 684 164 0 | b = u 97 1 1 | c = y	0.364 0.019 0.379 0.364 0.371 0.889 y Weighted Avg. 0.708 0.339 0.726 0.708 0.715 0.766 === Confusion Matrix === a b c <-- classified as 1752 512 49 | a = s 317 521 10 | b = u 61 2 36 | c = y
=== Run information === Scheme: weka.classifiers.bayes.NaiveBayes Relation: RenameNominalQualitativeMbQualitativeNom- weka.filters.unsupervised.attribute.Remove-R2- weka.filters.unsupervised.attribute.Remove-R2-3 Instances: 9588 Attributes: 2 AreaCode Qualitative Test mode: split 66.0% train, remainder test === Classifier model (full training set) === Naive Bayes Classifier Class Attribute s u y (0.71) (0.26) (0.03) ========== === Evaluation on test split ===	=== Run information === Scheme: weka.classifiers.bayes.NaiveBayes Relation: RenameNominalQualitativeMbQualitativeNom- weka.filters.unsupervised.attribute.Remove-R2- weka.filters.unsupervised.attribute.Remove-R1-2 Instances: 9588 Attributes: 2 Officer Qualitative Test mode: split 66.0% train, remainder test === Classifier model (full training set) === Naive Bayes Classifier Class Attribute s u y (0.71) (0.26) (0.03) ========== =	=== Run information === Scheme: weka.classifiers.bayes.NaiveBayes Relation: RenameNominalQualitativeMbQualitativeNom- weka.filters.unsupervised.attribute.Remove-R2 Instances: 9588 Attributes: 4 AreaCode SamMonth1 Officer Qualitative Test mode: split 66.0% train, remainder test === Classifier model (full training set) === Naive Bayes Classifier Class Attribute s u y (0.71) (0.26) (0.03) ========== ====

=== Summary === Correctly Classified Instances 2328 71.411 % Incorrectly Classified Instances 932 28.589 % Kappa statistic 0.2085 Mean absolute error 0.2526 Root mean squared error 0.3543 Relative absolute error 87.956 % Root relative squared error 93.7925 % Coverage of cases (0.95 level) 98.865 % Mean rel. region size (0.95 level) 70.2556 % Total Number of Instances 3260 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.893 0.72 0.752 0.893 0.816 0.714 s 0.31 0.104 0.513 0.31 0.386 0.75 u 0 0 0 0 0 0.825 y Weighted Avg. 0.714 0.538 0.667 0.714 0.68 0.727 === Confusion Matrix === a b c <-- classified as 2065 248 0 | a = s 585 263 0 | b = u 97 2 0 | c = y	Time taken to build model: 0.02 seconds === Evaluation on test split === === Summary === Correctly Classified Instances 2344 71.9018 % Incorrectly Classified Instances 916 28.0982 % Kappa statistic 0.1536 Mean absolute error 0.2496 Root mean squared error 0.3537 Relative absolute error 86.9251 % Root relative squared error 93.6345 % Coverage of cases (0.95 level) 98.865 % Mean rel. region size (0.95 level) 70.726 % Total Number of Instances 3260 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.942 0.825 0.736 0.942 0.826 0.705 s 0.193 0.056 0.55 0.193 0.286 0.751 u 0.01 0 0.5 0.01 0.02 0.85 y Weighted Avg. 0.719 0.6 0.681 0.719 0.661 0.721 === Confusion Matrix === a b c <-- classified as 2179 133 1 | a = s 684 164 0 | b = u	=== Evaluation on test split === === Summary === Correctly Classified Instances 2326 71.3497 % Incorrectly Classified Instances 934 28.6503 % Kappa statistic 0.348 Mean absolute error 0.2315 Root mean squared error 0.3553 Relative absolute error 80.6127 % Root relative squared error 94.0451 % Coverage of cases (0.95 level) 98.6503 % Mean rel. region size (0.95 level) 64.5706 % Total Number of Instances 3260 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.78 0.435 0.814 0.78 0.796 0.751 s 0.581 0.2 0.505 0.581 0.541 0.789 u 0.303 0.012 0.435 0.303 0.357 0.874 y Weighted Avg. 0.713 0.361 0.722 0.713 0.716 0.765 === Confusion Matrix === a b c <-- classified as 1803 478 32 | a = s 348 493 7 | b = u 64 5 30 | c = y

	97 1 1 | c = y	
=== Run information === Scheme: weka.classifiers.bayes.NaiveBayesUpdateable Relation: RenameNominalQualitativeMbQualitativeNom- weka.filters.unsupervised.attribute.Remove-R2- weka.filters.unsupervised.attribute.Remove-R2-3 Instances: 9588 Attributes: 2 AreaCode Qualitative Test mode: split 66.0% train, remainder test === Classifier model (full training set) === Naive Bayes Classifier Class Attribute s u y (0.71) (0.26) (0.03) ==========Time taken to build model: 0 seconds === Evaluation on test split === === Summary === Correctly Classified Instances 2328 71.411 % Incorrectly Classified Instances 932 28.589 % Kappa statistic 0.2085 Mean absolute error 0.2526	=== Run information === Scheme: weka.classifiers.bayes.NaiveBayesUpdateable Relation: RenameNominalQualitativeMbQualitativeNom- weka.filters.unsupervised.attribute.Remove-R2- weka.filters.unsupervised.attribute.Remove-R1-2 Instances: 9588 Attributes: 2 Officer Qualitative Test mode: split 66.0% train, remainder test === Classifier model (full training set) === Naive Bayes Classifier Class Attribute s u y (0.71) (0.26) (0.03) ========== Officer [total] 6884.0 2655.0 409.0 Time taken to build model: 0 seconds === Evaluation on test split === === Summary === Correctly Classified Instances 2344 71.9018 %	=== Run information === Scheme: weka.classifiers.bayes.NaiveBayesUpdateable Relation: RenameNominalQualitativeMbQualitativeNom- weka.filters.unsupervised.attribute.Remove-R2 Instances: 9588 Attributes: 4 AreaCode SamMonth1 Officer Qualitative Test mode: split 66.0% train, remainder test === Classifier model (full training set) === Naive Bayes Classifier Class Attribute s u y (0.71) (0.26) (0.03) ========== Time taken to build model: 0.02 seconds === Evaluation on test split === === Summary === Correctly Classified Instances 2326 71.3497 % Incorrectly Classified Instances 934 28.6503

Root mean squared error 0.3543 Relative absolute error 87.956 % Root relative squared error 93.7925 % Coverage of cases (0.95 level) 98.865 % Mean rel. region size (0.95 level) 70.2556 % Total Number of Instances 3260 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.893 0.72 0.752 0.893 0.816 0.714 s 0.31 0.104 0.513 0.31 0.386 0.75 u 0 0 0 0 0 0.825 y Weighted Avg. 0.714 0.538 0.667 0.714 0.68 0.727 === Confusion Matrix === a b c <-- classified as 2065 248 0 | a = s 585 263 0 | b = u 97 2 0 | c = y	Incorrectly Classified Instances 916 28.0982 % Kappa statistic 0.1536 Mean absolute error 0.2496 Root mean squared error 0.3537 Relative absolute error 86.9251 % Root relative squared error 93.6345 % Coverage of cases (0.95 level) 98.865 % Mean rel. region size (0.95 level) 70.726 % Total Number of Instances 3260 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.942 0.825 0.736 0.942 0.826 0.705 s 0.193 0.056 0.55 0.193 0.286 0.751 u 0.01 0 0.5 0.01 0.02 0.85 y Weighted Avg. 0.719 0.6 0.681 0.719 0.661 0.721 === Confusion Matrix === a b c <-- classified as 2179 133 1 | a = s 684 164 0 | b = u 97 1 1 | c = y	% Kappa statistic 0.348 Mean absolute error 0.2315 Root mean squared error 0.3553 Relative absolute error 80.6127 % Root relative squared error 94.0451 % Coverage of cases (0.95 level) 98.6503 % Mean rel. region size (0.95 level) 64.5706 % Total Number of Instances 3260 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.78 0.435 0.814 0.78 0.796 0.751 s 0.581 0.2 0.505 0.581 0.541 0.789 u 0.303 0.012 0.435 0.303 0.357 0.874 y Weighted Avg. 0.713 0.361 0.722 0.713 0.716 0.765 === Confusion Matrix === a b c <-- classified as 1803 478 32 | a = s 348 493 7 | b = u 64 5 30 | c = y
=== Run information === Scheme: weka.classifiers.lazy.LWL -U 0 -K -1 -A === Evaluation on test split ===	=== Run information === Scheme: weka.classifiers.lazy.LWL -U 0 -K -1 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\"" -W	=== Run information === Scheme: weka.classifiers.lazy.LWL -U 0 -K -1 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\"" -W

=== Summary === Correctly Classified Instances 2340 71.7791 % Incorrectly Classified Instances 920 28.2209 % Kappa statistic 0.2855 Mean absolute error 0.2202 Root mean squared error 0.3634 Relative absolute error 76.6635 % Root relative squared error 96.2035 % Coverage of cases (0.95 level) 94.4172 % Mean rel. region size (0.95 level) 55.818 % Total Number of Instances 3260 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.85 0.595 0.777 0.85 0.812 0.722 s 0.412 0.138 0.512 0.412 0.457 0.753 u 0.263 0.008 0.51 0.263 0.347 0.783 y Weighted Avg. 0.718 0.458 0.7 0.718 0.705 0.732 === Confusion Matrix === a b c <-- classified as 1965 331 17 | a = s 491 349 8 | b = u 72 1 26 | c = y	weka.classifiers.trees.DecisionStump Relation: RenameNominalQualitativeMbQualitativeNom-weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.attribute.Remove-R1-2 Instances: 9588 Attributes: 2 Officer Qualitative Test mode: split 66.0% train, remainder test === Classifier model (full training set) === Locally weighted learning ========== Using classifier: weka.classifiers.trees.DecisionStump Using linear weighting kernels Using all neighbours Time taken to build model: 0 seconds === Evaluation on test split === === Summary === Correctly Classified Instances 2344 71.9018 % Incorrectly Classified Instances 916 28.0982 % Kappa statistic 0.1536 Mean absolute error 0.2446 Root mean squared error 0.3532 Relative absolute error 85.1584 % Root relative squared error 93.4952 % Coverage of cases (0.95 level) 98.5583 % Mean rel. region size (0.95 level) 67.0654 % Total Number of Instances 3260	weka.classifiers.trees.DecisionStump Relation: RenameNominalQualitativeMbQualitativeNom-weka.filters.unsupervised.attribute.Remove-R2 Instances: 9588 Attributes: 4 AreaCode SamMonth1 Officer Qualitative Test mode: split 66.0% train, remainder test === Classifier model (full training set) === Locally weighted learning ========== Using classifier: weka.classifiers.trees.DecisionStump Using linear weighting kernels Using all neighbours Time taken to build model: 0.02 seconds === Evaluation on test split === === Summary === Correctly Classified Instances 2344 71.9018 % Incorrectly Classified Instances 916 28.0982 % Kappa statistic 0.1269 Mean absolute error 0.2588 Root mean squared error 0.3568 Relative absolute error 90.121 % Root relative squared error 94.4408 % Coverage of cases (0.95 level) 98.773 %

	=== Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.942 0.825 0.736 0.942 0.826 0.71 s 0.193 0.056 0.55 0.193 0.286 0.752 u 0.01 0 0.5 0.01 0.02 0.835 y Weighted Avg. 0.719 0.6 0.681 0.719 0.661 0.725 === Confusion Matrix === a b c <-- classified as 2179 133 1 | a = s 684 164 0 | b = u 97 1 1 | c = y	Mean rel. region size (0.95 level) 67.9243 % Total Number of Instances 3260 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.961 0.868 0.73 0.961 0.83 0.703 s 0.123 0.036 0.545 0.123 0.2 0.731 u 0.172 0.002 0.708 0.172 0.276 0.843 y Weighted Avg. 0.719 0.625 0.681 0.719 0.649 0.714 === Confusion Matrix === a b c <-- classified as 2223 86 4 | a = s 741 104 3 | b = u 81 1 17 | c = y
=== Run information === Scheme: weka.classifiers.rules.DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5" Relation: RenameNominalQualitativeMbQualitativeNom-weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.attribute.Remove-R2-3 Instances: 9588 Attributes: 2 AreaCode Qualitative	=== Run information === Scheme: weka.classifiers.rules.DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5" Relation: RenameNominalQualitativeMbQualitativeNom-weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.attribute.Remove-R1-2 Instances: 9588 Attributes: 2 Officer Qualitative	=== Run information === Scheme: weka.classifiers.rules.DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5" Relation: RenameNominalQualitativeMbQualitativeNom Instances: 9588 Attributes: 5 AreaCode SamDate SamMonth1 Officer

Test mode: split 66.0% train, remainder test === Classifier model (full training set) === Decision Table: Number of training instances: 9588 Number of Rules : 1570 Non matches covered by Majority class. Best first. Start set: no attributes Search direction: forward Stale search after 5 node expansions Total number of subsets evaluated: 1 Merit of best subset found: 71.882 Evaluation (for feature selection): CV (leave one out) Feature set: 1,2 Time taken to build model: 2.42 seconds === Evaluation on test split === === Summary === Correctly Classified Instances 2333 71.5644 % Incorrectly Classified Instances 927 28.4356 % Kappa statistic 0.2632 Mean absolute error 0.3096 Root mean squared error 0.3771 Relative absolute error 107.7976 % Root relative squared error 99.8326 % Coverage of cases (0.95 level) 99.6626 % Mean rel. region size (0.95 level) 93.7423 % Total Number of Instances 3260	Test mode: split 66.0% train, remainder test === Classifier model (full training set) === Decision Table: Number of training instances: 9588 Number of Rules : 120 Non matches covered by Majority class. Best first. Start set: no attributes Search direction: forward Stale search after 5 node expansions Total number of subsets evaluated: 1 Merit of best subset found: 70.672 Evaluation (for feature selection): CV (leave one out) Feature set: 1,2 Time taken to build model: 0.63 seconds === Evaluation on test split === === Summary === Correctly Classified Instances 2344 71.9018 % Incorrectly Classified Instances 916 28.0982 % Kappa statistic 0.1536 Mean absolute error 0.2571 Root mean squared error 0.3543 Relative absolute error 89.5285 % Root relative squared error 93.7894 % Coverage of cases (0.95 level) 99.2025 % Mean rel. region size (0.95 level) 74.5194 % Total Number of Instances 3260 === Detailed Accuracy By Class ===	Qualitative Test mode: split 66.0% train, remainder test === Classifier model (full training set) === Decision Table: Number of training instances: 9588 Number of Rules : 1570 Non matches covered by Majority class. Best first. Start set: no attributes Search direction: forward Stale search after 5 node expansions Total number of subsets evaluated: 13 Merit of best subset found: 71.882 Evaluation (for feature selection): CV (leave one out) Feature set: 1,5 Time taken to build model: 4.97 seconds === Evaluation on test split === === Summary === Correctly Classified Instances 2344 71.9018 % Incorrectly Classified Instances 916 28.0982 % Kappa statistic 0.1536 Mean absolute error 0.2571 Root mean squared error 0.3543 Relative absolute error 89.5285 % Root relative squared error 93.7894 % Coverage of cases (0.95 level) 99.2025 % Mean rel. region size (0.95 level) 74.5194 %

=== Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.861 0.629 0.77 0.861 0.813 0.688 s 0.38 0.13 0.507 0.38 0.434 0.748 u 0.202 0.006 0.526 0.202 0.292 0.766 y Weighted Avg. 0.716 0.48 0.694 0.716 0.698 0.706 === Confusion Matrix === a b c <-- classified as 1991 312 10 | a = s 518 322 8 | b = u 78 1 20 | c = y	TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.942 0.825 0.736 0.942 0.826 0.702 s 0.193 0.056 0.55 0.193 0.286 0.752 u 0.01 0 0.5 0.01 0.02 0.846 y Weighted Avg. 0.719 0.6 0.681 0.719 0.661 0.72 === Confusion Matrix === a b c <-- classified as 2179 133 1 | a = s 684 164 0 | b = u 97 1 1 | c = y	Total Number of Instances 3260 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.942 0.825 0.736 0.942 0.826 0.702 s 0.193 0.056 0.55 0.193 0.286 0.752 u 0.01 0 0.5 0.01 0.02 0.846 y Weighted Avg. 0.719 0.6 0.681 0.719 0.661 0.72 === Confusion Matrix === a b c <-- classified as 2179 133 1 | a = s 684 164 0 | b = u 97 1 1 | c = y
=== Run information === Scheme: weka.classifiers.trees.RandomForest -I 10 -K 0 -S 1 Relation: RenameNominalQualitativeMbQualitativeNom-weka.filters.unsupervised.attribute.Remove-R2-4 Instances: 9588 Attributes: 2 AreaCode Qualitative Test mode: split 66.0% train, remainder test	=== Run information === Scheme: weka.classifiers.trees.RandomForest -I 10 -K 0 -S 1 Relation: RenameNominalQualitativeMbQualitativeNom-weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.attribute.Remove-R1-2 Instances: 9588 Attributes: 2 Officer Qualitative Test mode: split 66.0% train, remainder test	=== Run information === Scheme: weka.classifiers.trees.RandomForest -I 10 -K 0 -S 1 Relation: RenameNominalQualitativeMbQualitativeNom-weka.filters.unsupervised.attribute.Remove-R2 Instances: 9588 Attributes: 4 AreaCode SamMonth1 Officer Qualitative

=== Classifier model (full training set) === Random forest of 10 trees, each constructed while considering 2 random features. Out of bag error: 0.2685 Time taken to build model: 11.14 seconds === Evaluation on test split === === Summary === Correctly Classified Instances 2311 70.8896 % Incorrectly Classified Instances 949 29.1104 % Kappa statistic 0.2908 Mean absolute error 0.2232 Root mean squared error 0.3622 Relative absolute error 77.7206 % Root relative squared error 95.8849 % Coverage of cases (0.95 level) 95.8589 % Mean rel. region size (0.95 level) 57.9755 % Total Number of Instances 3260 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.821 0.551 0.784 0.821 0.802 0.723 s 0.455 0.164 0.494 0.455 0.474 0.752 u 0.273 0.01 0.458 0.273 0.342	=== Classifier model (full training set) === Random forest of 10 trees, each constructed while considering 2 random features. Out of bag error: 0.2863 Time taken to build model: 0.52 seconds === Evaluation on test split === === Summary === Correctly Classified Instances 2338 71.7178 % Incorrectly Classified Instances 922 28.2822 % Kappa statistic 0.1572 Mean absolute error 0.2445 Root mean squared error 0.3533 Relative absolute error 85.1257 % Root relative squared error 93.5229 % Coverage of cases (0.95 level) 98.589 % Mean rel. region size (0.95 level) 67.6585 % Total Number of Instances 3260 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.935 0.814 0.737 0.935 0.824 0.709 s 0.205 0.062 0.537 0.205 0.297 0.751 u 0.01 0 0.5 0.01 0.02 0.838 y Weighted Avg. 0.717 0.594 0.678 0.717 0.663	Test mode: split 52.0% train, remainder test === Classifier model (full training set) === Random forest of 10 trees, each constructed while considering 3 random features. Out of bag error: 0.2885 Time taken to build model: 15.58 seconds === Evaluation on test split === === Summary === Correctly Classified Instances 3195 69.4263 % Incorrectly Classified Instances 1407 30.5737 % Kappa statistic 0.2927 Mean absolute error 0.2288 Root mean squared error 0.3957 Relative absolute error 79.5278 % Root relative squared error 104.6955 % Coverage of cases (0.95 level) 91.2429 % Mean rel. region size (0.95 level) 54.0055 % Total Number of Instances 4602 === Detailed Accuracy By Class === TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.778 0.502 0.791 0.778 0.785 0.685 s 0.506 0.196 0.477 0.506 0.491

0.852 y Weighted Avg. 0.709 0.434 0.699 0.709 0.703 0.734 === Confusion Matrix === a b c <-- classified as 1898 392 23 | a = s 453 386 9 | b = u 69 3 27 | c = y	0.724 === Confusion Matrix === a b c <-- classified as 2163 149 1 | a = s 674 174 0 | b = u 97 1 1 | c = y	0.712 u 0.343 0.015 0.41 0.343 0.374 0.785 y Weighted Avg. 0.694 0.407 0.697 0.694 0.696 0.695 === Confusion Matrix === a b c <-- classified as 2540 661 62 | a = s 585 607 7 | b = u 87 5 48 | c = y
=== Run information === Scheme: weka.classifiers.trees.RandomTree -K 0 -M 1.0 -S 1 Relation: RenameNominalQualitativeMbQualitativeNom-weka.filters.unsupervised.attribute.Remove-R2-4 Instances: 9588 Attributes: 2 AreaCode Qualitative Test mode: split 66.0% train, remainder test === Classifier model (full training set) === RandomTree ========== === Evaluation on test split === === Summary === Correctly Classified Instances 2340	=== Run information === Scheme: weka.classifiers.trees.RandomTree -K 0 -M 1.0 -S 1 Relation: RenameNominalQualitativeMbQualitativeNom-weka.filters.unsupervised.attribute.Remove-R2-weka.filters.unsupervised.attribute.Remove-R1-2 Instances: 9588 Attributes: 2 Officer Qualitative Test mode: split 66.0% train, remainder test === Classifier model (full training set) === RandomTree ========== Size of the tree : 121	=== Run information === Scheme: weka.classifiers.trees.RandomTree -K 0 -M 1.0 -S 1 Relation: RenameNominalQualitativeMbQualitativeNom Instances: 9588 Attributes: 5 AreaCode SamDate SamMonth1 Officer Qualitative Test mode: split 66.0% train, remainder test === Classifier model (full training set) === RandomTree ==========

71.7791 %
Incorrectly Classified Instances 920
28.2209 %
Kappa statistic 0.2855
Mean absolute error 0.2202
Root mean squared error 0.3634
Relative absolute error 76.6784 %
Root relative squared error 96.2039 %
Coverage of cases (0.95 level) 94.4172 %
Mean rel. region size (0.95 level) 55.818 %
Total Number of Instances 3260

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class
0.85 0.595 0.777 0.85 0.812 0.722 s
0.412 0.138 0.512 0.412 0.457 0.753 u
0.263 0.008 0.51 0.263 0.347 0.783 y
Weighted Avg. 0.718 0.458 0.7 0.718 0.705 0.732

=== Confusion Matrix ===

a b c <-- classified as
1965 331 17 | a = s
491 349 8 | b = u
72 1 26 | c = y

Time taken to build model: 0.02 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances 2344 71.9018 %
Incorrectly Classified Instances 916 28.0982 %
Kappa statistic 0.1536
Mean absolute error 0.2446
Root mean squared error 0.3532
Relative absolute error 85.1606 %
Root relative squared error 93.4966 %
Coverage of cases (0.95 level) 98.5583 %
Mean rel. region size (0.95 level) 67.0654 %
Total Number of Instances 3260

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class
0.942 0.825 0.736 0.942 0.826 0.71 s
0.193 0.056 0.55 0.193 0.286 0.752 u
0.01 0 0.5 0.01 0.02 0.835 y
Weighted Avg. 0.719 0.6 0.681 0.719 0.661 0.725

=== Confusion Matrix ===

a b c <-- classified as
2179 133 1 | a = s
684 164 0 | b = u
97 1 1 | c = y

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances 2261 69.3558 %
Incorrectly Classified Instances 999 30.6442 %
Kappa statistic 0.29
Mean absolute error 0.2194
Root mean squared error 0.4375
Relative absolute error 76.4084 %
Root relative squared error 115.8097 %
Coverage of cases (0.95 level) 75.3681 %
Mean rel. region size (0.95 level) 40.3476 %
Total Number of Instances 3260

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class
0.78 0.502 0.791 0.78 0.785 0.644 s
0.501 0.197 0.472 0.501 0.486 0.658 u
0.333 0.016 0.402 0.333 0.365 0.71 y
Weighted Avg. 0.694 0.408 0.697 0.694 0.695 0.649

=== Confusion Matrix ===

a b c <-- classified as
1803 470 40 | a = s
414 425 9 | b = u
61 5 33 | c = y