ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΜΑΤΙΚΗΣ
ΧΑΡΟΚΟΠΟΥ 89, 17671 ΑΘΗΝΑ -ΤΗΛ: 210-9549280, FAX: 210-9549281

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Μελέτη και σχεδίαση μεθόδων εξόρυξης γνώσης από Βιοιατρικά Δεδομένα»

Κωνσταντής Μαύρος

ΑΜ: 20723

Επιβλέπων: Βαρλάμης Ηρακλής

Αθήνα, Αύγουστος 2012

ΠΕΡΙΛΗΨΗ

Η μελέτη Βιοιατρικών δεδομένων μέσα από τις μεθόδους της εξόρυξης γνώσης αποτελεί ένα πολύ ενδιαφέρον αντικείμενο έρευνας. Επίσης, η εξόρυξη γνώσης από δεδομένα ογκολογίας είναι ένα αντικείμενο έρευνας-πρόκληση για τους ερευνητές. Στη παρούσα πτυχιακή εργασία παρουσιάζονται διάφορες τεχνικές εξόρυξης γνώσης όπως, Δέντρα Απόφασης, η Στατιστική Παλινδρόμηση, οι Ensemble Μέθοδοι και τα Support Vector Machines, για την πρόβλεψη της βιωσιμότητας ασθενών που έχουν προσβληθεί από το καρκίνο του μαστού. Τα δεδομένα που χρησιμοποιούνται παρέχονται από το SEER και είναι για τις χρονολογίες 1973-2008. Τα δεδομένα αυτά υποβλήθηκαν σε διάφορες τεχνικές προεπεξεργασίας έτσι ώστε να αποδόσουν τα καλύτερα δυνατά αποτελέσματα. Επιπλέον χρησιμοποιήθηκαν αυτοματοποιημένες τεχνικές επιλογής γνωρισμάτων για τη δημιουργία του τελικού συνόλου των δεδομένων με σκοπό την επιλογή των καταλληλότερων γνωρισμάτων..

ABSTRACT

The research on Biomedical Data through Data Mining methods is a very interesting research field. Furthermore, Data Mining on oncological data has been a challenging task for many reserachers. In this research paper, various Data Mining techniques are presented such as, Decision Trees, Logistic Regression, Ensemble Methods and Suport Vector Machines, to predict the breast cancer survaivability. Tha Dataset used in this research paper has been provided from SEER (Surveillance, Epidemiology and End Results) and refers to breast cancer cases from 1973 to 2008. The Dataset preprocessed with various techiques in order to generate the best possible results. Moreover, automated attribute selection techniques have been applied to the Dataset in order to use the best features of the dataset.

ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά, θα ήθελα να εκφράσω τις ευχαριστίες μου σε όλους όσους συνέβαλαν στην ολοκλήρωση της πτυχιακής εργασίας μου και γενικότερα των σπουδών μου στο Τμήμα Πληροφορικής και Τηλεματικής του Χαροκόπειου Πανεπιστημίου.

Ιδιαίτερα θα ήθελα να ευχαριστήσω των επιβλέποντα καθηγητή της πτυχιακής μου εργασίας κ. Βαρλάμη Ηρακλή για την υπομονή του, την άμεση βοήθεια και τη στήριξη που μου προσέφερε καθόλη τη διάρκεια της εκπόνησης της εργασίας μου.

Τέλος, θα ήθελα να ευχαριστήσω ιδιαίτερα την οικογένεια μου για όλη τη κατανόηση και την υποστήριξη που μου πρόσφεραν σε όλη τη διάρκεια των σπουδών μου, διότι χωρίς αυτή θα ήταν αδύνατο να τις πραγματοποιήσω.

ΠΕΡΙΕΧΟΜΕΝΑ

1. Εισαγωγή

Τα τελευταία χρόνια, κατα την επιχειρησιακή και την επιστημονική έρευνα έχει παρατηρηθεί μια εντυπωσιακή αύξηση του όγκου των δεδομένων που συλλέγονται για την πραγματοποίηση των ερευνών. Πολυεθνικές εταιρίες, όπως για παράδειγμα οι μεγάλες αλυσίδες πολυκαταστημάτων, χρησιμοποιούν terabytes δεδομένων, από τις αγορές που πραγματοποιούν οι πελάτες τους. Επιπλέον, διάφορα επιστημονικά πεδία όπως για παράδειγμα η Ιατρική και η Αστρονομία συλλέγουν δεδομένα από βάσεις δεδομένων, που έχουν τεράστιο μέγεθος. Όλα αυτά τα δεδομένα συλλέγονται με τρομερά υψηλές συχνότητες, της τάξης των GB/Hour.

Ωστόσο όλος αυτός ο όγκος των δεδομένων είναι άχρηστος εκτός και αν αυτά τα δεδομένα επεξεργαστούν κατάλληλα. Αυτή επεξεργασία και η ανάλυση είναι αρκετά δύσκολη λόγω του τεράστιου μεγέθους των δεδομένων. Επίσης οι σχέσεις μεταξύ των δεδομένων είναι εξαιρετικά πολύπλοκες και ανιχνεύονται πολύ δύσκολα.

Αυτό είχε ώς αποτέλεσμα τη δημιουργία ενός νέου επιστημονικού-ερευνητικού πεδίου, το οποίο θα μπορεί να επεξεργάζεται και να εξάγει συμπεράσματα με βάση αυτόν το μεγάλο όγκο των δεδομένων. Αυτό το πεδίο ονομάζεται Εξόρυξη Γνώσης από Δεδομένα (Data Mining).

Όσο η τεχνολογία θα εξελίσεται, όλο και περισσότεροι οργανισμοί, εταιρίες και επιστήμες, θα έχουν την ανάγκη να χρησιμοποιήσουν την Εξόρυξη Γνώσης καθώς ο όγκος των δεδομένων που θα διαχειρίζονται για την εξαγωγή συμπερασμάτων ολοένα και θα αυξάνεται.

1.1 Εξόρυξη Γνώσης απο Β.Δ. και Εξόρυξη Δεδομένων (Data Mining)

Η Εξόρυξη Δεδομένων είναι μια διαδικασία επιλογής, διερεύνησης και μοντελοποίησης μεγάλου όγκου δεδομένων με σκοπό τη εξαγωγή συμπερασμάτων

και συσχετίσεων μεταξύ τους, έτσι ώστε να προκύψει ένα καθαρό και συγχρόνως χρήσιμο αποτέλεσμα για τον αρμόδιο αναλυτή δεδομένων κάθε φορά σύμφωνα με τους Bellazi & Zupan (2008) και Giudici (2003).

Η επινόηση της Εξόρυξης Δεδομένων τοποθετείται περίπου στα μέσα του 1990 και σήμερα η έννοιά της έχει γίνει συνώνυμη με την έννοια της «Εξόρυξης Γνώσης από Βάσεις Δεδομένων» (Knowledge Discovery In Databases - KDD) η οποία σύμφωνα με τους Fayyad et al. (1996) και Bellazi & Zupan (2008), τονίζει περισσότερο τη διαδικασία ανάλυσης των δεδομένων παρά τις συγκεκριμένες μεθόδους ανάλυσης των δεδομένων. Η KDD είναι μια διεργασία η οποία αποτελείται από 5 στάδια (βλ. Εικόνα 1.1) ένα από τα οποία είναι και η εξόρυξη δεδομένων. Ενδιάμεσα σε αυτά τα 5 στάδια παργάγονται συγκεκριμένα προιόντα το οποία χρησιμοποιούνται για την πραγματοποίηση επόμενων σταδίων :

Αρχικά πρέπει να κατανοηθεί και να αξιοποιηθεί η αρχική γνώση και να αναγνώριστούν οι στόχοι που πρέπει να τεθούν.

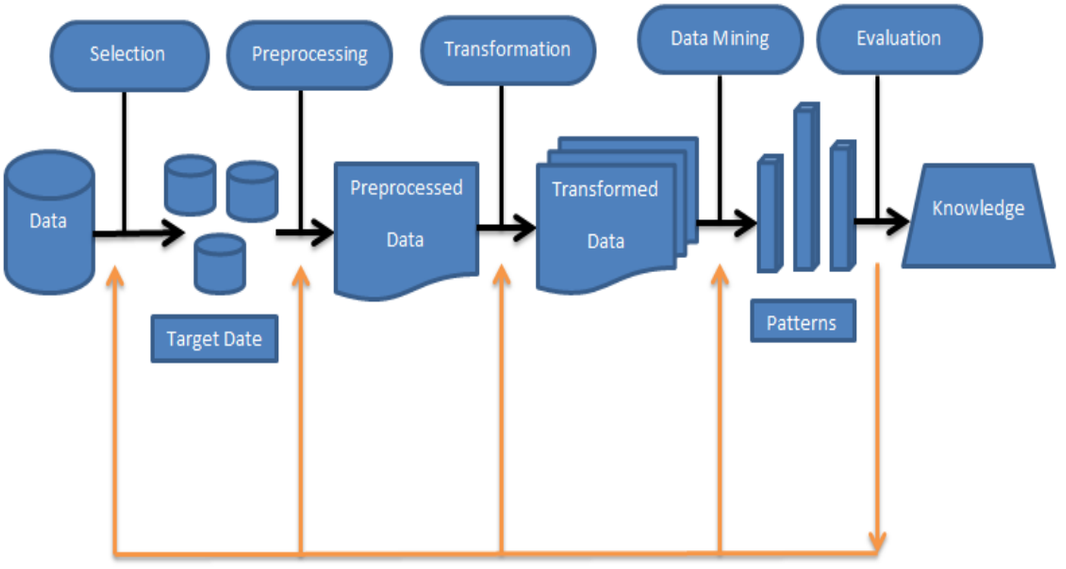
1. Στο πρώτο στάδιο πρέπει συγκεντρωθεί και να διαχωριστεί ένα συγκεκριμένο σύνολο δεδομένων πάνω στο οποίο θα πραγματοποιηθεί η εξόρυξη.
2. Στο δεύτερο στάδιο πραγματοποιείται ο καθαρισμός και η προεπεξεργασία των δεδομένων που έχουν επιλεγεί στο Στάδιο 1.
3. Στο τρίτο στάδιο πραγματοποιείται η μετατροπή των προηγούμενων δεδομένων με διάφορες τεχνικές μέσα από συγκεκριμένα προγράμματα για κάποιο σκοπό, όπως τη μείωση του μεγέθους του data set.

   Έπειτα από το τρίτο στάδιο πραγματοποιείται η συσχέτιση των στόχων που έχουν τεθεί στο πρώτο στάδιο, με μια συγκεριμένη μέθοδο εξόρυξης δεδομένων. Για παράδειγμα, κατηγοριοποίηση ή συσταδοποίηση (classification or clustering).

   Πρίν το τέταρτο στάδιο, την πραγματοποίηση της εξόρυξης γνώσης, επιλέγεται ο αλγόριθμος εξόρυξης γνώσης και η μέθοδος που θα χρησιμοποιηθεί για την αναζήτηση προτύπων δεδομένων. Επίσης ρυθμίζονται οι παράμετροι που πρέπει να χρησιμοποιηθούν.
4. Στο τέταρτο στάδιο πραγματοποιείται η εξόρυξη των δεδομένων.

Ενδιάμεσα στο τέταρτο και στο πέμπτο στάδιο γίνεται η μεταγλώττιση όλης της πληροφορίας που έχει εξαχθεί έπειτα από την πραγματοποίηση όλων των προηγούμενων σταδίων. Σε αυτό το στάδιο μπορεί να πραγματοποιηθεί και απεικόνιση των αποτελεσμάτων.

5. Στο πέμπτο και τελευταίο στάδιο προκύπτει η γνώση και αξιολογείται. Έπειτα η γνώση αυτή μπορεί να χρησιμοποιηθεί κατευθείαν για την επίλυση ενός ζητήματος (Fayyad et al., 1996)



Εικόνα 1.1. Διαδικασία εξόρυξης γνώσης από δεδομένα.

Οι τεχνικές Εξόρυξης Δεδομένων (Data Mining) είναι ευρύτατα διαδεδομένες σήμερα και εφαρμόζονται σε διάφορα ζητήματα εταιριών, επιστημονικά και ερευνητικά ζητήματα,όπως για παράδειγμα στην Ιατρική,ακόμη και σε κυβερνητικά ζητήματα και πιστεύεται πώς η εξόρυξη γνώσης από δεδομένα θα έχει σημαντική θετική επιρροή στη κοινωνία μας σύμφωνα με τους Chakrabarti et al. 2004. Η εξόρυξη δεδομένων αποτελεί ένα διεπιστημονικό πεδίο το οποίο έχει τις «ρίζες» του στη στατιστική, την τεχνητή νοημοσύνη (AI), τη μηχανική μάθηση και τις βάσεις δεδομένων (βλ. Εικόνα 1.2).

Εικόνα 1.2 Οι βάσεις της εξόρυξης δεδομένων.

Οι εργασίες που πραγματοποιούνται κατά την Εξόρυξη Δεδομένων χωρίζονται σε εργασίες για περιγραφή και πρόβλεψη. Η πρόβλεψη προϋποθέτει τη χρησιμοποίηση διαφόρων γνωστών μεταβλητών για την εκτίμηση μελλοντικών άγνωστων τιμών και η περιγραφή αφορά τη δημιουργία κατανοητών για τον άνθρωπο μοντέλων που θα περιγράφουν τα δεδομένα (Bellazi & Zupan, 2008). Γενικά η πρόβλεψη με την περιγραφή δεν έχουν μεγάλες διαφορές και οι στόχοι τους υλοποιούνται με διάφορες μεθόδους εξόρυξης δεδομένων. Οι κυριότερες μέθοδοι είναι οι εξής:

- Η κατηγοριοποίηση (classification) εκπαιδεύει μία συνάρτηση, η οποία κατηγοριοποιεί κάποια δεδομένα σε μια από δίαφορες κλάσεις που δημιουργούνται (Fayyad et al., 1996), (Hand, 1981), (Weiss & Kulikowski, 1991). Παραδείγματα μεθόδων κατηγοριοποίησης συναντόνται στη πρόγνωση μέσα από ιατρικά δεδομένα ,για τις τάσεις της οικονομίας κτλ.
- Η παλινδρόμηση (regression) εκπαιδεύει μια συνάρτηση η οποία αντιστοιχίζει κάποια δεδομένα σε μεταβλητές πρόβλεψης πραγματικών τιμών (Fayyad et al., 1996).
- Η συσταδοποίηση (clustering) είναι μια κοινή περιγραφική μέθοδος με την οποία αναζητούνται συστάδες (clusters), για τη περιγραφή των δεδομένων, έτσι ώστε τα σημεία της συστάδας να είναι όσο πιο όμοια μεταξύ τους και τα σημεία σε διαφορετικές συστάδες να είναι όσο το δυνατό λιγότερο όμοια

μεταξύ τους. Με τη συσταδοποίηση γίνεται κατανόηση του διαχωρισμού των δεδομένων και εξάγονται συμπεράσματα για τη κατανομή.

Όπως έχει αναφερθεί η Εξόρυξη Δεδομένων βρίσκει πολλές εφαρμογές τα τελευταία χρόνια σε διάφορους τομείς της κοινωνίας. Ένας πολύ σημαντικός τομέας που έχει άμεση εφαρμογή το Data Mining είναι και η Ιατρική καθώς εκεί βρίσκεται συγκεντρωμένος τεράστιος όγκος δεδομένων. Στις επόμενες ενότητες εστιάζουμε σε αυτό το κομμάτι της εξόρυξης δεδομένων.

1.2 Εξόρυξη Γνώσης από Ιατρικά δεδομένα

Ο ρόλος της πληροφορικής έχει εδραιωθεί πλέον στα περισσότερα συστήματα υγείας – ιατρικής περίθαλψης ανα το κόσμο. Η χρησιμοποίηση ηλεκτρονικών υπολογιστών στα περισσότερα νοσοκομεία ,αλλά και σε υπόλοιπους οργανισμούς που έχουν σχέση με την ιατρική περίθαλψη των ανθρώπων έδωσε τη δυνατότητα να αποθηκευτεί μεγάλος όγκος ιατρικών δεδομένων και να υπάρχει εύκολη πρόσβαση σε αυτά (medical databases) (Bhatnagar et al. 2006). Τα δεδομένα αυτά, που αποθηκεύονται πλέον σε ψηφιακή μορφή, αφορούν εγγραφές ασθενών στο αρχείο, τις ασθένειες του κάθε ανθρώπου, τα φάρμακα που το χορηγούνται, τη θεραπεία που έχει ακολουθηθεί, δημογραφικά στοιχεία κτλ.

Ωστόσο, όλος αυτός ο όγκος των ιατρικών δεδομένων παρότι είναι πολύ χρήσιμος, παρουσιάζει δυσκολίες η μελέτη τους έτσι ώστε να εξαχθεί κάποια χρήσιμη πληροφορία για να πάρουμε μια απόφαση. Με τις υπάρχουσες μεθόδους ανάλυσης είναι εξαιρετικά δύσκολο να υπάρξει κάποια γνώση, γι’ αυτό και είναι αναγκαία η υλοποίηση μεθόδων μέσα από υπολογιστικά συστήματα για να πραγματοποιηθεί μία σωστή ανάλυση των δεδομένων (Lavrac, 1999).

Αυτό, έχει σαν αποτέλεσμα να είναι προτιμότερη η χρήση των τεχνικών εξόρυξης γνώσης για τα ιατρικά δεδομένα. Η εξόρυξη από ιατρικά δεδομένα είναι ένα από τα πιο ενδιαφέροντα και δύσκολα πεδία της εξόρυξης γνώσης. Τα δεδομένα που χρειάζονται για την εφαρμογή της εξόρυξης είναι τόσα πολλά, περίπλοκα και

ετερογεννή που καθιστούν την εξόρυξη μία πρόκληση για κάθε αναλυτή (Delen, 2009).

Διάφορες τεχνίκες χρησιμοποιούνται για την λήψη ιατρικών αποφάσεων και πιο συγκεκριμένα για διαγνώσεις και προγνώσεις καρκίνου. Οι επεξηγηματικές και οι επαναληπτικές μέθοδοι είναι οι τεχνικές που χρησιμοποιούνται πιο συχνά κατά τη διαδικασία της εξόρυξης γνώσης από ιατρικά δεδομένα. Οι πιο σημαντικές δυσκολίες που συναντούν οι ερευνητές στο τομέα της εξόρυξης γνώσης στο τομέα της ιατρικής είναι (Delen, 2009):

1. **Η ετερογένεια των ιατρικών δεδομένων**. Τα ιατρικά δεδομένα χωρίς να έχουν υποστεί κάποια προ-επεξεργασία είναι ογκώδη και ετερογενή. Τα δεδομένα αυτά αποθηκεύονται ύστερα από εξετάσεις του ασθενούς, από εικόνες (π.χ. ακτινογραφίες) και εργαστηριακά δεδομένα. Ο συνδυασμός αυτών των δεδομένων μπορεί να είναι αναγκαίος για την πρόγνωση, τη διάγνωση και τη περίθαλψη ενός ασθενούς, παρότι είναι τελείως διαφορετικά μεταξύ τους. Γι' αυτό το λόγο δεν μπορούν να αγνοηθούν (Cios & Moore, 2002). Τα κυριότερα προβλήματα που σχετίζονται με την ετερογένεια των δεδομένων είναι:
    - Το μέγεθος και η περιπλοκότητα των δεδομένων.
    - Η ερμηνεία του κάθε ιατρού
    - Οι μετρικές όπως το sensitivity και το specificity που μετρούν την πιθανότητα επιτυχίας ή λάθους.
    - Η δυσκολία να χαρακτηριστούν μαθηματικά αυτά τα δεδομένα.
    - Η κανονικοποίηση των δεδομένων. Είναι μια μέθοδος μετασχηματισμού των δεδομένων, όπου οι συνεχείς μεταβλητές διαμορφώνονται έτσι ώστε να βρίσκονται στο διάστημα [0,1]. Μια τεχνική κανονικοποίησης είναι η min-max normalization.

2. **Τα ηθικά, νομικά και κοινωνικά ζητήματα (Cios & Moore, 2002).** Τα ιατρικά δεδομένα που συλλέγονται στις βάσεις δεδομένων αφορούν ανθρώπινα θέματα υγείας γι' αυτό το λόγο υπάρχει ένα μεγάλο ηθικό και νομικό πλαίσιο έτσι ώστε να καλύπτει τη προσβολή του ασθενούς ή την άσκοπη χρήση των δεδομένων του. Τα κύρια σημεία αυτών των ζητημάτων μπορούν να χωρισθούν σε 5 κατηγορίες :

- Η ιδιοκτησία των δεδομένων.
- Οι νόμοι που διέπουν αυτά τα δεδομένα.
- Η ιδιωτικότητα και η ασφάλεια των προσωπικών δεδομένων.
- Τα αναμενόμενα προνόμια.
- Τα ζητήματα διαχείρισης των δεδομένων.

Όπως έχει αναφερθεί τα ιατρικά δεδομένα βρίσκουν άμεση εφαρμογή στην εξόρυξη γνώσης λόγω του μεγάλου όγκου τους και τις πολυπλοκότητάς τους. Ωστόσο ένας τομέας της ιατρικής που παρουσιάζει ιδιαίτερο ερευνητικό ενδιαφέρων σε σχέση με την εξόρυξη δεδομένων είναι η Ογκολογία. Σε αυτό το τομέα έχει παρατηρηθεί μια αρκετά μεγάλη αύξηση στο ερευνητικό ενδιαφέρον σύμφωνα με τους Delen et al. (2004) διότι είναι ένα δύσκολο ζήτημα για ανάλυση. Υπάρχουν πάρα πολλά είδη καρκίνου, που εμφανίζονται σε διάφορα μέρη του ανθρώπινου σώματος σε κάθε ηλικία.

Επίσης υπάρχουν μικρές λεπτομέρειες που πρέπει να λάβει υπόψη του ένας ερευνητής όπως για παράδειγμα το μέγεθος του όγκου ή το αν έχει ακολουθήσει θεραπεία με χημειοθεραπείες ο ασθενής είτε η διάρκεια ζωής του ασθενούς. Ακόμη είναι αναγκαίο να υπάρχει διακριτικότητα αλλά και σεβασμός των ερευνητών προς τους ασθενείς λόγω των προσωπικών δεδομένων. Επομένως σε αυτό το πεδίο ανάλυσης υπάρχει και πρόβλημα ετερογένειας και ζητήματα ήθους, σεβασμού και νομιμότητας. Ωστόσο το θέμα της εξόρυξης γνώσης από δεδομένα ογκολογίας θα αναλυθεί περισσότερο στα επόμενα κεφάλαια καθώς αποτελεί το κεντρικό θέμα της εργασίας.

1.3 Σκοπός της διπλωματικής εργασίας

Ο καρκίνος του μαστού είναι μια από τις πρώτες αιτίες θανάτου ανά το κόσμο (World Health Organization, 2010) και η δεύτερη αιτία θανάτου για τις γυναίκες στην Αμερική (National Cancer Institute, 2010). Σκοπός της παρούσας διπλωματικής εργασίας είναι η πρόγνωση της δυνατότητας επιβίωσης ασθενών που έχουν προσβληθεί από τον καρκίνο του μαστού μέσω των τεχνικών της εξόρυξης

δεδομένων (data mining) και η εφαρμογή νέων τεχνικών πάνω σε αυτά τα δεδομένα με σκοπό την καλύτερη ανάλυση. Το ζήτημα της διπλωματικής εργασίας αντιμετωπίστηκε με τη μέθοδο της κατηγοριοποίησης (classification) δοκιμάζοντας διάφορους αλγορίθμους, που προσφέρει το «ελεύθερο» λογισμικό Weka, όπως ο C4.5 (j48), η Στατιστική Παλινδρόμηση (Logistic Regression), τα Διανύσματα Υποστήριξης (SVM - support vector machines), καθώς και η συνδυαστική μέθοδος (ensemble) Vote. Για τις δοκιμές χρησιμοποιήθηκαν δεδομένα από το SEER (Surveillance, Epidemiology and End Results[1]) το οποίο παρέχει πληροφορίες και και στατιστικά για όλες τις περιπτώσεις καρκίνου στην Αμερική.

[1] SEER(Surveillance , Epidemiology and End Results) : http://seer.cancer.gov/

2. Υπόβαθρο

2.1 Γενικά για τον καρκίνο του μαστού

Σύμφωνα με το American Cancer Society (2012) ο καρκίνος του μαστού είναι ο πιό συνήθης τύπος καρκίνου για τις γυναίκες στις ΗΠΑ. Η πιθανότητα να προσβληθεί μια γυναίκα, κατά τη διάρκεια της ζωής της, από καρκίνο του μαστού είναι 1/8. Υπολογίζεται ότι προέκυψαν 226.870 νέες περιπτώσεις γυναικών πού προσβλήθηκαν από το συγκεκριμένο τύπο καρκίνου για το έτος 2012 σύμφωνα με το American Cancer Society. Ακόμη τα θανάσιμα κρούσματα του καρκίνου του μαστού υπολογίζονται στις 39,510. Ο συγκεκριμένος τύπος καρκίνου εμφανίζεται και σε άντρες αλλά σε πολύ μικρότερο ποσοστό. Για το 2012, εμφανίστηκαν 2,140 κρούσματα καρκίνου του μαστού σε άντρες, δηλαδή το 1% όλων των κρουσμάτων (American Cancer Society, 2012).

Ο καρκίνος του μαστού είναι ένας κοκοήθης όγκος ο οποίος δημιουργείται από διάφορα κύτταρα του μαστού τα οποία δεν αναπτύσσονται φυσιολογικά (Delen et al., 2004, Aragones et al., 2003). Οι επιστήμονες έχουν εντοπίσει διάφορες αιτίες που μπορεί να αυξήσουν τις πιθανότητες για δημιουργία το καρκίνο του μαστού οι οποίες είναι (Aragones et al., 2003):

- Η ηλικία
- Η προιστορία άλλων μελών της οικογένειας σε σχέση με αυτό το καρκίνο
- Η παχυσαρκία
- Οι γυναίκες που δεν έχουν παιδιά.

Σύμφωνα με το Centers for Disease Control and Prevention (CDC), (2008) , το 5-10% των καρκίνων οφείλονται σε κάποιες ανωμαλίες που «κληρονομούν» οι άνθρωποι από τους γονείς τους. Επίσης το 90% των περιπτώσεων του καρκίνου του μαστού έχει παρατηρηθεί ότι εμφανίζεται λόγω γενετικών ανωμαλιών που παρατηρούνται με την αύξηση της ηλικίας των ανθρώπων. Οι τρόποι αντιμετώπισης του καρκίνου του μαστού διακρίνονται σε τοπικές και συστηματικές. Παραδείγματα τοπικών μεθόδων

είναι η χειρουργική επέμβαση και η ακτινοβολίες και οι συστηματικές μέθοδοι είναι οι χημειοθεραπείες και οι ορμονοθεραπείες (Delen et al., 2004). Η ανάλυση όλων των στοιχείων του συνόλου των δεδομένων απόκτα ιδιαίτερο ενδιαφέρων με τη τεχνική του data mining.

2.2 Τεχνικές κατηγοριοποίησης σε δεδομένα καρκίνου του μαστού

Όπως έχει αναφερθεί η εξόρυξη γνώσης από δεδομένα ογκολογίας αποτελεί ένα δύσκολο αλλά παράλληλα και ενδιαφέρον κομμάτι έρευνας λόγω του τεράστιου όγκου των δεδομένων και των πολλών χαρακτηριστικών που πρέπει να ληφθούν υπόψη κατά τη διάρκεια της έρευνας.

Κατά αυτόν το τρόπο η εξόρυξη γνώσης από δεδομένα καρκίνου του μαστού είναι διαδεδομένη λόγω της συχνότητας εμφάνισης του συγκεκριμένου τύπου καρκίνου (μεγάλος αριθμός κρουσμάτων άρα και μεγάλο dataset) και των αρκετών χαρακτηριστικών που πρέπει να υποστούν μια προεπεξεργασία για την επιλογή των καταλληλότερων. Τα δεδομένα για τη παρούσα πτυχιακή εργασία παρέχονται από το SEER όπως έχει αναφερθεί, και αφορούν τις περιπτώσεις καρκίνου του μαστού από το 1973 έως το 2008.

Το SEER (Surveillance, Epidemiology and End Results[2]) είναι ένα πρόγραμμα του National Cancer Institute[3] που λειτουργεί για να παρέχει πληροφορίες και στατιστικά στοιχεία για όλους τους τύπους καρκίνου, με σκοπό τη μείωση της θνησιμότητας από τις διάφορες μορφές καρκίνου στις Η.Π.Α. Το SEER συλλέγει δεδομένα από περιπτώσεις καρκίνου από διάφορες πηγές και τοποθεσίες σε όλες τις Η.Π.Α. Η καταγραφή των δεδομένων στο αρχείο του SEER έχει αρχίσει από το 1973 με ένα μικρό αριθμό εγγραφών και συνεχίζεται μέχρι και σήμερα σύμφωνα με το National Cancer Institute δημιουργώντας μια μεγάλη βάση δεδομένων με πολλά στοιχεία. Όλα

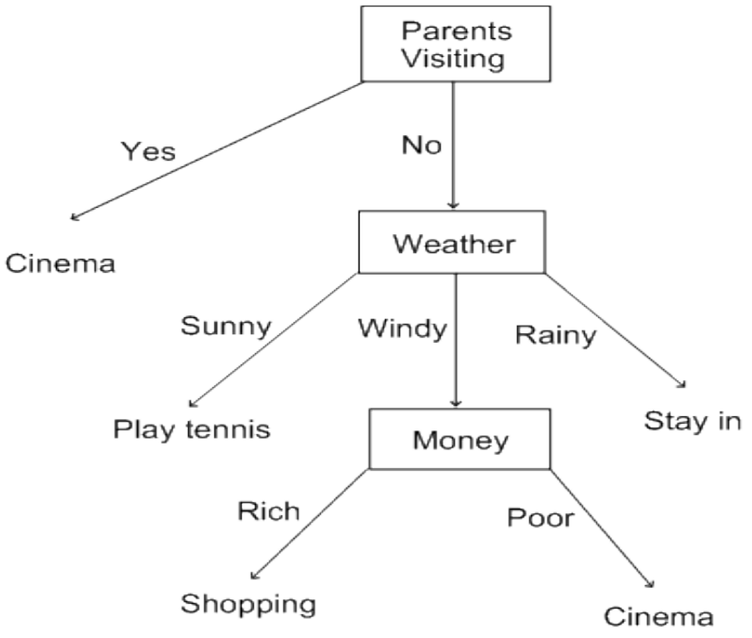
[2] SEER(Surveillance , Epidemiology and End Results) : http://seer.cancer.gov/
[3] National Cancer Institute : http://www.cancer.gov/

αυτά τα αρχεία του SEER έχουν ως σκοπό την έρευνα σε διάφορους τομείς όπως για τη διάγνωση, την πρόληψη ακόμη και την βιωσιμότητα των ασθενών που έχουν προσβληθεί από καρκίνο ένα θέμα το οποίο εξετάζεται στη συγκεκριμένη πτυχιακή εργασία.

Για την εξόρυξη γνώσης από τα δεδομένα του μαστού του SEER στη παρούσα πτυχιακή εργασία, χρησιμοποιήθηκε η μέθοδος της κατηγοριοποίησης (classification). Η μέθοδος αυτή έχει χρησιμοποιηθεί στης περισσότερες έρευνες της υπάρχουσας βιβλιογραφίας σχετικά με την εξόρυξης γνώσης από δεδομένα ογκολογίας. Συγκεκριμένα έχουν χρησιμοποιηθεί οι παρακάτω τεχνικές :

- Δέντρα απόφασης (Decision Trees)
- Μέθοδοι Παλινδρόμησης
- Support Vector Machines (SVM)
- Ensemble Μέθοδοι.

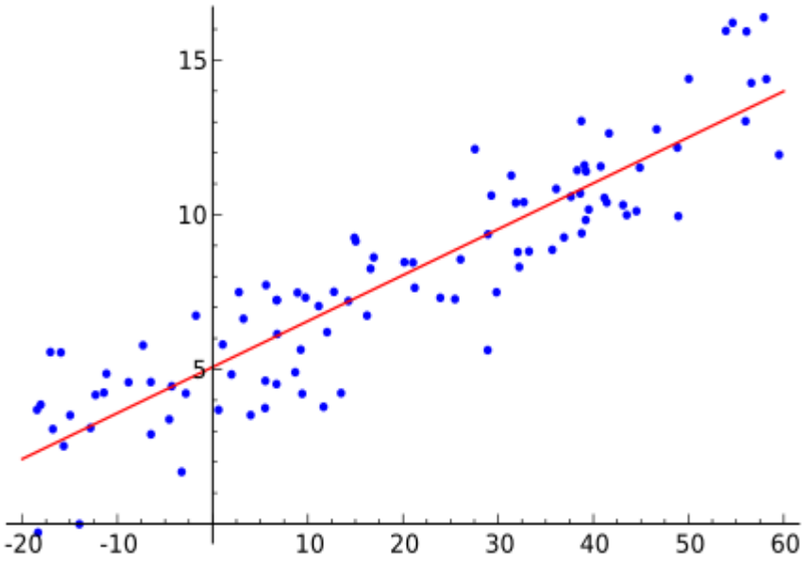
Τα **δέντρα απόφασης (Decision Trees)** (βλ. Εικόνα 2.1) είναι μία από τις πιο γνωστές τεχνικές κατηγοριοποίησης η οποία χρησιμοποιείται ευρέως στην εξόρυξη γνώσης. Κατά την εφαρμογή ενός δέντρου απόφασης κατασκευάζεται ένα δέντρο του οποίου τα φύλλα αναπαριστούν την κατηγοριοποίηση, όπου κάθε φύλλο είναι και μία κλάση και οι διακλαδώσεις αναπαριστούν τους διαχωρισμούς που πραγματοποιούνται κάθε φορά για να γίνει η κατηγοριοποίηση (Cha & Tappert, 2009). Κάθε εσωτερικός κόμβος υποδηλώνει την επαλήθευση ενός ή περισσοτέρων γνωρισμάτων ενός dataset με δευτερεύων δέντρο απόφασης για κάθε δυνατό αποτέλεσμα του τέστ που πραγματοποιείται (Quinlan, 1998). Η συγκεκριμένη μέθοδος χρησιμοποιεί κάποια δεδομένα εκπαίδευσης από περιπτώσεις με τις οποίες το δέντρο απόφασης δημιουργήθηκε, αρχικά για γενίκευση και αξιολόγηση της αξιοπιστίας των κανόνων που εξάγονται από το δέντρο απόφασης και κατά δεύτερο λόγο για να βελτιώσει τη συλλογή των κανόνων σε έναν ολοκληρωμένο κανόνα ο οποίος θα δώσει μια λύση (Quinlan, 1998). Οι λύσεις και οι κανόνες που θα προκύψουν από ένα δέντρο



Εικόνα 2.1 Ένα απλό Δέντρο Απόφασης(Πηγή : http://www.doc.ic.ac.uk/).

απόφασης είναι εύκολα κατανοήτες ακόμη και από έναν που δεν είναι ειδικός στο τομέα αυτό, έτσι οι τεχνική των δέντρων απόφασης είναι αρκετά διαδεδομένη στο τομέα της εξόρυξης γνώσης (Apte & Weiss, 1997). Υπάρχουν πολλοί αλγόριθμοι για τα δέντρα απόφασης με τους πιο διαδεδομένους να είναι ο C4.5 και ο ID3. Στη παρούσα πτυχιακή εργασία χρησιμοποιείται ο αλγόριθμος J48 μια εξέλιξη του C4.5.

Οι **Μέθοδοι Παλινδρόμησης** είναι από τα βασικά εργαλεία της στατιστικής. Έχουν ως σκοπό τη δημιουργία μοντέλων πρόβλεψης και σχετίζουν τη τιμή μιας εξαρτημένης συνεχούς μεταβλητής με τις τιμές από μία ομάδα ανεξάρτητων μεταβλητών (Dzeroski et al., 2004). Η ανάλυση με τις μεθόδους παλινδρόμησης χρησιμοποιείται σε ευρεία κλίμακα για προβλέψεις και η χρήση τους έχει άμεση σχέση με τη μηχανική μάθηση. Ακόμη η χρήση μεθόδων παλινδρόμησης είναι αρκετά εύκολη ωστόσο είναι ταυτόχρονα και από τις μεθόδους που έχουν τις λιγότερες δυνατότητες (Abernethy, 2010). Στον τομέα της εξόρυξης γνώσης υπάρχουν διάφοροι αλγόριθμοι παλινδρόμησης με τους πιο διαδεδομένους να είναι ο Linear Regression (βλ. Εικόνα 2.2) , ο Logistic Regression και ο Simple Logistic.



Εικόνα 2.2. Μια γραμμική παλινδρόμηση με μία ανεξάρτητη μεταβλητή (Πηγή : http://en.wikipedia.org/wiki/Linear_regression)

Τα **Support Vector Machines (SVM)** είναι μια δημοφιλής μέθοδος μηχανικής μάθησης η οποία χρησιμοποιείται στην κατηγοριοποίηση και την παλινδρόμηση. Έχουν μελετηθεί εντατικά και έχουν αξιολογηθεί με διάφορες τεχνικές και μάλιστα έχει παρατηρηθεί ότι τα υπολογιστικά αποτελέσματα τους έχουν πλεονέκτημα απέναντι σε αυτά των υπόλοιπων ανταγωνιστικών αλγορίθμων. Τα SVM βασίζονται στο στοιχείο του Structural Risk Minimization (Vapnik, 1995). Η ιδέα για το Structural Risk Minimization είναι να υπάρχει μία υπόθεση h για την οποία μπορούμε να εγγυηθούμε το μικρότερο δυνατό σφάλμα (Joachims, 1998). Το πραγματικό λάθος της h είναι η πιθανότητα η h να είναι λανθασμένη σε ένα τυχαίο τεστ εκπαίδευσης. Τα SVM βρίσκουν την υπόθεση h, η οποία ελαχιστοποιεί ένα άνω όριο του πραγματικού λάθους μέσα από τον αποτελεσματικό έλεγχο της VC-διάστασης [4] της H (Joachims, 1998).

Τα SVM μπορούν να θεωρηθούν ως μέθοδοι για τη κατασκευή ενός ειδικού κανόνα, που ονομάζεται γραμμικός κατηγοριοποιητής (linear classifier) που κατά έναν τρόπο

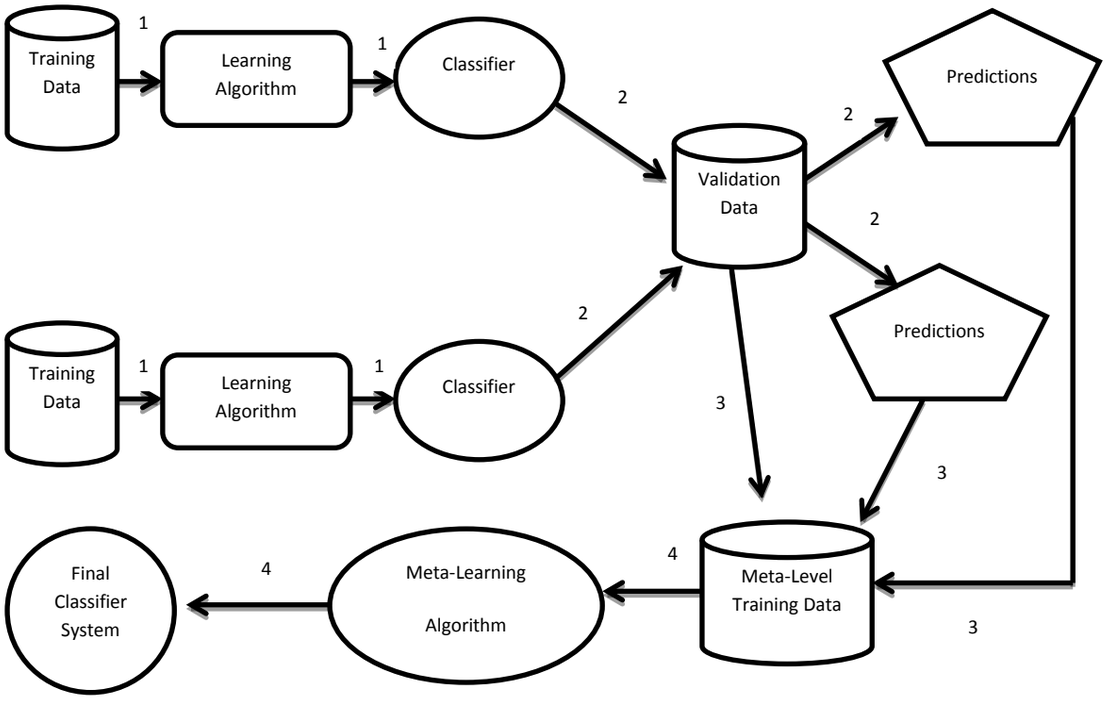
[4] VC-Dimension(Vapnik–Chervonenkis-Dimension)– Μετρα τη πολυπλοκότητα ενός αλγορίθμου κατηγοριοποίησης.

θα παράγει ταξινομητές που θα εγγυούνται θεωρητικά καλύτερα αποτελέσματα στις προγνώσεις (Fradkin & Muchnik, 2000).Παρόλο που τα SVM παράγουν γραμμικούς κανόνες, προσαρμόζονται εύκολα και με μια αλλαγή στις ιδιότητες τους, που ονομάζεται «Kernel-trick», μπορούν να παράγουν και μη γραμμικούς κανόνες (Fradkin & Muchnik, 2000).

Τα Support Vectors Machines χρησιμοποιούνται σε διάφορες ερευνητικές εργασίες εξόρυξης γνώσης όπως σε γονιδιακές αναλύσεις, στην ανάλυση της επιβιωσιμότητας ασθενών με καρκίνο, σε ταξινομητές κειμένων και σε πολλά άλλα. Στη συγκεκριμένη έρευνα χρησιμοποιήθηκε ο αλγόριθμος libSVM, που θα αναλυθεί στα επόμενα κεφάλαια.

Οι **Ensemble μέθοδοι**, τώρα, είναι αλγόριθμοι μάθησης που χρησιμοποιούν ένα σύνολο από επιμέρους ταξινομητές σε πρώτο επίπεδο. Στη συνέχεια χρησιμοποιούν τις παρεχόμενες πληροφορίες από κάθε ένα ταξινομητή και μια επιπλέον διαδικασία απόφασης (π.χ. μια μέθοδο voting ή έναν επιπλέον ταξινομητή) (Dietterich, 2000) για να ταξινομήσουν τα δεδομένα. Οι αλγόριθμοι των ensemble μεθόδων ανήκουν στη κατηγορία των meta-learning αλγορίθμων. Ως meta-learning ορίζεται ως η μάθηση από πρότερη γνώση (Chan & Stolfo, 1993). Αυτό γίνεται με την μάθηση των προγνώσεων των ταξινομητών που χρησιμοποιεί ένας meta-learning αλγόριθμος, από ένα συγκεκριμένο σύνολο δεδομένων. Τα στάδια μίας meta-learning διαδικασίας είναι τα παρακάτω (βλ. Εικόνα 2.3) (Chan et al., 2000):

1. Οι ταξινομητές εκπαιδεύονται με το αρχικό σετ δεδομένων εκπαίδευσης.
2. Οι προγνώσεις προκύπτουν από τους ταξινομητές μάθησης σε ένα ξεχωριστό σετ δεδομένων επαλήθευσης.
3. Δημιουργείται ένα meta-level σετ δεδομένων εκπαίδευσης το οποίο προκύπτει από το σετ επαλήθευσης (validation set) και τις προγνώσεις των ταξινομητών μάθησης.
4. Ο τελευταίος ταξινομητής (meta-classifier) εκπαιδευέται από το meta-level training set



Εικόνα 2.3. Τα στάδια του Meta-Learning

Σύμφωνα με θεωρητικές και εμπειρικές έρευνες μία καλή ensemble μέθοδος είναι αυτή όπου ο κάθε ταξινομητής, που συμμετέχει στον ensemble ταξινομητή, παρέχει μεγάλη ακρίβεια και που τα λάθη τους είναι σε διαφορετικά σημεία στο διάστημα που δίνεται ως είσοδος (Maclin & Opitz, 1999). Ο τελικός ταξινομήτης (ensemble classifier) έχει γενικά μεγαλύτερη ακρίβεια από τους ταξινομητές που αποτελείται. Υπάρχουν διάφοροι αλγόριθμοι για τις ensemble μεθόδους. Από τους πιο διαδεδομένους είναι ο Bagging, ο Boosting και ο Vote.

Όλες οι παραπάνω μέθοδοι κατηγοριοποίησης καθώς και η τεχνική της κατηγοριοποίησης έχουν χρησιμοποιηθεί σε πολλές έρευνες για την εξόρυξη γνώσης από δεδομένα ογκολογίας.

2.3 Υπάρχουσες έρευνες σε ογκολογικά δεδομένα

Διάφορες έρευνες έχουν γίνει κατα καιρούς στον τομέα της εξόρυξης γνώσης από δεδομένα ογκολογίας.

Οι Delen et al. (2004) παρουσίασαν μία έρευνα στο τομέα του Data Mining με σκοπό τη πρόγνωση της επιβιωσιμότητας ασθενών που έχουν προσβληθεί από το καρκίνο του μαστού. Στη συγκεκριμένη έρευνα χρησιμοποιήθηκαν δεδομένα από το SEER1 που αφορούσαν δεδομένα ασθενών με καρκίνο του μαστού στην Αμερική για την περίοδο 1973-2000.

Το αρχείο του SEER για ασθενείς με καρκίνο του μαστού περιείχε 433,272 εγγραφές και 72 μεταβλητές. Ωστόσο οι Delen et al. (2004) δημιούργησαν και μια ακόμη κατηγορική μεταβλητή STR (Survival Time Recode), με τιμές 0 ή 1, η οποία αντιπροσωπεύει τον αριθμό των ετών και των μηνών επιβίωσης ενός ασθενούς από τη στιγμή που διαγνώστηκε ο καρκίνος. Με βάση αυτή τη μεταβλήτη έκαναν μια προ-επεξεργασία των δεδομένων, καθώς ο τεράστιος όγκος των δεδομένων κρύβει παγίδες όπως οι πολλές ελλειπείς τιμές και οι μεταβλητές που δεν χρειάζονται για τη πραγματοποίηση της έρευνας. Έτσι μετά τη πραγματοποίηση της προεπεξεργασίας κατέληξαν σε ένα σύνολο από δεδομένα που αποτελούνταν από 202,932 εγγραφές και 17 γνωρίσματα. Οι Delen et al. (2004) χρησιμοποίησαν τρείς μεθόδους κατηγοριοποίησης για την εξόρυξη γνώσης: Τεχνητά Νευρωνικά Δίκτυα, Δέντρα Απόφασης και συγκεκριμένα τον αλγόριθμο C5 και τη Logistic Regression.

Οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση των αλγορίθμων είναι η ακρίβεια (accuracy), η ευαισθησία (sensitivity) και η ιδιαιτερότητα (specificity). Επίσης κατά την εκτέλεση των αλγορίθμων και την εκπαίδευση των δεδομένων χρησιμοποιήθηκε 10-fold cross validation προκειμένου να μειωθούν τα σφάλματα που σχετίζονται με τη τυχαια δειγματοληψία. Σύμφωνα με τα αποτελέσματα των Delen et al. (2004) ο αλγόριθμος C5 λειτούργησε καλύτερα σε σχέση με τους υπόλοιπους δύο πετυχαίνοντας accuracy= 0.9362. Δεύτερος καλύτερος αλγόριθμος αποδείχτηκε αυτός των Νευρωνικών Δικτύων με accuracy= 0.9121. Λιγότερο αποδοτικός φαίνεται να είναι ο αλγόριθμος της Logistic Regression με accuracy= 0.8920.

Ακόμη μία έρευνα που έγινε για την εξόρυξη γνώσης από δεδομένα καρκίνου του μαστού πραγματοποιήθηκε από τους Bellaachia & Guven (2006). Σκοπός της έρευνας είναι η ανάλυση της επιβιωσιμότητας των ασθενών με καρκίνο του μαστού μέσω των τεχνικών του data mining. Σε αυτή την έρευνα χρησιμοποιήθηκαν δεδομένα από το SEER που αφορούσαν δεδομένα ασθενών με καρκίνο του μαστού στην Αμερική για την περίοδο 1973-2002 δηλάδη το σύνολο των δεδομένων είναι πιό πρόσφατο και πιο μεγάλο σε σχέση με την έρευνα των Delen et al. (2004). Ο σύνολικός αριθμός των εγγραφών στο συγκεκριμένο αρχείο φτάνει τις 482,052.

Σε αντίθεση με τους Delen et al. (2004), οι Bellaachia & Guven (2006) για τη προέπεξεργασία των δεδομένων εκτός από τη κατηγορική μεταβλητή STR (Survival Time Recode), με τιμές 0 ή 1, πρόσθεσαν και άλλες δύο, την Cause Of Death (COD) που αντιπροσωπεύει την αιτία θανάτου ενός ασθενή και την Vital Status Recode (VSR), για το αν είναι εν ζωή ο ασθενής ή όχι. Έτσι, έπειτα από τη προεπεξεργασία των δεδομένων προέκυψαν 151,886 εγγραφές με 16 γνωρίσματα, αριθμός αρκετά μικρότερος σε σχέση με την έρευνα των Delen et al. (2004). Οι Bellaachia & Guven (2006) υποστηρίζουν πως στην έρευνα των Delen et al. (2004) υπάρχουν ασάφειες και πως οι μεταβλητές COD και VSR είναι απαραίτητες για την εξαγωγή της γνώσης. Οι μέθοδοι που επέλεξαν οι Bellaachia & Guven για την κατηγοριοποίηση είναι τα Δεντρα Απόφασης με τον αλγόριθμο C4.5 τα Τεχνητά Νευρωνικά Δίκτυα και τον αλγόριθμο Naïve Bayes. Για την αξιολόγηση των αποτελεσμάτων χρησιμοποιήθηκαν τρείς μετρικές το accuracy, το precision και το recall τα οποία προκύπτουν από τον πίνακα των True Positives (TP) και False Positives (FP).

Πιο αποδοτικός αλγόριθμος αποδείχθηκε ο C4.5 με accuracy= 86.7%, έπειτα ο αλγόριθμος των Νευρωνικών Δικτύων με accuracy= 86.5% και ο λιγότερο αποδοτικός ο Naives Bayes με accuracy= 84.5%. Σύμφωνα με τους Bellaachia & Guven (2006) οι αρκετά μεγάλες διαφόρες που υπάρχουν στα αποτελέσματα μεταξύ της δικής το υ έρευνας και σε αυτή των Delen et al.(2004), οφείλεται στο διαφορετικό αρχείο που χρησιμοποιήθηκε από το SEER και στην αγνόηση σημαντικών γνωρισμάτων όπως το COD και το VSR.

Ακόμη μια έρευνα που αφορά την εξόρυξη γνώσης από δεδομένα ασθενών με καρκίνο του μαστού πραγματοποιήθηκε από τους Rajesh & Anand, (2012). Η

συγκεκριμένη έρευνα είχε ώς σκοπό την την διάγνωση του καρκίνου του μαστού ενός ασθενούς με την τεχνική της κατηγοριοποίησης. Τα δεδομένα που χρησιμοποιήθηκαν προέρχονται από το αρχείο του SEER για τις χρονολογίες 1973-2008. Το αρχείο περιέχει 630,218 εγγραφές και 124 γνωρίσματα-μεταβλητές, δηλαδή αριθμός πολύ μεγαλύτερος από τα δεδομένα που χρησιμοποιήθηκαν στις έρευνες των Delen et al.(2004) και Bellaachia & Guven (2006). Ωστόσο χρησιμοποιήθηκαν μόνο 1430 εγγραφές.

Κατά τη φάση της προεπεξεργασίας των δεδομένων οι Rajesh & Anand, (2012) αφαίρεσαν γνωρίσματα τα οποία σχετίζονται με δημογραφικά στοιχεία, γνωρίσματα των οποίων οι ελλιπείς τιμές ξεπερνούσαν το 60% και τέλος γνωρίσματα που περιείχαν διπλότυπα. Έτσι ο συνολικός αριθμός των γνωρισμάτων περιορίστηκε στα 15 από τα 124 που ήταν αρχικά. Από τις 1430 εγγραφές προέκυψαν τελικά 1183 λόγω των ελλιπών τιμών σε συγκεκριμένα γνωρίσματα. Στη συνέχεια οι Rajesh & Anand, (2012) προχώρησαν στη φάση της κατηγοριοποίησης. Αρχικά, στο στάδιο της εκπαίδευσης οι 1430 εγγραφές χωρίστηκαν σε 3 ομάδες των 500 εγγραφών έτσι ώστε να δοκιμαστούν διάφοροι κανόνες και αλγόριθμοι. Συγκρίνοντας το Error Rate διαφόρων αλγορίθμων κατέληξαν το συμπέρασμα ότι ο πιο αποδοτικός αλγόριθμος για την κατηγοριοποίηση των δεδομένων είναι ο C4.5, που είχε το μικρότερο Error Rate = 0.0599 με τα νευρωνικά δίκτυα να ακολουθούν με Error Rate = 0.0739. Έτσι οι Rajesh & Anand, (2012) αποφάσισαν να χρησιμοποιήσουν για τη κατηγοριοποίηση τον αλγόριθμο C4.5.Τα αποτελέσματα της κατηγοριοποίησης αξιολογήθηκαν σύμφωνα με τις γνωστές μετρικές accuracy, sensitivity και specificity. Έτσι τα αποτελέσματα του αλγόριθμου C4.5 είχαν ως εξής : accuracy= 0.922 , sensitivity= 0.4666 και specificity= 0.9736

Άλλη μια έρευνα που αφορά εξόρυξη γνώσης από δεδομένα ογκολογίας διεξήχθει από τους Lin et al.(2012). Σκοπός και αυτής της έρευνας είναι η εξόρυξη γνώσης από δεδομένα ασθενών με καρκίνο του μαστού με στόχο την πρόβλεψη της βιωσιμότητας. Σύμφωνα με τις προηγούμενες έρευνες και εδώ τα δεδομένα που χρησιμοποιήθηκαν προέρχονται από το SEER. Τα δεδομένα αφορούν τις χρονολογίες 1973-2008. Το μέγεθος των δεδομένων σύμφωνα με τους Lin et al.(2012) είναι το διπλάσιο σε σχέση με το μέγεθος των δεδομένων που χρησιμοποίησαν οι Delen et al.(2004).

Κατά στο στάδιο της προεπεξεργασίας των δεδομένων οι Lin et al.(2012) ακολούθησαν παρόμοια διαδικασία με αυτή των Delen et al.(2004). Το σύνολο των δεδομένων χωρίζεται σε δύο μεγάλες κατηγορίες, στους ασθενής που επιβίωσαν και σε αυτούς που δεν βρίσκονται στη ζωή. Τις εγγραφές που αφορούν ασθενείς που δεν επιβίωσαν για 5 χρόνια αφού διαγνώστηκε ο καρκίνος και η αιτία θανάτου τους είναι διαφορετική από το καρκίνο του μαστού αφαιρούνται από το σύνολο των δεδομένων. Επίσης οι Lin et al.(2012) αποφάσισαν να αφαιρέσουν κάποια δημογραφικά γνωρίσματα και επέλεξαν να χρησιμοποιήσουν τα γνωρίσματα που χρησιμοποίησαν και οι Delen et al.(2004) στην έρευνά τους. Οι μέθοδοι που χρησιμοποίησαν για την κατηγοριοποίηση των δεδομένων είναι τα δέντρα απόφασης με τον αλγόριθμο J48 και η Logistic Regression.

Έρευνες	**Τεχνικές**	**Accuracy**	**Sensitivity**	**Specificity**	**Σύνολο Δειγμάτων**
Delen et al.(2004)	C5	**93.62%**	**96.02%**	**90.66%**	202.932
	ANN	91.21%	94.37%	87.48%	
	Logistic	89.20%	87.86%	90.17%	
Bellaachia & Guven (2006)	C4.5	**86.70%**	87.22	-	151.886
	Naïve Bayes	84.5%	85.20%	-	
	ANN	86.50%	**90,10%**	-	
Lin et al.(2012)	C4.5	91.07%	**99.61%**	**91.82%**	215.375
	Logistic	**91.27%**	98.93%	83.54%	
Rajesh & Anand, (2012)	C4.5	92.20%	46.66%	97.36%	1.304

Πίνακας 1. Συγκεντρωτικός πίνακας των υπάρχουσων ερευνών

Για την αξιολόγηση των αποτελεσμάτων των αλγορίθμων χρησιμοποιήθηκαν οι μετρικές accuracy, sensitivity, specificity καθώς και η καμπύλη ROC χρησιμοποιείται για την αξιολόγηση των μοντέλων που προέκυψαν αφού χρησιμοποιήθηκαν οι διάφορες μέθοδοι της εξόρυξης γνώσης και δείχνει τη σχέση μεταξύ sensitivity και specificity σύμφωνα με τους Lin et al.(2012). Σύμφωνα με τα αποτελέσματα ο J48 και η Logistic Regression δεν έχουν κάποια ουσιαστική διαφορά στην ακρίβειά τους. Το accuracy για τον J48 είναι 91,07 για τη Logistic Regression 91,27. Ωστόσο σύμφωνα με τη καμπύλη ROC η Logistic Regression είναι πιο αποδοτική σε σχέση με τον αλγόριθμο των δέντρων απόφασης στη πρόβλεψη για την βιωσιμότητα των ασθενών με καρκίνο του μαστού (Logistic Regression= 0.829 και J48= 0.717). Τα συγκεντρωτικά αποτελέσματα φαίνονται στο Πίνακας 1.

2.4 Μέθοδοι εξόρυξης γνώσης που δεν έχουν εφαρμοστεί σε δεδομένα ογκολογίας.

Όπως προκύπτει από τις περισσότερες έρευνες που έχουν διεξαχθεί στο τομέα της εξόρυξης γνώσης από δεδομένα ογκολογίας, οι πιο δημοφιλείς μέθοδοι που χρησιμοποιούνται αφορούν τα δέντρα απόφασης, τα τεχνητά νευρωνικά δίκτυα και τις διάφορες στατιστικές μεθόδους όπως η Logistic Regression.

Ωστόσο υπάρχουν και άλλες μέθοδοι οι οποίες σύμφωνα με έρευνες έχουν αποδειχθεί πιο αποδοτικές από ότι οι συνηθισμένες μέθοδοι. Δύο από αυτές είναι τα Support Vector Machines (SVM) και οι Ensemble μέθοδοι.

Τα SVM θεωρούνται μέθοδοι που οπωσδήποτε πρέπει να δοκιμαστούν καθώς φαίνεται πως τα αποτελέσματά τους προσφέρουν μεγαλύτερη ακρίβεια σε σχέση με της συνηθισμένες μεθόδους εξόρυξης γνώσης (Ghosh et al, 2008).

Οι Ensemble μέθοδοι, έχουν αρχίσει να γίνονται αρκετά δημοφιλής με την αύξηση του ενδιαφέροντος πάνω στο τομέα της Εξόρυξης Γνώσης. Αυτές οι μέθοδοι έχουν το πλεονέκτημα ότι μπορούν να χρησιμοποιήσουν για την κατηγοριοποίηση ενός

συνόλου δεδομένων διάφορους ταξινομητές και να συλλέξουν πληροφορίες από αυτούς καταλήγοντας σε ένα τελικό ταξινομητή που θα προσφέρει αποτελέσματα με μεγαλύτερη ακρίβεια.

Όπως είναι λογικό αυτές οι δυο μέθοδοι παρουσιάζουν ιδιαίτερο ενδιαφέρον και για αυτό αποτελούν και αυτές ένα αντικείμενο μελέτης και σύγκρισης μαζί με της υπόλοιπες μεθόδους, όπως τα δέντρα απόφασης και της Logistic Regression, της συγκεκριμένης πτυχιακής εργασίας.

3. Σχεδίαση μοντέλου και επεξεργασία των δεδομένων

3.1 Κατανόηση των δεδομένων.

Το σύνολο των δεδομένων που χρησιμοποιήθηκαν στη παρούσα πτυχιακή εργασία παρέχεται από το SEER (Surveillance, Epidemiology and End Results) ενός προγράμματος του National Cancer Institute. Το SEER είναι μια από της μεγαλύτερες και πιο αλοκληρωμένες πηγές πληροφόρησης για δεδομένα όλων των τύπων καρκίνου από τους οποίους έχουν προσβληθεί οι ασθενείς σε ολόκληρες τις Η.Π.Α.

Το αρχείο του προγράμματος SEER για το 2012, περιέχει δεδομένα που συλλέγονται και δημοσιοποιούνται, από περιπτώσεις ασθενών που έχουν προσβληθεί από διάφορους τύπους καρκίνου, την βιωσιμότητα των ασθενών αφού έχει γίνει διάγνωση του καρκίνου, η θνησιμότητα λόγω του καρκίνου και άλλα. Όλες οι εγγραφές που υπάρχουν μέσα στο αρχείο του SEER αφορούν 17 περιοχές των Η.Π.Α και αντιστοιχούν περίπου το 28% του πληθυσμού των Ηνωμένων Πολιτειών. Οι βάσεις δεδομένων του SEER παρέχουν πληροφορίες για περισσότερες από 4 εκατομμύρια περιπτώσεις καρκίνων σε αρχικό στάδιο και δεν εξαπλώνονται και περιπτώσεις καρκίνων με μετάσταση. Σύμφωνα με το SEER 170,000 καινούριες περιπτώσεις εισάγονται στο αρχείο τους κάθε χρόνο. Το αρχείο του SEER περιέχει επίσης δημογραφικά στοιχεία ασθενών, λεπτομέρειες για την ασθένειά του όπως σε ποιό σημείο διαγνώστηκε ο καρκίνος για πρώτη φορά, τη μορφολογία του καρκίνου του ασθενούς, τους τρόπους θεραπείας του κτλ. Τα δεδομένα θνησιμότητας των ασθενών του αρχείου SEER παρέχονται από το National Center for Health Statistics. Το πρόγραμμα SEER συγκαταλέγεται στα κορυφαία αρχεία περιπτώσεων καρκίνου σε όλο το κόσμο, όσον αφορά τη ποιότητα και την αξιοπιστία των πληροφοριών που προσφέρει.

Οι γεωγραφικές περιοχές που συμπεριλαμβάνονται στο πρόγραμμα SEER επιλέχθηκαν με βάση την ικανότητά τους να λειτουργούν και να διατηρούν έναν υψηλής ποιότητας σύστημα αναφοράς με βάση το πληθυσμό. Υπολογίζεται ότι οι βάσεις δεδομένων που παρέχονται από το SEER είναι συμπληρωμένες κατα 98% για κάθε μία από τις 17 περιοχές των Η.Π.Α που καταγράφονται στο αρχείο. Ο

πληθυσμός που καλύπτεται από το SEER είναι συγκρίσιμος με τον γενικό πληθυσμό των Ηνωμένων Πολιτείων όσον αφορά τα μέτρα της φτώχειας και της εκπαίδευσης. Ο πληθυσμός του SEER φαίνεται να είναι πιο αστικός και υπάρχει μεγαλύτερος αριθμός ασθενών που έχουν γεννηθεί στο εξωτερικό παρά στις Η.Π.Α (SEER, 2012).

Σε αυτή την έρευνα χρησιμοποιήθηκε το αρχείο του προγράμματος SEER που αφορά τις χρονολογίες 1973-2008 και αφορά αρχεία 17 περιοχών των Η.Π.Α. Τα δεδομένα έχουν χωρισθεί σε εννέα κατηγορίες και κάθε μία αφορά ένα συγκεκριμένο τύπο καρκίνου. Τα συγκεκριμένα αρχεία είναι αρχεία κειμένου τύπου ASCII.

Για τη συγκεκριμένη έρευνα από τις εννέα κατηγορίες καρκίνου επιλέχθηκε το αρχείο BREAST.txt που αφορά το καρκίνο του μαστού. Το αρχικό σέτ δεδομένων περιέχει ***630,218 εγγραφές (instances) και 124 γνωρίσματα (attributes).*** Κάθε εγγραφή αφορά έναν ασθενή και τα χαρακτηριστικά του καρκίνου του μαστού από τον οποίο έχει προσβληθεί. Λόγω του πολύ μεγάλου μεγέθους των δεδομένων, οπότε και μπορεί να υπάρχουν ελλειπείς τιμές και πολλά γνωρίσματα που δεν έχουν σημασία για την έρευνα, είναι αναγκαίο να πραγματοποιηθεί μια προεπεξεργασία των δεδομένων (data preprocessing) με στόχο να υπάρξουν καλύτερα και πιο ακριβή αποτελέσματα.

Επίσης από τα 124 γνωρίσματα του αρχικού συνόλου των δεδομένων, επιλέχθηκαν τελικά 17 πιο σημαντικά γνωρίσματα σύμφωνα και με τις έρευνες των Delen et al.(2004) και Bellaachia & Guven (2006). Τα γνωρίσματα αυτά είναι τα παρακάτω και πιο συγκεντρωτικά φαίνονται στον Πίνακας 2 που ακολουθεί :

- **MaritalStatusAtDX**: περιγράφει την οικογενειακή κατάσταση του ασθενούς
- **Race_Ethnicity**: περιγράφει την εθνικότητα του ασθενούς
- **AgeAtDiagnosis**: παρέχει πληροφορίες για την ηλικία του κάθε ασθενούς.
- **PrimarySite**: περιγράφει σε πιό σημείο εμφανίστηκε.
- **Histology (ICD_O_2)**: Περιέχει κωδικούς που σχετίζονται με τη μορφολογία του όγκου κάθε ασθενούς
- **Behavior (ICD_O_2):** Περιέχει στοιχεία για την εξέλιξη του όγκου του ασθενούς

- **Grade:** περιγράφει το διαχωρισμό των κωδικών διαφόρων κυττάρων
- **EOD_TumorSize**: περιγράφει τις διαστάσεις του όγκου του ασθενούς
- **EOD_Extension**: περιγράφει την εξάπλωση του όγκου περετέρο από το αρχικό σημείο που εμφανίστηκε.
- **EOD_LymphNodeInvolv**: περιγράφει το μεγαλύτερο αριθμό των λεμφαδένων που έχουν προσβληθεί από καρκίνο
- **RegionalNodesPositive**: καταγράφει τους τοπικούς λεμφαδένες που εξετάστηκαν από παθολόγο και βρέθηκαν να έχουν μετάσταση
- **RXSumm_Radiation**: Η μέθοδος της ακτινοθεραπείας που χρησιμοποιήθηκε για θεραπεία του ασθενους σε αρχικό στάδιο
- **RXSumm_SurgeryType**: Καταγράφει πληροφορίες από τις εγχειρίσεις που έχει υποβληθεί στο πρώτο στάδιο θεραπείας είτε λόγω καρκίνου είτε άλλης ασθένειας.
- **SEERHistoricStageA**: παρέχει πληροφορίες για τις διάφορες καταστάσεις των όγκων.
- **SEEROtherCauseOfDeathClassification**: Χρησιμοποιείται για να δούμε την πιθανότητα αν μια ομάδα ασθενών επέζησε μετά από άλλη ασθένεια και όχι καρκίνο.Σε αυτή η περίπτωση ο θάνατος από άλλη ασθένεια θεωρείται γεγονός.

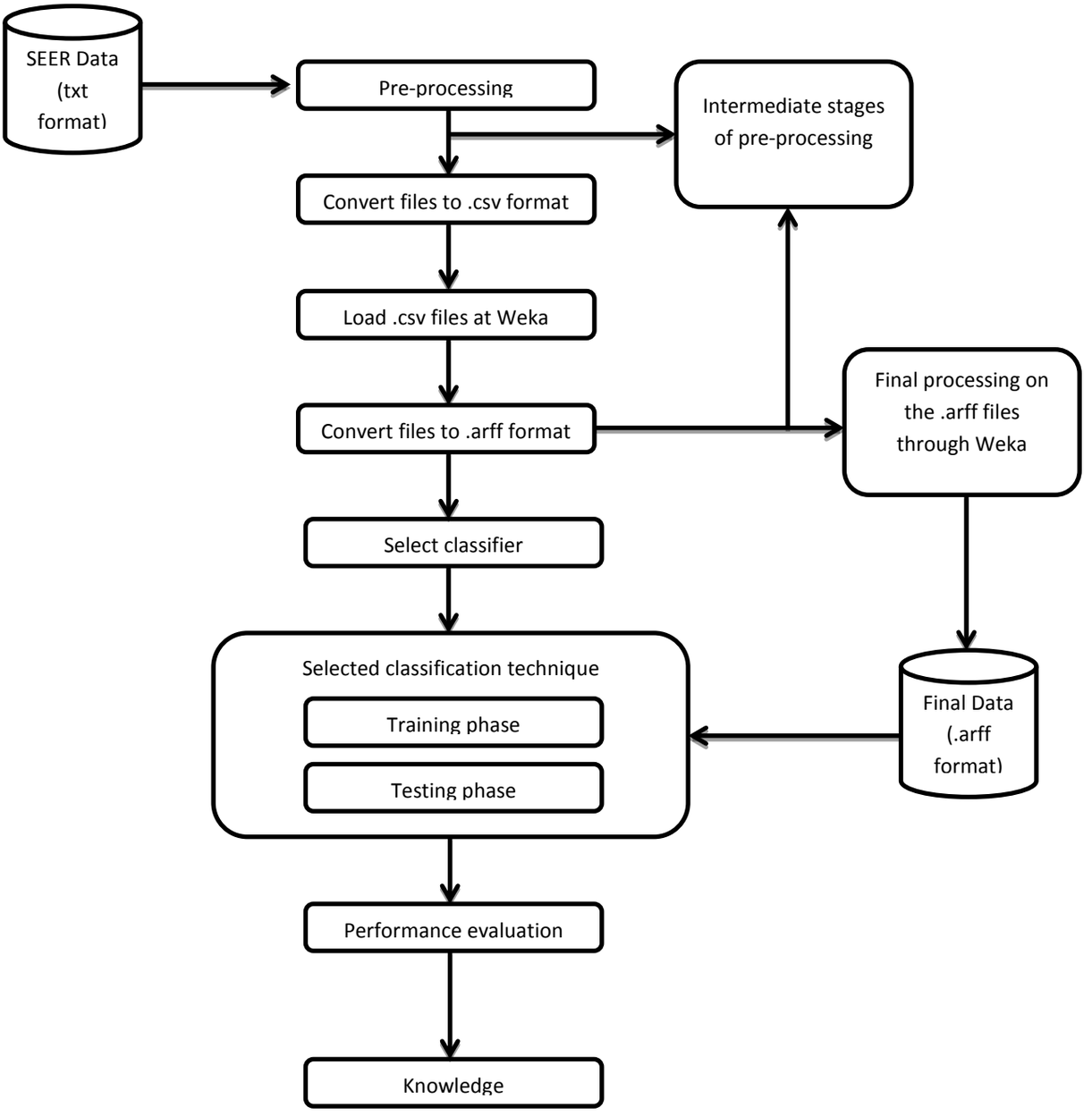
Attributes	Τύπος κάθε γνωρίσματος	Τιμές κάθε γνωρίσματος
MaritalStatusA□DX	Κατηγορικό	6
Race_Et□nicity	Κατηγορικό	29
AgeAtDiagnosis	Αριθμητικό	95
PrimarySite	Κατηγορικό	9
Histology (ICD_O_2)	Κατηγορικό	86
Behavior (ICD_O_2)	Κατηγορικό	2
Grade	Κατηγορικό	5
EOD_TumorSize	Κατηγορικό	179
EOD_Extension	Κατηγορικό	3
EOD_LymphNodeInvolv	Κατ□γ□ρικό	10

RegionalNodesPositive	Κατηγορικό	61
RXSumm_Radiation	Κατηγορικό	9
RXSumm_SurgeryType	Κατηγορικό	25
SEE□HistoricStageA	Κατηγορικό	5
NumberOfPrimaries	Αριθμητικό	8
SEEROtherCauseOfDeathClassification	Κατηγορικό	3
SurvivalBinary	Κατηγορ□κ	2

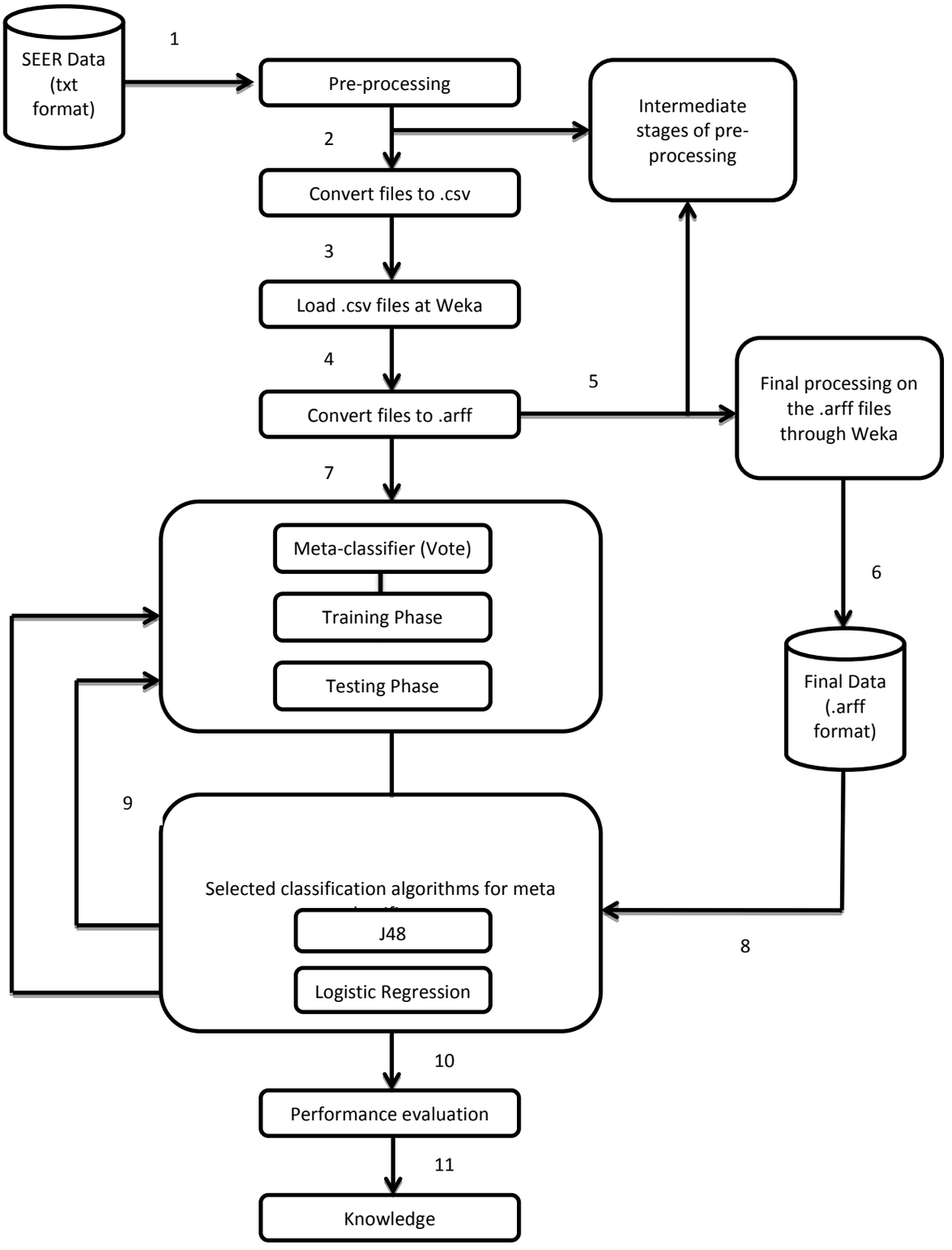
Πίνακας 2. Πίνακας γνωρισμάτων του Dataset.

Η σχεδίαση του μοντέλου για την εξόρυξης γνώσης από τα δεδομένα πρέπει να γίνει πολύ προσεκτικά έτσι ώστε να υπάρχουν αποδοτικά αποτελέσματα. Στη συγκεκριμένη έρευνα η σχεδίαση του μοντέλου αποτελείται από έντεκα συνολικά στάδια τα οποία φαίνονται στην **Εικόνα 3.1** και είναι τα παρακάτω:

1. Αρχικά γίνεται η εισαγωγή του συνόλου των δεδομένων από το SEER
2. Έπειτα ξεκινά το στάδιο της προ-επεξεργασίας το οποίο περιέχει και τα στάδια 3, 4, 5 και 6.
3. Στο τρίτο στάδιο μετατρέπεται το αρχικό .txt αρχείο σε αρχείο .csv.
4. Το τέταρτο στάδιο περιλαμβάνει τη φόρτωση του αρχείου στο πρόγραμμα Weka.
5. Στο πέμπτο στάδιο μετατρέπεται το .csv αρχείο σε αρχείο .arff.
6. Κατά το έκτο στάδιο γίνονται οι τελικές διεργασίες της προ-επεξεργασίας πάνω στο .arff αρχείο.
7. Στο έβδομο στάδιο, το οποίο πραγματοποιείται παράλληλα με το έκτο, γίνεται η επιλογή του κατάλληλου ταξινομητή
8. Στο όγδοο στάδιο έχουμε την παραγωγή του τελικού αρχείου προς κατηγοριοποίηση.
9. Κατά το ένατο στάδιο εφαρμόζονται οι διάφορες τεχνικές κατηγοριοποίησης και γίνεται η εκπαίδευση των δεδομένων για την εξαγωγή της γνώσης.
10. Στο δέκατο στάδιο γίνεται η αξιολόγηση των αποτελεσμάτων
11. Στο τελευταίο στάδιο εξετάζεται η γνώση που προέκυψε.



Εικόνα 3.1. Η σχεδίαση του μοντέλου εξόρυξης γνώσης για τους ταξινόμητες των Decision Trees και των Functions



Εικόνα 3.2. Η σχεδίαση του μοντέλου εξόρυξης γνώσης για τους meta-classifiers

Η παραπάνω σχεδίαση του μοντέλου αφορά ταξινομητές όπως τα δέντρα απόφασης και τους συναρτησιακούς ταξινομητές. Για τους meta-classifiers ακολουθήθηκε ένας διαφορετικός σχεδιασμός που οφείλεται κυρίως στον διαφορετικό τρόπο επιλογής της κατάλληλη τεχνικής κατηγοριοποίησης. Σύμφωνα, λοιπόν, με τους meta-classifiers επιλέγεται αρχικά ένας meta-ταξινόμητης, πχ στη συγκεκριμένη πτυχιακή εργασία ο Vote. Ωστόσο σύμφωνα με τη λειτουργία, των meta-classifiers, απαιτείται και επιλογή κάποιων περισσοτέρων «βοηθητικών» αλγορίθμων σύμφωνα με τους οποίους ο τελικός ταξινομητής (Vote) θα πραγματοποιήσει τη κατηγοριοποίηση. Για αυτό και χρειάζεται ένας λίγο διαφορετικός σχεδιασμός για τους meta-classifiers, ο οποίος φαίνεται στην Εικόνα 3.2

3.2 Προ-επεξεργασία των δεδομένων.

Τα δεδομένα που συλλέγονται για την σχεδίαση μεθόδων εξόρυξης γνώσης παρουσιάζουν διάφορες δυσκολίες οι οποίες μπορούν να δημιουργήσουν προβλήματα σε όλη τη διαδικασία. Τέτοιες δυσκολίες είναι:

- Τα ελλειπή δεδομένα: δηλαδή οι τιμές που λείπουν από γνωρίσματα του dataset ακόμη και έλλειψη ολόκληρων γνωρισμάτων που είναι σημαντικά για την έρευνα. Αιτίες των ελλειπών τιμών μπορεί να είναι πολλά πεδία τα οποία δεν συμπληρώνονται ως μη σημαντικά ή η διαγραφή των παλαιότερων δεδομένων για την εξοικονόμηση χώρου.
- Τα δεδομένα που περιέχουν θόρυβο: δηλαδή να υπάρχουν λάθη και ακραίες τιμές στα dataset και γενικότερα αν υπάρχουν διακυμάνσεις στις τιμές διαφόρων μεταβλητών. Αιτίες θορύβου μπορεί να είναι τα τυχαία λάθη, προβλήματα στη μετάδοση των δεδομένων ακόμη και οι περιορισμοί της τεχνολογίας.
- Οι ασυνέπειες στα δεδομένα: δηλαδή να υπάρχουν διαφορές σε κωδικούς που αντιπροσωπεύουν ονόματα. Αιτίες αυτού του προβλήματος μπορεί να είναι η έλλειψη μοναδικού αναγωριστικού ή τα χαρακτηριστικά που προκύπτουν από τις τιμές των άλλων. (Yang et al, 2010)

Για αυτό και οι ερευνητές προβαίνουν σε μία σειρά από διεργασίες οι οποίες θα εξασφαλίσουν την ποιότητα των δεδομένων που θα χρησιμοποιηθούν στην έρευνα. Τέτοιες διεργασίες είναι οι εξής:

- Ο καθαρισμός των δεδομένων: δηλαδή η συμπλήρωση τιμών που λείπουν, η διαχείρισει του θορύβου που είναι πιθανό να δημιουργείται και η επίλυση των ασυνεπειών
- Η ολοκλήρωση των δεδομένων: δηλαδή η ολοκλήρωση των βάσεων δεδομένων ή των αρχείων έτσι ώστε να επιλυθεί ο πλεονασμός των δεδομένων.
- Ο μετασχηματισμός των δεδομένων με τις τεχνικές της κανονικοποίησης και της άθροισης.
- Η μείωση των δεδομένων : δηλαδή η μείωση του συνόλου των δεδομένων με τρόπο έτσι ώστε να παραχθούν αποδοτικότερα αποτελέσματα.
- Η διακριτοποίηση των δεδομένων (discretization): δηλαδή η μείωση μέρους των αριθμητικών δεδομένων με απόδοση ιδιαίτερων τιμών.

Όλες αυτές οι διεργασίες που πρέπει να πραγματοποιηθούν για τη βελτίωση του συνόλου των δεδομένων αποτελούν το στάδιο της προ-επεξεργασίας των δεδομένων στην εξόρυξη γνώσης.

Η προ-επεξεργασία των δεδομένων αποτελεί ένα από τα σημαντικότερα βήματα της εξόρυξης γνώσης. Αφού έχει πραγματοποιηθεί η προεπεξεργασία των δεδομένων, προκύπτει ένα νέο σύνολο δεδομένων αρκετά μικρότερο από το αρχικό το οποίο θα προσφέρει όμως πολύ καλύτερα αποτελέσματα. Όπως στις περισσότερες έρευνες για την εξόρυξη γνώσης ο περισσότερος χρόνος αφιερώνεται στο συγκεκριμένο βήμα, μέχρι και 80% του συνολικού χρόνου της έρευνας (Yang et al, 2010), έτσι και στη παρούσα πτυχιάκη εργάσια δώθηκε πολύ μεγάλη έμφαση στο καθαρισμό και την προετοιμασία των «ακατέργαστων» δεδομένων που παρέχονται από το SEER.

Όπως είναι γνωστό στόχος της συγκεκριμένης έρευνας είναι η ανάπτυξη μοντέλων εξόρυξης γνώσης για τη πρόβλεψη της βιωσιμότητας των ασθενών που έχουν προσβληθεί από καρκίνο του μαστού. Έτσι εκτός από τα 124 γνωρίσματα του dataset

δημιουργήθηκε ένα επιπλέον το οποίο ονομάζεται SurvivalBinary έτσι ώστε να δημιουργηθεί ένα κατώφλι που να διαχωρίζει τους ασθενής που επιβίωσαν και αυτούς που δεν επιβίωσαν. Το γνώρισμα SurvivalBinary είναι κατηγορικό και παίρνει δύο τιμές 0 και 1, όπου στο ***0*** αντιστοιχούν οι ασθενείς που ***δεν επιβίωσαν*** και στο ***1*** ασθενείς που ***επιβίωσαν*** (survived, not survived).

Για αυτόν το διαχωρισμό ακολουθήθηκε μια συγκεκριμένη στρατηγική η οποία είναι διαφορετική σε σχέση με την έρευνα των Delen et al.(2004) καθώς διαπιστώθηκε πώς οι συγκεκριμένοι ερευνητές παρέλειψαν να συμπεριλάβουν για τη δημιουργία του κατωφλίου, δύο πολύ σημαντικά γνωρίσματα του dataset το VitalStatusRecode (VSR), που αφορά το αν βρίσκεται εν ζωή ένας ασθενής ή έχει πεθάνει και το CauseOfDeath (COD) που υποδηλώνει την αιτία θανάτου ενός ασθενή. Ωστόσο η στρατηγική που ακολουθήθηκε για το διαχωρισμό είναι παρόμοια με αυτή των Bellaachia & Guven (2006). Έτσι χρησιμοποιήθηκαν τρία πολύ απαραίτητα γνωρίσματα, το SurvivalTimeRecode (STR) που αφορά τους πόσους μήνες καί χρόνια επιβίωσε ένας ασθενής έπειτα από τη διάγνωση, το VitalStatusRecode (VSR) και το CauseOfDeath (COD). Το κατώφλι που τέθηκε με την εξαρτημένη μεταβλητή SurvivalBinary είναι οι 60 μήνες. Δηλαδή όσοι ασθενείς επιβίωσαν για περισσότερους από 60 μήνες έπειτα από τη διάγνωση του καρκίνου και βρίσκονται εν ζωή κατηγοριοποιούνται ως survived (1) και όσοι ασθενείς επιβίωσαν για λιγότερο από 60 μήνες και η αιτία θανάτου είναι ο καρκίνος του μαστού κατηγοριοποιούνται ως not survived (0). Οι υπόλοιπες εγγραφές του συνόλου των δεδομένων που δεν αντιστοιχούν στο παραπάνω έλεγχο, δηλαδή εγγραφές ασθενών που επιβίωσαν έπειτα από τη διάγνωση λιγότερο από 60 μήνες και έχουν καταγραφεί ώς εν ζωή ασθενείς (STR<60 και VSR=alive) ή εγγραφές ασθενών που επιβίωσαν έπειτα από τη διάγνωση λιγότερο από 60 μήνες και η αιτία θανάτου τους δεν είναι ο καρκίνος του μαστού (STR<60 και COD </> Breast Cancer), αγνοούνται και αφαιρούνται τελείως από το dataset.Παρακάτω φαίνεται ο ψευδοκώδικας για τη δημιουργία του κατωφλίου:

If (STR > 60) and (VSR = alive)

Then “Η εγγραφή κατηγοριοποιείται σαν survived”

Else_if (STR) < 60 and (COD = Breast Cancer)

Then “Η εγγραφή κατηγοριοποιείται σαν not survived”

Else

“Η εγγραφή αγνοείται”

Για την υλοποίηση του κατωφλίου ήταν αναγκαίο να αναπτυχθεί κώδικας σε Java με τη βοήθεια του προγράμματος NetBeans 6.8 έτσι ώστε να «φιλτραριστούν» όλα τα 630,218 δεδομένα και να προκύψει ένα νεό σύνολο δεδομένων το οποίο θα περιέχει λιγότερες εγγραφές και θα είναι σαφώς πιο αποδοτικό. Όπως είναι γνωστό τα δεδομένα που παρέχονται από το SEER βρίσκονται σε αρχεία της μορφής .txt. Αυτός ο τύπος των αρχείων όμως δεν υποστηρίζεται από το Weka, πάνω στο οποίο βασίζεται η έρευνα. Έτσι τα αρχεία αφου «φιλτραριστούν» μέσα από το κώδικα της Java αποθηκεύονται σε μορφή .csv έτσι ώστε να είναι δυνατή η επεξεργασία τους μέσα από το Weka. Αυτό αποτελεί και το δεύτερο στάδιο της σχεδίασης. Εν συνεχεία το .csv αρχείο που δημιουργήθηκε σύμφωνα με τον διαχωρίσμο που έγινε με τη δημιουργία του κατωφλίου φορτώθηκε στο πρόγραμμα Weka έτσι ώστε να μετατραπεί σε αρχείο .arff που αποτελεί τη βασική μορφή αρχείων που υποστηρίζει το Weka. Αυτά τα βήματα αποτελούν το 3^{ο} και το 4^{ο} στάδιο της σχεδίασης αντίστοιχα όπως φαίνεται και στην **Εικόνα 3.1.**

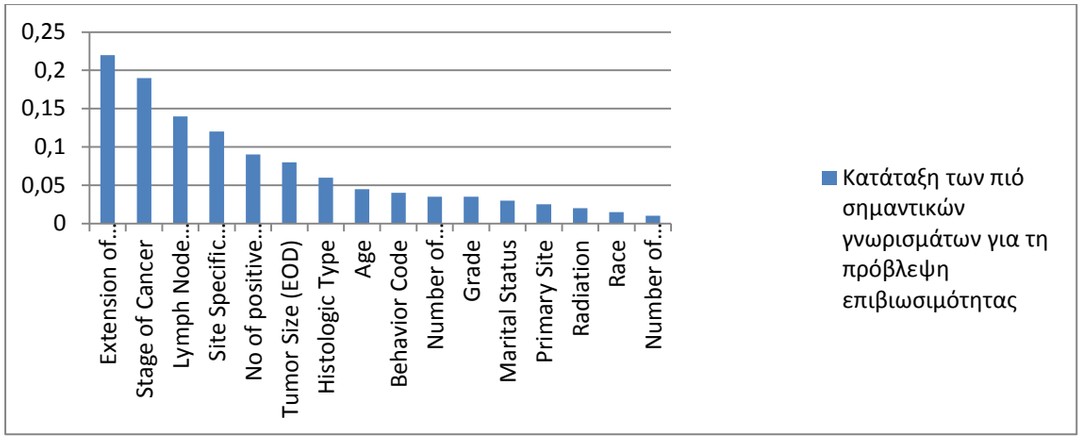
Έπειτα από τη υλοποίηση του κατωφλίου και την εισαγωγή του νέου κατηγορικού γνωρίσματος SurvivalBinary το νέο σύνολο που προέκυψε αποτελέιται από 242,558 και 125 γνωρίσματα. Από τις 242,558 εγγραφές που προέκυψαν οι 165,235 κατηγοριοποιούνται ως survived και οι 77,323 κατηγοριοποιούνται ώς not survived όπως φαίνεται και στο Πίνακας 3

	Σύνολο εγγραφών	Ποσοστό
Survived (1)	165,235	68,12%
Not Survived (0)	77,323	32,88%
Total	242,558	100%

Πίνακας 3. $1_η$ κατηγοριοποίηση των εγγραφών σύμφωνα με το κατώφλι

Στο πέμπτο στάδιο της σχεδίασης γίνονται οι διεργασίες για την υλοποίηση του τελικού .arff αρχείου το οποίο θα χρησιμοποιηθεί και για τη φάση της κατηγοριοποίησης.Τα γνωρίσματα (attributes) του συγκεκριμένου dataset φτάνουν τα 125. Ωστόσο πολλά από αυτά δεν έχουν καμία απολύτως χρησιμότητα καθώς ορισμένα περιέχουν πολλές ελλειπείς τιμές (missing values) και άλλα δεν αποτελούν κριτήριο για αυτή την έρευνα.

Πιο συγκεκριμένα αποφασίστηκε να αφαιρεθούν τελείως από το dataset όσα attributes περιέχουν ελλειπείς τιμές μεγαλύτερες από 80%. Σύμφωνα με τη ανάλυση το σύνολο των δεδομένων του SEER περιέχει γνωρίσματα όπως τα SiteSpecific πεδία και τα Override πεδία τα οποία έχουν έως και 100% ελλειπείς τιμές. Έτσι εντοπίστηκαν 46 πεδία τα οποία περιείχαν ελλιπείς τιμές από 80% και περισσότερο και αφαιρέθηκαν από το dataset. Έπειτα από αυτή τη διαφοροποίηση, το σύνολο των δεδομένων πλέον αποτελείται από 242,558 εγγραφές και 74 γνωρίσματα. Ωστόσο σύμφωνα με τους Bellaachia & Guven (2006) και Delen et al.(2004) από τα 74 γνωρίσματα που απομένουν υπάρχουν 57 τα οποία πρέπει να αφαιρεθούν καθώς θα επηρεάσουν αρνητικά τα αποτελέσματα της έρευνας. Σύμφωνα με τους Bellaachia & Guven (2006) τα 16 γνωρίσματα που απομένουν είναι πολύ σημαντικά και πρέπει να να χρησιμοποιηθούν στη φάση της κατηγοριοποίησης. Έτσι δημιουργήθηκε ένα διάγραμμα κατάξης των 16 γνωρισμάτων σύμφωνα με τη σημαντικότητα τους(βλ. Διάγραμα 1).



Διάγραμα 1. Κατάταξη των 16 πιό σημαντικών γνωρισμάτων του Dataset [Πηγη: Bellaachia & Guven (2006)]

Παρόλα αυτά χρησιμοποιήθηκε και ένας ακόμα τρόπος για την επίλογη των τελικών γνωρισμάτων. Αυτός ο τρόπος είναι το Attribute Selection μέσα από το πρόγραμμα Weka. Σύμφωνα με αυτή τη μέθοδο επιλέγεται ένας αλγόριθμος για την αξιολόγηση των γνωρισμάτων και ένας αλγόριθμος αναζήτησης. Αυτές οι μέθοδοι επιλέχτηκαν να δοκιμαστούν σε ένα τυχαίο δείγμα 33496 εγγραφών του συνόλου των δεδομένων που περιέχει 630,218 εγγραφές. Συγκεκριμένα επιλέχθηκαν τρείς τεχνικές attribute selection με τις παρακάτω μεθόδους:

- **CfsSubsetEval και GeneticSearch**: ο CfsSubsetEval είναι ένας αλγόριθμος για την αξιολόγηση των γνωρισμάτων. Στην ουσία αξιολογεί υποσύνολα των γνωρισμάτων. Αξιολογεί την ικανότητα πρόβλεψης κάθε γνωρίσματος σε συνδυασμό με το βαθμό συσχέτισης μεταξύ τους (Hall & Holmes, 2000). Η μέθοδος GeneticSearch χρησιμοποιεί έναν αλγόριθμο που ακολουθεί τυφλή στρατηγική αναζήτησης και δεν χρειάζεται γνώση των ιδιοτήτων της συνάρτησης για τη βελτιστοποίηση της (Payne & Glen, 1993).
- **InfoGain και Ranker**: η InfoGain μέθοδος αξιολόγησης που υπολογίζει την εντροπία της κλάσης των γνωρισμάτων. Όσο η εντροπία μειώνεται αντιπροσωπεύει τη πληροφορία που παράγεται από τα γνωρίσματα για τη κλάση και

ονομάζεται κέρδος πληροφορίας (information gain) και το κέρδος αυτό θέτεται σάν μία τιμή σε κάθε γνώρισμα (Hall & Holmes, 2000). Η Ranker χρησιμοποιεί έναν αλγόριθμο που ταξινομέι τα γνωρίσματα ανάλογα με την αξιολόγησή τους.

- **Chi-squared και Ranker:** Η Chi-squared μέθοδος αξιολογεί το κάθε ένα γνώρισμα ξεχωριστά, μετρώντας το στατιστικό X^2 με βάση το γνώρισμα-κλάση και στη συνέχεια τα γνωρίσματα ταξινομούνται ανάλογα με τη μεγαλύτερη τιμή του X^2 που τους έχει αποδοθεί (Liu et al., 2002)

Έτσι, έπειτα από την χρησιμοποίηση των παραπάνω τεχνικών προκύπτουν νέα σέτ γνωρισμάτων. Για τις μεθόδους **CfsSubsetEval και GeneticSearch** προκύπτουν 30 γνωρίσματα και για τις **InfoGain και Ranker** και **Chi-squared και Ranker** 20 γνωρίσματα.Ο πίνακας κατηγοριοποίησης των εγγραφών με βάση το κατώφλι φαίνεται παρακάτω(βλ. Πίνακας 4). Το σύνολο των δεδομένων και τα νέα γνωρίσματα θα

	Σύνολο εγγραφών	**Ποσοστό**
Survived (1)	22,941	68,48%
Not Survived (0)	10,555	32,52%
Total	33,496	100%

Πίνακας 4. Κατηγοριοποίηση των εγγραφών του Attribute Selection με βάση το κατώφλι.

ταξινομηθούν με τους αλγορίθμους J48, Logistic Regression και Vote και τα αποτελέσματά τους θα παρουσιαστούν στο $5^{ο}$ κεφάλαιο.

Αφού πραγματοποιήθηκε η τελική επιλογή των γνωρισμάτων που θα χρησιμοποιηθούν για τη φάση της κατηγοριοποίησης παρατηρήθηκε ότι υπήρχαν ακόμη ορισμένα γνωρίσματα, συγκεκριμένα τέσσερα, το οποία περιείχαν ακόμη έναν

αριθμό από ελλιπείς τιμές που κυμαινόντουσαν περίπου στο 24-32%. Ωστόσο επειδή ο αριθμός των ελλιπών τιμών στα συγκεκριμένα γνωρίσματα είναι αρκετά μικρός αποφασίστηκε να μην διαγραφούν ολόκληρα τα γνωρίσματα αλλά είτε να αφαιρεθούν οι ελλιπείς τιμές είτε να αντικατασταθούν, με τεχνικές μέσα από το πρόγραμμα Weka οι οποίες θα διευκρινιστούν στο επόμενο κεφάλαιο. Έτσι ύστερα και από αυτή την επεξεργασία δημιουργήθηκε το τελικό σύνολο των δεδομένων το οποίο θα αποτελείται από 165,725 εγγραφές και 17 συνολικά γνωρίσματα τα οποία φαίνονται στο Πίνακας 2. Ο διαχωρισμός σύμφωνα με το κατώφλι που έχει δημιουργηθεί φαίνεται στο Πίνακας 5.

	Σύνολο εγγραφών	**Ποσοστό**
Survived (1)	129,032	77,85%
Not Survived (0)	36,693	22,15%
Total	165,725	100%

Πίνακας 5. $2_{η}$ κατηγοριοποίηση των εγγραφών σύμφωνα με το κατώφλι

Παρατηρώντας το τελικό σύνολο των εγγραφών που προέκυψε με την επεξεργασία είναι φανερό πώς τα αποτελέσματα είναι πάρα πολύ κοντά σε σχέση με αυτά που προέκυψαν στην έρευνα των Bellaachia & Guven (2006) έπειτα από τη προ-επεξεργασία που είχαν πραγματοποιήσει (βλ.Πίνακας 6), αν λάβουμε υπόψη οτι στη παρούσα πτυχιακή εργασία χρησιμοποιήθηκε πιό πρόσφατο dataset από το SEER (1973-2008). Οι Bellaachia & Guven (2006) στην έρευνά τους κατέληξαν σε ένα τελικό σύνολο 151,886 εγγραφών έναν αριθμό αρκετά κόντα στον αριθμό τον εγγραφών της παρούσας έρευνας (165,725). Σύμφωνα με το κατώφλι που δημιούργησαν οι Bellaachia & Guven (2006), ο αριθμός των εγγραφών που καταχωρήθηκε ως survived ήταν 116,738 (76,8%) και ως not survived 35,148 (23,2%), αριθμοί οι οποίοι είναι και αυτοί αντίστοιχοι με τους αριθμούς των εγγραφών που διαχωρίστηκαν σύμφωνα με το κατώφλι της πτυχιακής εργασίας [survived: 129,032 (77,85%) και not survived: 36,693 (22,15%)].

	Σύνολο εγγραφών	Ποσοστό
Survived (1)	116,738	76,80%
Not Survived (0)	35,148	23,20%
Total	151,886	100%

Πίνακας 6. Κατηγοριοποίηση εγγραφών σύμφωνα με το κατώφλι .[Πηγή: Bellaachia & Guven (2006)]

Ωστόσο, ο αριθμός των εγγραφών που παρουσίασαν οι Delen et al.(2004) στην έρευνά τους, για την κατηγοριοποίηση τους σε survived και not survived είναι τελείως διαφορετικός από αυτόν που διαμορφώθηκε στην πτυχιακή έρευνα (βλ. Πίνακας 7). Οι Delen et al.(2004) κατέληξαν σε ένα τελικό dataset το οποίο αποτελείται από 202,932 εγγραφές, αριθμός αρκετά μεγαλύτερος από αυτόν που παρουσιάστηκε στη παρούσα πτυχιακή εργασία (165,725), παρόλο που το αρχικό dataset που χρησιμοποίησαν ήταν πολύ μικρότερο (433,272<<630,218). Επίσης, η μεγάλη διαφορά εντοπίζεται στον διαχωρισμό που πραγματοποίησαν σύμφωνα με το κατώφλι που χρησιμοποιήσαν, καθώς οι εγγραφές που ταξινομήθηκαν ως survived είναι 93,273 ενώ αυτές που ταξινομήθηκαν ως not survived 109,659. Δηλαδή, αποτελέσματα τελείως αντίθετα από αυτά της πτυχιακης εργασίας και της έρευνας των Bellaachia & Guven (2006). Αυτό οφείλεται, όπως έχει αναφερθεί, στη λάθος προσέγγιση που έγινε κατά τη φάση της προ-έπεξεργασίας που και παρέλειψαν να λάβουν υπόψην τους δύο πολύ σημαντικά γνωρίσματα για τη δημιουργία του κατωφλίου.

	Σύνολο εγγραφών	Ποσοστό
Survived (1)	93,273	46,00%
Not Survived (0)	109,659	54,00%
Total	202,932	100%

Πίνακας 7. Κατηγοριοποίηση εγγραφών σύμφωνα με το κατώφλι. [Πηγή: Delen et al.(2004)]

Έτσι ολοκληρώνεται και το έκτο στάδιο της σχεδίασης της τεχνικής της εξόρυξης γνώσης, φτάνωντας στο έβδομο στάδιο όπου και γίνεται η επιλογή του κατάλληλου

ταξινομητή κάθε φορά. Το συγκεκριμένο στάδιο λειτουργεί παράλληλα με τα ενδιάμεσα στάδια του pre-processing των δεδομένων. Στη συγκεκριμένη έρευνα έχουν επιλεγεί τρείς διαφορετικοί ταξινομητές:

- Τα δέντρα απόφασης
- Οι meta-classifiers και
- Οι συναρτησιακοί ταξινομητές.

Στο επόμενο στάδιο, το όγδοο, γίνεται η εξαγωγή του τελικού συνόλου των δεδομένων σε τύπο αρχείου .arff. Στη συνέχεια, τα δεδομένα αυτά θα περάσουν στη διαδικασία της κατηγοριοποίησης, που αποτελεί και το ένατο στάδιο της σχεδίασης, ανάλογα με τις τεχνικές που έχουν επιλεχθεί. Τέλος η αξιολόγηση των αποτελεσμάτων και η γνώση που προέκυψε αποτελούν το δέκατο και το ενδέκατο στάδιο της σχεδίασης αντίστοιχα.

4. Υλοποίηση

4.1 Μοντέλα πρόβλεψης.

Για την απόκτηση γνώσης από το dataset που παρέχεται από το SEER και αφορά δεδομένα ασθενών που έχουν προσβληθεί από καρκίνο του μαστού αποφασίστηκε να χρησιμοποιηθέι η τεχνική της κατηγοριοποίησης (classification). Για να γίνει αυτό χρησιμοποιήθηκαν οι παρακάτω τέσσερεις μέθοδοι :

- Δέντρα απόφασης
- Functions
- Support Vector Machines
- Meta-classifiers.

Πιο συγκεκριμένα οι αλγόριθμοι που χρησιμοποιήθηκαν για κάθε μέθοδο είναι ο j48 για τα δέντρα απόφασης, ο libSVM που αποτελεί μία υλοποίηση για τα Support Vector Machines για το πρόγραμμα Weka, η Logistic Regression για τις Functions και τέλος ο Vote με συνδυασμό J48 και Logistic Regression για τους meta-classifiers.

Ο αλγόριθμος **J48** που υπάρχει στις βιβλιοθήκες του Weka είναι ένας αλγόριθμος για την κατασκευή δέντρων απόφασης. Τα δέντρα απόφασης είναι τεχνικές κατηγοριοποίησης οι οποίες έγιναν αρκετά δημοφιλείς, με την ανάπτυξη της εξόρυξης γνώσης στο πεδίο των πληροφοριακών συστημάτων. Ο J48 βασίζεται στον αλγόριθμο C4.5 ο οποίος είναι «απόγονος» των αλγορίθμων CLS και ID3 του Quinlan, (1993). Όπως οι CLS και ID3 έτσι και ο C4.5 κατασκευάζει ταξινομητές που παριστάνονται ως δέντρα απόφασης, ωστόσο μπορεί να κατασκευάσει και ταξινομητές με ένα πιο κατανοητό σύνολο κανόνων (Ghosh et al., 2008). Ο C4.5 για ένα δεδομένο σύνολο δεδομένων S με n εγραφές, πρώτα δημιουργεί ένα αρχικό δέντρο με την στρατηγική διαίρει και βασίλευε όπως φαίνεται παρακάτω σύμφωνα με τους Ghosh et al., (2008) :

1. Αν όλες οι εγγραφές του S ανήκουν στην ίδια κλάση ή το S είναι πολύ μικρό, τότε το δέντρο είναι φύλλο και χαρακτηρίζεται από τη πιο συχνή κλάση του S.
2. Αλλιώς, πραγματοποιεί έναν έλεγχο-γνωρίσματος με δύο ή περισσότερα αποτελέσματα. Στη συνέχεια τοποθετεί ως ρίζα του δέντρου το συγκεκριμένο έλεγχο με κάθε διακλάδωση να αποτελεί ένα αποτέλεσμα αυτού του ελέγχου, χωρίζοντας το S σε μικρότερα υποσύνολα S_1, S_2,……. ανάλογα με το αποτέλεσμα για κάθε εγγραφή του S, και στη συνέχεια πραγματοποιεί το ίδιο για τα υπόλοιπα υποσύνολα.

Υπάρχουν διάφοροι έλεγχοι οι οποίοι μπορούν να πραγματοποιηθούν στο δεύτερο βήμα. Σύμφωνα με τους Ghosh et al., (2008), ο C4.5 χρησιμοποιεί δύο ευριστικές μεθόδους για να κατατάξει τους πιθανούς ελέγχους. Η μία μέθοδος είναι το κέρδος πληροφορίας (information gain), η οποία μειώνει τη συνολική εντροπία των υποσυνόλων S_i (ωστόσο είναι αρκετά μεροληπτική μέθοδος όταν ο έλεγχος περιέχει πολλά αποτελέσματα). Η άλλη μέθοδος διαιρεί το κέρδος πληροφορίας με τη πληροφορία που παρέχεται από τα αποτελέσματα του ελέγχου. Ο αλγόριθμος C4.5 μπορεί να χειρίζεται και αριθμητικά και κατηγορικά γνωρίσματα, που είναι ένα πολύ σημαντικό πλεονέκτημα καθώς στη συγκεκριμένη έρευνα χρησιμοποιούνται και οι δύο τύποι γνωρισμάτων. Επίσης για να αποφύγει την υπερ-εκπαίδευση (overfitting), ο C4.5 βελτιστοποιεί το αρχικό δέντρο χρησιμοποιώντας έναν αλγόριθμο βελτιστοποίησης. Ο αλγόριθμος βελτιστοποίησης βασίζεται στον υπολογισμό του ρυθμού λάθους (error rate) που σχετίζεται με ένα σύνολο από N εγγραφές, από τις οποίες E δεν ανήκουν στη πιο συχνή κατηγορία-κλάσση (Ghosh et al., 2008). Η βελτιστοποίηση πραγματοποιείται από τη ρίζα μέχρι τα φύλλα του δέντρου.

Μία άλλη τεχνική κατηγοριοποίησης που χρησιμοποιήθηκε είναι αυτή με τον αλγόριθμο που αφορά το στατιστικό μοντέλο **Logistic Regression.** Η συγκεκριμένη παλινδρόμηση αποτελεί μία γενίκευση της γραμμικής παλινδρόμησης (Hastie et al, 2001). Επίσης είναι ένα από τα πιο διαδεδομένα στατιστικά μοντέλα για την επεξεργασία δεδομένων με πολλαπλές μεταβλητές, ιδιαίτερα στο τομέα της ιατρικής σύμφωνα με τους Bilge et al, (2009). Το συγκεκριμένο στατιστικό μοντέλο είναι

διαδεδομένο για την εξόρυξη γνώσης από ιατρικά δεδομένα για τους παρακάτω λόγους σύμφωνα με τους (Ostir & Uchida, 2000):

- Αρχικά τα μοντέλα στατιστικής παλινδρόμησης, βοηθούν στην ερμηνεία και τη συγκέντρωση των γεγονότων, γεγονός που επιτρέπει στους ερευνητές να εξάγουν συμπεράσματα για τα υπό μελέτη φαινόμενα.
- Επίσης η στατιστική παλινδρόμηση είναι πολύ χρήσιμη όταν το αποτέλεσμα τις έρευνας είναι δυαδικό και όχι συνεχές και μπορεί να δώσει μια λογική απάντηση σε ένα ερευνητικό ερώτημα. Για παράδειγμα, στη παρούσα έρευνα η μεταβλητή στην οποία κατηγοριοποιούνται οι εγγραφές (SurvivalBinary) είναι κατηγορική και συγκερκιμένα δυαδική, παίρνει τιμές 0,1.
- Τέλος, σε μία έρευνα τα γνωρίσματα μπορεί να είναι είτε κατηγορικά είτε αριθμητικά ή ένας συνδυασμός και των δύο. Η στατιστική παλινδρόμηση έχει το πλεονέκτημα να μπορεί να χείρίζεται και τους δύο τύπους γνωρισμάτων ταυτόχρονα.

Ένας άλλος αλγόριθμος που χρησιμοποιήθηκε είναι ο **libSVM** (library for Support Vector Machines), ο οποίος είναι ένας αλγόριθμος υλοποίησης των SVM και χρησιμοποιείται ευρέως σε διάφορα πεδία της εξόρυξης γνώσης. Τα Support Vector Machines είναι διάσημες μέθοδοι μηχανικής μάθησης για κατηγοριοποίηση των δεδομένων (Chang & Lin, 2011). Μπορούν να προσφέρουν μια από τις πιο ακριβείς και ισχυρές μεθόδους κατηγοριοποίησης σε σχέση με πολλούς άλλους διαφορετικούς αλγορίθμους (Ghosh et al., 2008). Μια τυπική χρήση του libSVM περιλαμβάνει δύο στάδια, πρώτα την εκπαίδευση ενός συνόλου δεδομένων έτσι ώστε να αποκτηθεί ένα μοντέλο και κατά δεύτερο λόγο τη χρησιμοποίηση του μοντέλου για τη πρόβλεψη πληροφοριών από τον έλεγχο ενός συνόλου δεδομένων (Chang & Lin, 2011). Ο libSVM χρησιμοποιεί δύο μορφές των Support Vector Machines για κατηγοριοποίηση :

1. C-Support Vector Classification
2. *ν*-Support Vector Classification

Επίσης, όπως αρχικά έχει αναφερθεί, για τη συγκεκριμένη έρευνα χρησιμοποιήθηκε και ensemble μέθοδος για την κατηγοριοποίηση των δεδομένων. Ο αλγόριθμος που χρησιμοποιήθηκε για αυτή τη μέθοδο είναι ο **Vote.** Ο αλγόριθμος Vote συνδυάζει διάφορους αλγορίθμους ταξινόμησης για την εξαγωγή κάποιου αποτελέσματος. Στην έρευνα, οι αλγόριθμοι που χρησιμοποιήθηκαν για την υλοποίηση του Vote είναι ο J48 και ο Logistic Regression. Οι αλγόριθμοι J48 και ο Logistic Regression επεξεργάζονται το σύνολο των δεδομένων και καταλήγουν σε διάφορεςς προβλέψεις. Οι προβλέψεις αυτές ύστερα θα αξιολογηθούν από τον αλγόριθμο Vote για να καταλήξει αυτός στο τελικό αποτέλεσμα. Ωστόσο για την επιλογή των κατάλληλων προβλέψεων από τον συνδυασμό των δύο παραπάνω αλγορίθμων ο Vote χρησιμοποιεί τις παρακάτω τεχνικές.

- Average of probabilities
- Product of probabilities
- Majority Voting
- Minimum probability
- Maximum probability και
- Median.

Έπειτα από την εφαρμογή αυτών των τεχνικών ο Vote είναι έτοιμος να κάνει την τελική πρόγνωση.

Ένα άλλο σημαντικό μέρος της έρευνας είναι η μείωση των εγγραφών του συνόλου των δεδομένων αλλά και η εξάλλειψη των ελλειπών τιμών. Για την υλοποίηση αυτής της επεξεργασίας χρησιμοποιήθηκαν οι δύο παρακάτω λειτουργίες του προγράμματος Weka.

1. Η RemoveWithValues και
2. η ReplaceMissingValues

Η εντολή RemoveWithValues αφαιρεί όλες τις εγγραφές που έχουν ελλειπείς τιμές σε συγκεκριμένα γνωρίσματα, τα οποία τα επιλέγει ο χρήστης. Επίσης αφαιρούνται και οι εγγραφές οι οποίες έχουν γνωρίσματα με ελλειπείς τιμές που είναι γειτονικά στο

αρχικά επιλεγμένο γνώρισμα. Έτσι με τη συγκεκριμένη εντολή είναι εφικτή η μείωση του μεγέθους του αρχικού συνόλου των δεδομένων.

Η εντολή ReplaceMissingValues αφορά και αυτή την επεξεργασία ελλειπών τιμών των γνωρισμάτων ωστόσο δεν αφαιρεί τις εγγραφές που παρουσιάζουν ελλειπείς τιμές σε συγκεκριμένα γνωρίσματα. Η συγκεκριμένη εντολή αντικαθιστά τις ελλειπείς τιμές των αριθμητικών και κατηγορικών γνωρισμάτων, με τη μεσαία τιμή καθώς και με τον αριθμό που εμφανίζεται περισσότερο στο σύνολο των τιμών των δεδομένων εκπαίδευσης.

4.2 Μετρικές για την αξιολόγηση των αποτελεσμάτων.

Αφού διεξαχθεί μία έρευνα και παραχθούν κάποια συγκεκριμένα αποτελέσματα είναι απαραίτητο τα αποτελέσματα αυτά να αξιολογηθούν με διάφορες διαθέσιμες μετρικές. Υπάρχουν πολλά μέτρα για την αξιολόγηση της απόδοσης των τεχνικών που χρησιμοποιήθηκαν, από τα οποία τα πιο σημαντικά φαίνονται παρακάτω :

- Το accuracy. Το συγκεκριμένο μέτρο είναι ένα από τα πιο συνηθισμένα για την αξιολόγηση αποτελεσμάτων για έρευνες σχετικές με την εξόρυξη δεδομένων. Ορίζεται ως το σύνολο των ορθά κατηγοριοποιημένων δεδομένων πρός το συνολικό αριθμό των δεδομένων και ο τύπος είναι ο παρακάτω και προκύπτει από το confusion matrix (βλ. Πίνακας 8):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

  Όπου TP= True Positives, TN= True Negatives, FP= False Positives, FN= False Negatives.

- Το precision. Δωθέντος του συνόλου των δεδομένων των ασθενών με καρκίνο του μαστού που έχει κατηγοριοποιηθεί, η precision(ακρίβεια) ορίζεται ώς οι εγγραφές που έχουν κατηγοριοποιηθεί ορθά από το

ταξινομητή είναι πραγματικά ορθές και δίνετε από το παρακάτω τύπο που προκύπτει από το confusion matrix (βλ. Πίνακας 8) :

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

- Το recall. Όμοια, δωθέντος του συνόλου των δεδομένων των ασθενών με καρκίνο του μαστού που έχει κατηγοριοποιηθεί, η recall (ανάκληση) ορίζεται ως ο αριθμός των ορθών εγγραφών που έχει κατορθώσει να βρει ο ταξινομητής και δίνεται από τον παρακάτω τύπο όπου όμοια και αυτός προκύπτει από το confusion matrix (βλ. Πίνακας 8):

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Όσο μεγαλύτερη είναι η ανάκληση τόσο λιγότερες ορθές εγγραφές θα έχουν ταξινομηθεί λάθος. Το recall αναφέρεται και ως Sensitivity.

- Το F-Measure. Το μέτρο F-Measure χρησιμοποιεί και το precision και το recall για τον υπολογισμό του αποτελέσματος (Gaussier & Goutte, 2005) και είναι το αρμονικό μέσο αυτών των δύο μετρικών και δίνεται από το παρακάτω τύπο:

$$\mathrm{F-Measure} = 2 \times \frac{Recall \times Precision}{TP + FN}$$

ή σύμφωνα με τους τύπους (2) και (3):

$$\mathrm{F-Measure} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

Το F-Measure είναι ένα σύνθετο μέτρο που είναι καλύτερο για αλγορίθμους με μεγαλύτερο recall και συναντά δυσκολίες με αλγορίθμους που έχουν μεγαλύτερη specificity (Japkowicz et al, 2006). Γενικά μία υψηλή τιμή στο F-Measure σημαίνει ότι τα precision και recall είναι αρκετά μεγάλα.

- Το specificity. Το συγκεκριμένο μέτρο, ορίζεται ως το ποσοστό των λανθασμένων εγγραφών οι οποίες έχουν κατηγοριοποιηθεί από το ταξινομητή ώς ορθές και δίνεται από τον παρακάτω τύπο (Kapp et al, 2005):

$$Specificity = \frac{TN\ (True\ Negatives)}{TN(True\ Negatives) + FP\ (False\ Positives)} \quad (2)$$

Στη συγκεκριμένη έρευνα τα μέτρα που χρησιμοποιήθηκαν για την αξιολόγηση των αποτελεσμάτων είναι το Accuracy, το Precision, το Recall ή Sensitivity και το Specifity.

Επίσης για την καλύτερη αξιοπιστία και επίτευξη καλύτερων αποτελσμάτων κατά τη φάση της κατηγοριοποίησης χρησιμοποιήθηκε η μέθοδος percentage split με ποσοστό 66%. Σύμφωνα με αυτή τη μέθοδο, ένα μέρος του συνολικού dataset χρησιμοποιείται ώς σέτ εκπαίδευσης και το υπόλοιπο μέρος χρησιμοποιείται για επαλήθευση. Για παράδειγμα αν έχουμε 100 εγγραφές και χρησιμοποιήσουμε percentage split 66% οι πρώτες 66 εγγραφές θα χρησιμοποιηθούν για εκπαίδευση και οι εγγραφές από τη θέση 67 και πάνω θα χρησιμοποιηθούν για επαλήθευση. Η συγκεκριμένη μέθοδος χρησιμοποιήθηκε για όλους τους αλγορίθμους της έρευνας.

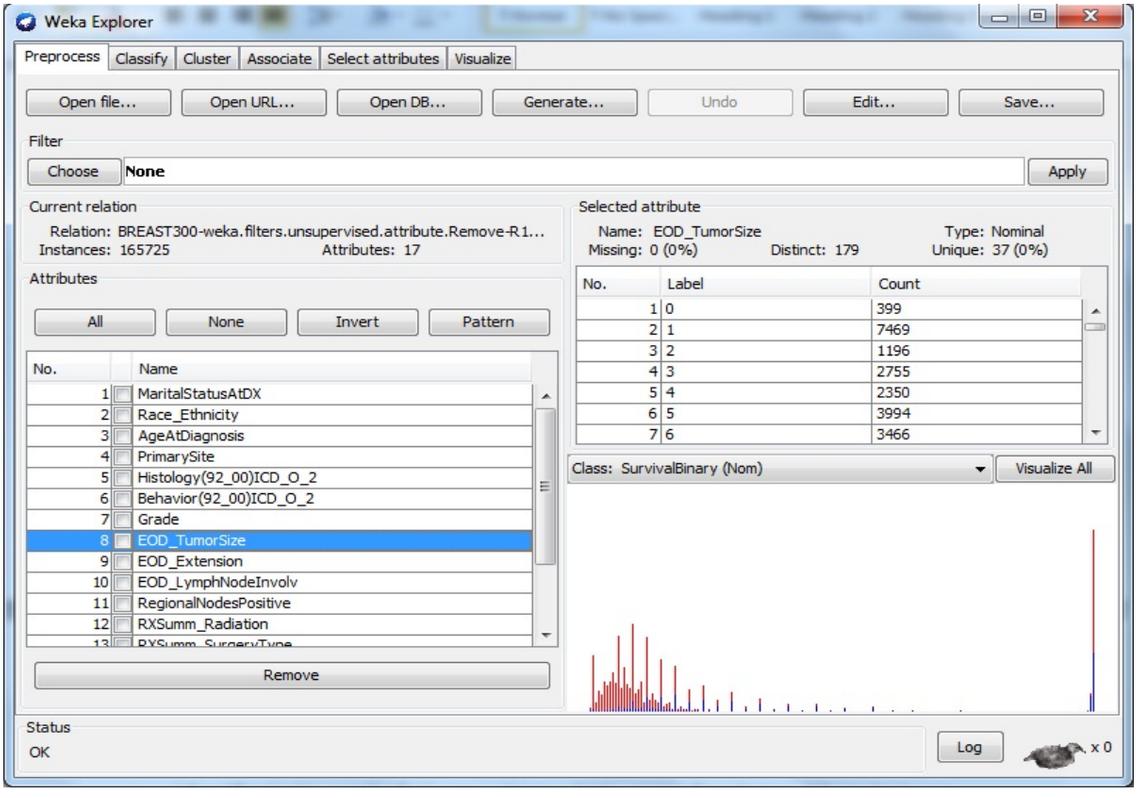
	Actual Class	
Predicted Class	TP (True Positives)	FP (False Positives)
	FN (False Negatives)	TN (True Negatives)

Πίνακας 8. Confusion Matrix

4.3 Το πρόγραμμα Weka.

Το πρόγραμμα Weka (Waikato Environment for Knowledge Analysis) είναι ένα ανοιχτού κώδικα λογισμικό το οποίο δημιουργήθηκε το 1992 εξαιτίας της αισθητής ανάγκης για ένα πρόγραμμα το οποίο θα επιτρέπει στους ερευνητές να

χρησιμοποιήσουν τεχνολογικά εξελιγμένες τεχνικές μηχανικής μάθησης (Frank et al, 2009). Σήμερα το Weka είναι αναγνωρισμένο ως ένα από τα πιό σημαντικά συστήματα για τις μεθόδους εξόρυξης δεδομένων και μηχανικής μάθησης και ένα από τα πιό σημαντικά πλεονεκτήμτά του είναι η ελευθερία που δίνεται στους χρήστες να επεξεργάζονται το πηγαίο κώδικα του προγράμματος με σκοπό τη δημιουργία νέων τεχνικών και αλγορίθμων έτσι ώστε να ανανεώνεται το πρόγραμμα (Piatetsky-Shapiro, 2005). Η κύρια μορφή αρχείων που υποστηρίζεται από το Weka είναι η .arff ωστόσο υποστηρίζονται και αρχεία της μορφής .csv. Στην αρχική οθόνη του Weka (βλ. Εικόνα 4.1) παρουσιάζονται πολλές χρήσιμες πληροφορίες για το αρχείο που χρησιμοποιείται στην έρευνα όπως ο αριθμός των εγγραφών, των γνωρισμάτων, των ελλειπών τιμών καθώς και επίσης η ανάλυση των τιμών και το υ τύπο υ κάθε γνωρίσματος.



Εικόνα 4.1. Αρχική οθόνη του προγράμματος Weka.

Το Weka περιλαμβάνει πολλούς αλγορίθμους για τεχνικές κατηγοριοποίησης (classification), παλινδρόμησης (regression), συσταδοποίησης (clustering) και

κανόνων συσχέτισης (association rules). Μερικοί από τους πιο σημαντικούς αλγορίθμους που περιέχονται στο Weka είναι :

- J48 και ID3 (Decision Trees)
- Naïve Bayes
- Logistic και Linear Regression
- Meta-classifiers όπως ο Vote, ο Bagging και ο adaBoost.
- k-NN για τα νευρωνικά δίκτυα.

Επίσης το συγκεκριμένο πρόγραμμα προσφέρει ένα μεγάλο σύνολο τεχνικών για την προ-επεξεργασία των δεδομένων, που είναι και η πιο σημαντική φάση της εξόρυξης γνώσης. Μερικές τεχνικές που πραγματοποιούνται για preprocessing των δεδομένων φαίνονται παρακάτω :

- NumericToNominal: η οποία μετατρέπει τα τα αριθμητικά γνωρίσματα σε κατηγορικά
- NominalToBinary: η οποία μετατρέπει ένα κατηγορικό γνώρισμα σε δυαδικό.
- Descritize: η οποία πραγματοποιεί διακριτοποίηση των δεδομένων
- RandomSubset: η συγκεκριμένη διαλέγει έναν τυχαίο αριθμό γνωρισμάτων
- RemoveWithValues: που αφαιρεί εγγραφές οι οποίες περιέχουν ελλειπείς τιμές σε συγκεκριμένα γνωρίσματα.

Τέλος, το πρόγραμμα Weka προσφέρει ένα ακόμη πολύ χρήσιμο πεδίο για την προεπεξεργασία των δεδομένων, το οποίο ονομάζεται Select Attributes και αφορά τη επιλόγη κατάλληλων γνωρισμάτων για την επίτευξη καλύτερων αποτελεσμάτων πρόβλεψης. Το συγκεκριμένο πεδίο δίνει τη δυνατότητα στους χρήστες να χρησιμοποιήσουν διάφορους αλγορίθμους, οι οποίοι μπορούν να επιλέξουν τα καταλληλότερα attributes κάθε φορά (Frank et al, 2009).

5. Αποτελέσματα

5.1 Αποτελέσματα αξιολόγησης με βάση την αρχική επιλογή γνωρισμάτων

Στη παρούσα έρευνα τα μοντέλα εξόρυξης γνώσης που χρησιμοποιήθηκαν, αξιολογήθηκαν με βάση συγκεκριμένα μέτρα απόδοσης. Τα μέτρα αυτά είναι η πιστότητα (accuracy), η ακρίβεια (precision), η ανάκληση ή ευαισθησία (recall or sensitivity) και η ιδιαιτερότητα (specificity). Όλοι οι αλγόριθμοι αξιολογήθηκαν με τα ίδια μέτρα απόδοσης. Τα αποτελέσματα αναπαριστούνται σε πίνακες. Οι πίνακες των αποτελεσμάτων παρέχουν πληροφορίες όπως τα ποσοστά των μέτρων αξιολόγησης, την κατηγορία συμφωνα με την οποία ταξινομήθηκαν οι εγγραφές (class), την τεχνική τις επαλήθευσης πουχρησιμοποιήθηκε (percentage split) και το confusion matrix (βλ. Πίνακας 8). Το confusion matrix είναι ένας πίνακας για την αναπαράσταση των αποτελεσμάτων της κατηγοριοποίησης (Delen et al, 2004). Το πάνω αριστερό μέρος του πίνακα αναπαριστά τα True Positives (TP) τα οποία αφορούν τις εγγραφές που έχουν κατηγοριοποιηθεί σωστά ώς ορθές. Το πάνω δεξιό μέρος αναπαριστά τα False Positives (FP) τα οποία αναφέρονται στις εγγραφές οι οποίες ταξινομήθηκαν ως ορθές ενώ ήταν λανθασμένες. Το κάτω αριστερά κελί του πίνακα περιέχει τα False Negatives (FN) που αφορούν τις εγγραφές που ταξινομήθηκαν λανθασμένα ενώ ήταν ορθές και το κάτω δεξιό κελί του πίνακα περιέχει τα True Negatives τα οποία αναφέρονται στις εγγραφές που ταξινομήθηκαν που δεν ήταν ορθές και ταξινομήθηκαν λανθασμένα. Αρχικά τα δεδομένα εκπαιδεύτηκαν με τον αλγόριθμο των δέντρων απόφασης J48. Σύμφωνα με την αξιολόγηση που πραγματοποιήθηκε τα αποτελέσματα είναι αρκετά ενθαρυντικά. Ο J48, λοιπόν, υλοποίησε την κατηγοριοποίηση των δεδομένων με accuracy= 87.8%, precision= 87.3%, recall= 87.8% και specificity= 95,7% όπως φαίνεται και στο Πίνακας 9.

Percentage Split	Class	Decision Trees (J48)					
		Confusion Matrix		Accuracy	Precision	Recall	Specificity
66%	0	7.453	5.981	-	79.8%	59.9%	-
	1	1.882	42.030	-	89.4%	95.7%	-
Weighted Average				87.8%	87,3%	87.8 %	95.7%

Πίνακας 9. Αποτελέσματα εκτέλεσης αλγορίθμου J48

Τα αποτελέσματα είναι χαμηλότερα από αυτά που προέκυψαν στην έρευνα των Delen et al., (2004). Για τον αλγόριθμο J48 οι Delen et al., (2004) προέβλεψαν ότι ταξινόμηση πραγματοποιήθηκε με accuracy= 93,6%, recall= 96,0% και specificity= 95,7% όπως φαίνεται στο Πίνακας 10. Λαμβάνοντας υπόψη όμως πώς οι Delen et al., (2004) δεν πραγματοποίησαν τις απαραίτητες ενέργειες για τη σωστή προεπεξεργασία των γεγονότων οδηγούμαστε στο συμπέρασμα πώς τα αποτελέσματά τους δεν είναι άμεσα συγκρίσιμα με τα δικά μας.

Fold cross validation	Decision Trees (J48)		
	Accuracy	Recall	Specificity
10	93.6%	96,0%	95.7%

Πίνακας 10. Αποτελέσματα της εκτέλεσης του J48 της έρευνας των Delen et al, (2004)

Παρόλα αυτά, συγκρίνοντας τα αποτελέσματα της παρούσας έρευνας για τον J48 με αυτά στην έρευνα των Bellaachia & Guven (2006) (βλ. Πίνακας 11), που πραγματοποίησαν παρόμοια διαδικασία προεπεξεργασίας, διακρίνεται μία βελτίωση στην αποδοτικότητα του αλγορίθμου σε ολα τα μέτρα αξιολόγησης. Οι Bellaachia & Guven (2006) πρόεβλεψαν accuracy= 86.7%, recall= 87.0% και precision= 86.2%.

Fold cross validation	Decision Trees (J48)		
	Accuracy	Recall	Precision
10	86.7%	87,0%	86,2 %

Πίνακας 11. Αποτελέσματα της εκτέλεσης του J48 της έρευνας των Bellaachia & Guven (2006)

Στη συνέχεια για την εκπαίδευση των δεδομένων χρησιμοποιήθηκε ο αλγόριθμος της στατιστικής παλινδρόμησης (Logistic Regression). Σύμφωνα με τα μέτρα αξιολόγησης που χρησιμοποιήθηκαν ο αλγόριθμος Logistic Regression παρουσίασε accuracy= 88,3%, recall= 88,4%, precision= 88,0% και specificity= 96,4% (βλ. Πίνακας 12).

Percentage Split	Class	Functions (Logistic Regression)					
		Confusion Matrix		Accuracy	Precision	Recall	Specificity
66%	0	7.432	5.002	-	82.7%	59.8%	-
	1	1.551	42.361	-	89.4%	96.5%	-
Weighted Average				88.3%	88,0%	88.4%	96.4%

Πίνακας 12. Αποτελέσματα εκτέλεσης αλγορίθμου Logistic Regression

Τα αποτελέσματα αυτής της εκτέλεσης είναι και αυτά μικρότερα από τα αποτελέσματα που είχαν προβλέψει οι Delen et al., (2004). Οι Delen et al., (2004), προέβλεψαν για τον αλγόριθμο Logistic Regression accuracy= 89,2%, recall= 90,1% και specificity= 87,8% (βλ. Πίνακας 13). Ωστόσο για τους λόγους πού έχουν αναφερθεί και στις προηγούμενες ενότητες τα αποτελέσματα τους δεν είναι άμεσα συγκρίσιμα με τα δικά μας. Οι Bellaachia & Guven (2006) στην έρευνα τους δεν χρησιμοποίησαν τον αλγόριθμο Logistic Regression, αλλά προτίμησαν να χρησιμοποιήσουν τον αλγόριθμό Naive Bayes, για τον οποίο προέβλεψαν accuracy= 84,5%, recall= 85,2%, precision= 83,6% (βλ. Πίνακας 14).

Fold cross validation	Logistic Regression		
	Accuracy	Recall	Specificity
10	89,2%	90,1%	87,8%

Πίνακας 13. Αποτελέσματα της εκτέλεσης του Logistic Regression της έρευνας των Delen et al, (2004)

Fold cross validation	Naives Bayes		
	Accuracy	Recall	Precision
10	84,5%	85,2%	83,6 %

Πίνακας 14. Αποτελέσματα της εκτέλεσης του Naive Bayes της έρευνας των Bellaachia & Guven (2006)

Όπως έχει αναφερθεί οι Delen et al., (2004) και οι Bellaachia & Guven, (2006) χρησιμοποίησαν και ένα τρίτο αλγόριθμο ταξινόμησης ο οποίος αφορά τα τεχνητά νευρωνικά δίκτυα (βλ. Πίνακας 15 και Πίνακας 16).

Fold cross validation	Artificial Neural Networks (ANN)		
	Accuracy	Recall	Specificity
10	89,2%	90,1%	87,8%

Πίνακας 15. Αποτελέσματα της εκτέλεσης των ANN της έρευνας των Delen et al, (2004)

Fold cross validation	Artificial Neural Networks (ANN)		
	Accuracy	Recall	Precision
10	86,5%	85,2%	83,6 %

Πίνακας 16. Αποτελέσματα της εκτέλεσης των ANN της έρευνας των Bellaachia & Guven (2006)

Ωστόσο στη παρούσα πτυχιακή εργασία αποφασίστηκε να χρησιμοποιηθούν δύο αλγόριθμοι η οποίοι εφαρμόζονται για πρώτη φορά στην εξόρυξη γνώσης από δεδομένα καρκίνου του μαστού. Ο πρώτος αλγόριθμος είναι ο libSVM που αφορά τα

Support Vector Machines. Ο libSVM σύμφωνα με τα μέτρα αξιολόγησης που χρησιμοποιήθηκαν παρουσίασε accuracy= 87.9%, recall= 87.9%, precision= 87,3% και specificity= 96.2% όπως φαίνεται και στο Πίνακας 17.

Percentage Split	Class	Support Vector Machines (libSVM)					
		Confusion Matrix		Accuracy	Precision	Recall	Specificity
10%	0	11.803	9.315	-	79.4%	55.9%	-
	1	3.068	78.203	-	89.4%	96.2%	-
Weighted Average				87.9%	87,3%	87.9%	96.2%

Πίνακας 17. Αποτελέσματα εκτέλασης αλγορίθμου libSVM

Ο δεύτερος αλγόριθμος είναι ο Vote, ο οποίος είναι ένας συνδυαστικός αλγόριθμος μεθόδων (ensemble). Οι ταξινομητές που χρησιμοποιήθηκαν στον αλγόριθμο Vote είναι ο Logistic Regression και ο J48. Τα αποτελέσματα της αξιολόγησης φαίνονται στο. Ο Vote είχε accuracy= 88.4%, recall= 88.4%, precision= 88.0%% και specificity= 96.5% (βλ. Πίνακας 18)

Percentage Split	Class	Ensemble Methods [Vote(J48,Logistic)]					
		Confusion Matrix		Accuracy	Precision	Recall	Specificity
66%	0	7.404	5.030	-	83.1%	59.5%	-
	1	1.502	42.410	-	89.4%	96.6%	-
Weighted Average				88.4%	88,0%	88.4%	96.5%

Πίνακας 18. Αποτελέσματα εκτέλεσης αλγορίθμου Vote.

Εξετάζοντας τα αποτελέσματα αξιολόγησης διαπιστώθηκε πώς όλοι οι αλγόριθμοι που χρησιμοποιήθηκαν για την κατηγοριοποίηση του συνόλου των δεδομενων στη παρούσα πτυχιακή εργασία είχαν καλύτερα αποτελέσματα από τους αλγορίθμους που χρησιμοποιήθηκαν στην έρευνα των Bellaachia & Guven, (2006) με γνωμονα πάντα την παρόμοια τακτική που ακολουθήθηκε στη φάση της προεπεξεργασίας των δεδομένων. Στην έρευνά τους πιο αποδοτικός αλγόριθμος αποδήχθηκε ο J48, αν τους κατατάξουμε ανάλογα με την πιστότητα τους, με accuracy= 86,7% ποσοστό που είναι μικρότερο από αυτό που πέτυχε ο λιγότερο αποδοτικός αλγόριθμος της πτυχιακής εργασίας που είναι ο J48 με accuracy= 87,8%.

Έτσι έπειτα από την κατάταξη των αλγορίθμων που χρησιμοποιήθηκαν στην πτυχιακή εργασία, ανάλογα με την πιστότητα τους (accuracy), προκύπτει ότι ο πιο αποδοτικός αλγόριθμος είναι ο Vote με accuracy= 88,4%, δεύτερος πιο αποδοτικός ο

αλγόριθμος της στατιστικής παλινδρόμησης με accuracy= 88,3%, έπεται ο libSVM με accuracy= 87,9% και τέλος ο J48 με accuracy= 87,8% (βλ. Πίνακας 19).

Αλγόριθμος	Accuracy	Precision	Recall	Specificity
Vote	**88,4%**	**88,0%**	**88.4%**	96.5%
Logistic Regression	88,3%	**88,0%**	**88.4%**	96.4%
libSVM	87,9%	87,3%	87.9%	96.2%
J48	87,8%	87,3%	87.8 %	95.7%

Πίνακας 19. Κατάταξη των αλγορίθμων κατηγοριοποίησης σύμφωνα με την πιστότητα τους (accuracy)

5.2 Αποτελέσματα με βάση το Attribute Selection

1. **CfsSubsetEval – GeneticSearch**

   Αρχικά έγινε η επιλογή των γνωρισμάτων με τις μεθόδους CfsSubsetEval και GeneticSearch, όπου και προέκυψαν 32 γνωρίσματα. Στη συνέχεια εφαρμόστηκαν οι αλγόριθμοι J48, Logistic Regression και Vote. Πρώτα χρησιμοποιήθηκε ο αλγόριθμος J48 τα αποτελέσματα που έδωσε φαίνονται στο Πίνακας 20

Percentage Split	Class	Decision Trees (J48)					
		Confusion Matrix		Accuracy	Precision	Recall	Specificity
66%	0	2.369	1.158	-	81.3%	67.2%	-
	1	544	7.318	-	86.3%	93.1%	-
Weighted Average				85.05%	84,3%	85.1 %	88.2%

Πίνακας 20. Πίνακας αποτελεσμάτων για τον αλγόριθμο J48 που χρησιμοποιήθηκε ύστερα από attribute selection με CfsSubsetEval – GeneticSearch

Στη συνέχεια εφαρμόστηκε ο αλγόριθμος Logistic Regression και τα αποτελέσματα φαίνονται στον Πίνακας 21.

Percentage Split	Class	Functions (Logistic Regression)					
		Confusion Matrix		Accuracy	Precision	Recall	Specificity
66%	0	2.441	1.056	-	80.5%	70.1%	-
	1	600	7.262	-	87.3%	92.4%	-
Weighted Average				85.45%	85.2%	85.5 %	87.3%

Πίνακας 21. Πίνακας αποτελεσμάτων για τον αλγόριθμο Logistic Regression που χρησιμοποιήθηκε ύστερα από attribute selection με CfsSubsetEval – GeneticSearch

Τέλος, χρησιμοποιήθηκε ο αλγόριθμος Vote ο οποίος χρησιμοποιεί παράλληλα τους J48 και Logistic Regression και τα αποτελέσματα εμφανίζονται στον Πίνακας 22.

Percentage Split	Class	Meta [Vote(Logistic Regression – J48]					
		Confusion Matrix		Accuracy	Precision	Recall	Specificity
66%	0	2.396	1.131	-	81.9%	67.9%	-
	1	529	7.333	-	86.6%	93.3%	-
Weighted Average				85.42%	85.2%	85.4 %	88.6%

Πίνακας 22. Πίνακας αποτελεσμάτων για τον αλγόριθμο Vote που χρησιμοποιήθηκε ύστερα από attribute selection με CfsSubsetEval – GeneticSearch

Η ίδια τακτική χρησιμοποιήθηκε και για τις υπόλοιπες τεχνικές επιλογής γνωρισμάτων και τα αποτελέσματα τους φαίνονται συγκεντρωμένα στους παρακάτω πίνακες.

2. **InfoGain – Ranker**

Percentage Split	Class	Decision Trees (J48)					
		Confusion Matrix		Accuracy	Precision	Recall	Specificity
66%	0	2.613	914	-	81.7%	74.1%	-
	1	584	7.278	-	88.8%	92.6%	-
Weighted Average				86.84%	86.6%	86.8 %	88.8%

Πίνακας 23. Πίνακας αποτελεσμάτων για τον αλγόριθμο J48 που χρησιμοποιήθηκε ύστερα από attribute selection με InfoGain – Ranker.

Percentage Split	Class	Functions (Logistic)					
		Confusion Matrix		Accuracy	Precision	Recall	Specificity
66%	0	2.342	1.185	-	77.3%	66.4%	-
	1	689	7.173	-	85.8%	91.2%	-
Weighted Average				83.54%	83.2%	83.5 %	85.8%

Πίνακας 24. Πίνακας αποτελεσμάτων για τον αλγόριθμο Logistic Regression που χρησιμοποιήθηκε ύστερα από attribute selection με InfoGain – Ranker.

Percentage Split	Class	Meta [Vote(Logistic Regression – J48]					
		Confusion Matrix		Accuracy	Precision	Recall	Specificity
66%	0	2.555	972	-	82.3%	72.4%	-
	1	550	7.312	-	88.3%	93.0%	-
Weighted Average				86.63%	86,4%	86.6 %	88.2%

Πίνακας 25. Πίνακας αποτελεσμάτων για τον αλγόριθμο Vote που χρησιμοποιήθηκε ύστερα από attribute selection με InfoGain – Ranker.

3. Chi-squared - Ranker

Percentage Split	Class	Decision Trees (J48)					
		Confusion Matrix		Accuracy	Precision	Recall	Specificity
66%	0	2.581	946	-	81.9%	73.2%	-
	1	572	7.290	-	88.5%	92.7%	-
Weighted Average				86.67%	86.5%	86.7 %	88.5%

Πίνακας 26. Πίνακας αποτελεσμάτων για τον αλγόριθμο J48 που χρησιμοποιήθηκε ύστερα από attribute selection με Chi-squared – Ranker.

Percentage Split	Class	Functions (Logistic)					
		Confusion Matrix		Accuracy	Precision	Recall	Specificity
66%	0	2.388	1.139	-	77.0%	67.7%	-
	1	715	7.147	-	86.3%	90.9%	-
Weighted Average				83.72%	83.4%	83.7 %	86.2%

Πίνακας 27. Πίνακας αποτελεσμάτων για τον αλγόριθμο Logistic Regression που χρησιμοποιήθηκε ύστερα από attribute selection με Chi-squared – Ranker.

Percentage Split	Class	Meta [Vote(Logistic Regression – J48]					
		Confusion Matrix		Accuracy	Precision	Recall	Specificity
66%	0	2.521	1006	-	82.7%	71.5%	-
	1	527	7.335	-	87.9%	93.3%	-
Weighted Average				86.53%	86,3%	86.5 %	87.9%

Πίνακας 28. Πίνακας αποτελεσμάτων για τον αλγόριθμο Vote που χρησιμοποιήθηκε ύστερα από attribute selection με Chi-squared – Ranker.

Κατατάσοντας τους αλγορίθμους με βάση το accuracy βλέπουμε ότι ο πιο αποδοτικός αλγόριθμος είναι ο J48 με βάση τις μεθόδους InfoGain-Ranker για την επιλογή γνωρισμάτων. Δεύτερος πιο αποδοτικός είναι ο J48 για τις μεθόδους Chi-squared – Ranker και τρίτος αποτελεσματικότερος ο Vote για τις InfoGain-Ranker. Λιγότερο αποδοτικός αλγόριθμος εμφανίστηκε ο Logistic Regression των InfoGain-Ranker. Συγκεντρωτικά τα αποτελέσματα των αλγορίθμων των παραπάνω μεθόδων, με βάση το accuracy φαίνονται στο Πίνακας 29.

Συγκρίνοντας τώρα τα αποτελέσματα των αλγορίθμων έπειτα από την εφαρμογή των νέων τεχνικών επιλογής γνωρισμάτων με αυτά του Πίνακας 19, όπου περιέχονται τα 17 γνωρίσματα, παρατηρήθηκε ότι, οι αλγόριθμοι που εφαρμόστηκαν στα δεδομένα με τα νέα γνωρίσματα είναι έλαχιστα λιγότερο αποδοτικοί από αυτούς του Πίνακας 19. Ο πιο αποτελεσματικός αλγόριθμος του Πίνακας 29 είναι ο J48(InfoGain-Ranker) με accuracy= 86.84% σε σχέση με τον με τον πιο αποδοτικό αλγόριθμο του Πίνακας 19 που είναι ο Vote με accuracy= 88,4% και τον λιγότερο αποδοτικό αλγόριθμο του Πίνακας 19 που είναι ο J48 με accuracy= 87,7%. Ωστόσο, συγκρίνοντας τον J48(InfoGain-Ranker) με accuracy= 86.84% (βλ. Πίνακας 29), με τον πιο αποδοτικό αλγόριθμο των Bellaachia & Guven, (2006) που είναι ο J48 με

accuracy= 86,7% (βλ. Πίνακας 11), παρατηρούμε ότι ο J48(InfoGain-Ranker) είναι ελάχιστα καλύτερος. Επίσης είναι σημαντικό να σημειωθεί πως η εφαρμογή των αλγορίθμων έπειτα από την επιλογή των γνωρισμάτων με τις μεθόδους CfsSubsetEval-GeneticSearch, InfoGain-Ranker και Chi-squared-Ranker έγινε σε ένα δείγμα 33.496 εγγραφών του συνόλου των δεδομένων. Αυτό μπορεί να συνεπάγεται και μικρές αποκλίσεις, είτε θετικές είτε αρνητικές, στην αποδοτικότητα των αλγορίθμων που χρησιμοποιήθηκαν.

Μέθοδοι επιλογής γνωρισμάτων	**Αλγόριθμος**	**Accuracy**
CfsSubstEval-GeneticSearch	J48	85.05%
	Logistic Regression	85.45%
	Vote	85.42%
InfoGain-Ranker	J48	**86.84**
	Logistic Regression	83.54%
	Vote	86.63%
Chi-squared-Ranker	J48	86.67
	Logistic Regression	83.72
	Vote	86.53

Πίνακας 29. Συγκεντρωτικός πίνακας αποτελεσμάτων για τις μεθόδους επιλογής γνωρισμάτων.

6. Συμπεράσματα – Μελλοντικές Επεκτάσεις

6.1 Συμπεράσματα

Στη συγκεκριμένη πτυχιακή εργασία πραγματοποιήθηκε μία έρευνα στο τομέα της εξόρυξης γνώσης από ιατρικά δεδομένα που είχε σκοπό την ανάπτυξη διαφόρων μεθόδων πρόβλεψης για την δυνατότητα επιβίωσης των ασθενών που έχουν προσβληθεί από το καρκίνο του μαστού. Πιο συγκεκριμένα χρησιμοποιήθηκαν τέσσερεις μέθοδοι εξόρυξης γνώσης: Decision Trees με τον αλγόριθμο J48, Support Vector Machines όπου χρησιμοποιήθηκε μία υλοποίηση τους, ο libSVM, η στατιστική παλινδρόμηση (Logistic Regression) και τέλος οι ensemble μέθοδοι με τον αλγόριθμο Vote. Αυτές οι μέθοδοι εφαρμόστηκαν πάνω σε ένα σύνολο δεδομένων ασθενών με καρκίνο του μαστού του SEER, το οποίο περιείχε 630,218 εγγραφές και 124 γνωρίσματα και είναι διαθέσιμο έπειτα από ειδική άδεια που πρέπει να διατεθεί στον ερευνητή.

Ωστόσο, για να χρησιμοποιηθούν τα δεδομένα αυτά από τις μεθόδους εξόρυξης γνώσης έπρεπε πρώτα να υποστούν μια προεπεξεργασία, η οποία και απαιτούσε το περισσότερο χρόνο για την υλοποίηση της συγκεκριμένης έρευνας. Συμφωνα με τη προεπεξεργασία που πραγματοποιήθηκε, δημιουργήθηκε ένα κατώφλι, με μία δυαδική μεταβλητή-κλάση, για να χωριστούν οι ασθενείς σε αυτούς που επιβίωσαν και σε αυτούς που δεν επιβίωσαν αφού είχαν προσβληθεί από το καρκίνο του μαστού. Έτσι, σύμφωνα με το κατώφλι, οι ασθενείς που επιβίωσαν 5 χρόνια (60 μήνες), αφού είχε διαγνωσθεί σε αυτούς ο καρκίνος, και βρίσκονται ακόμα εν ζωή ταξινομούνται στη κατηγορία αυτών που επιβίωσαν και όσοι επιβίωσαν για λιγότερο από 5 χρόνια και η αιτία θανάτου τους ήταν ο καρκίνος του μαστού ταξινομούνται στη κατηγορία αυτών που δεν επιβίωσαν. Οι υπόλοιπες εγγραφές αφαιρέθηκαν από το σύνολο των δεδομένων. Πολύ σημαντικό ρόλο στην προεπεξεργασία των δεδομένων έχει και η επιλογή των κατάλληλων γνωρισμάτων των δεδομένων. Έτσι οι μέθοδοι εξόρυξης γνώσης πραγματοποιήθηκαν στο ίδιο σύνολο εγγραφών αλλά με διαφορετικό σετ γνωρισμάτων κάθε φορά. Πιο συγκεκριμένα χρησιμοποιήθηκαν τέσσερα σετ γνωρισμάτων. Αρχικά χρησιμοποιήθηκε ένα σετ γνωρισμάτων παρόμοιο με αυτό που

χρησιμοποίησαν οι Bellaachia & Guven (2006) και Delen et al.(2004) στις έρευνες τους όπου αποτελούνταν από 17 γνωρίσματα και στη συνέχεια χρησιμοποιήθηκαν τεχνικές επιλογής γνωρισμάτων (attribute selection) μέσα από το πρόγραμμα Weka.

Για αυτό το λόγο τα αποτελέσματα χωρίστηκαν σε δύο κατηγορίες. Στη πρώτη κατηγορία είναι τα αποτελέσματα από τα 17 γνωρίσματα που υπάρχουν και στις παραπάνω έρευνες και στην δεύτερη κατηγορία είναι αυτά από τα γνωρίσματα που επιλέχθηκαν ύστερα από την εφαρμογή του attribute selection. Για τη πρώτη κατηγορία, κατατάσσοντας τους αλγορίθμους που χρησιμοποιήθηκαν για τη πρόβλεψη με βάση το accuracy πιο αποδοτικός αλγόριθμος αποδείχθηκε ο Vote με accuracy= 88.4%, δεύτερος ο Logistic Regression με 88.3% και ακολούθησαν ο libSVM και ο J48 με 87.9% και 87.8% αντίστοιχα. Για τη δεύτερη κατηγορία πιο αποδοτικός αλγόριθμος αποδείχθηκε ο J48 με accuracy= 86,84%, αφού πρώτα εφαρμόστηκε η μέθοδος InfoGain-Ranker για την επιλογή των γνωρισμάτων. Τα αποτελέσματα είναι αρκετά ενθαρυντικά καθώς και σε σχέση με άλλε παρόμοιες έρευνες τα αποτελέσματα των αλγορίθμων είναι αποδοτικότερα.

Έτσι, όπως φαίνεται στην παρούσα πτυχιακή εργασία μπορούν να αναπτυχθούν μοντέλα πρόγνωσης μέσα από τις τεχνικές της εξόρυξης δεδομένων τα οποία μπορούν να προβλέψουν με ακρίβεια περιπτώσεων ασθενών που έχουν προσβληθεί από διάφορους τύπους καρκίνου. Παρόλα αυτά, είναι αναγκαίο ύστερα από την εφαρμογή της εξόρυξης γνώσης και την εξαγωγή των αποτελεσμάτων, τα αποτελέσματα να αξιολογηθούν από ειδικούς στο τομέα της ιατρικής για να υπάρξουν ασφαλέστερα συμπεράσματα.

6.2 Μελλοντικές Επεκτάσεις

Όπως έχει αναφερθέι και στην Ενότητα της Σχεδίασης για την επιλογή των γνωρισμάτων χρησιμοποιήθηκε και η τεχνική attribute ή feature selection μέσα από το πρόγραμμα Weka. Τα αποτελέσματα ήταν αρκετά καλά, αλλά όχι καλύτερα από την αρχική επιλογή των γνωρισμάτων που είχε αποφασιστεί. Ο τομέας, λοιπόν του feature selection έχει ακόμα πολλά περιθώρια βελτίωσης για μελλοντικές έρευνες. Υπάρχουν πάρα πολλές διαθέσιμες τεχνικές και συνδυασμοί αυτών μέσα από το πρόγραμμα Weka αλλά και μέσα από άλλα παρόμοια προγράμματα. Επομένως θα

έιχε ιδαίτερο ενδιαφέρον μελλοντικά άλλοι ερευνητές να χρησιμοποιήσουν την τεχνική του feature selection πάνω σε ιατρικά δεδομένα.

Επιπλέον, στη παρούσα έρευνα χρησιμοποιήθηκε για πρώτη φορά στα δεδομένα αυτά ensemble μέθοδος και συγκεκριμένα τον αλγόριθμο Vote. Γενικότερα, οι ensemble μέθοδοι δεν έχουν εφαρμοστεί σε μεγάλο αριθμό ερευνών εξόρυξης γνώσης από ιατρικά δεδομένα. Έτσι θα είναι πολύ σημαντικό μελλοντικά οι ερευνητές να ασχοληθούν με τη βελτίωση αυτών των μεθόδων, αφενός γιατί στις περισσότερες έρευνες χρησιμοποιούνται σχεδόν οι ίδιες μέθοδοι κάθε φορά, αφετέρου οι αλγόριθμοι των ensemble μεθόδων φαίνονται πιο αποδοτικοί σε ιατρικά δεδομένα σε σχέση με τους συνηθισμένους αλγορίθμους που χρησιμοποιούνται.

7. Βιβλιογραφία

Abernethy, M. (2010) Data mining with WEKA, Part 1: Introduction and regression. http://www.ibm.com/developerworks/opensource/library/os-weka1/index.html

American Cancer Society (ACS), (2012) : http://www.cancer.org/

Anand, S., Rajesh, K. (2012). Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm. International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 2, April 2012

Apte, C., Weiss, S. (1997) Data mining with decision trees and decision rules, Future Generation Computer Systems 13 , pp. 197-210

Aragones, J., Gomez-Ruiz, J., Ramos-Jimenez, G., Munoz-Perez, J., Alba-Conejo, E. (2003). A combined neural network and decision trees model for prognosis of breast cancer relapse, Artificial Intelligence in Medicine 27, pp. 45–63

Bellaachia, A., Guven, E. (2006), Predicting Breast Cancer Survivability Using Data Mining Techniques. 9th Workshop on Mining Scientific and Engineering Datasets in conjunction with the 6th SIAM International Conference on Data Mining (SDM 2006), Saturday, April 22.

Bellazzi, R. and Zupan, B. (2008) Predictive data mining in clinical medicine: Current issues and guidelines, pp. 81-97

Bilge, U., Eken, C., Kartal, M., Eray, O. (2009). Artificial neural network, genetic algorithm, and logistic regression applications for predicting renal colic in emergency settings, Int. J. Emerg. Med. 2:99–105 DOI 10.1007/s12245-009-0103-1

Breast Cancer statistics from Centers for Disease Control and Prevention, http://www.cdc.gov/cancer/breast/statistics/.

Cha, S.H. & Tappert, C., (2009) A Genetic Algorithm for Constructing Compact Binary Decision Trees, Journal Of Pattern Recognition Research 1, pp.1-13

Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., Piatetsky-Shapiro, G., Wang, W., (2004) Data Mining Curriculum: A Proposal, Intensive Working Group of ACM SIGKDD Curriculum Committee

Chan, P. and Stolfo, S. (1993) Meta-learning for multistrategy and parallel learning. In Proc. Second Intl. Work. Multistrategy Learning, pp. 150–165,

Chan, P., Stolfo, S., Prodromidis, A., (2000) Meta-learning distributed data mining systems: Issues and approaches.

Chang C., Lin C., J., (2011) LIBSVM: A Library for Support Vector Machines, ACM Transactions on Intelligent Systems and Technology, Vol. 2, No. 3, Article 27

Cios, K., Moore, W., (2002) Uniqueness of medical data mining, Artificial Intelligence in Medicine 26 pp. 1–24.

Delen, D., (2009) Analysis of cancer data: a data mining approach, Expert Systems, February, Vol. 26, No. 1

Delen, D., Walker, G., Kadam, A., (2004) Predicting breast cancer survivability: A comparison of three data mining methods , Artificial Intelligence in Medicine

Dzeroski, S., Todorovski, L., Ljubic, L., (2004) Inducing Polynomial Equations for Regression, J.-F. Boulicaut et al. (Eds.): ECML, LNAI 3201, pp. 441–452

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., (1996) Data mining and knowledge discovery in databases, Commun. ACM 39, 24–26.

Fradkin, D. and Muchnik, I, (2000) Support Vector Machines for Classification, Mathematics Subject Classification. 62H30.

Frank, E., Holmes, G., Pfahringer, B., Reutemann P., Witten I., Hall, M. (2009) The WEKA Data Mining Software: An Update, ACM SIGKDD Explorations Newsletter, Volume 11 Issue 1

Gaussier, E., Goutte, C., (2005) A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation, D.E. Losada and J.M. Fernandez-Luna (Eds.): ECIR 2005, LNCS 3408, pp. 345–359.

Ghosh, J., Wu, X., Kumar V., Quinlan, J.R., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B., Yu, P., Zhou, Z.H., Steinbach, M., Hand, D., Steinberg, D., (2008) Top 10 algorithms in data mining, Knowl Inf Syst pp. 14–37.

Giudici, P., (2003) Applied Data Mining Statistical Methods for Business and Industry Wiley & Sons.

Hall, M., Holmes, G. (2000) Benchmariking Attribute Selection Techniques For Data Mining. Working Paper Series ISSN 1170-487X.

Hand, D. J. (1981) Discrimination and Classification.Chichester, U.K.: Wiley.

Hastie, T., Tibshirani, R., Friedman, J. (2001) The elements of statistical learning. New York, NY: Springer-Verlag;

Japkowicz, N., Sokolova, M., Szpakowicz, S., (2006) Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation, A. Sattar and B.H. Kang (Eds.): AI 2006, LNAI 4304, pp. 1015–1021

Joachims, T., (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Lecture Notes in Computer Science, Volume 1398/1998, 137-142

Kapp, E., Schütz, F., Connolly, L., Chakel J., Meza J., Miller C., Fenyo F., Eng J., Adkins, J., Omenn, G, Simpson, R. (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms:Sensitivity and specificity analysis, Proteomics, pp. 3475–3490.

Lavrac, N., (1999) Selected techniques for data mining in medicine , Artificial Intelligence in Medicine 16 3–23

Lin, Y.S., Wang, K.M., Makond, B., Wu, W., Wang, K.J., (2012) Optimal Data mining method for predicting breast cancer survivability, International Journal of Innovative Management, Information & Production, Volume 3

Liu, H., Li, J., Wong, L. (2002) A Comparative Study on Feature Selection and Classi_cation Methods Using Gene Expression Pro_les and Proteomic Patterns. Genome Informatics pp. 13-51

Maclin, R. & Opitz, D., (1999) Popular Ensemble Methods: An Empirical Study, Journal of Artificial Intelligence Research 11, pp.169-198

National Cancer Institute (2010), A snapshot of breast cancer, retrieved September 22, http://www.cancer.gov/aboutnci/servingpeople/snapshots/breast.pdf .

National Center for Health Statistics, http://www.cdc.gov/nchs/

Ostir GV, Uchida T. (2000) Logistic regression: A nontechnical review. Am JPhys Med Rehabil, pp. 565–572.

Payne, A.W.R, Glen, R.C. (1993) Molecular recognition using a binary genetic search algorithm. J. Mol. Graphics, Vol. 11, June, 1993.

Piatetsky-Shapiro, G. (2005) KDnuggets news on SIGKDD serviceaward. http://www.kdnuggets.com/news/2005/n13/2i.html.

Quinlan J., (1993) C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann.

Vapnik, V. (1995) The Nature of Statistical Learning Theory. Springer, New York.

Wasan, S. K., Bhatnagar, V. and Kaur, H., (2006) The impact of data mining techniques on medical diagnostics , Data Science Journal, Volume 5, 19 October, 2006.

Weiss, S. I., and Kulikowski, C. (1991) Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems. San Francisco, Calif.: Morgan Kaufmann.

World Health Organization (2010), Quick cancer facts, http://www.who.int/cancer/en.

Yang, Q., Zhang, S., Zhang, C., (2003) Data preparation for data mining, Applied Artificial Intelligence: An International Journal, 17:5-6, pp. 375-381. http://dx.doi.org/10.1080/713827180