ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
Μεταπτυχιακό Πρόγραμμα Σπουδών M.I.S.

«Αναζήτηση Γνώσης σε Νοσοκομειακά Δεδομένα»

Μεταπτυχιακός Φοιτητής: Τορτοπίδης Γεώργιος
Μηχανικός Η/Υ & Πληροφορικής

Επιβλέπων Καθηγητής: Βλαχάβας Ιωάννης
Καθηγητής Τμ. Πληροφορικής Α.Π.Θ.

Εξεταστής Καθηγητής: Μαργαρίτης Κων/νος
Καθηγητής Τμ. Εφηρμ. Πληροφορικής Παν. Μακεδονίας

Θεσ/νίκη – Φεβρουάριος 2004

Δομή Παρουσίασης

- Εισαγωγή - Σκοπός Διπλωματικής Εργασίας.
- Ανακάλυψη Γνώσης – Εξόρυξη σε Δεδομένα.
- Εργαλείο Ανακάλυψης Γνώσης.
- Νοσοκομειακά Δεδομένα.
- Εφαρμογή μεθόδων Ανακάλυψης Γνώσης στα Δεδομένα.
- Συμπεράσματα – Μελλοντική Προσπάθεια.

Εισαγωγή - Σκοπός Διπλωματικής Εργασίας.

Εισαγωγή - Σκοπός Διπλωματικής Εργασίας

Γεγονότα

- Το νοσοκομειακό περιβάλλον συγκεντρώνει τεράστιες ποσότητες πληροφοριών (Βάσεις Δεδομένων) λόγω εισαγωγής της πληροφορικής και νέων τεχνολογιών.

- Η Ανακάλυψη Γνώσης από Κλινικά δεδομένα αποτελεί κρίσιμη εφαρμογή. Η ιατρική από μόνη της δεν καταφέρνει να αποδείξει τη γνώση, η οποία ουσιαστικά αποτελεί παράγωγο της εμπειρίας και της συχνότητας εμφάνισης στα περιστατικά των ασθενών

⬇ Σκοπός Εργασίας ⬇

- Πειραματική Εφαρμογή Μεθόδων Ανακάλυψης Γνώσης σε Βάσεις Δεδομένων (KDD) κλινικών στοιχείων ασθενών Νοσοκομείου, σε συνεργασία με τους ειδικούς του τομέα (ιατρούς).

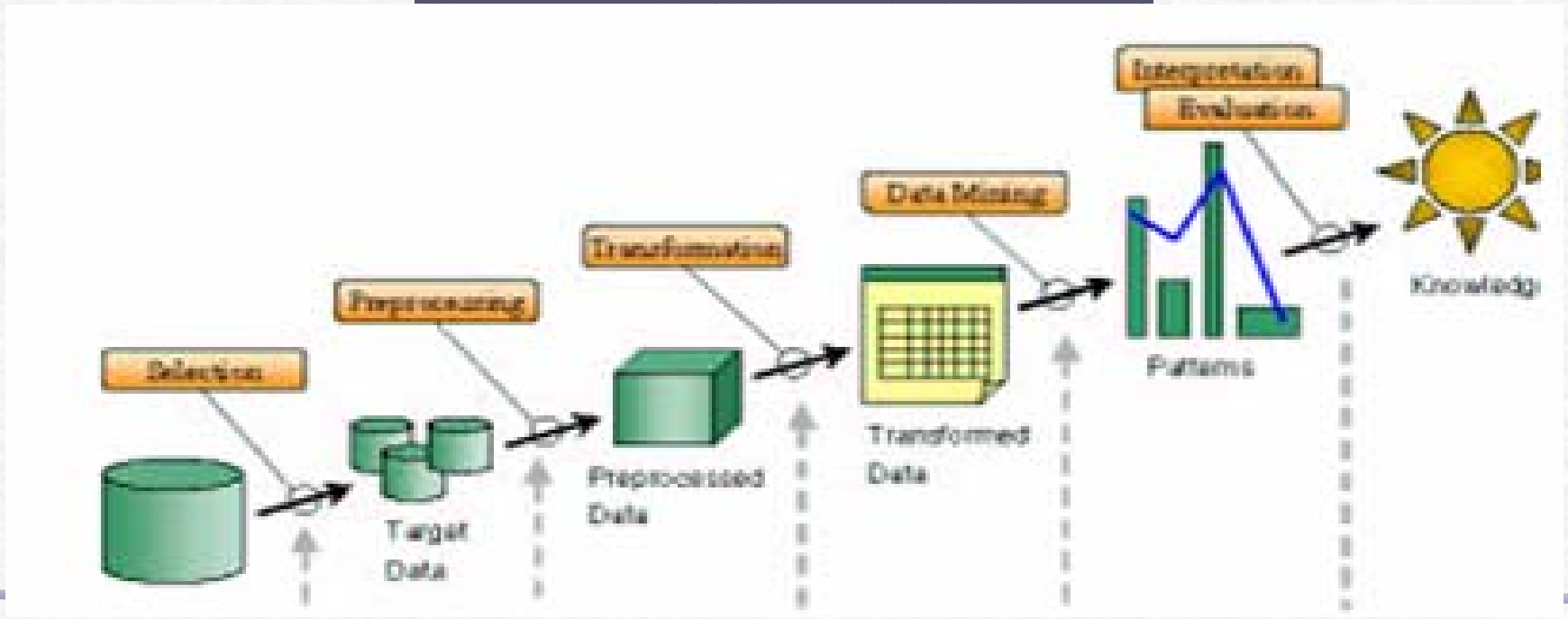
- Αξιολόγηση της προκύπτουσας γνώσης από τους ειδικούς του τομέα και πιθανή αξιοποίηση της προς βελτίωση των παρεχόμενων υπηρεσιών υγείας.

Ανακάλυψη Γνώσης – Εξόρυξη σε Δεδομένα.

Ανακάλυψη Γνώσης – KDD (Knowledge Discovery)

- Η Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery in Databases) είναι μια σύνθετη διαδικασία για τον προσδιορισμό έγκυρων, νέων, χρήσιμων και κατανοητών σχέσεων-προτύπων σε δεδομένα.
- Απαιτείται συνήθως η συνδρομή ενός ειδικού του τομέα εφαρμογής (π.χ. ιατρού)
- Πρόκειται για επαναληπτική και αλληλεπιδραστική διαδικασία, στη διάρκεια των επιμέρους σταδίων της οποίας, ο ειδικός καλείται να πάρει συγκεκριμένες αποφάσεις.

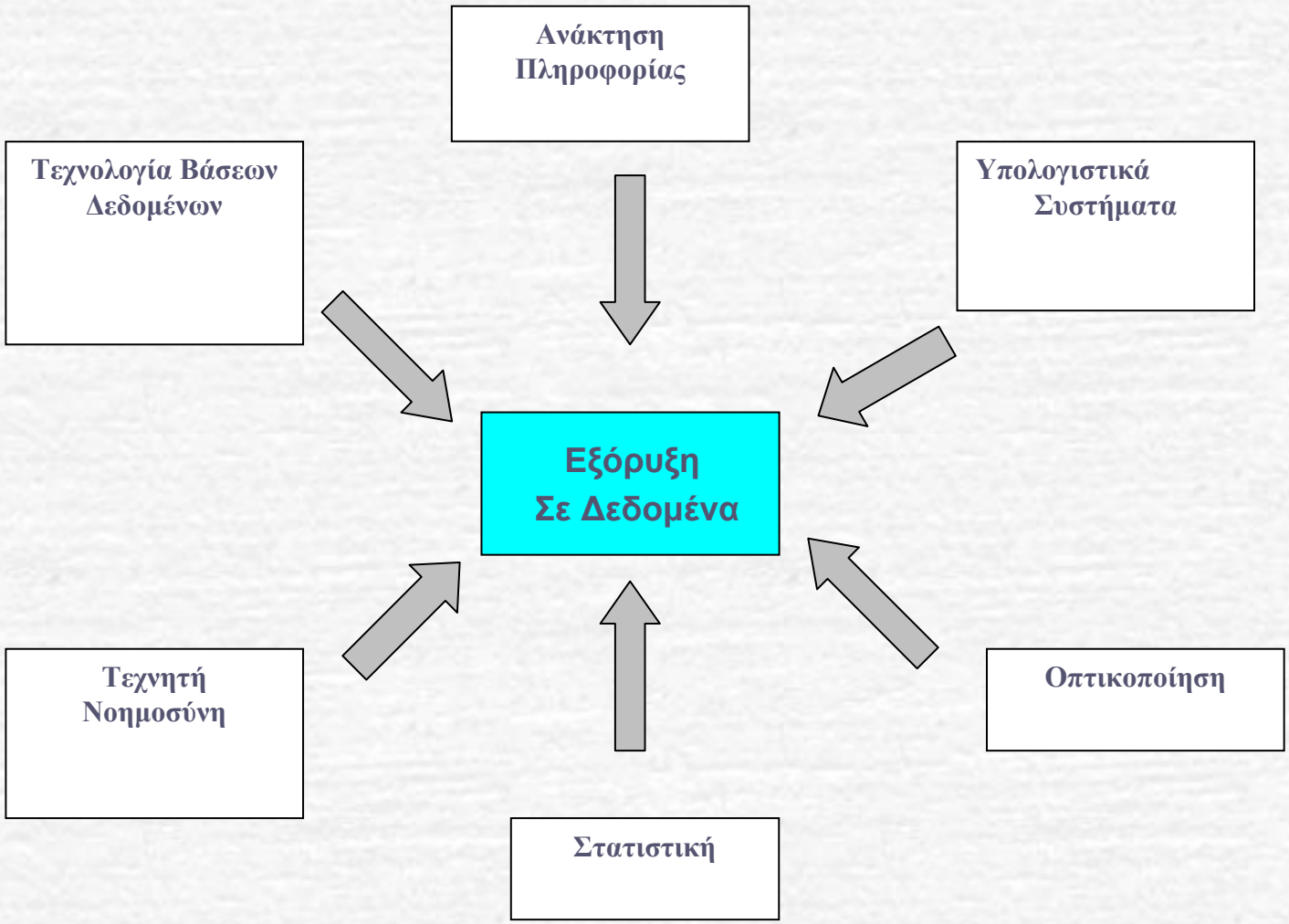
Τα Στάδια της Εξόρυξης Γνώσης



Εξόρυξη σε Δεδομένα (Data Mining)

- Η Εξόρυξη σε δεδομένα (Data Mining) αποτελεί συγκεκριμένο βήμα στη διαδικασία Ανακάλυψης Γνώσης από μεγάλες Βάσεις Δεδομένων (KDD).
- Αποτελεί περίπου το 20% της συνολικής KDD διαδικασίας.
- Το στάδιο της Εξόρυξης σε Δεδομένα αφορά ουσιαστικά την εφαρμογή αλγορίθμων Μηχανικής Μάθησης σε Βάσεις Δεδομένων.
  - Αυτή η βάση δεδομένων συνήθως έχει σχεδιαστεί για άλλο σκοπό & περιέχει ελλιπή ή λανθασμένα στοιχεία.
  - Αντίθετα, στη μηχανική μάθηση τα δεδομένα είναι και σωστά κωδικοποιημένα και σωστά επιλεγμένα.
- Η Μηχανική Μάθηση ορίζεται σαν κάθε Αυτόματη Διαδικασία μάθησης και αποτελεί τμήμα της Τεχνητής Νοημοσύνης.
- Μπορεί να διακριθεί σε:
  - **Μάθηση με Επίβλεψη (Supervised Learning)** – Στο σύστημα δίνονται παραδείγματα αντικειμένων μιας κατηγορίας και αυτό προσπαθεί να βρει τις κοινές ιδιότητες της κατηγορίας. Έτσι μπορεί μετέπειτα να προβλέψει κάποιο χαρακτηριστικό (εξαρτημένη μεταβλητή)
  - **Μάθηση χωρίς Επίβλεψη (Unsupervised Learning)** – Το σύστημα προσπαθεί να ανακαλύψει μόνο του ομάδες και συσχετίσεις με βάση μόνο τις ιδιότητες τους.

Εξόρυξη σε Δεδομένα



Είδη Γνώσης που Προκύπτουν

- **Πρότυπα Πληροφόρησης (Informative Patterns):** Περιγράφουν συσχετίσεις μεταξύ των δεδομένων, τις οποίες ο ειδικός του τομέα (ιατρός) δεν γνώριζε (Μάθηση χωρίς Επίβλεψη). Αντιπροσωπευτικές μέθοδοι είναι:
  - **Ομαδοποίηση (Clustering)**
  - **Σειριακά Πρότυπα (Sequential Patterns)**
  - **Κανόνες Συσχέτισης (Association Rules)**

- **Πρότυπα Πρόβλεψης (Predictive Patterns):** Προβλέπουν την τιμή ενός πεδίου μιας εγγραφής με βάση τις τιμές των υπολοίπων πεδίων (Μάθηση με Επίβλεψη). Αντιπροσωπευτικές μέθοδοι είναι:
  - **Κατηγοριοποίηση (Classification)**
  - **Δένδρα Κατηγοριοποίησης (Classification Trees)**
  - **Εμπειρική Σχέση Μεταβλητών (Regression)**
    - **Νευρωνικά Δίκτυα (Neural Networks)**

- Στην εργασία επικεντρωθήκαμε σε 3 από αυτές τις τεχνικές.

Κατηγοριοποίηση

- Μέθοδος αναζήτησης προτύπων πρόβλεψης όπου η τιμή του πεδίου που πρέπει να προβλεφθεί δεν εκφράζεται με συνεχή τιμή αλλά με κατηγορία (categorical-nominal).
- Οι πιο κοινές μέθοδοι κατηγοριοποίησης είναι τα Δένδρα Κατηγοριοποίησης.
- Ορίζουν μια σειρά ερωτήσεων και καταλήγουν στην πρόβλεψη της κατηγορίας στην οποία ανήκει η υπό εξέταση εγγραφή. Ποσοτικό χαρακτηριστικό που προσδίδει βαρύτητα στη μέθοδο είναι η ακρίβεια πρόβλεψης που εκφράζει την ικανότητα της μεθόδου να προβλέψει την τιμή του προς εξέταση πεδίου.
- Ευκολία Ερμηνείας – Πληθώρα Αλγορίθμων (id3, C4,5 κ.τ.λ)

Ομαδοποίηση (Clustering)

- Πρότυπα πληροφόρησης που προκύπτουν με ομαδοποίηση των εγγραφών της βάσης δεδομένων, έτσι ώστε εγγραφές που ανήκουν στην ίδια ομάδα να έχουν κοινά χαρακτηριστικά.
- Ο ειδικός του τομέα θα καθορίσει τη σημασία που έχει κάθε μια από τις ομάδες που προκύπτουν.
- Η ομαδοποίηση μπορεί να χρησιμοποιηθεί και σαν εισαγωγή ή πρώτο βήμα πριν από την εφαρμογή κάποιας άλλης τεχνικής εξόρυξης από δεδομένα ή μοντελοποίησης.

Κανόνες Συσχέτισης (Association Rules)

- Πρότυπα Πληροφόρησης της μορφής «Εάν Χ τότε Υ» όπου Χ και Υ είναι εκφράσεις που συνδέουν τιμές των πεδίων των εγγραφών της βάσης δεδομένων.

- Αυτό που προσδίδει βαρύτητα και αξία σε κανόνες τέτοιας μορφής είναι 2 ποσοτικά μεγέθη:

  - **Εμπιστοσύνη :** Είναι ο λόγος των εγγραφών που ικανοποιούν το Χ και το Υ προς τις εγγραφές που ικανοποιούν μόνο το Χ. Συχνά ονομάζεται και Accuracy (Ακρίβεια) του κανόνα. Είναι μια σημαντική ένδειξη του **πόσο ενδιαφέρον** είναι ένας κανόνας.

  - **Υποστήριξη :** Είναι ο λόγος των εγγραφών που ικανοποιούν το Χ και το Υ προς το σύνολο των εγγραφών. Ουσιαστικά καθορίζει το πόσο **σημαντικό** είναι το πρότυπο που εντοπίστηκε κυρίως για τον τελικό χρήστη.

Εργαλείο Ανακάλυψης Γνώσης.

Εργαλείο Ανακάλυψης Γνώσης

- WEKA: Open source Εργαλείο υλοποιημένο σε JAVA
- Αποτέλεσμα πειραματικής εργασίας Πανεπιστημίου Waikato της Ν.Ζηλανδίας.
- Διατίθεται δωρεάν (freeware) στο Διαδίκτυο.
- Παροχή δυνατοτήτων Προεπεξεργασίας, Καθαρισμού και Διακριτοποίησης των δεδομένων με χρήση Φίλτρων (Filters)
- Πληθώρα Αλγορίθμων και Μεθόδων όπως:

  - **Classification:** Decision Trees, Bayes, Logistic Regression, J4.8, ID3 κ.τ.λ
  - **Prediction:** Linear regression, Model Tree Generator κ.τ.λ.
  - **Clustering:** EM, Cobweb, Simple k-Means, FastestFirst κ.τ.λ.
  - **Association:** Apriori κ.τ.λ.

Νοσοκομειακά Δεδομένα

Νοσοκομειακά Δεδομένα Εφαρμογής

- Δημογραφικά και Κλινικά δεδομένα Διαβητικών Ασθενών Εξωτερικού Διαβητολογικού Ιατρείου Α' Παθολογικής Κλινικής Νοσοκομείου Παπαγεωργίου.
- Εξαγωγή Κλινικών Δεδομένων από Βάση Δεδομένων PARADOX 7.0
- Εξαγωγή Δημογραφικών Δεδομένων από Σύστημα Διαχείρισης Επιχειρησιακών Πόρων SAP R/3 του Νοσοκομείου και Βάσης Δεδομένων ORACLE 8.0.6.2.0
- Σύνθεση αυτών και δημιουργία Αποθήκης Δεδομένων (Data Warehouse).
- Μετατροπή αυτών σε μορφή κατάλληλη για εισαγωγή στο WEKA.

Σύνθεση Δεδομένων

Δημογραφικά Χαρακτηριστικά

- Καταγωγή
- Επάγγελμα
- Φύλο
- Ηλικία

Κλινικά Χαρακτηριστικά

- Τιμές Σακχαρώδους Διαβήτη
- Χοληστερίνη
- Τριγλυκερίδια
- Ινσουλίνη
- Βάρος
- Ύψος
- Κληρονομικότητα
- Υπέρταση
- Δίαιτα Αντιμετώπισης
- Κάπνισμα

Εφαρμογή μεθόδων Ανακάλυψης Γνώσης στα Δεδομένα

Στάδια KDD – Εφαρμογή στα Δεδομένα

- **Προεπεξεργασία**
  - Αντιμετώπιση ελλιπών δεδομένων
  - Αντιμετώπιση κενών πεδίων
- **Μετασχηματισμός**
  - Βασικές Μετατροπές στον τύπο των Δεδομένων
  - Διακριτοποίηση τιμών πεδίων π.χ. το πεδίο CHOL (Τιμές Χοληστερίνης) διακριτοποιείται ως:

Διακριτή Τιμή	Επεξήγηση (Τιμές Χοληστερίνης)
0	<120 (Ελάχιστες τιμές)
1	120-200 (Επιθυμητές Τιμές)
2	201-239 (Οριακές Τιμές)
3	>240 (Παθολογικές Τιμές)

Στάδια KDD – Εφαρμογή στα Δεδομένα

- **Δημιουργία Αποδεκτού Αρχείου** προς είσοδο στο WEKA

  Το WEKA δέχεται ένα συγκεκριμένου τύπου αρχείο το οποίο καλείται arff (Attribute-Relation File Format). Αποτελείται από 2 ανεξάρτητα τμήματα. Το **τμήμα πληροφοριών επικεφαλίδας** που περιέχει την λίστα χαρακτηριστικών των εγγραφών και το **τμήμα πληροφοριών δεδομένων** στο οποίο περιέχει τις εγγραφές πάνω στις οποίες θα εφαρμοστούν οι τεχνικές εξόρυξης.

Τμήμα Πληροφοριών Επικεφαλίδας

```
@relation Patien_SAKXARO.csv

@attribute POLI {1,2,3}
@attribute EPAGGELMA {1,2,3}
@attribute ILIKIA {1,2,3,4,5,6,7,8,9}
@attribute FILO {Γ,Α}
@attribute YPERTASI {0,1}
@attribute IDDM_NIDDM {IDDM,NIDDM}
@attribute ZAKX_DIAB_ numeric
@attribute ZAKX_DIAITA {1,2,3}
@attribute ZAX_DIAB_R {<150,>200,150-200}
@attribute DISDI_CHOL {0,1,2,3}
@attribute DISDI_HDL {1,2,3}
@attribute KLHRON_DIA {0,1}
@attribute KLHRON_STE {0,1}
```

Τμήμα Πληροφοριών Δεδομένων

```
@ DATA

1,1,7,Γ,1,5,NIDDM,26,1,?,?,150-200,?,2,2,2,3,1,1,0,0,1,0
1,1,8,Γ,1,3,NIDDM,35,3,?,?,150-200,?,2,1,1,4,1,0,0,0,1,0
1,1,8,Γ,0,0,NIDDM,5,2,?,?,<150,?,1,1,1,2,1,0,0,0,0,10
1,1,8,Γ,0,15,NIDDM,2,3,?,?,<150,?,1,1,1,4,0,0,0,0,0,20
3,1,1,Α,1,1,IDDM,21,2,?,?,>200,?,2,2,1,2,0,0,0,0,0,0
3,1,2,Α,0,2,IDDM,10,1,?,?,<150,?,1,1,1,2,0,0,0,0,0,0
3,1,2,Α,1,1,IDDM,6,1,?,?,<150,?,3,2,1,5,0,0,0,0,0,0
1,1,3,Α,1,0,IDDM,4,1,?,?,<150,?,2,2,2,2,0,0,0,0,0,0
1,1,4,Α,0,0,NIDDM,2,1,?,?,>200,?,1,1,1,2,1,0,0,0,0,0
2,1,4,Α,1,7,NIDDM,4,2,?,?,150-200,?,2,2,1,3,0,0,0,0,0,0
2,1,4,Γ,0,17,NIDDM,26,1,?,?,150-200,?,1,1,1,2,0,0,0,0,0,0
3,1,4,Α,1,0,NIDDM,10,2,?,?,150-200,?,3,2,1,2,1,0,0,0,0,0
2,1,5,Α,1,5,NIDDM,23,2,?,?,>200,?,2,2,1,2,1,0,0,0,0,0
1,1,5,Α,1,10,NIDDM,10,2,?,?,>200,?,3,2,2,2,1,0,0,0,0,0
1,1,5,Γ,1,2,NIDDM,4,3,?,?,150-200,2,1,1,1,3,0,0,0,0,0,0
```

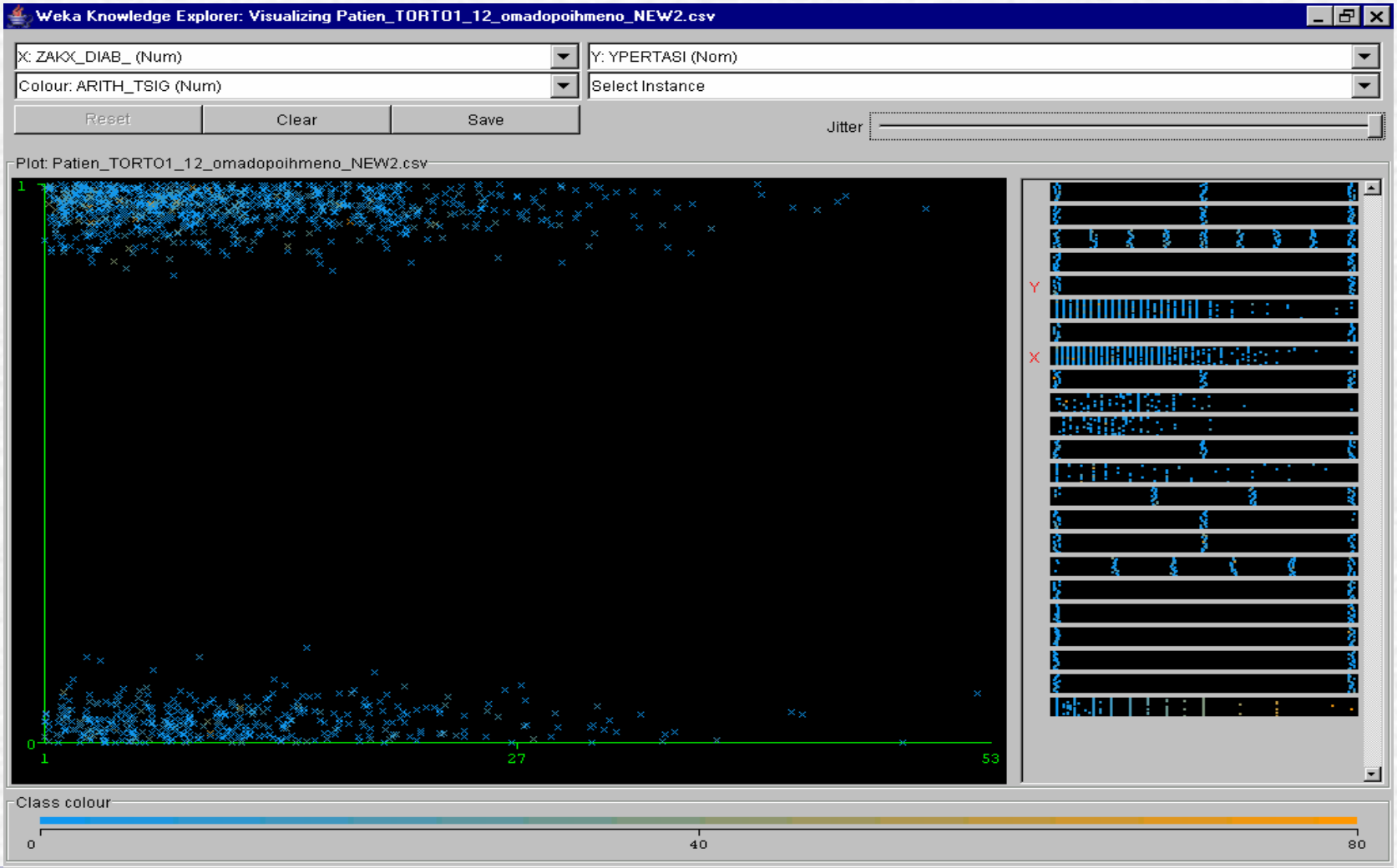
KDD – Προέλεγχος

KDD – Προέλεγχος

Στατιστικός Έλεγχος και Αρχικές Παρατηρήσεις

- Ισομερές δείγμα όσο αφορά το φύλο (FILO).
- Το μεγαλύτερο μέρος των ασθενών βρίσκεται στις υψηλές ηλικιακές ομάδες (άνω των 45 ετών -HLIKIA).
- Οι Ινσουλινοεξαρτώμενοι (IDDM-NIDDM) ασθενείς είναι σαφώς λιγότεροι από το συνολικό δείγμα.
- Δεν εμφανίζεται κάποια συσχέτιση μεταξύ Σακχαρώδους Διαβήτη και δεικτών Χοληστερίνης (CHOL) και τριγλυκεριδίων (TG).
- Παρατηρούμε την πιθανή ύπαρξη κάποιας αλληλεξάρτησης μεταξύ Διαβήτη & Υπέρτασης (YPER) όπως φαίνεται : ➔

KDD – Προέλεγχος



Κατηγοριοποίηση (Classification)

- **Εφαρμογή Αλγορίθμου ID3**

  - Δείγμα 1296 ασθενών.

  - Κλάση Πρόβλεψης το χαρακτηριστικό Ινσουλινοεξαρτώμενου ή μη ασθενούς (IDDM/NIDDM).

  - Διακριτοποίηση όλων των χαρακτηριστικών λόγω φύσης ορθής λειτουργίας αλγορίθμου.

  - Ακρίβεια Πρόβλεψης **91,81%**

  - Η χρήση ινσουλίνης απαιτείται κύριως στους ασθενείς μικρών ηλικιών

- **Εφαρμογή Μεθόδου Δημιουργίας Πίνακα Κανόνων**

ILIKIA	DISDI_CHOL	BMI_CODE	KLHRON_DIA	KLHRON_DIS	IDDM_NIDDM
2	2	5	1	0	IDDM
2	1	5	0	1	IDDM
2	1	5	1	0	IDDM
2	3	5	0	0	IDDM
2	2	4	1	0	IDDM
2	3	3	1	0	IDDM
1	2	3	1	0	IDDM
1	2	2	0	0	IDDM

- **Η χρήση ινσουλίνης απαιτείται κυρίως στους ασθενείς μικρών ηλικιών**

Κατηγοριοποίηση (Classification)

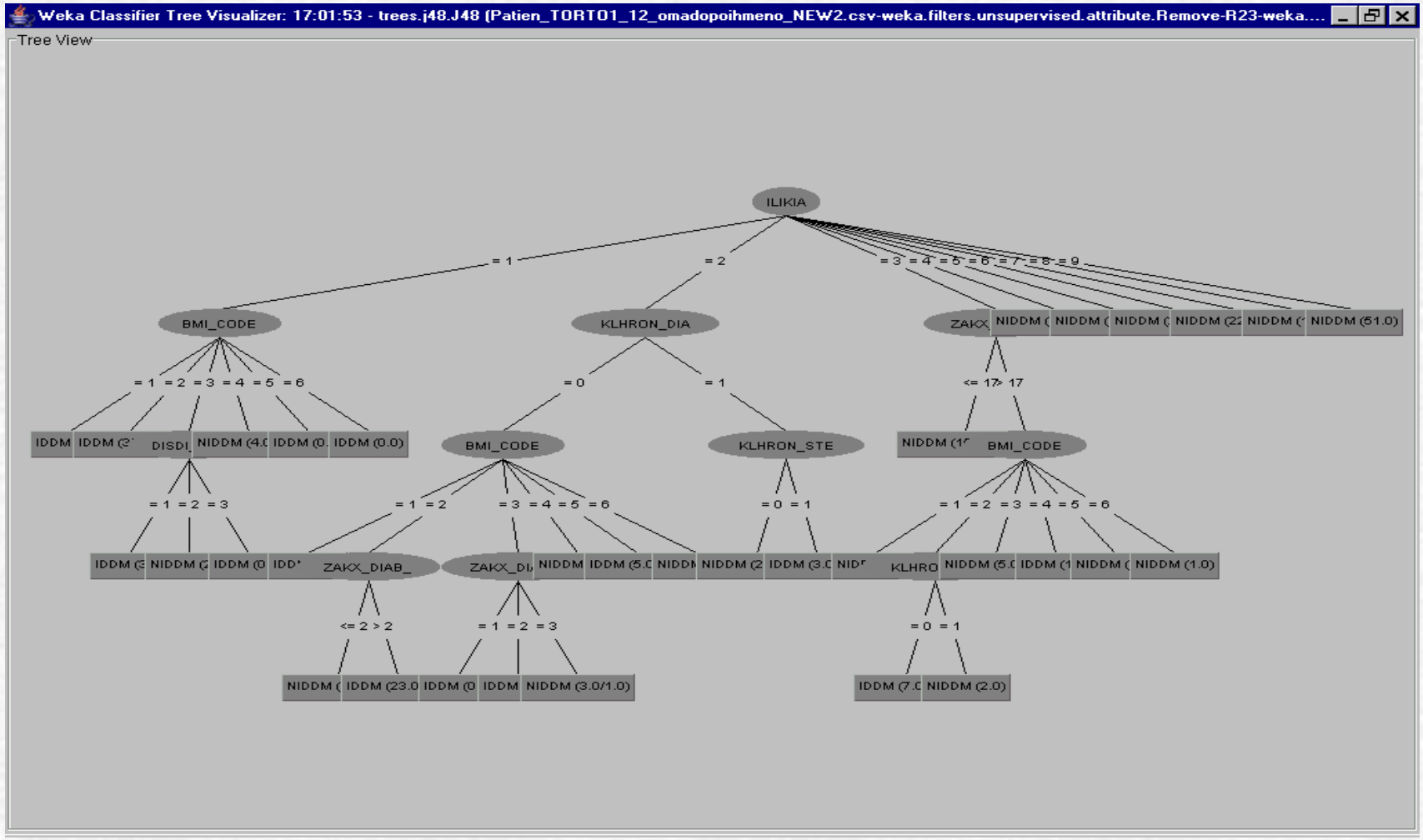
Εφαρμογή Αλγορίθμου J4.8 (C4.5)

- Κλάση πρόβλεψης το χαρακτηριστικό IDDM/NIDDM.
- Δείγμα 1296 ασθενών.
- Προκύπτει Decision Tree 44 κόμβων.
- Ακρίβεια πρόβλεψης **96,6%**
- Βασικό χαρακτηριστικό κατασκευής του δένδρου αποτελεί η ηλικία (HLIKIA).
- Επαλήθευση (Cross Validation) για βελτίωση του αποτελέσματος με Αριθμό Επαναλήψεων Folds = 10.
- Πίνακας λαθών (Confusion Matrix):

a	b	<-- classified as
72	35	a = IDDM
8	**1181**	b = NIDDM

- **Η χρήση ινσουλίνης απαιτείται κυρίως στους ασθενείς μικρών ηλικιών**

Κατηγοριοποίηση (Classification)



Decision Tree του Αλγορίθμου J4.8

Ομαδοποίηση (Clustering)

- **Εφαρμογή Μεθόδου Simple-kMeans**

  - Επαναληπτικός Αλγόριθμος
  - Απαιτεί ύπαρξη Χαρακτηριστικού Συνεχών Τιμών
  - Εφαρμόζεται για Δημιουργία 2 Ομάδων (συγκεκριμένη εφαρμογή)
  - Τα κέντρα των ομάδων δίνονται παρακάτω, από όπου μπορούν να εξαχθούν τα ποιοτικά και ποσοτικά χαρακτηριστικά κάθε ομάδας:

Cluster 0

Mean/Mode: 6 A 1 6.9656 NIDDM 11.0325 2 >200 3 2 1 2 0 0 0 0 0

Cluster 1

Mean/Mode: 6 Γ 0 6.3256 NIDDM 11.3798 2 150-200 1 1 1 2 0 0 0 0 0

Clustered Instances (Πληθυσμιακή Κατανομή

Ομάδα 0 **830 (64%) Πληθυσμός 1ης Ομάδας** – Υψηλός Διαβήτης – Επιπλέον Προβλήματα

Ομάδα 1 **466 (36%) Πληθυσμός 2ης Ομάδας** - Μέτρια Επίπεδα Διαβήτη με ευκολία Αντιμετώπισης

- **Οι Γυναίκες πιο σταθερές και σε χαμηλά επίπεδα Διαβήτη**

Κανόνες Συσχέτισης (Association Rules)

- **Εφαρμογή Αλγορίθμου Apriori**

  - Απαιτεί πεδία Διακριτών τιμών – Διακριτοποίηση
  - Επιλέγεται Υποστήριξη (support) = 0.75
  - Προκύπτουν οι παρακάτω κανόνες (ουσιαστικοί) :

FILO=Γ DISDI_TG=2 424	**==> YPERTASI=1 424**	**conf:(1)**
FILO=Γ 693	**==> IDDM_NIDDM=NIDDM 652**	**conf:(0.94)**
FILO=Γ YPERTASI=1 491	**==> IDDM_NIDDM=NIDDM 461**	**conf:(0.94)**
FILO=Γ DISDI_HDL=1 511	**==> IDDM_NIDDM=NIDDM 478**	**conf:(0.94)**
FILO=Γ YPERTASI=1 491	**==> DISDI_TG=2 424**	**conf:(0.86)**

- Σταθερότητα ως προς τον γυναικείο πληθυσμό

- Μη ύπαρξη εξάρτησης Διαβήτη και Χοληστερίνης, TG, HDL.

Συμπεράσματα – Μελλοντική Προσπάθεια

Συμπεράσματα – Μελλοντικές Προσπάθειες

Συμπεράσματα

- Ο τομέας της υγείας αποτελεί πηγή μεγάλου όγκου ποιοτικών & ποσοτικών δεδομένων.
- Επιτακτική η ανάγκη ύπαρξης ειδικού τομέα για αξιολόγηση προτύπων.
- Η εφαρμογή των μεθόδων εξόρυξης σε δεδομένα οδήγησε σε επαλήθευση των εμπειρικών γνώσεων των ιατρών σε ότι αφορά τον σακχαρώδη διαβήτη.
- Εκδηλώθηκε ενδιαφέρον από μέρους των ιατρών για αύξηση των εγγραφών του αρχείου δεδομένων (εισαγωγή επιπλέον ασθενών) καθώς επίσης και η αντικατάσταση συγκεκριμένων χαρακτηριστικών με άλλα τα οποία οι γιατροί τα θεωρούν στενά συνδεδεμένα με τον διαβήτη.
- Προσοχή στην προσαρμογή της εμπειρικής γνώσης των ειδικών του τομέα στην μορφή ενός προτύπου.

Μελλοντική Εφαρμογή

- Ενημέρωση αρχείου δεδομένων με περισσότερες εγγραφές και εφαρμογή επιπλέον μεθόδων εξόρυξης σε δεδομένα.
- Αντικατάσταση χαρακτηριστικών με άλλα πιο ποιοτικά όπως αυτό ζητήθηκε από τους γιατρούς.
- Η εφαρμογή των μεθόδων σε περισσότερες εγγραφές με πιο ποιοτικά χαρακτηριστικά, θα οδηγήσει σε ισχυρότερους κανόνες και πρότυπα δηλαδή στην εξαγωγή νέας γνώσης, άρα σε βελτίωση στις παρεχόμενες υπηρεσίες υγείας.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
Μεταπτυχιακό Πρόγραμμα Σπουδών M.I.S.

«Αναζήτηση Γνώσης σε Νοσοκομειακά Δεδομένα»

ΣΑΣ ΕΥΧΑΡΙΣΤΩ