ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΘΕΣΣΑΛΟΝΙΚΗ

Predicting the Choice of Contraceptive Method using Classification

ΠΑΠΑΔΟΠΟΥΛΟΣ ΧΡΗΣΤΟΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Νικόλαος Σαμαράς

ΕΞΕΤΑΣΤΗΣ: Κωνσταντίνος Παπαρίζος

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εξόρυξη Γνώσης(Data Mining)

- **Τι είναι Data Mining**
  1. Ορίζουμε την **εξόρυξη γνώσης (data mining)** σαν τη διαδικασία της χρήσης μιας ή περισσοτέρων τεχνικών εκμάθησης υπολογιστών για την αυτόματη ανάλυση και εξαγωγή γνώσης από δεδομένα που περιέχονται σε μια βάση δεδομένων.
  2. Είναι διαδικασία μιας πολύπλευρης ανάλυσης δεδομένων και εξαγωγής χρήσιμων πληροφοριών, πληροφορίες που μπορούν να χρησιμοποιηθούν για να αυξήσουν τα έσοδα, να μειώσουν τις δαπάνες ή και τα δύο.
  3. Η διαδικασία εξόρυξης γνώσης είναι ένα από τα πολλά εργαλεία που χρησιμοποιούνται στην ανάλυση δεδομένων. Επιτρέπει στους χρήστες την ανάλυση δεδομένων από πολλές διαφορετικές σκοπιές, την ταξινόμηση, το συνοψισμό και τελικά την εξαγωγή των σχέσεων που προσδιορίζονται από τη διαδικασία

Εξόρυξη Γνώσης(Data Mining)

- **Ιστορική εξέλιξη του Data Mining**
  - Η σχεδίαση τεχνικών του σημερινού data mining χρονολογείται γύρω στην δεκαετία του 1950.
  - Κατά την δεκαετία του 1960, η τεχνητή νοημοσύνη και ο κλάδος της στατιστικής εφάρμοσαν νέους αλγορίθμους όπως η ανάλυση παλινδρόμησης. Ο όρος data mining πρωτοχρησιμοποιήθηκε αυτή την δεκαετία.
  - Στις αρχές της δεκαετίας του 1990, ο όρος Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (KDD) χρησιμοποιήθηκε για πρώτη φορά και δημιουργήθηκα το πρώτο εργαστήριο KDD.
  - Κατά τη δεκαετία 1990, το data mining εξελίχθηκε από μία ενδιαφέρουσα τεχνολογία σε μία πρότυπη πρακτική από εταιρείες.

Εξόρυξη Γνώσης(Data Mining)

- **Λειτουργία της εξόρυξης δεδομένων**
  - Η πληροφοριακή τεχνολογία έχει εξελιχθεί σε δύο ξεχωριστά συστήματα, της συναλλαγής και της ανάλυσης.
  - Η εξόρυξη γνώσης έρχεται για να παίξει το ρόλο του συνδέσμου μεταξύ των δύο.
  - Το λογισμικό εξόρυξης γνώσης, αναλύει τις σχέσεις και τα μοτίβα στα αποθηκευμένα δεδομένα συναλλαγής.
  - Διάφοροι τύποι λογισμικού ανάλυσης είναι διαθέσιμοι:
    1. στατιστική ανάλυση
    2. μηχανική εκμάθηση
    3. νευρωνικά δίκτυα

Εξόρυξη Γνώσης(Data Mining)

- **Λειτουργία της εξόρυξης δεδομένων**
  - Γενικά, επιδιώκεται οποιοσδήποτε από τους τέσσερις τύπους σχέσεων:
  - **Κατηγορίες (Classes)**
    - Τα αποθηκευμένα δεδομένα χρησιμοποιούνται για να εντοπίσουν πληροφορίες για προκαθορισμένες ομάδες. Παραδείγματος χάριν, μια αλυσίδα εστιατορίων θα μπορούσε να εξαγάγει τις καταναλωτικές συνήθειες των πελατών, που καθορίζονται από τις επισκέψεις αυτών και να τις αναλύσει, λειτουργώντας με βάση το τι πραγματικά η ανάλυση αυτή επιτάσσει. Αυτές οι πληροφορίες θα μπορούσαν να ενδεχομένως να χρησιμοποιηθούν για να αυξήσουν την κατανάλωση των σπεσιαλιτέ της ημέρας.
  - **Ομάδες (Clusters)**
    - Τα δεδομένα ομαδοποιούνται σύμφωνα με λογικές σχέσεις ή καταναλωτικές προτιμήσεις. Παραδείγματος χάριν, τα δεδομένα μπορούν να εξαχθούν για να προσδιορίσουν τους τομείς της αγοράς ή τις καταναλωτικές συγγένειες.

Εξόρυξη Γνώσης(Data Mining)

- **Λειτουργία της εξόρυξης δεδομένων**
  - **Σχέσεις (Associations)**
    - Τα δεδομένα μπορούν να εξαχθούν για να προσδιορίσουν τις σχέσεις. Το παράδειγμα μπύρα-πάνες είναι ένα παράδειγμα του συνειρμικού-σχεσιακού data mining.
  - **Σειριακά μοτίβα (Sequential Patterns)**
    - Τα δεδομένα εξάγονται ώστε να προβλεφθούν τα μοτίβα και οι τάσεις συμπεριφοράς. Παραδείγματος χάριν, ένας πωλητής εξοπλισμού ειδών εξοχής, θα μπορούσε να προβλέψει την πιθανότητα πώλησης ενός σακιδίου πλάτης, βασισμένος στην αγορά από έναν πελάτη, υπνόσακου και παπουτσιών πεζοπορίας.

Εξόρυξη Γνώσης(Data Mining)

- **Λειτουργία της εξόρυξης δεδομένων**
  - Η εξόρυξη γνώσης περιλαμβάνει πέντε σημαντικά στάδια:
    1. **Συλλογή των Δεδομένων** (Data Warehousing, Web Crawling, …)
    2. **Καθαρισμός Δεδομένων** (Data Cleaning) – Επεξεργασία – εμπλουτισμός των δεδομένων
    3. **Feature Extraction**: Επιλογή των σημαντικών γνωρισμάτων
    4. **Εφαρμογή Μοντέλων/Αλγορίθμων Εξόρυξης Δεδομένων**
    5. **Απεικόνιση/αξιολόγηση** των αποτελεσμάτων

Εξόρυξη Γνώσης(Data Mining)

- **Λειτουργία της εξόρυξης δεδομένων**
  - Τα διαφορετικά επίπεδα ανάλυσης που είναι διαθέσιμα:
    - **Τεχνητά νευρωνικά δίκτυα (Artificial neural networks):** Μη γραμμικά προβλεπτικά μοντέλα που μαθαίνουν μέσω της εκπαίδευσης και μοιάζουν στη δομή με τα βιολογικά νευρικά δίκτυα.
    - **Γενετικοί αλγόριθμοι (Genetic algorithms):** Οι τεχνικές βελτιστοποίησης που χρησιμοποιούν διαδικασίες επεξεργασίας όπως ο γενετικός συνδυασμός, η μεταλλαγή (combination), και η φυσική επιλογή σε ένα μοτίβο, βασίστηκαν στις έννοιες της φυσικής εξέλιξης.
    - **Δέντρα απόφασης (Decision Trees):**Δέντρο-διαμορφωμένες δομές που αντιπροσωπεύουν τα σύνολα αποφάσεων. Αυτές οι αποφάσεις παράγουν τους κανόνες για την ταξινόμηση ενός συνόλου δεδομένων. Μια απλή δομή όπου οι μη τερματικοί κόμβοι αντιπροσωπεύουν τα αποτελέσματα των αποφάσεων. Τα δέντρα αποφάσεων έχουν διάφορα πλεονεκτήματα, όπως το ότι είναι εύκολο να τα καταλάβουμε, μπορούν να μετασχηματιστούν σε κανόνες και πειραματικά έχει αποδειχθεί ότι λειτουργούν πολύ καλά.

Εξόρυξη Γνώσης(Data Mining)

- **Λειτουργία της εξόρυξης δεδομένων**
  - Τα διαφορετικά επίπεδα ανάλυσης που είναι διαθέσιμα:
    - **Μέθοδος κοντινότερων γειτόνων (Nearest neighbor method)**: Μια τεχνική που ταξινομεί κάθε εγγραφή σε ένα σύνολο δεδομένων βασισμένο σε έναν συνδυασμό των ταξινομήσεων των εγγραφών *K* και του πιο κοντινού "συγγενή" με το *K* σε ένα ιστορικό σύνολο δεδομένων.
    - **Επαγωγή κανόνα (Rule induction)**: Η εξαγωγή των χρήσιμων if-then κανόνων από τα δεδομένα, βασιζόμενα στη στατιστική σημασία.
    - **Απεικόνιση στοιχείων (Data visualization)**: Η οπτική ερμηνεία των σύνθετων σχέσεων στα πολυδιάστατα δεδομένα. Τα εργαλεία γραφικής αναπαράστασης χρησιμοποιούνται για να επεξηγήσουν τις σχέσεις των δεδομένων

Δέντρα Απόφασης(Decision Trees)

- **Δέντρα απόφασης**
  - Τα δέντρα απόφασης είναι μια δημοφιλής δομή για καθοδηγούμενη εκμάθηση.
  - Σε αυτό το κεφάλαιο θα δούμε πιο αναλυτικά τον αλγόριθμο C4.5, που χρησιμοποιείται για την κατασκευή δέντρων απόφασης.
  - Θα εφαρμόσουμε αυτόν τον αλγόριθμο σε ένα παράδειγμα βάσης δεδομένων προώθησης πιστωτικών καρτών, ώστε να προσπαθήσουμε να τον κατανοήσουμε.

Δέντρα Απόφασης(Decision Trees)

- **Ένας αλγόριθμος κατασκευής δέντρων απόφασης**
  - Τα δέντρα απόφασης κατασκευάζονται χρησιμοποιώντας μόνο εκείνα τα γνωρίσματα που είναι σε θέση να διακρίνουν τις έννοιες προς εκμάθηση.
  - Για να χτίσουμε ένα δέντρο απόφασης, πρέπει αρχικά να επιλέξουμε ένα υποσύνολο περιπτώσεων από το σύνολο των δεδομένων που θα χρησιμοποιηθούν στην εκπαίδευση (**υποσύνολο δεδομένων εκπαίδευσης - training set**).
  - Αυτό το υποσύνολο (**δεδομένα ελέγχου - test set**) χρησιμοποιείται έπειτα από τον αλγόριθμο για να κατασκευάσει το δέντρο απόφασης.
  - Τα υπόλοιπα δεδομένα, τα δεδομένα training set, χρησιμοποιούνται στην εξέταση της ακρίβειας του κατασκευασμένου δέντρου. Εάν το δέντρο απόφασης ταξινομεί τις περιπτώσεις σωστά, η διαδικασία ολοκληρώνεται. Εάν μια περίπτωση είναι ανακριβώς ταξινομημένη, η περίπτωση προστίθεται στο επιλεγμένο υποσύνολο των training set και ένα νέο δέντρο κατασκευάζεται

Δέντρα Απόφασης(Decision Trees)

- **Ένας αλγόριθμος κατασκευής δέντρων απόφασης**
  - Τα βήματα του αλγορίθμου είναι τα ακόλουθα:
    1. Έστω T είναι το σύνολο στιγμιότυπων εκπαίδευσης, το training set
    2. Επιλέγουμε ένα χαρακτηριστικό που διαφοροποιεί καλύτερα τις περιπτώσεις που περιλαμβάνονται στο T.
    3. Δημιουργούμε έναν κόμβο στο δέντρο του οποίου η αξία είναι το επιλεγμένο χαρακτηριστικό. Δημιουργούμε θυγατρικούς δεσμούς από αυτόν τον κόμβο, όπου κάθε σύνδεση αντιπροσωπεύει μια μοναδική αξία για τα επιλεγμένα χαρακτηριστικά. Χρησιμοποιούμε τις τιμές των θυγατρικών δεσμών για να υποδιαιρέσουμε περαιτέρω τα στιγμιότυπα σε δευτερεύουσες κλάσεις.

Δέντρα Απόφασης(Decision Trees)

- **Ένας αλγόριθμος κατασκευής δέντρων απόφασης**
  - Τα βήματα του αλγορίθμου είναι τα ακόλουθα:

    4. Για κάθε δευτερεύουσα κλάση που δημιουργήθηκε στο βήμα 3:
       - Εάν τα στιγμιότυπα στη δευτερεύουσα κλάση ικανοποιούν προκαθορισμένα κριτήρια ή εάν το σύνολο των υπολοίπων επιλογών γνωρισμάτων γι' αυτή τη διαδρομή του δέντρου είναι μηδέν, καθορίζουμε την κατηγοριοποίηση των καινούργιων στιγμιότυπων που ακολουθούν αυτή τη διαδρομή αποφάσεων.
       - Εάν η δευτερεύουσα κλάση δεν ικανοποιεί τα προκαθορισμένα κριτήρια, και υπάρχει τουλάχιστον ένα γνώρισμα για να υποδιαιρέσει περαιτέρω τη διαδρομή του δέντρου, αφήστε το *T* να είναι το τρέχον σύνολο των στιγμιότυπων της δευτερεύουσας κλάσης και επιστρέφουμε στο βήμα 2.

Δέντρα Απόφασης(Decision Trees)

- **Παράδειγμα Δέντρου Απόφασης**
  - Έστω ότι έχουμε μία βάση με δεδομένα προώθησης πιστωτικών καρτών
  - Θέλουμε να αναπτύξουμε ένα μοντέλο πρόβλεψης με χαρακτηριστικό εξόδου το *life insurance promotion*.
  - Κατά συνέπεια, τα γνωρίσματα εισόδου περιορίζονται στις income *range, credit card insurance, sex* και *age*.

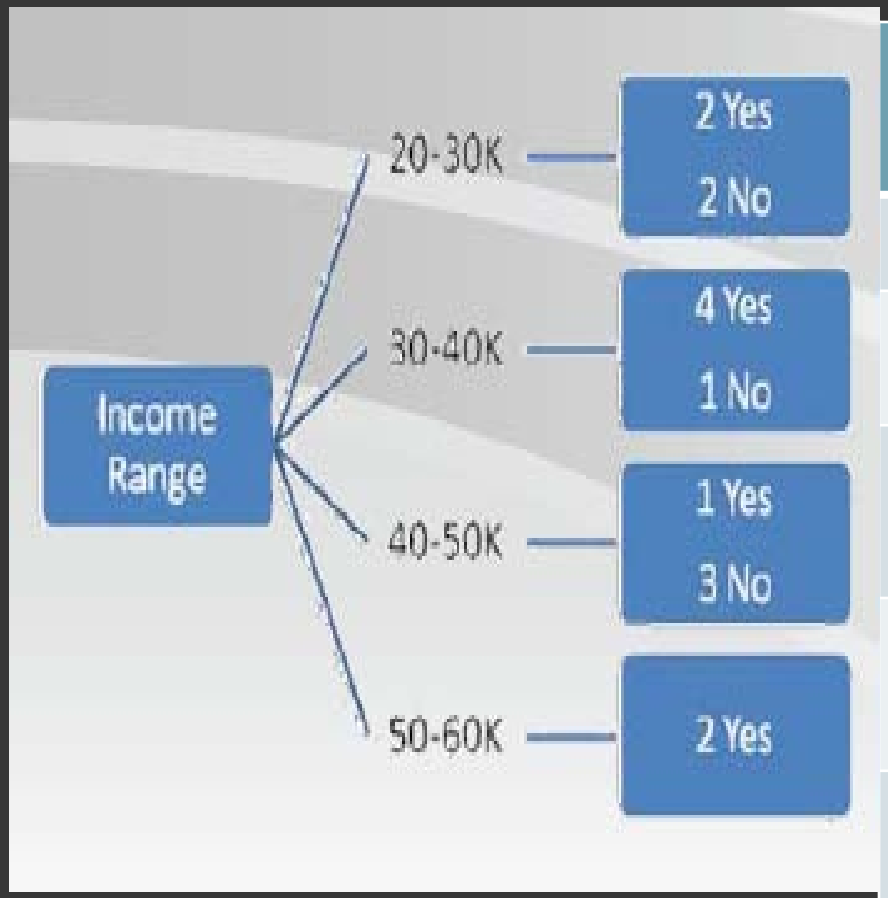
Δέντρα Απόφασης(Decision Trees)

- **Παράδειγμα Δέντρου Απόφασης**
  - Βάση δεδομένων προώθησης πιστωτικών καρτών

income range	*life insurance promotion*	*credit card insurance*	*sex*	*age*
40-50 χιλ	Όχι	Όχι	Άντρας	45
30-40 χιλ	Ναι	Όχι	Γυναίκα	40
40-50 χιλ	Όχι	Όχι	Άντρας	42
30-40 χιλ	Ναι	Ναι	Άντρας	43
50-60 χιλ	Ναι	Όχι	Γυναίκα	38
20-30 χιλ	Όχι	Όχι	Γυναίκα	55
30-40 χιλ	Ναι	Ναι	Άντρας	35
20-30 χιλ	Όχι	Όχι	Άντρας	27
30-40 χιλ	Όχι	Όχι	Άντρας	43
30-40 χιλ	Ναι	Όχι	Γυναίκα	41
40-50 χιλ	Ναι	Όχι	Γυναίκα	43
20-30 χιλ	Ναι	Όχι	Άντρας	29
50-60 χιλ	Ναι	Όχι	Γυναίκα	39
40-50 χιλ	Όχι	Όχι	Άντρας	55
20-30 χιλ	Ναι	Ναι	Γυναίκα	19

Δέντρα Απόφασης(Decision Trees)

Παράδειγμα δέντρου απόφασης



•Η Εικόνα δείχνει το τμήμα του δέντρου που δημιουργήθηκε στο βήμα 3 του αλγορίθμου, υπό την προϋπόθεση ότι το εύρος εισοδήματος είναι επιλεγμένο σαν κόμβος ανωτάτου επιπέδου.

•Οι συνολικές μετρήσεις yes και no για το γνώρισμα εξόδου φαίνονται στη βάση του κάθε κλάδου του τμήματος του δέντρου.

•Για να αξιολογήσουμε αυτή την επιλογή, κάνουμε πρώτα την τιμή της κάθε διαδρομής του τμήματος του δέντρου την πιο συχνή κλάση.

•Έχουμε δύο στιγμιότυπα από κάθε κλάση που ακολουθούν τον κλάδο που δίνεται από το *income range = 50-60K*. Άρα, μπορούμε να επιλέξουμε τη *life insurance promotion = no ή life insurance promotion = yes* σαν τιμή της διαδρομής.

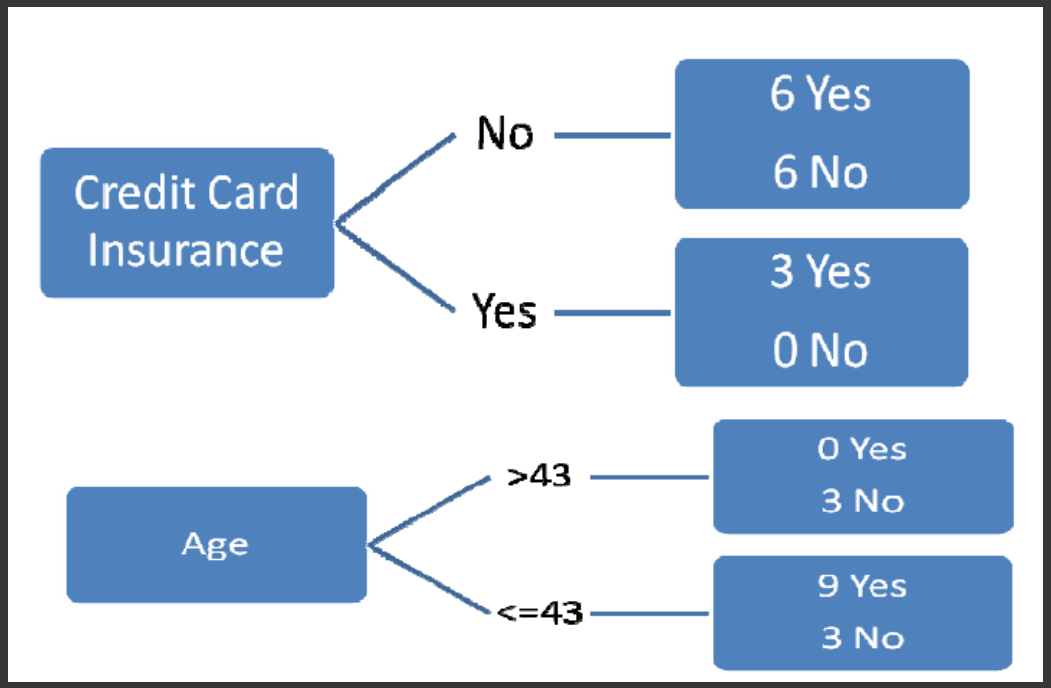
•Για να σπάσουμε το δεσμό, κλίνουμε προς την πιο συχνή κλάση, η οποία είναι η *life insurance promotion = yes*. Για τον κλάδο που δείχνει *income range = 30-40K* επιλέγουμε *life insurance promotion = yes* σαν τιμή της διαδρομής.

Για *income range = 40-50K* επιλέγουμε *life insurance promotion = no*, και για *income range = 50-60K* επιλέγουμε *life insurance promotion =yes*.

Δέντρα Απόφασης(Decision Trees)

Παράδειγμα δέντρου απόφασης

- Συνεχίζουμε με τα υπόλοιπα γνωρίσματα



Συγκρίνοντας τα αποτελέσματα παρατηρούμε ότι το γνώρισμα *age* προσφέρει το καλύτερο αποτέλεσμα ανάμεσα των πιθανών επιλογών γνωρίσματος.

Κατά συνέπεια το γνώρισμα *age* γίνεται γνώρισμα επιλογής και εκτελούμε το 3ο βήμα του αλγορίθμου.

Δέντρα Απόφασης(Decision Trees)

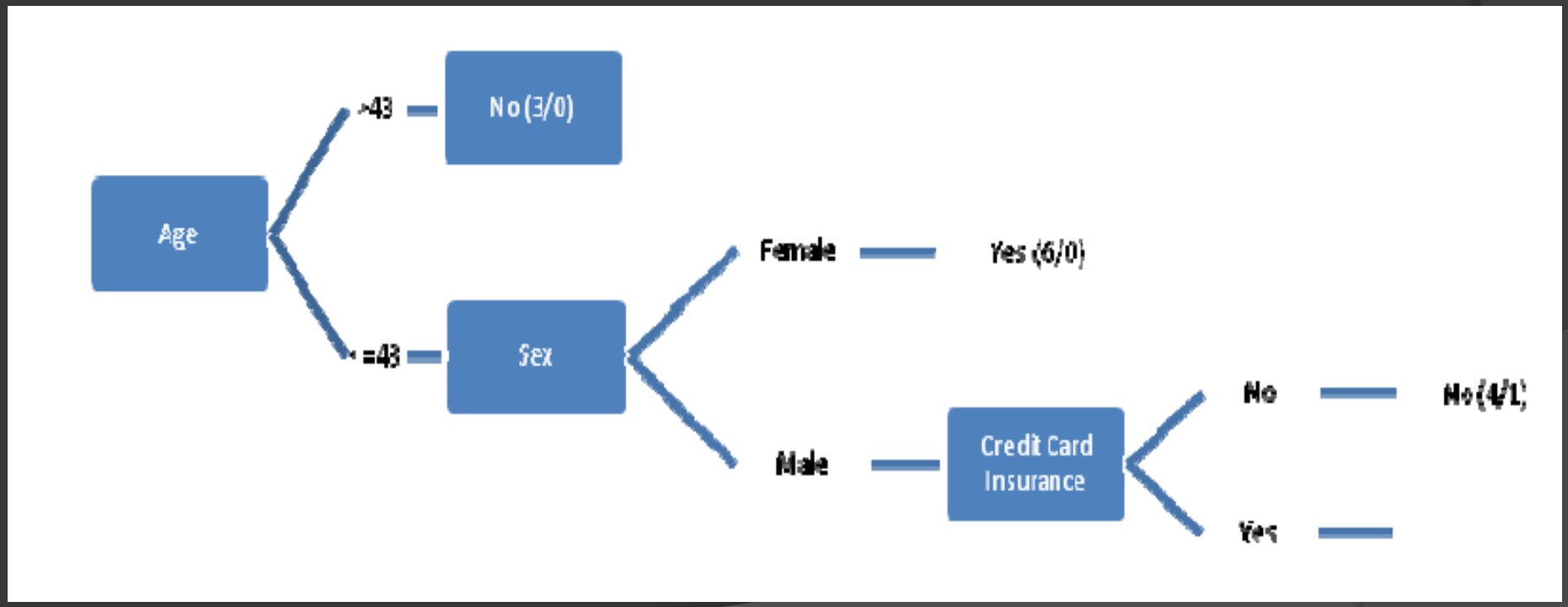
- **Παράδειγμα Δέντρου Απόφασης**
  - Το βήμα 4 του αλγορίθμου απαιτεί να εξετάσουμε τον κάθε κλάδο του τμήματος του δέντρου για να καθορίσουμε εάν θα συνεχίσουμε τη διαδικασία κατασκευής του δέντρου.
  - Ο αλγόριθμος δηλώνει δύο πιθανότητες για τον τερματισμό μιας διαδρομής του δέντρου:
    - Πρώτον, εάν τα στιγμιότυπα που ακολουθούν ένα συγκεκριμένο κλάδο ικανοποιούν ένα προκαθορισμένο κριτήριο, όπως είναι μια ελάχιστη ακρίβεια κατηγοριοποίησης του συνόλου εκπαίδευσης, ο κλάδος γίνεται μια τερματική διαδρομή.
    - Μια δεύτερη πιθανότητα για τον τερματισμό μιας διαδρομής του δέντρου είναι η απουσία ενός γνωρίσματος για τη συνέχιση της διαδικασίας διαχωρισμού του δέντρου. Για να είμαστε σίγουροι, εάν έχει επιλεγεί ένα κατηγοριακό γνώρισμα, οι τιμές του είναι σε θέση να διαιρέσουν το δέντρο μια μόνο φορά. Όμως, ένα αριθμητικό γνώρισμα μπορεί να χρησιμοποιηθεί για να διαχωρίσει τα δεδομένα πολλές φορές.

Δέντρα Απόφασης(Decision Trees)

- **Παράδειγμα Δέντρου Απόφασης**
  - Για το παράδειγμά μας των στιγμιότυπων εκπαίδευσης που ακολουθούν τον κλάδο που έχει *age > 43*, έχουν όλα μια τιμή no για την προσφορά ασφάλειας ζωής. Κατά συνέπεια, τερματίζουμε αυτή τη διαδρομή και ονομάζουμε τον κόμβο *life insurance promotion = no*.
  - Στη συνέχεια παίρνουμε τη διαδρομή με *age <= 43*. Αυτή διαδρομή δείχνει 9 στιγμιότυπα που έχουν *yes* για το γνώρισμα εξόδου και 3 στιγμιότυπα που έχουν *no* για το γνώρισμα εξόδου.
  - Καθώς υπάρχει τουλάχιστον ένα ακόμη γνώρισμα που πρέπει να εφαρμοστεί, είμαστε σε θέση να συνεχίσουμε την κατασκευή του δέντρου.
  - Το βήμα 4β του αλγορίθμου μας λέει ότι τα στιγμιότυπα που ακολουθούν αυτή τη διαδρομή αντιστοιχίζονται σαν νέες τιμές του T.
  - Μετά την αντιστοίχηση του T, τα βήματα του αλγορίθμου 2, 3, και 4 επαναλαμβάνονται.
  - Αυτή η διαδικασία συνεχίζεται μέχρις ότου όλες οι διαδρομές να συναντήσουν κριτήρια τερματισμού ή όλες οι πιθανότητες επιλογών γνωρισμάτων να έχουν εξαντληθεί.

Δέντρα Απόφασης(Decision Trees)

- **Παράδειγμα Δέντρου Απόφασης**
  - Ένα δέντρο αποφάσεων τριών κόμβων για τη βάση δεδομένων προώθησης πιστωτικών καρτών



Δέντρα Απόφασης(Decision Trees)

- **Κανόνες δέντρων αποφάσεων**
  - Οι κανόνες έχουν την τάση να είναι πιο ελκυστικοί από τα δέντρα.
  - Για να καταδείξουμε τη διαδικασία απλοποίησης των κανόνων, θεωρήστε το δέντρο αποφάσεων της πρηγούμενης διαφάνειας.
  - Ένας κανόνας που δημιουργήθηκε ακολουθώντας μια διαδρομή του δέντρου φαίνεται εδώ:

    IF *Age <= 43 & Sex = Male & Credit Card Insurance = No* **THEN** *Life Insurance Promotion = No*
  - Οι συνθήκες γι' αυτό τον κανόνα καλύπτουν 4 από τα 15 στιγμιότυπα των δεδομένων εκπαίδευσης με ακρίβεια 75%.

Δέντρα Απόφασης(Decision Trees)

- **Άλλες μέθοδοι κατασκευής δέντρων αποφάσεων**
  - Ο C4.5 είναι ο πιο πρόσφατος αλγόριθμος κατασκευής δέντρων αποφάσεων του Quinlan.
  - Ο ID3 (Quinlan, 1986) έχει μελετηθεί εκτενώς και είναι ο πρόδρομος του C4.5.
  - Ο CART (Breiman et al., 1984) παρουσιάζει ιδιαίτερο ενδιαφέρον καθώς πολλά εμπορικά προϊόντα υλοποιούν παραλλαγές του αλγορίθμου. Επιπλέον, ήταν το πρώτο σύστημα που εισήγαγε τα **δέντρα παλινδρόμησης (regression trees)**. Ουσιαστικά, τα δέντρα παλινδρόμησης παίρνουν τη μορφή των δέντρων αποφάσεων, όπου οι κόμβοι είναι αριθμητικές αντί για κατηγοριακές τιμές.

Δέντρα Απόφασης(Decision Trees)

- **Δημιουργία Συνδυαστικών Κανόνων**
  - Η **ανάλυση ομοιότητας (affinity analysis)** είναι η γενική διαδικασία καθορισμού των πραγμάτων που πάνε μαζί.
  - Μια τυπική εφαρμογή είναι η ανάλυση του καλαθιού αγοράς, της οποίας ο σκοπός είναι να καθορίσουμε εκείνα τα προϊόντα που είναι πιθανόν να αγοραστούν από έναν πελάτη κατά τη διάρκεια μιας αγοράς.
  - Η έξοδος της ανάλυσης καλαθιού αγοράς είναι ένα σύνολο συνδυασμών για τη συμπεριφορά πελάτη - αγορών.
  - Οι συνδυασμοί δίνονται στη μορφή ενός ειδικού συνόλου κανόνων, γνωστών σαν συνδυαστικών κανόνων.
  - Οι συνδυαστικοί κανόνες χρησιμοποιούνται για να βοηθήσουν στον καθορισμό κατάλληλων στρατηγικών προώθησης προϊόντων.

Δέντρα Απόφασης(Decision Trees)

- **Δημιουργία Συνδυαστικών Κανόνων**
  - **Εμπιστοσύνη και υποστήριξη**
    - Ας υποθέσουμε ότι θέλουμε να καθορίσουμε εάν υπάρχουν ενδιαφέρουσες σχέσεις που πρέπει να βρεθούν στις τάσεις αγοράς ενός πελάτη, μεταξύ των παρακάτω τεσσάρων ειδών ενός παντοπωλείου :
      1. Γάλα
      2. Τυρί
      3. Ψωμί
      4. Αβγά
    - Πιθανοί συνδυασμοί περιλαμβάνουν τα παρακάτω :

Δέντρα Απόφασης(Decision Trees)

- **Δημιουργία Συνδυαστικών Κανόνων**
  - **Εμπιστοσύνη και υποστήριξη**
    1. Εάν οι πελάτες αγοράζουν γάλα, αγοράζουν επίσης ψωμί.
    2. Εάν οι πελάτες αγοράζουν ψωμί, αγοράζουν επίσης γάλα.
    3. Εάν οι πελάτες αγοράζουν γάλα και αβγά, αγοράζουν οι επίσης τυρί και ψωμί.
    4. Εάν οι πελάτες αγοράζουν γάλα, τυρί και αβγά, αγοράζουν επίσης ψωμί.
  - Ο κάθε συνδυαστικός κανόνας σχετίζεται με μια **εμπιστοσύνη (confidence)**.
  - Γι' αυτό τον κανόνα η εμπιστοσύνη είναι η πιθανότητα υπό συνθήκες αγοράς ψωμιού, δεδομένης της αγοράς γάλατος.

Δέντρα Απόφασης(Decision Trees)

- **Δημιουργία Συνδυαστικών Κανόνων**
  - **Εμπιστοσύνη και υποστήριξη**
    - Κατά συνέπεια, εάν ένας συνολικός αριθμός 10.000 συναλλαγών πελατών σχετίζεται με αγορά γάλακτος, και 5.000 από αυτές τις ίδιες συναλλαγές σχετίζονται επίσης με αγορά ψωμιού, η εμπιστοσύνη της αγοράς ψωμιού, δεδομένης της αγοράς γάλακτος, είναι 5.000/10.000 = 50%.
    - Μια σημαντική πληροφορία που η τιμή εμπιστοσύνης ενός κανόνα δεν παρέχει είναι το ποσοστό όλων των συναλλαγών που περιέχουν τις τιμές γνωρισμάτων που βρίσκονται σ' ένα σχετιζόμενο κανόνα.
    - Αυτό το στατιστικό μέγεθος είναι γνωστό σαν **υποστήριξη (support)** για ένα κανόνα.
    - Η υποστήριξη είναι απλά το ελάχιστο ποσοστό των στιγμιότυπων στη βάση δεδομένων που περιέχει όλα τα στοιχεία που εμφανίζονται σ' ένα συγκεκριμένο συνδυαστικό κανόνα.

Ανάλυση και επεξεργασία DataSet

- **Dataset Επιλογής Μεθόδου Αντισύλληψης**
  - **Περιγραφή των δεδομένων**
    - Αυτό το σύνολο των δεδομένων είναι ένα υποσύνολο του 1987 National Indonesia Contraceptive Prevalence Survey.
    - Τα δείγματα είναι έγγαμες γυναίκες οι οποίες κατά τη διάρκεια της συνέντευξης δε γνωρίζανε αν ήταν έγγειες και κάποιες βέβαια δεν ήταν.
    - Το ζητούμενο είναι να προβλεφθεί όπως αναφέρθηκε και στην εισαγωγή η επιλογή της μεθόδου αντισύλληψης βασιζόμενη στο κοινωνικοοικονομικό προφίλ των ερωτηθέντων.

Ανάλυση και επεξεργασία DataSet

- **Dataset Επιλογής Μεθόδου Αντισύλληψης**
  - **Why interested in Contraceptive Method Choice**
    - Η Ινδονησία είναι μία χώρα με πολύ μεγάλο πληθυσμό. Το 2005 η Ινδονησία βρισκόταν στην τέταρτη θέση παγκοσμίως για τον πληθυσμό της. Παρόλο την μεγάλη έκταση της είναι πολύ πυκνοκατοικημένη. Κι αυτό εξαιτίας του πολύ γρήγορου ρυθμού γέννησης, που οδηγεί σε μεγάλες αριθμητικά οικογένειες.
    - Ο αριθμός των παιδιών σε μία οικογένεια συνήθως επηρεάζει το κοινωνικοοικονομικό επίπεδο της οικογένειας. Μεγαλύτερος αριθμός παιδιών συνεπάγεται και μικρότερο επίπεδο (στις περισσότερες των περιπτώσεων).
    - Η κυβέρνηση της Ινδονησίας έχει προωθήσει ένα Σχέδιο Οικογενειακού Προγραμματισμού όπου έχει συστηθεί κάθε οικογένεια να έχει το πολύ δύο παιδιά προωθώντας τη χρήση της αντισύλληψης.

Ανάλυση και επεξεργασία DataSet

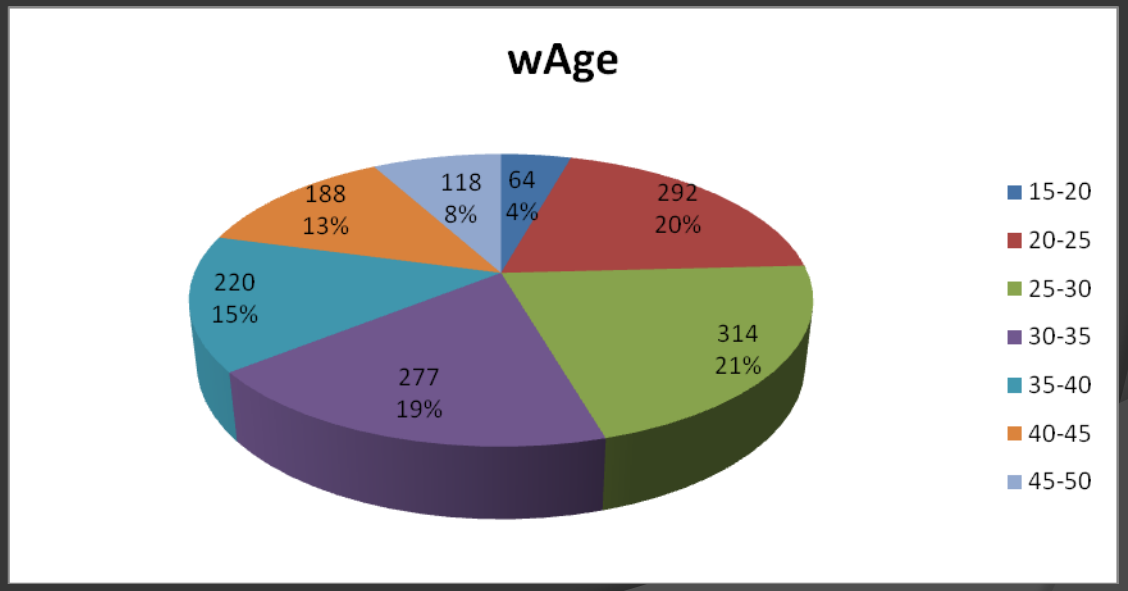
- **Dataset Επιλογής Μεθόδου Αντισύλληψης**
  - **Why interested in Contraceptive Method Choice**

name	description	type	values
Cont	Contraceptive Method	class label	1=No-use 2=Long-term 3=Short-term
wAge	wife's age	numeric	
wEdu	wife's education	categorical	1=low, 2, 3, 4=high
hEdu	husband's education	categorical	1=low, 2, 3, 4=high
nChild	number of children	numeric	
wReligion	wife's religion	binary	0=Non-Islam, 1=Islam
wWork	is the wife working?	binary	0=Yes, 1=No
hOcc	husband's occupation	categorical	1, 2, 3, 4
stand	standard of living	categorical	1=low, 2, 3, 4=high
media	exposure to media	binary	0=Good, 1=Not good

Ανάλυση και επεξεργασία DataSet

- Dataset Επιλογής Μεθόδου Αντισύλληψης
  - Στατιστική Ανάλυση Input χαρακτηριστικών
    - Ηλικία Γυναίκας

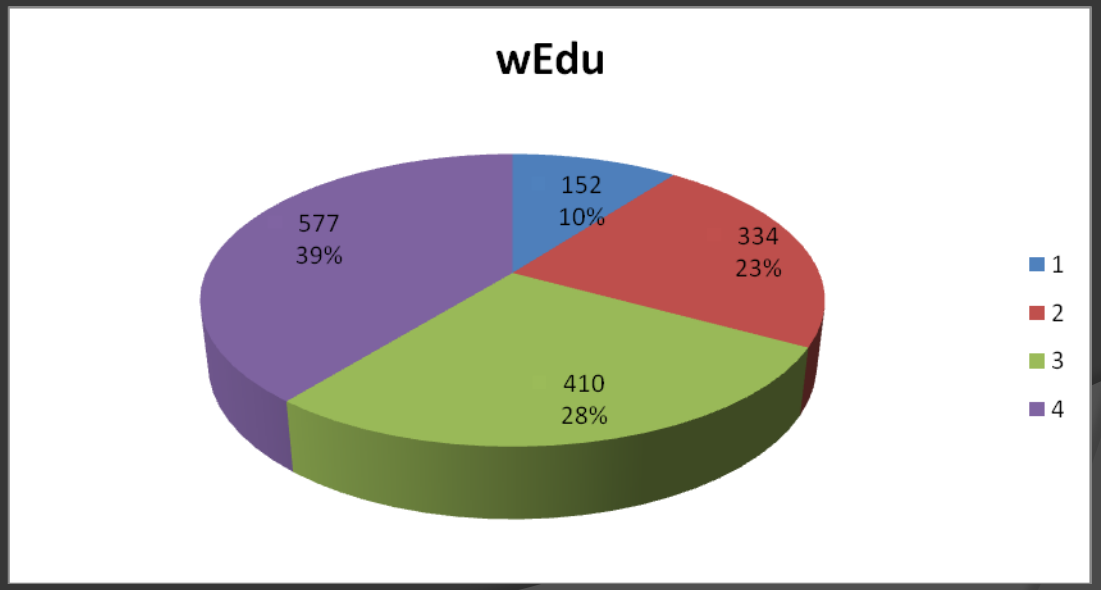
Τιμές	Πλήθος	Ποσοστό
15-20	64	4,34%
20-25	292	19,82%
25-30	314	21,32%
30-35	277	18,81%
35-40	220	14,94%
40-45	188	12,76%
45-50	118	8,01%
	1473	100,00%



Ανάλυση και επεξεργασία DataSet

- Dataset Επιλογής Μεθόδου Αντισύλληψης
  - Στατιστική Ανάλυση Input χαρακτηριστικών
    - Μορφωτικό επίπεδο γυναίκας

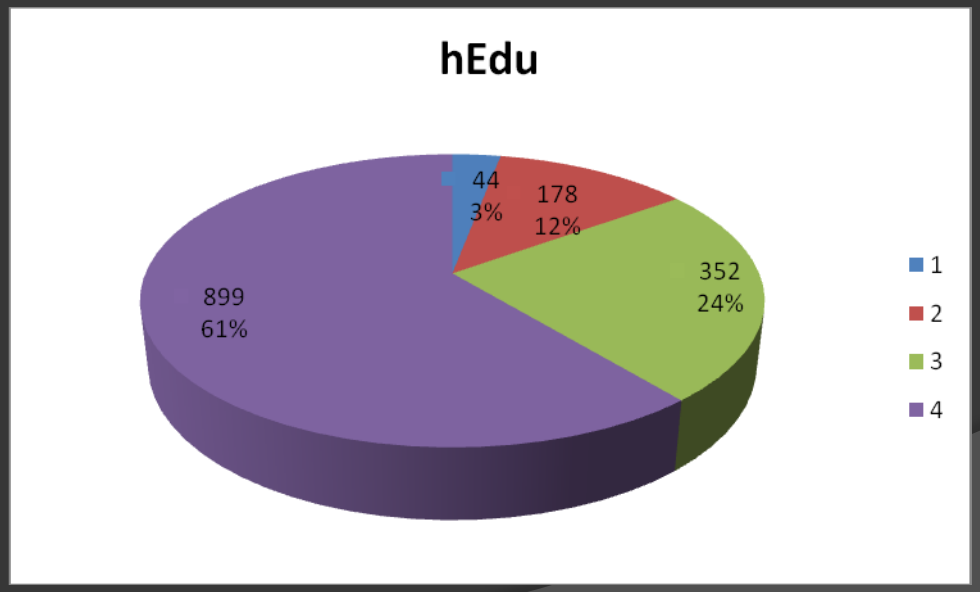
Τιμές	Πλήθος	Ποσοστό
1	152	10,32%
2	334	22,67%
3	410	27,83%
4	577	39,17%
	1473	100,00%



Ανάλυση και επεξεργασία DataSet

- Dataset Επιλογής Μεθόδου Αντισύλληψης
  - Στατιστική Ανάλυση Input χαρακτηριστικών
    - Μορφωτικό επίπεδο άντρα

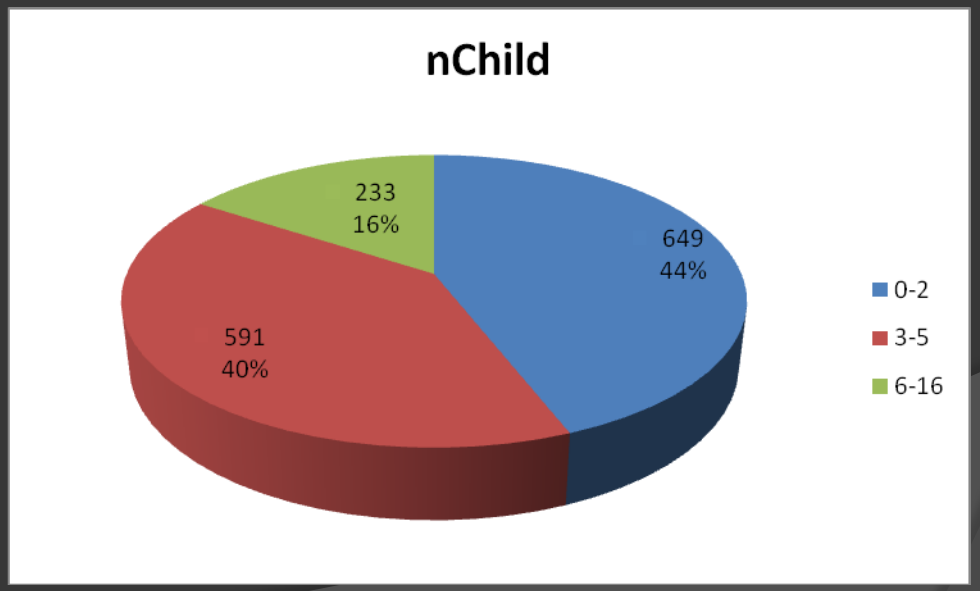
Τιμές	Πλήθος	Ποσοστό
1	44	2,99%
2	178	12,08%
3	352	23,90%
4	899	61,03%
	1473	100,00%



Ανάλυση και επεξεργασία DataSet

- Dataset Επιλογής Μεθόδου Αντισύλληψης
  - Στατιστική Ανάλυση Input χαρακτηριστικών
    - Αριθμός τέκνων

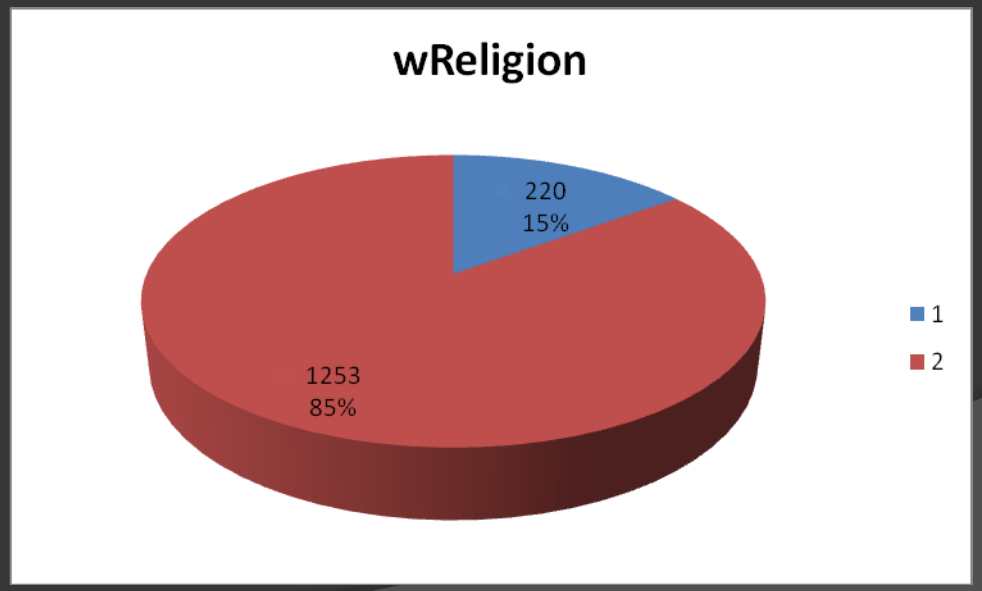
Τιμές	Πλήθος	Ποσοστό
0-2	649	44,06%
3-5	591	40,12%
6-16	233	15,82%
	1473	100,00%



Ανάλυση και επεξεργασία DataSet

- Dataset Επιλογής Μεθόδου Αντισύλληψης
  - Στατιστική Ανάλυση Input χαρακτηριστικών
    - Θρήσκευμα γυναίκας

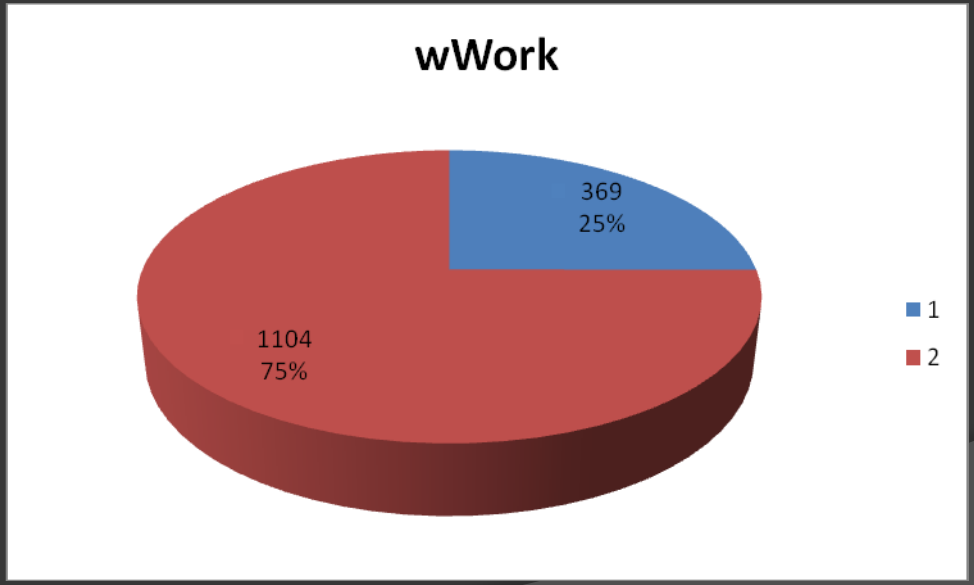
Τιμές	Πλήθος	Ποσοστό
0	220	14,94%
1	1253	85,06%
	1473	100,00%



Ανάλυση και επεξεργασία DataSet

- Dataset Επιλογής Μεθόδου Αντισύλληψης
  - Στατιστική Ανάλυση Input χαρακτηριστικών
    - Εργάζεται ή όχι η γυναίκα

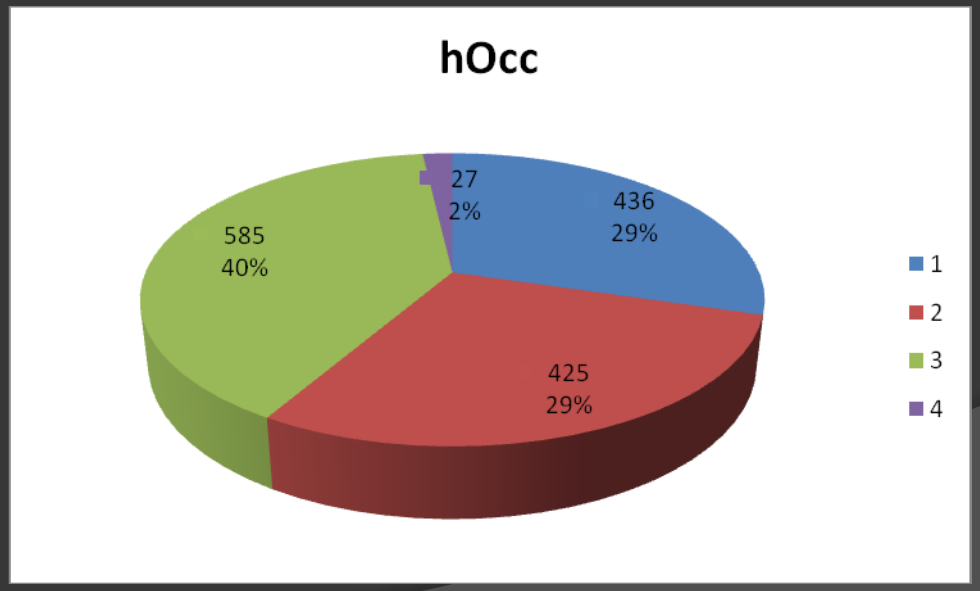
Τιμές	Πλήθος	Ποσοστό
0	369	25,05%
1	1104	74,95%
	1473	100,00%



Ανάλυση και επεξεργασία DataSet

- Dataset Επιλογής Μεθόδου Αντισύλληψης
  - Στατιστική Ανάλυση Input χαρακτηριστικών
    - Επάγγελμα άνδρα

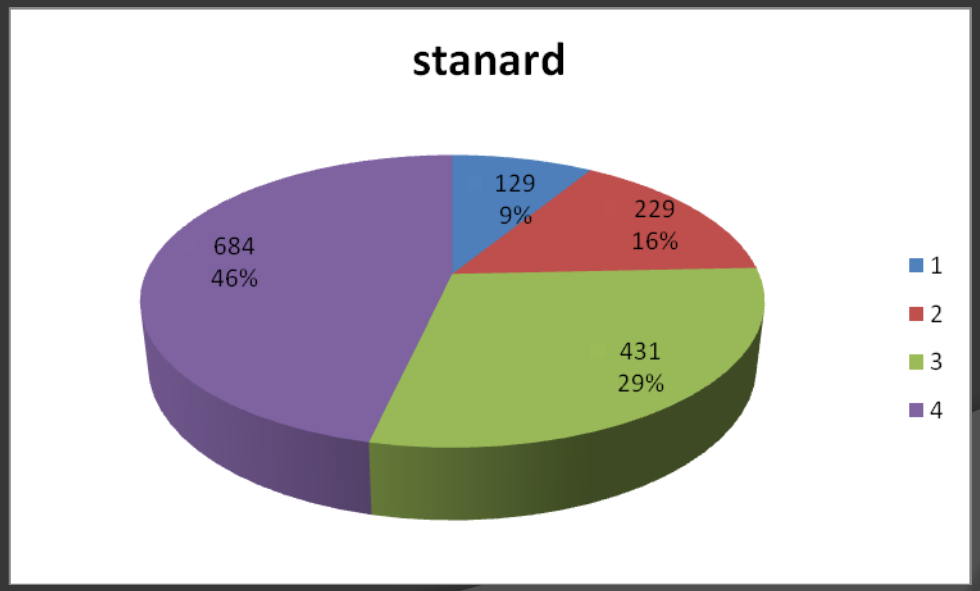
Τιμές	Πλήθος	Ποσοστό
1	436	29,60%
2	425	28,85%
3	585	39,71%
4	27	1,83%
	1473	100,00%



Ανάλυση και επεξεργασία DataSet

- Dataset Επιλογής Μεθόδου Αντισύλληψης
  - Στατιστική Ανάλυση Input χαρακτηριστικών
    - Standard Of Living

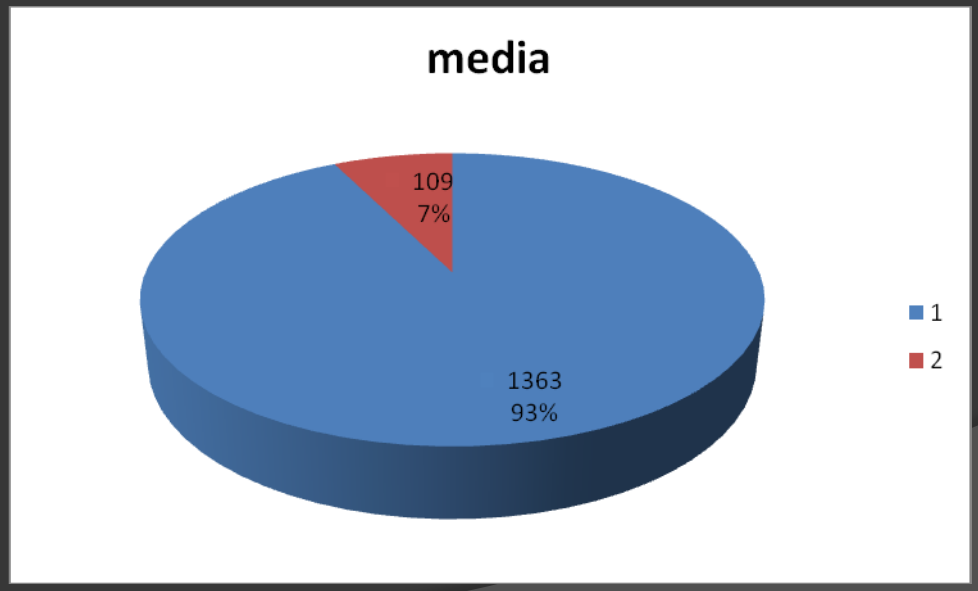
Τιμές	Πλήθος	Ποσοστό
1	129	8,76%
2	229	15,55%
3	431	29,26%
4	684	46,44%
	1473	100,00%



Ανάλυση και επεξεργασία DataSet

- Dataset Επιλογής Μεθόδου Αντισύλληψης
  - Στατιστική Ανάλυση Input χαρακτηριστικών
    - Media Exposure

Τιμές	Πλήθος	Ποσοστό
0	1363	92,60%
1	109	7,40%
	1472	100,00%



Φιλοσοφία Πειραμάτων

- **Περιγραφή της διαδικασίας**
- Για να κατασκευάσουμε ένα δέντρο το οποίο θα θεωρείται αξιόπιστο, θα πρέπει να πειραματιστούμε σε διάφορα δέντρα και συγκρίνοντας τα αποτελέσματα που θα μας δώσουν να αποφασίσουμε σε πιο θα καταλήξουμε, πιο θα θεωρηθεί ότι μας δίνει τα καλύτερα αποτελέσματα.
- Για το σκοπό αυτό κάθε φορά που “τρέχουμε” ένα δέντρο αλλάζουμε τα training και test data, αλλά και κάποιες από τις παραμέτρους δημιουργίας.
- Τα διάφορα σενάρια τα οποία εκτελέσαμε έγιναν με την εφαρμογή CTree. Η εφαρμογή διατίθεται δωρεάν από τον παρακάτω σύνδεσμο: *http://www.geocities.com/adotsaha/CTree/CTree.zip*
- Σε κάθε σειρά πειραμάτων, επιλέγουμε τις ίδιες παραμέτρους όσον αφορά το αν θα γίνεται κλάδεμα ή όχι στο δέντρο μας και πότε θα σταματάει η δημιουργία κόμβων στο δέντρο.
- Στην προκειμένη περίπτωση επιλέξαμε τις αρχικές τιμές.

Φιλοσοφία Πειραμάτων

- **Αναλυτική παρουσίαση σεναρίων**
  - **Σενάριο 1ο**
    - Επιλέξαμε να χρησιμοποιήσουμε τυχαία το 70% του συνόλου, δηλαδή 1004 στιγμιότυπα να είναι δεδομένα εκπαίδευσης και το υπόλοιπο τυχαίο 30%, δηλαδή 469 να είναι δεδομένα ελέγχου.
      - ***Classification Tree Model***

Number of Training observations	1004
Number of Test observations	469
Number of Predictors	9
Class Variable	ContMU
Number of Classes	3
Majority Class	1
% MissClassified if Majority Class is used as Predicted Class	39%

Φιλοσοφία Πειραμάτων

- **Αναλυτική παρουσίαση σεναρίων**
  - **Σενάριο 1ο**
    - ***Tree information***

Total Number of Nodes	489
Number of Leaf Nodes	297
Number of Levels	14
% Missclasssified	
On Training Data	19,62%
On Test Data	51,17%
Time Taken	
Data Processing	10 Sec
Tree Growing	1 Min : 34 Sec
Tree Pruning	41 Sec
Tree Drawing	49 Sec
Classification using final tree	43 Sec
Rule Generation	52 Sec
Total	4 Min : 50 Sec

Φιλοσοφία Πειραμάτων

- **Αναλυτική παρουσίαση σεναρίων**
  - **Σενάριο 1ο**
    - ***Confusion Matrix***

Training Data

True Class \ Predicted Class	1	2	3	
1	37 1	15	40	426
2	34	177	26	237
3	54	28	259	341
	45 9	220	325	1004

Test Data

True Class \ Predicted Class	1	2	3	
1	119	30	54	203
2	28	34	34	96
3	56	38	76	170
	203	102	164	469

Φιλοσοφία Πειραμάτων

- **Αναλυτική παρουσίαση σεναρίων**
  - **Σενάριο 1ο**
    - ***Rule Summary Table***

Rules 33

Rule ID	Class	Length	Support	Confidence	Capture
0	1	0	100,0%	42,4%	100,0%
1	1	1	6,4%	96,9%	14,6%
2	1	1	3,1%	64,5%	4,7%
3	1	2	8,8%	84,1%	17,4%
4	1	1	6,4%	71,9%	10,8%
5	1	1	2,0%	55,0%	2,6%
6	1	1	44,2%	52,9%	55,2%
7	1	2	4,1%	90,2%	8,7%
8	1	1	7,4%	68,9%	12,0%
9	1	3	2,5%	76,0%	4,5%
10	1	1	8,6%	55,8%	11,3%
11	1	2	1,2%	91,7%	2,6%
12	1	1	25,1%	52,4%	31,0%
13	1	1	12,1%	55,4%	15,7%
14	1	1	28,6%	51,9%	35,0%
15	1	2	5,1%	66,7%	8,0%
16	1	1	10,4%	64,4%	15,7%
17	3	2	12,3%	58,5%	21,1%
18	1	2	4,1%	92,7%	8,9%
19	1	2	4,2%	83,3%	8,2%
20	1	1	6,7%	53,7%	8,5%
21	1	2	3,1%	93,5%	6,8%
22	1	1	25,5%	61,7%	37,1%
23	1	2	4,5%	84,4%	8,9%
24	1	1	20,5%	55,3%	26,8%
25	3	1	0,8%	62,5%	1,5%
26	1	2	7,9%	79,7%	14,8%
27	2	3	3,7%	62,2%	9,7%
28	3	2	11,5%	53,9%	18,2%
29	3	2	22,4%	53,3%	35,2%
30	1	2	3,6%	88,9%	7,5%
31	1	1	1,6%	75,0%	2,8%
32	1	2	12,1%	76,9%	21,8%
33	3	2	19,1%	51,0%	28,7%

Φιλοσοφία Πειραμάτων

- **Αναλυτική παρουσίαση σεναρίων**
  - **Σενάριο 1ο**
    - ***Rule Text***

Rule0		ContMU = 1
Rule1	IF	nChild < 1
	THEN	ContMU = 1
Rule2	IF	hEdu = 1
	THEN	ContMU = 1
Rule3	IF	nChild < 2
	AND	wAge >= 28
	THEN	ContMU = 1
Rule4	IF	wAge >= 47
	THEN	ContMU = 1
Rule5	IF	hOcc = 4
	THEN	ContMU = 1

Φιλοσοφία Πειραμάτων

- **Αναλυτική παρουσίαση σεναρίων**
  - **Σενάριο 2ο**
    - Για την εκτέλεση του δεύτερου σεναρίου χρειαστήκαμε τη βοήθεια του πρώτου.
    - Σύμφωνα λοιπόν με τα αποτελέσματα του πρώτου βρέθηκε πως η ρίζα του δέντρου ήταν το μορφωτικό επίπεδο της γυναίκας.
    - Ταξινομήσαμε λοιπόν τα δεδομένα μας ως προς αυτό το χαρακτηριστικό και στη συνέχεια επιλέξαμε τυχαία το 70% του συνόλου ως δεδομένα εκπαίδευσης και το υπόλοιπο τυχαίο 30 % ως δεδομένα ελέγχου.

Φιλοσοφία Πειραμάτων

- **Αναλυτική παρουσίαση σεναρίων**
  - **Σενάριο 3ο**
    - Τα επόμενο σενάριο, έγινε με την λογική της ιεραρχίας που προέκυψε από τα πρώτο πείραμα.
    - Η κατανομή δηλαδή των training και test data, θα γίνει σύμφωνα με την ιεραρχία των χαρακτηριστικών που το πρώτο πείραμα μας έδωσε.
    - Στο πιο σημαντικό χαρακτηριστικό, το ποσοστό εμφάνισης θα είναι 50%, στο επόμενο 35%, στο επόμενο 15%.
    - Σύμφωνα με τα πείραμα 1, η ιεραρχία που προκύπτει είναι:
      1. Επίπεδο μόρφωσης γυναίκας
      2. Αριθμός τέκνων
      3. Επίπεδο μόρφωσης άνδρα

Φιλοσοφία Πειραμάτων

- **Αναλυτική παρουσίαση σεναρίων**
  - **Σενάριο 3ο**
    - Στην πράξη ήταν να επιλογή των χαρακτηριστικών να γίνει κατά 50% (δηλαδή 737 εγγραφές) από το μορφωτικό επίπεδο της γυναίκας. Αυτές τις 737 εγγραφές τις χωρίσαμε σύμφωνα με τη στατιστική ανάλυση.
    - Η στατιστική ανάλυση έδειξε πως το 10,32% των γυναικών, δηλαδή 152 γυναίκες έχουν μορφωτικό επίπεδο 1, 22,67% δηλαδή 334 επίπεδο 2, 27,83% δηλαδή 410 επίπεδο 3 και 39,17% δηλαδή 577 επίπεδο 4.
    - Πρακτικά αυτό σημαίνει ότι επιλέχθηκαν 50% από 152 δηλαδή 76 γυναίκες, οι οποίες έχουν επίπεδο μόρφωσης 1, 50% από 334 δηλαδή 167 επιπέδου 2, 50% από 410 δηλαδή 205 επιπέδου 3 και 50% από 577 δηλαδή 289 επιπέδου 4 άσχετα με τα υπόλοιπα χαρακτηριστικά τους. Έχουμε έτσι ένα σύνολο 737 δεδομένων, που είναι το 50% των συνολικών δεδομένων.
    - Από τις 76 γυναίκες μορφωτικού επιπέδου 1 τα 2/3 θα μπουν ως training data (51) και το 1/3 ως test data (25). Το ίδιο θα ισχύσει και για τις υπόλοιπες γυναίκες των αντίστοιχων μορφωτικών επιπέδων.
    - Την πρακτική που ακολουθήσαμε για το πρώτο χαρακτηριστικό την εφαρμόζουμε και στα άλλα χαρακτηριστικά και κάθε φορά στο υπόλοιπο του συνόλου δεδομένων

Φιλοσοφία Πειραμάτων

- **Αναλυτική παρουσίαση σεναρίων**
  - **Σενάριο 4ο**
    - Το σενάριο αυτό είναι το ίδιο με το πρώτο μόνο που έχουμε πλέον το 75% του συνόλου δεδομένων ως δεδομένα εκπαίδευσης και το 25% ως δεδομένα ελέγχου.

Επίλογος

- **Στατιστική αποτίμηση σφαλμάτων**
  - Η αξιολόγηση των μοντέλων γίνεται με την εφαρμογή τους στα δεδομένα ελέγχου (test set).
  - Το πιο γενικό μέτρο της απόδοσης του μοντέλου και αυτό που θα χρησιμοποιήσουμε εμείς για να καταλήξουμε στο καλύτερο μοντέλο είναι το **ποσοστό σφάλματος κατηγοριοποιητή (classifier error rate)** το οποίο δίνεται από τον ακόλουθο τύπο**:**

$$classifier\ error\ rate\ (E) = \frac{\#\ of\ test\ set\ errors}{\#\ of\ test\ set\ instances}$$

Επίλογος

- **Στατιστική αποτίμηση σφαλμάτων**
  - Παρατηρούμε ότι το μικρότερο σφάλμα κατηγοριοποιητή το έχει το 4ο σενάριο, τα αποτελέσματα μας δείχνουν δηλαδή ότι αυτό μελλοντικά θα μας δίνει τις καλύτερες προβλέψεις.

Σενάριο	% clsassifier error rate
1	51,17
2	50,58
3	53,77
4	47,75