Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Τμ. Μηχανικών Πληροφορικής και Υπολογιστών

Προετοιμασία των δεδομένων

Βασικά Στάδια κατά την Προετοιμασία των δεδομένων

- Καθαρισμός των Δεδομένων (**cleaning**)
  - Ελλιπείς τιμές (missing values)
  - Μη σωστά δεδομένα – ακραίες τιμές (outliers)
  - Δεδομένα με προβλήματα συσχέτισης(Inconsistent data)
  - Διπλοεγγραφές
- Σύνδεση δεδομένων από διαφορετικές πηγές (βάσεις δεδομένων, αρχεία,…) (**integration**)
- Μετασχηματισμός των δεδομένων (**transformation**)

Ελλιπείς τιμές (missing values)

1	Gender	Race	Birth_Ye...	Marital_...	Years_o...	Hours_P...	Preferre...	Preferre...	Preferre...	Read_N...	Online_...	Online_...	Facel
2	M	White	1972	M	8	1	Firefox	Google	Yahoo	Y	N	N	Y
3	M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N	Y
4	F	African A...	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y		Y
5	F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N	N
6	M	White	1954	M	2	3	Internet ...	Bing	Hotmail	Y	Y	N	Y
7	M	African A...	1982	D	15	4	Internet ...	Google	Yahoo	Y	N	Y	N
8	M	African A...	1981	D	11	2	Firefox	Google	Yahoo		Y		Y
9	M	White	1977	S	3	3	Internet ...	Yahoo	Yahoo	Y			Y
10	F	African A...	1969	M	6	2	Firefox	Google	Gmail	N	Y	N	N
11	M	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y		Y	Y
12	F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	N	N	Y

Ελλιπείς τιμές

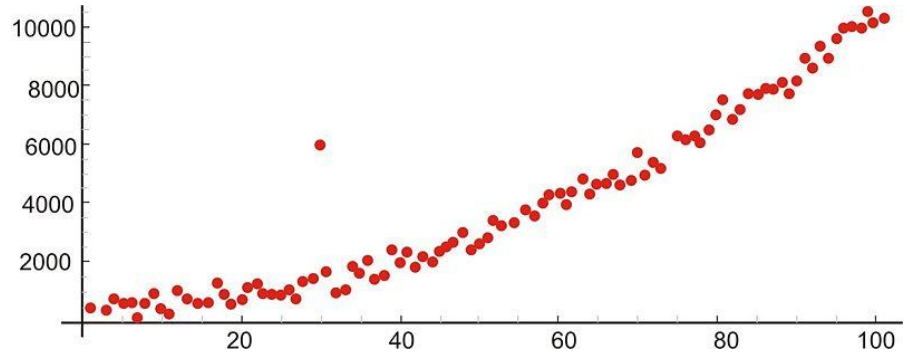
Ελλιπείς τιμές (missing values)

- Αιτία
  - Από λάθος κατά την εισαγωγή των δεδομένων.
  - Κάποια δεδομένα επειδή θεωρούνται π.χ. ευαίσθητα δεν θα έπρεπε να εμφανιστούν.
- Λύσεις
  - Διαγραφή των εγγραφών με ελλιπείς στοιχεία (μόνο στην περίπτωση που είναι μικρό το πλήθος και δεν επηρεάζει τα συνολικά δεδομένα).
  - Συμπλήρωση των τιμών μέσω υπολογισμού τους με βάση στατιστικούς δείκτες (π.χ. μέσος όρος,..).
  - Διατήρηση των εγγραφών ως έχουν.

Μη σωστά δεδομένα (outliers)

Μη σωστά δεδομένα (outliers)

- Μη ομαλές είναι οι τιμές που δεν ακολουθούν την κατανομή των άλλων τιμών.
- Αυτές οι τιμές ονομάζονται ακρότατα ή ακραίες τιμές (outliers)
- Τα ακρότατα εμφανίζονται σε περιπτώσεις λάθους εισαγωγής δεδομένων ή λόγω κάποιου περίεργου γεγονότος το οποίο χρήζει περαιτέρω ανάλυσης (π.χ. στις περιπτώσεις οικονομικής απάτης).



Ανακριβείς τιμές (Inconsistent data)

- Τα ανακριβή (inconsistent) δεδομένα, παρατηρούνται
  - Όταν έχουμε δεδομένα από διαφορετικές πηγές και τα ενώνουμε σε ένα αρχείο, όπου παρατηρούνται περιπτώσεις να μην όλα τα δεδομένα σωστά.
  - Επίσης το αρχείο το οποίο έχουμε για να κάνουμε την ανάλυση να σχεδιάστηκε για άλλο σκοπό.
  - Όταν οι τιμές σε ένα πεδίο είναι λάθος π.χ. μεταγενέστερη ημερομηνία εγγραφής.

Διπλοεγγραφές

Υπάρχουν περιπτώσεις να υπάρχει 2 ή περισσότερες φορές μια εγγραφή.

- Οι διπλοεγγραφές εμφανίζονται συνήθως όταν ενώνουμε αρχεία από διαφορετικές πηγές. Αν για παράδειγμα ένας φοιτητής έχει 2 διευθύνσεις τότε τα δεδομένα του συγκεκριμένου φοιτητή είναι πολύ πιθανό να εμφανιστούν 2 φορές, αν δεν έχει ληφθεί υπόψιν αυτός ο περιορισμός.
- Σε αυτές τις περιπτώσεις πραγματοποιούμε καθαρισμό των δεδομένων (Data cleaning) με διαγραφή των διπλό-εγγραφών.

Μετασχηματισμός των δεδομένων και μείωση του πλήθους.

Υπάρχουν 2 κατηγορίες μετασχηματισμών

- Μετατροπή δεδομένων σε άλλους τύπους

- Μείωση αριθμού δεδομένων ή πεδίων
  - Δειγματοληψία
  - Μείωση στηλών
  - Μέτρα ομοιότητας/απόστασης

Μετασχηματισμός των δεδομένων (transformation)

Μέσω των μετασχηματισμών επιτυγχάνεται η μετατροπή των δεδομένων από μια μορφή σε μια άλλη.

Για παράδειγμα μπορούμε να μετατρέψουμε ένα τακτικό (ordinal) δεδομένο σε δυαδικό (binary).

ordinal

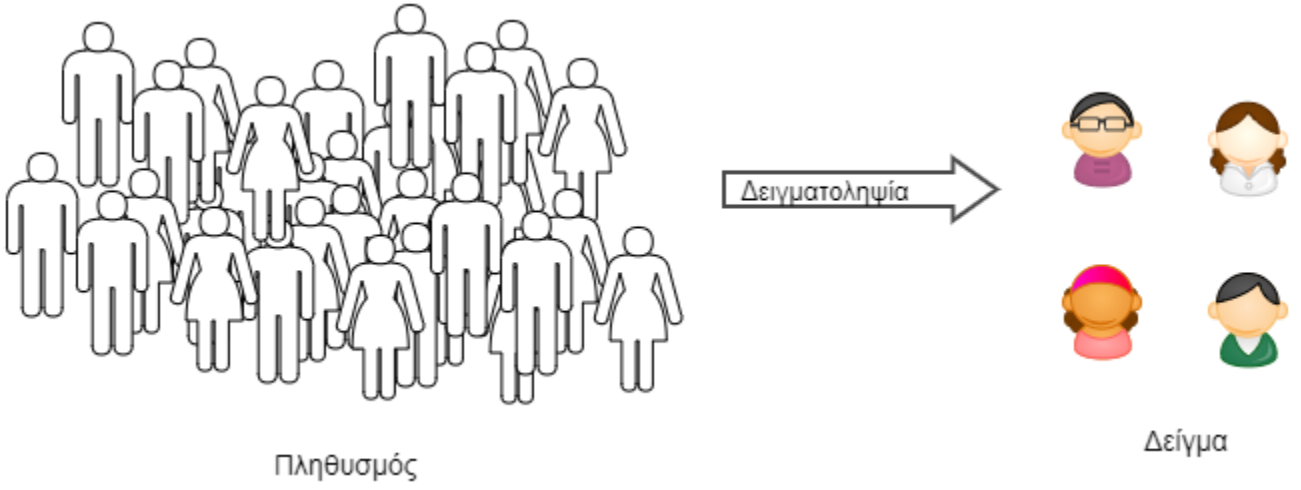
Θερμοκρασία
Κρύο
Μέτρια
Ζέστη

binary

Εκδρομή
False
True
True

Δειγματοληψία

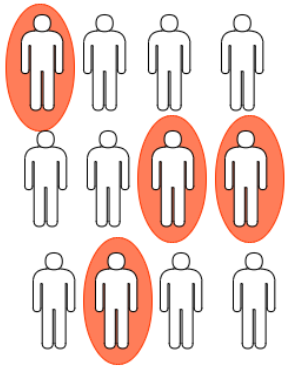
- Σε περιπτώσεις όπου υπάρχει ένα πολύ μεγάλο σύνολο δεδομένων μπορούμε να πάρουμε ένα αντιπροσωπευτικό δείγμα προκειμένου να πραγματοποιήσουμε την ανάλυσή μας.
- Το δείγμα αν έχει εξαχθεί σωστά οδηγεί στα ίδια αποτελέσματα σε σχέση με τη χρησιμοποίηση όλου του δείγματος.



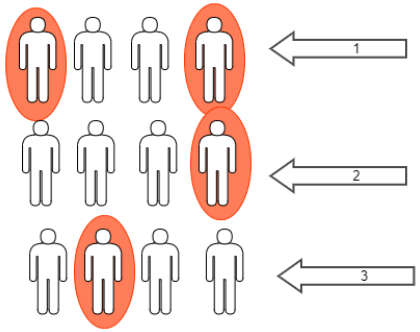
Δειγματοληψία

Μέθοδοι δειγματοληψίας

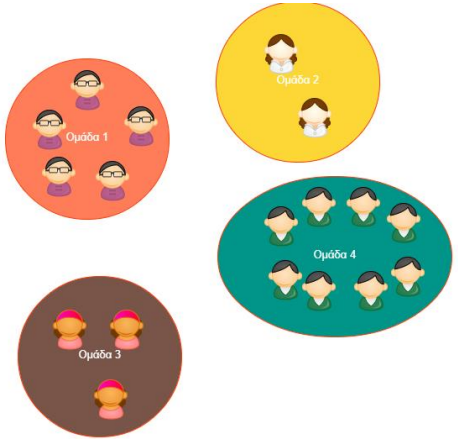
- **Απλή:** Επιλέγονται τυχαία δείγματα από το σύνολο των δεδομένων.



- **Στρωματοποιημένη δειγματοληψία (Stratified):** Δημιουργούνται υποσύνολα με βάση κάποιο συγκεκριμένο χαρακτηριστικό, από τα οποία επιλέγονται τυχαίες εγγραφές.



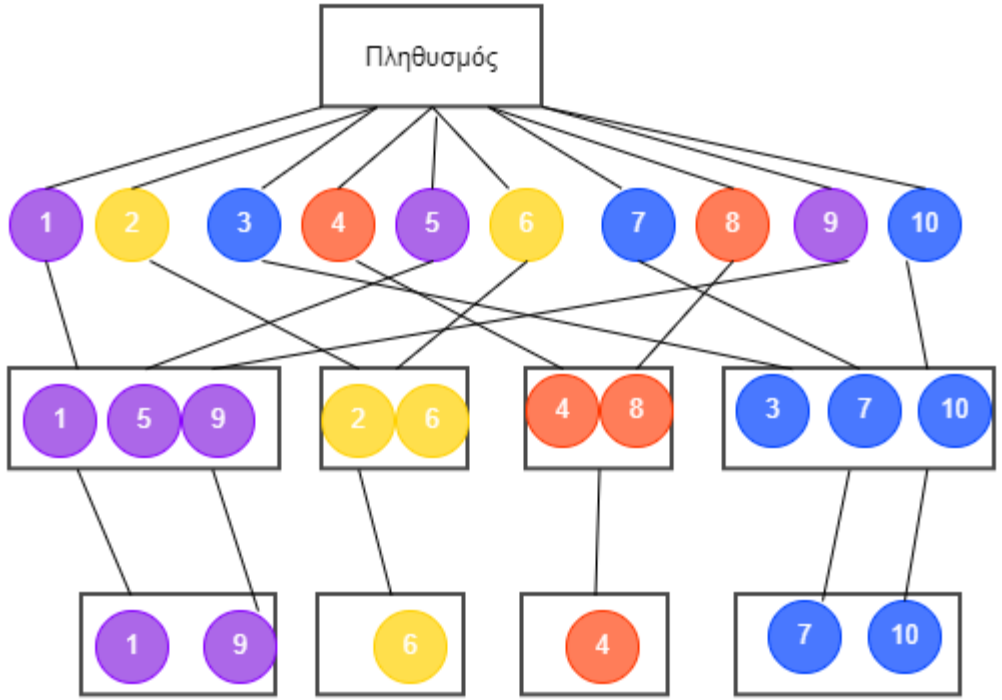
- **Δειγματοληψία ομάδων (Cluster):** Το σύνολο των δεδομένων διαχωρίζεται σε clusters με βάση κάποιο χαρακτηριστικό και στη συνέχεια ένα τυχαίο δείγμα από τα cluster επιλέγεται και αναλύεται.

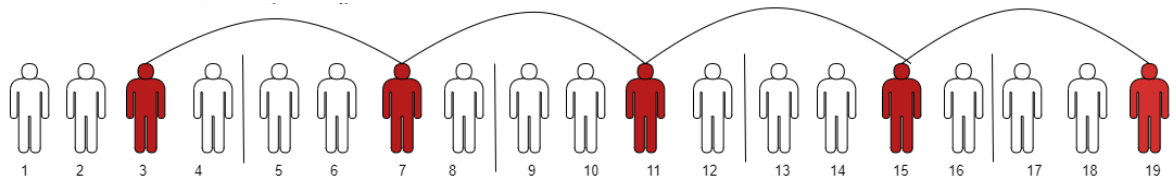


Δειγματοληψία

Μέθοδοι δειγματοληψίας

- **Πολύ-επίπεδη (Multistage):** Σε αυτή τη μεθοδολογία το δείγμα σπάει σε clusters και στη συνέχεια από τα αρχικά clusters, δημιουργούνται νέα και στη συνέχεια από τα νέα clusters επιλέγονται δεδομένα και αναλύονται. Η διαδικασία της δημιουργίας νέων clusters μπορεί να συνεχιστεί μέχρι το επιθυμητό στάδιο.
- **Systematic sampling:** Το δείγμα δημιουργείται ορίζοντας κάποια τιμή με βάση την οποία θα επιλέγουμε τα δεδομένα. Έστω ότι επιλέγουμε από όλο το δείγμα το δεδομένο στη θέση 3 και λέμε ότι θέλουμε τα στοιχεία που βρίσκονται κάθε φόρα 4(x) θέσεις μετά το επιλεγμένο. Τότε θα έχουμε
- **1.Στάδιο 3**
- **2.Στάδιο 3+x=3+4=7**
- **2.Στάδιο 7+x=7+4=11……**





Μείωση στηλών

Η μείωση των πεδίων που θα επεξεργαστούμε γίνεται με 2 τρόπους.

- Δημιουργία νέων χαρακτηριστικών.

  Ανάλυση Πρωταρχικών Συνιστωσών – Principal Component Analysis (PCA), όπου μπορούμε για παράδειγμα 2 μεταβλητές να τις ενοποιήσουμε σε μια νέα μεταβλητή.

- Επιλογή συγκεκριμένων πεδίων.
  - Πίνακας συσχετίσεων, όπου μπορούμε να ελέγξουμε τη συσχέτιση των μεταβλητών και αν δούμε ότι υπάρχει υψηλή συσχέτιση τότε να μην συμπεριλάβουμε στην ανάλυση τις συσχετιζόμενες μεταβλητές.

Μέτρηση Ομοιότητας

Ένας τρόπος καθορισμού του βαθμού ομοιότητας δύο αντικειμένων αποτελεί η μέτρηση της απόστασης μεταξύ των αντικειμένων.

Οι μέθοδοι που θα χρησιμοποιήσουμε και οι οποίοι θα παρουσιαστούν στα επόμενα κεφάλαια είναι

- Η Ευκλείδεια Απόσταση και
- Η Απόσταση Manhattan

ΤΕΛΟΣ