Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Τμ. Μηχανικών Πληροφορικής και Υπολογιστών

Καθαρισμός Δεδομένων

Καθαρισμός των Δεδομένων (cleaning)

Σε αυτή την ενότητα θα συζητήσουμε σχετικά με τα παρακάτω προβλήματα που εντοπίζονται κατά την ανάλυση δεδομένων και περιλαμβάνονται στο στάδιο του καθαρισμού των δεδομένων:

- Ελλιπείς τιμές (missing values)
- Μη σωστά δεδομένα (outliers)
- Διπλοεγγραφές

Ελλιπείς τιμές (missing values)

- Ελλιπείς τιμές έχουμε όταν παρατηρούμε να λείπουν τιμές από τα αρχικά μας δεδομένα.

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)

Προκειμένου να διορθώσουμε τις εγγραφές που παρουσιάζουν ελλιπείς τιμές μπορούμε να εφαρμόσουμε μια από τις παρακάτω μεθόδους

Αντικατάσταση τιμής με βάση

- τον ΜΟ των υπολοίπων τιμών
- Τη μικρότερη τιμή
- Τη μεγαλύτερη τιμή
- Να βάλουμε μια δική μας τιμή
- Να βάλουμε το μηδέν.

missing_values

Δεδομένα με ακραίες τιμές(outliers)

Μέσω της διαδικασίας «Outlier detection» μπορούμε να εντοπίσουμε αντικείμενα τα οποία παρουσιάζουν συμπεριφορά διαφορετική από την αναμενόμενη.

Σε αυτά τα αντικείμενα συνήθως εστιάζει ένας αναλυτής για περαιτέρω ανάλυση γιατί ενδεχομένως

- Να δείχνουν μια οικονομική απάτη
- Να ανιχνεύουν κάποιο φαινόμενο σε μια ανάλυση ιατρικών δεδομένων
- Να ανιχνεύουν παραβιάσεις στην ασφάλεια ενός δικτύου.

Βέβαια υπάρχουν και περιπτώσεις που μπορεί απλά να προστέθηκαν από λάθος στο δείγμα και να πρέπει να απομακρυνθούν για διορθωθεί τα δεδομένα.

outlier

Διπλοεγγραφές

Προκειμένου να απομακρύνουμε τις διπλοεγγραφές από το δείγμα μας χρησιμοποιούμε το αντικείμενο με τίτλο «Remove Duplicates»