

ΕΡΓΑΣΙΑ ΓΙΑ ΤΟ ΜΑΘΗΜΑ:  
PRIVACY PRESERVING DATA PUBLISHING PROJECT

ΤΙΤΛΟΣ ΕΡΓΑΣΙΑΣ:  
ΨΗΦΙΑΚΗ ΑΣΦΑΛΕΙΑ ΚΑΙ ΙΔΙΩΤΙΚΟΤΗΤΑ

ΕΚΠΟΝΗΘΗΚΕ ΑΠΟ ΤΗ ΦΟΙΤΗΤΡΙΑ:  
ΜΟΝΑΝΤΕΡΟΥ ΑΓΓΕΛΙΚΗ, [amonanteou@uth.gr](mailto:amonanteou@uth.gr)   
ΑΜ: M012022010, ΕΞΑΜΗΝΟ ΣΠΟΥΔΩΝ: 10

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:   
ΣΠΑΘΟΥΛΑΣ ΓΕΩΡΓΙΟΣ

ΛΑΜΙΑ ΔΕΚΕΜΒΡΙΟΣ 2022

**Περιεχόμενα**

[ΕΙΣΑΓΩΓΗ 3](#__RefHeading___Toc2531_1102466801)

[ΑΝΑΓΝΩΡΙΣΗ ΠΕΔΙΩΝ 3](#__RefHeading___Toc2533_1102466801)

[ΥΠΟΛΟΓΙΖΟΝΤΑΣ ΤΙΣ ΠΙΘΑΝΟΤΗΤΕΣ 3](#__RefHeading___Toc2537_1102466801)

[*α) Για κάποιον από τους κατοίκους της πόλης γνωρίζετε ότι είναι άντρας, 70 ετών και White: British, και ότι πρόσφατα νοσηλεύτηκε στο συγκεκριμένο νοσοκομείο. Τι συμπεράσματα μπορείτε να βγάλετε για τις πιθανότητες να πάσχει από κάποια συγκεκριμένη ασθένεια?* 3](#__RefHeading___Toc257_1273455314)

[*β) Υπάρχει κάποιος στην λίστα των κατοίκων της πόλης ο οποίος, εάν ξέρετε ότι επισκέφθηκε το νοσοκομείο, μπορείτε να είστε σίγουροι για τον λόγο?* 4](#__RefHeading___Toc259_1273455314)

[GENERALIZATION 4](#__RefHeading___Toc2539_1102466801)

# ΕΙΣΑΓΩΓΗ

Στην παρούσα εργασία υλοποιηθούμε ένα πρόγραμμα, ικανό να διαχειριστεί το πλήθος των εγγραφών (αρχεία των 10,000 εγγραφών) σε Python.

# **ΑΝΑΓΝΩΡΙΣΗ ΠΕΔΙΩΝ**

Τα πεδία που αποτελούν το σύνολο QID είναι τα πεδία που ενδεχομένως μπορούν να αποκαλύψουν το άτομο που σχετίζεται με τα δεδομένα.Αυτά είναι:

* όνομα και επώνυμο (όνομα) (από citizens\_list.csv)
* φύλο (και από τα δύο αρχεία)
* ηλικία (και από τα δύο αρχεία)
* φυλή (και από τα δύο αρχεία)

Set T:

ασθένεια (από hospital\_data.csv)

# ΥΠΟΛΟΓΙΖΟΝΤΑΣ ΤΙΣ ΠΙΘΑΝΟΤΗΤΕΣ

## *α) Για κάποιον από τους κατοίκους της πόλης γνωρίζετε ότι είναι άντρας, 70 ετών και White: British, και ότι πρόσφατα νοσηλεύτηκε στο συγκεκριμένο νοσοκομείο. Τι συμπεράσματα μπορείτε να βγάλετε για τις πιθανότητες να πάσχει από κάποια συγκεκριμένη ασθένεια?*

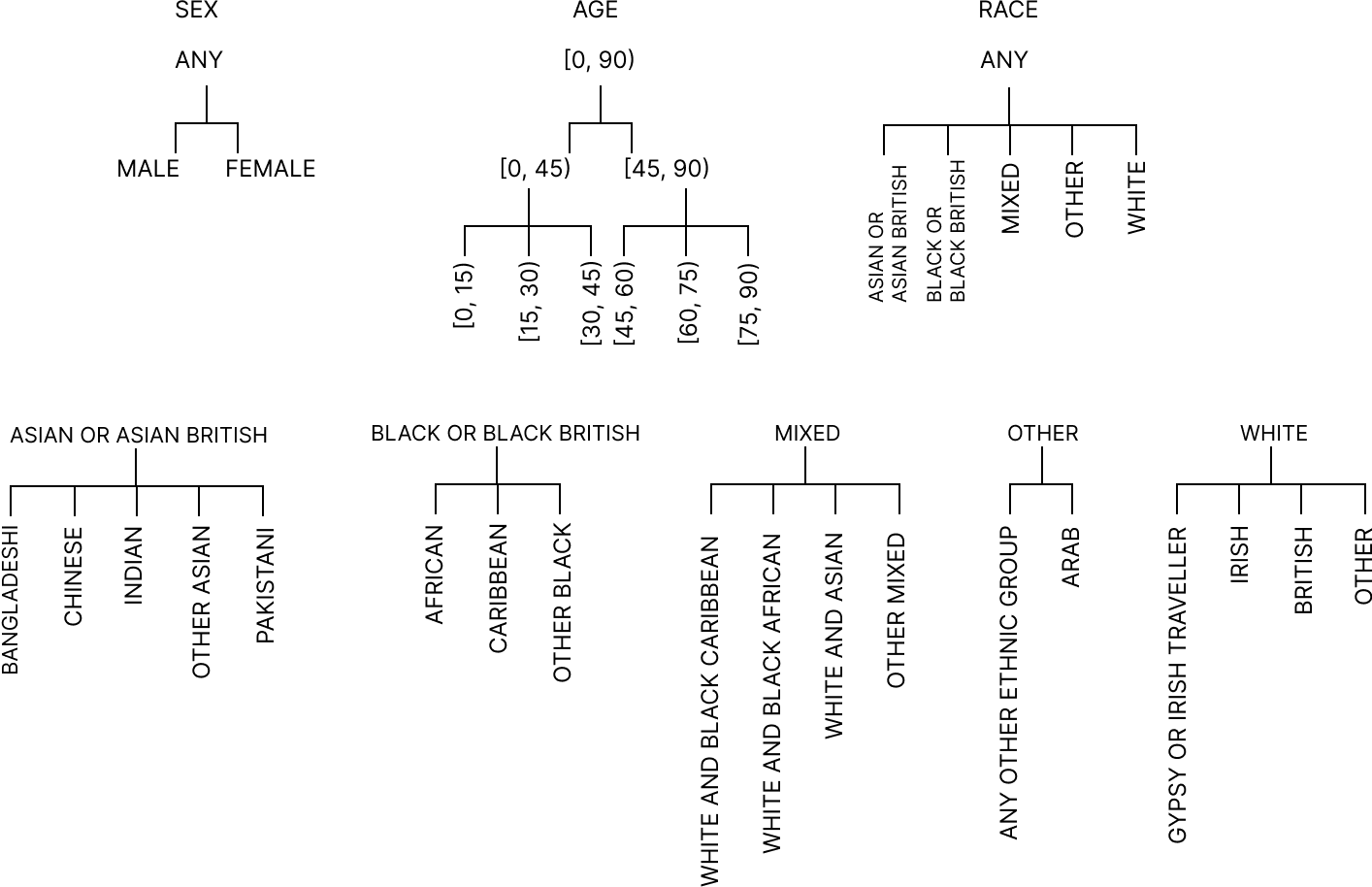
Για να υπολογίσουμε τις πιθανότητες μιας συγκεκριμένης ασθένειας για τον πολίτη (άνδρας, 70 ετών και λευκός Βρετανός) που περιγράφεται στο ερώτημα α, θα πρέπει να φιλτράρουμε το αρχείο hospital\_data.csv και υπολογίζουμε την πιθανότητα από τις εγγραφές που εμφανίζουν αυτά τα δεδομένα (καταχωρήσεις με αντίστοιχο φύλο, ηλικία και φυλή). Δηλαδή μετράμε πόσες φορές εμφανίζεται η κάθε ασθένεια και διαιρούμε με το συνολικό αριθμό ασθενείων για να βρούμε την πιθανότητα.

## *β) Υπάρχει κάποιος στην λίστα των κατοίκων της πόλης ο οποίος, εάν ξέρετε ότι επισκέφθηκε το νοσοκομείο, μπορείτε να είστε σίγουροι για τον λόγο?*

Μπορούμε να ξέρουμε ότι 708 εγγραφές επισκέφτηκαν το νοσοκομείο για συγκεκριμένο λόγο.

# **GENERALIZATION**

α.

**

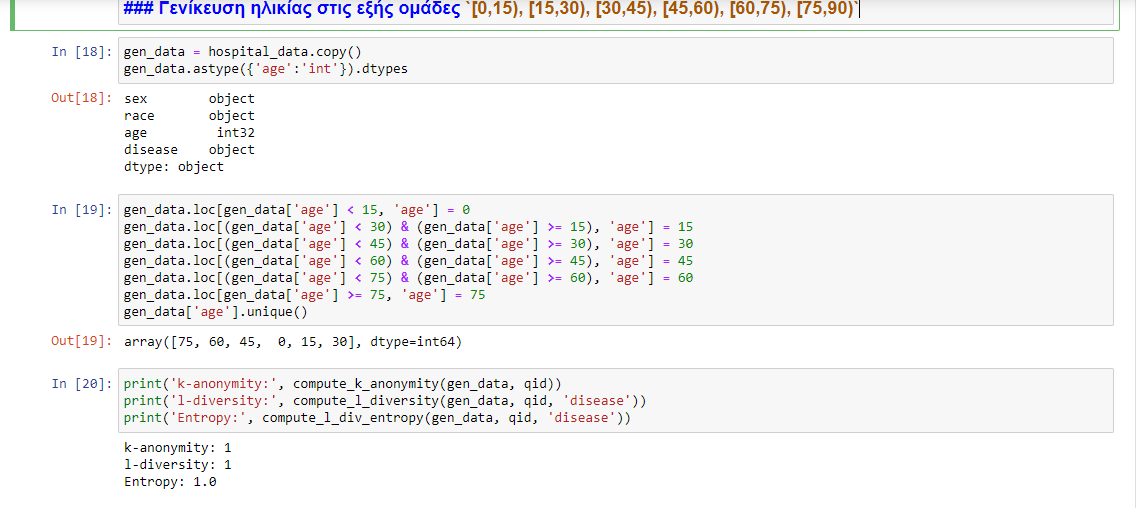
Ελέγχηθηκε ο πίνακας και ισχύουν οι ακόλουθες τιμές που ικανοποιούν το k = 1 για k-anonymity, το l = 1 για l- diversity και τέλος l = 1 για entropy l -diversity.



β)Πήραμε τους παρακάτω συνδυασμούς:

1. Γενίκευση ηλικίας στις εξής ομάδες [0,15), [15,30), [30,45), [45,60), [60,75), [75,90)

Παρατηρούμε ότι δεν υπάρχει διαφορά στα k = 1, l = 1 και entropy = 1

2. Γενίκευση φυλής με πρώτο επίπεδο ηλικίας

3. Γενίκευση ηλικίας σε 2ο επίπεδο

4. Γενικευση ηλικιας και φυλης σε 2ο επιπεδο

Παρατηρούμε ότι το καλύτερο είναι το τεευταίο γιατί έχει μεγαλύτερες τιμες στα k, l, entropy