

ΟΝΟΜΑ: ΑΓΓΕΛΟΣ ΝΙΚΟΛΑΟΣ  
ΕΠΩΝΥΜΟ: ΠΟΤΑΜΙΑΝΟΣ  
ΑΜ: 1084537  
ΕΤΟΣ ΣΠΟΥΔΩΝ: 3ο

### ΑΣΚΗΣΗ 1

**Βήμα 1:** Υπολογισμός εντροπίας της κατηγορίας του στόχου:

Entropy(Heartattack)=

Entropy(4,2)=

$$-(\frac{2}{3} \log_2 \frac{2}{3}) - (\frac{1}{3} \log_2 \frac{1}{3}) =$$

$$1.6 * \frac{1}{3} + \frac{2}{3} * 0.6 =$$

$$\frac{1}{3} * \frac{16}{10} + \frac{2}{3} * \frac{6}{10} =$$

$$\frac{16}{30} + \frac{12}{30} = \frac{28}{30} \text{ ή } \sim 0.93$$

**Βήμα 2:** Υπολογισμός εντροπίας για κάθε χαρακτηριστικό ως προς Heartattack:

A) Entropy(Heartattack, chestpain)=

$$\frac{3}{6} * E(3,0) + \frac{3}{6} * E(1,2) =$$

$$\frac{1}{2} * (-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3}) =$$

$$\frac{1}{6} * (1.6) + \frac{2}{6} * (0.6) =$$

$$\frac{1}{6} * \frac{16}{10} + \frac{2}{6} * \frac{6}{10} =$$

$$\frac{16}{60} + \frac{12}{60} = \frac{28}{60}$$

B) Entropy(Heartattack, male)=

$$\frac{4}{6} * E(2,2) + \frac{2}{6} * E(2,0) =$$

$$\frac{2}{3} \text{ ή } \sim 0.67$$

Γ) Entropy(Heartattack, smokes)=

$$\frac{4}{6} * E(3,1) + \frac{2}{6} * E(1,1) =$$

$$\frac{4}{6} * (-\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4}) + \frac{2}{6} =$$

$$\frac{12}{24} * (0.4) + \frac{4}{24} (2) + \frac{8}{24} =$$

$$\frac{12}{24} * \frac{4}{10} + \frac{4}{24} * \frac{20}{10} + \frac{80}{240} =$$

$$\frac{208}{240} \text{ ή } \sim 0.86$$

Δ) Entropy(Heartattack, exercises)=

$$\frac{4}{6} * E(2,2) + \frac{2}{6} * E(2,0) =$$

$$\frac{2}{3} \text{ ή } \sim 0.67$$

**Βήμα 3:** Υπολογισμός «κέρδους πληροφορίας» για να επιλέξουμε το κατάλληλο χαρακτηριστικό:

A)  $\text{Gain}(\text{Heartattack}, \text{chestpain}) = 56/60 - 28/60 = 28/60$  ή  $\sim 0,46$

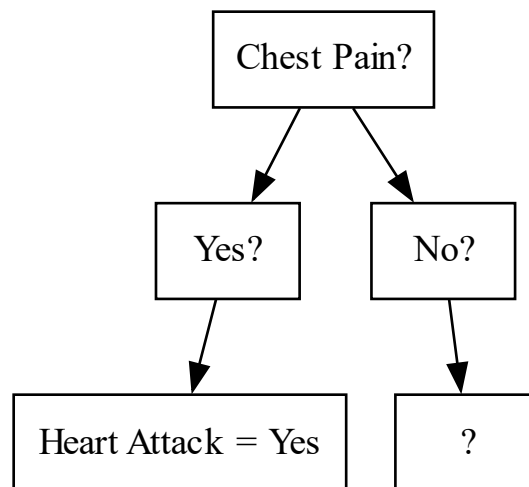
B)  $\text{Gain}(\text{Heartattack}, \text{male}) = 25/30 - 20/30 = 5/30$  ή  $\sim 0,16$

Γ)  $\text{Gain}(\text{Heartattack}, \text{smokes}) = 28/30 - 208/240 = 224/240 - 208/240 = 16/240$  ή  $\sim 0,06$

Δ)  $\text{Gain}(\text{Heartattack}, \text{exercises}) = 25/30 - 20/30 = 5/30$  ή  $\sim 0,16$

Επιλέγω το Chestpain.

Παρατηρώ ότι εφόσον chestpain=yes τότε και το heartattack=yes ή διαφορετικά παρατηρώ ότι το chestpain yes είναι ένα κλαδί με εντροπία 0 αφού  $E(3,0)$  και επομένως θεωρείται φύλλο. Γνωρίζω ακόμα, πως κάθε κλαδί που ξεκινά από την ρίζα και καταλήγει σε φύλλο, θεωρείται ολοκληρωμένο. Επομένως καταλήγω στον ακόλουθο Δένδρο Αναζήτησης:



**Βήμα 5:** Υπολογισμός εντροπίας chestpain no:

$$\begin{aligned} \text{Entropy}(\text{chestpain no}) &= E(1,2) = \\ &= -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = \\ &= \frac{1}{3} * 1.6 + \frac{2}{3} * 0.6 = \\ &= \frac{1}{3} * \frac{16}{10} + \frac{2}{3} * \frac{6}{10} = \\ &= \frac{16}{30} + \frac{12}{30} = \frac{28}{30} \end{aligned}$$

**Βήμα 6:** Υπολογισμός εντροπίας για κάθε χαρακτηριστικό ως προς Chestpain no:

A)  $\text{Entropy}(\text{chestpain no}, \text{male}) = \frac{2}{3} * E(0,2) + \frac{1}{3} * E(1,0) = 0$

B)  $\text{Entropy}(\text{chestpain no}, \text{smokes}) = \frac{2}{3} * E(1,1) + \frac{1}{3} * E(1,0) = \frac{2}{3} = 0,67$

Γ)  $\text{Entropy}(\text{chestpain no}, \text{exercises}) = \frac{2}{3} * E(2,0) + \frac{1}{3} * E(1,0) = 0$

**Βήμα 7:** Υπολογισμός «κέρδους πληροφορίας» για να επιλέξουμε το κατάλληλο χαρακτηριστικό:

A)  $\text{Gain}(\text{chestpain no, male}) = 28/30 - 0 = 28/30$

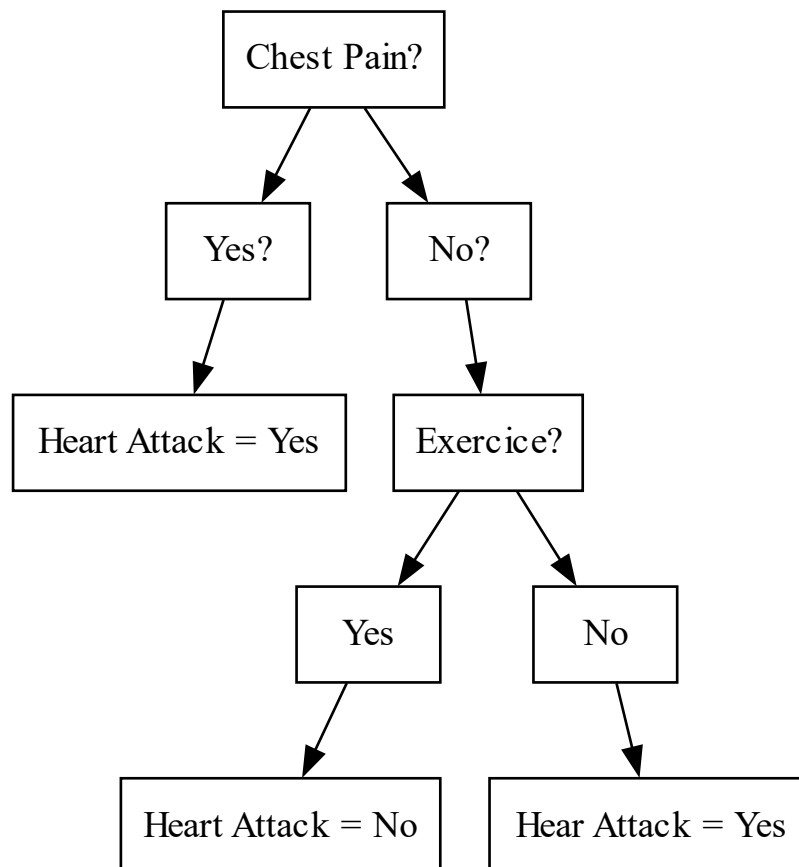
B)  $\text{Gain}(\text{chestpain no, smokes}) = 28/30 - 2/3 = 28/30 - 20/30 = 8/30$

Γ)  $\text{Gain}(\text{chestpain no, exercises}) = 28/30 - 0 = 28/30$

Παρατηρούμε ότι έχουμε έχουμε ισοπαλία κέρδους πληροφορίας. Θεωρώ ως καλύτερο κριτήριο την σωματική άσκηση «exercises» αφού μετά από μελέτη κατέληξα πως μόνο το 23.2% της Αμερικής γυμνάζεται επαρκώς (πηγή: <https://www.cdc.gov/nchs/fastats/exercise.htm>). Αυτό σημαίνει ότι αυτό το κλαδί του δένδρου μας απαντάει σχεδόν 80/20, δίνοντας μας κοντινότερη προσέγγιση στο ως προς αν θα πάθει καρδιακή ανακοπή κάποιος, έναντι της φυλετικού κριτηρίου «male» που θα μας έδινε μια προσέγγιση κοντά στο 50/50.

Παρατηρώ ακόμα ότι εφόσον exercise=yes τότε και το heartattack=no ή διαφορετικά παρατηρώ ότι το exercise=yes είναι ένα κλαδί με εντροπία 0 αφού  $E(2,0)$  και επομένως θεωρείται φύλλο καθώς και πως exercise=no τότε και το heartattack=yes ή διαφορετικά παρατηρώ ότι το exercise=no είναι ένα κλαδί με εντροπία 0 αφού  $E(1,0)$ . Γνωρίζω ακόμα, πως κάθε κλαδί που ξεκινά από την ρίζα και καταλήγει σε φύλλο, θεωρείται ολοκληρωμένο.

Επομένως καταλήγω στον ακόλουθο τελικό Δένδρο Αναζήτησης με την χρήση graphviz:



## ΑΣΚΗΣΗ 2

Μετατρέπω το παραπάνω δένδρο αναζήτησης σε σύνολο κανόνων απόφασης:

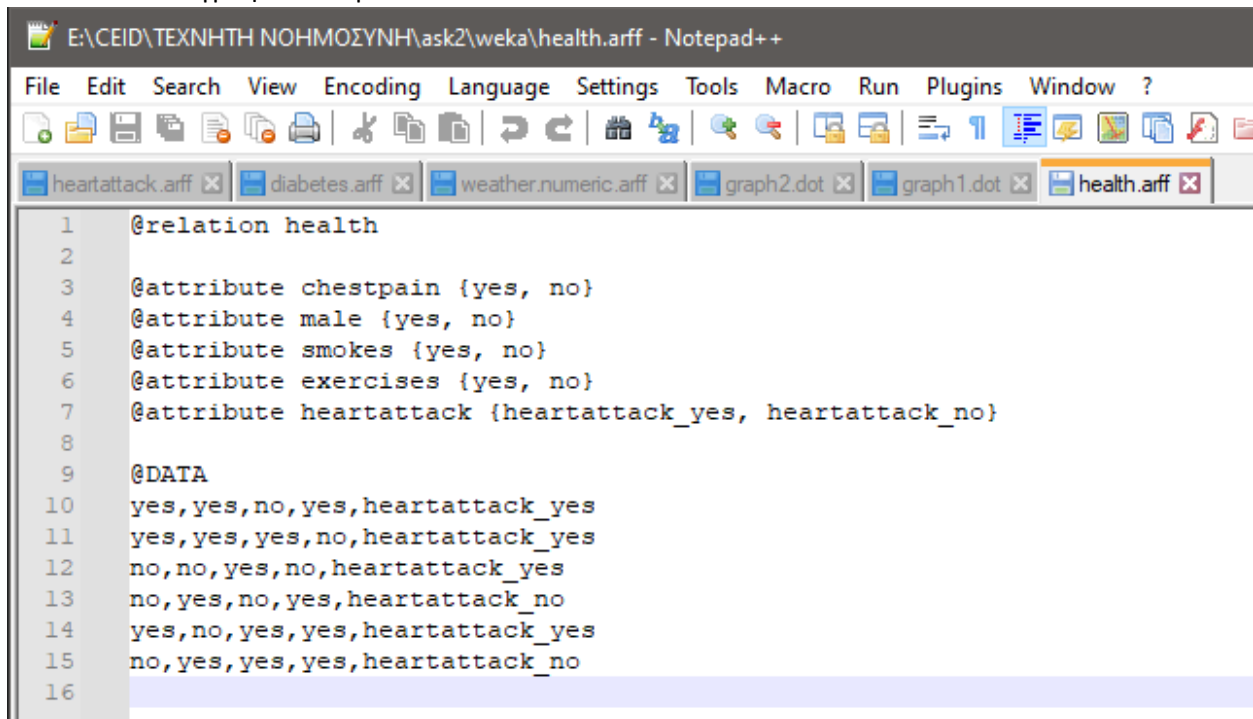
**If** chestpain=yes **then** Heartattack=yes  
**If** chestpain=no **and** exercise=yes **then** heartattack=no  
**If** chestpain=no **and** exercise=no **then** heartattack=yes

Επιπλέον για το male μπορούμε να διευρύνουμε αυτούς τους κανόνες

**If** chestpain=no **and** male=yes **then** heartattack=no  
**If** chestpain=no **and** exercise=no **then** heartattack=no

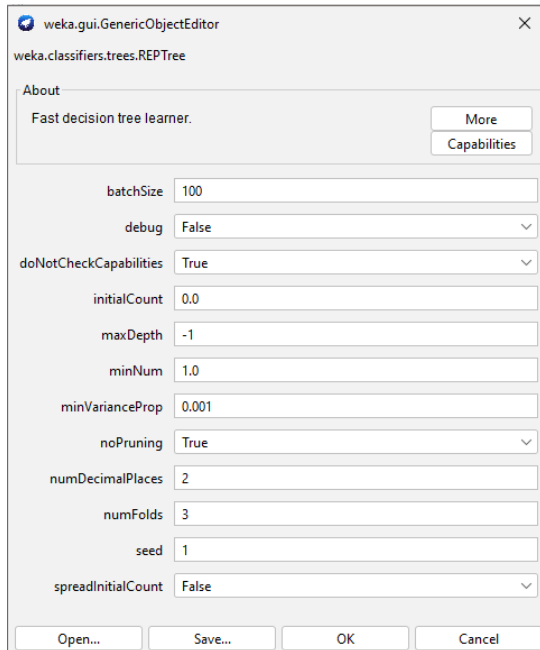
## ΑΣΚΗΣΗ 3

Αφού εγκατέστησα το weka 3.8.6, δημιούργησα το .arff αρχείο, το οποίο και ονόμασα health.  
Στο health.arff έγραψα τα παρακάτω.



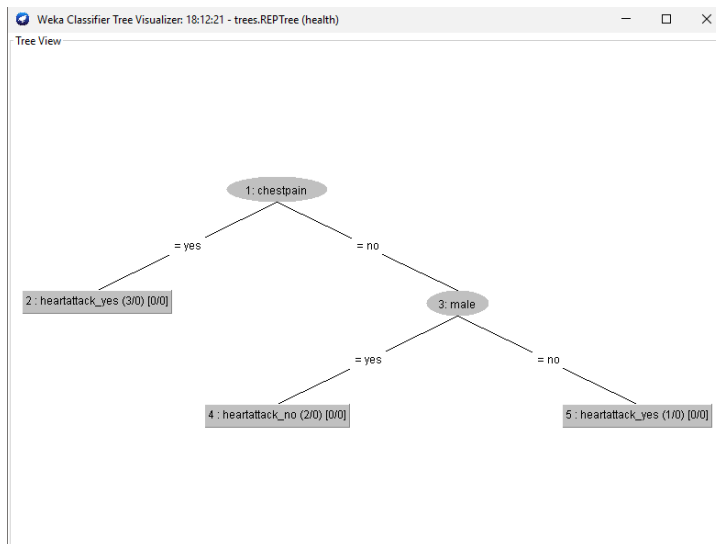
```
1 @relation health
2
3 @attribute chestpain {yes, no}
4 @attribute male {yes, no}
5 @attribute smokes {yes, no}
6 @attribute exercises {yes, no}
7 @attribute heartattack {heartattack_yes, heartattack_no}
8
9 @DATA
10 yes,yes,no,yes,heartattack_yes
11 yes,yes,yes,no,heartattack_yes
12 no,no,yes,no,heartattack_yes
13 no,yes,no,yes,heartattack_no
14 yes,no,yes,yes,heartattack_yes
15 no,yes,yes,yes,heartattack_no
16
```

Στην συνέχεια, άνοιξα το αρχείο μέσω του weka explores και στο tab classify επέλεξα τον classifier: classifiers→trees→REPTree όπως και αναγράφετε στο link της άσκησης. Τροποποίησα τις ρυθμίσεις ακολούθως:



Αλλαγές: minNum από 2.0 σε 1.0 και noPruning από False σε True, αφού δεν γινόταν ολοκληρωμένη ανάπτυξη του δέντρου αλλά έφτανε μόνο μέχρι το επίπεδο 1. Τα υπόλοιπα τα άφησα στις default τιμές τους.

Έπειτα, επέλεξα use training set→Start→δεξί κλικ στο 18:12:21 – tress.REPTree→Visualize tree



Παρατηρώ ότι γίνεται τυχαία επιλογή του male έναντι της προσωπικής μου επιλογής exercises.

Στιγμιότυπο classifies output:

```
Classifier output
    male
    exercise
    smokes
    heartattack
Test mode: evaluate on training data

=== Classifier model (full training set) ===

REPTree
=====
chestpain = yes : heartattack_yes (3/0) [0/0]
chestpain = no
| male = yes : heartattack_no (2/0) [0/0]
| male = no : heartattack_yes (1/0) [0/0]

Size of the tree : 5

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

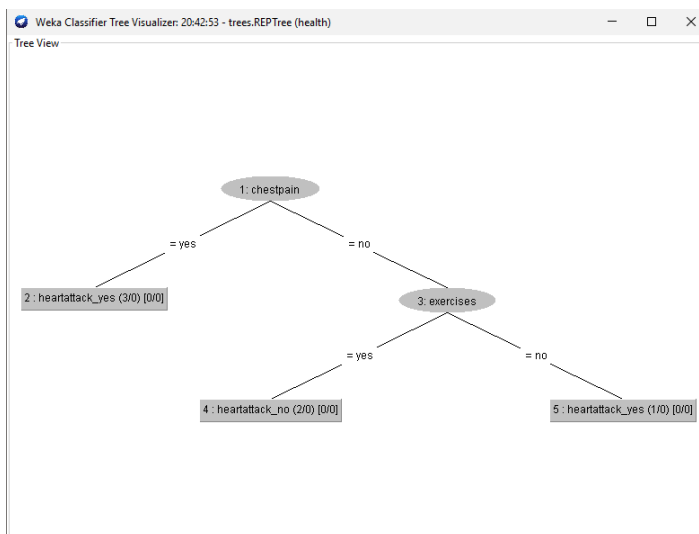
Correctly Classified Instances      6      100 %
Incorrectly Classified Instances    0        0 %
Kappa statistic                    1
Mean absolute error                 0
Root mean squared error             0
Relative absolute error             0 %
Root relative squared error         0 %
Total Number of Instances          6

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC   ROC Area  PRC Area  Class
1.000   0.000   1.000   1.000   1.000   1.000   1.000   1.000   heartattack_yes
1.000   0.000   1.000   1.000   1.000   1.000   1.000   1.000   heartattack_no
Weighted Avg.   1.000   0.000   1.000   1.000   1.000   1.000   1.000   1.000

=== Confusion Matrix ===
a b  <-- classified as
4 0 | a = heartattack_yes
0 2 | b = heartattack_no
```

Αντιστρέφοντας ωστόσο την σειρά των @attribute exercise με την @attribute male λαμβάνω το παρακάτω δένδρο απόφασης:



Επομένως συμπεραίνουμε ότι το weka επιλέγει τα attributes που θα χρησιμοποιήσει για την κατασκευή του δέντρου απόφασης βάση α) Του κέρδους πληροφορίας και β) της σειράς εμφάνισης τους στο .arff αρχείο.

Όλα τα αρχεία κώδικα που χρησιμοποιήθηκαν για την εργασία:

<https://github.com/nickpotamianos/aiexercise2.git>