

**ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ «ΕΦΑΡΜΟΣΜΕΝΗΣ ΟΙΚΟΝΟΜΙΚΗΣ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ»

ΚΑΤΕΥΘΥΝΣΗ

«ΔΙΟΙΚΗΣΗ, ΑΝΑΛΥΤΙΚΗ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΕΠΙΧΕΙΡΗΣΕΩΝ»

Master of Science in

Business Administration, Analytics and Information Systems

**Data Science (R)**

**Assignment 2**

**Κωνσταντίνος Κουτσομπίνας**

Αθήνα 21/06/2025

Περιεχόμενα

[Άσκηση 3 3](#_Toc201397262)

[a) Training και test set 3](#_Toc201397263)

[b) Εκπαίδευση και βασικά χαρακτηριστικά 3](#_Toc201397264)

[c) Περιγραφή τερματικού κόμβου 3](#_Toc201397265)

[d) Γραφική αναπαράσταση δέντρου 3](#_Toc201397266)

[e) Confusion Matrix 3](#_Toc201397267)

[f) Cross Validation 3](#_Toc201397268)

[g) Γραφική αναπαράσταση αποτελεσμάτων cross validation 3](#_Toc201397269)

[h) Βέλτιστο δέντρο 3](#_Toc201397270)

[i) Pruning 3](#_Toc201397271)

[j) Διαφορά στο train 3](#_Toc201397272)

[k) Διαφορά στο test 3](#_Toc201397273)

# Άσκηση 3

## Training και test set

Χωρίζουμε τα δεδομένα σε train (800 samples) και test (270 samples)

1. set.seed(42)

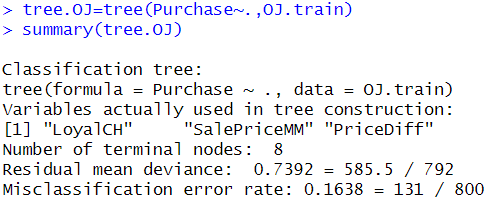
2. train=sample(1:nrow(OJ), 800)

3. OJ.train=OJ[train,]

4. OJ.test=OJ[-train,]

## Εκπαίδευση και βασικά χαρακτηριστικά

Εκπαιδεύουμε ένα δέντρο στο training set:



Βλέπουμε ότι το δέντρο έχει **8 τερματικούς κόμβους και 16.38%(131/800) misclassification error rate**

## Περιγραφή τερματικού κόμβου

Παρατηρούμε τους κόμβους του δέντρου:

A screenshot of a computer code

AI-generated content may be incorrect.

* Ένας τερματικός κόμβος είναι ο κόμβος 4.
* Στον κόμβο αυτό υπάρχουν 64 παρατηρήσεις, εκ των οποίων το 100% είναι της κλάσης ΜΜ (minute maid orange juice) συνεπώς το predicted class σε αυτόν τον κόμβο είναι MM.
* Στον κόμβο αυτό καταλήγουν οι παρατηρήσεις που έχουν LoyalCH < 0.48285 and LoyalCH < 0.064156 δηλαδή απλά **LoyalCH < 0.064156.**
* Αυτό το αποτέλεσμα είναι λογικό καθώς το loyalCH εκφράζει το brand loyalty του καταναλωτή για τη κλάση CH (Citrus Hill orange juice). Συνεπώς μικρό loyalty προς την κλάση ch είναι λογικό να σημαίνει ότι ο καταναλωτής ανήκει στην κλάση MM.
* Επιπλέον αξίζει να σημειωθεί ότι στη ρίζα του δέντρου (κόμβος 1) φαίνεται ότι οι παρατηρήσεις μας είναι **imbalanced** καθώς στις 800 παρατηρήσεις του training set έχουμε 61.5% CH και 38.5% MM.

## Γραφική αναπαράσταση δέντρου

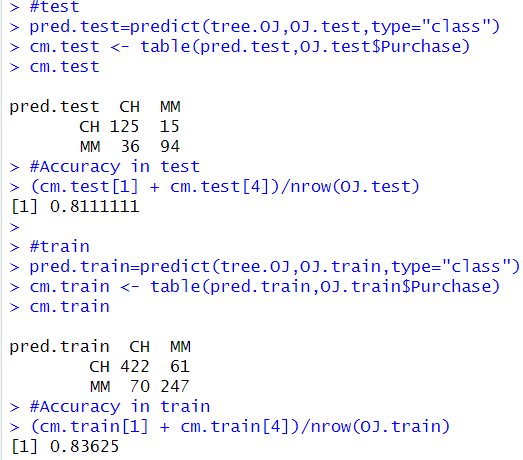
A diagram of a number of numbers

AI-generated content may be incorrect.

* Βλέπουμε τους 8 τερματικούς κόμβους που αναφέραμε και πριν.
* Το δέντρο μας δεν είναι ιδιαίτερα μεγάλο.
* Παρατηρούμε ότι μόνο 3 μεταβλητές από τις 17 (!) χρησιμοποιούνται για τα split του δέντρου.
* Επίσης φαίνεται ότι υπάρχουν 4 τερματικοί κόμβοι με προβλεπόμενη κλάση MM και 4 με CH.
* Μία ακόμα σημαντική παρατήρηση είναι ότι υπάρχουν φύλλα-αδέρφια (τα σημειωμένα στο παρακάτω διάγραμμα), τα οποία έχουν τα ίδια labels. Δηλαδή έχουν δημιουργηθεί splits τα οποία δεν αλλάζουν τη πρόβλεψη του εξεταζόμενου δείγματος, είτε ικανοποιείται η συνθήκη είτε όχι. A screenshot of a computer

  AI-generated content may be incorrect.

## Confusion Matrix



Παρατηρούμε ότι το accuracy στο test set είναι 81.1% (δηλ. error 18.9%) ενώ στο train set είναι 83.6% (δηλ. error 16.4%, όπως είδαμε και στο [ερώτημα b](#_Εκπαίδευση_και_βασικά))

Το αποτέλεσμα αυτό είναι λογικό καθώς το δέντρο εκπαιδεύτηκε στο train set, συνεπώς περιμένουμε κάποιο overfit στις λεπτομέρειες και στο θόρυβο του train set, που οδηγεί στην καλύτερη απόδοση στα γνωστά δεδομένα.

Ωστόσο και οι 2 αποδόσεις είναι αρκετά καλές και αρκετά κοντά, συνεπώς το σφάλμα γενίκευσης είναι σχετικά μικρό.

## Cross Validation

Εφαρμόζουμε cross validation

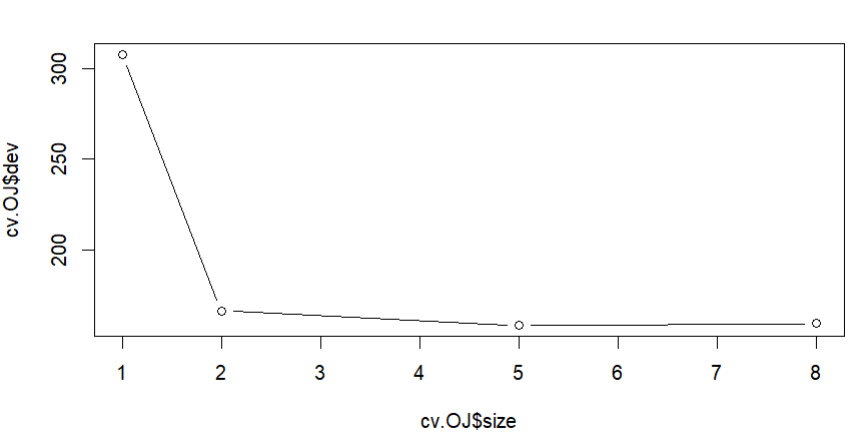
1. #f) cv.tree

2. set.seed(42)

3. cv.OJ=cv.tree(tree.OJ,FUN=prune.misclass)

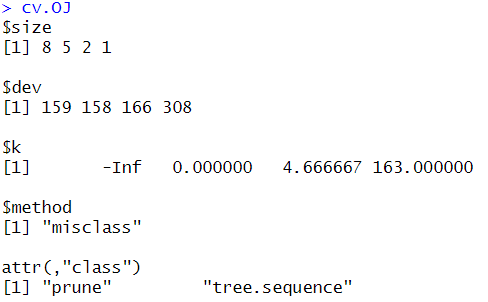
## Γραφική αναπαράσταση αποτελεσμάτων cross validation

Βλέπουμε το σφάλμα σε σχέση με το πλήθος τερματικών κόμβων του δέντρου, όπως προέκυψε από το cross validation



## Βέλτιστο δέντρο

Παρατηρούμε και στο προηγούμενο διάγραμμα, αλλά και στις λεπτομέρειες του cross validation:



ότι το μικρότερο σφάλμα (dev=158) εμφανίζεται για πλήθος τερματικών κόμβων (size) ίσο με 5

## Pruning

«Κλαδεύουμε» το δέντρο μας σε πλήθος τερματικών κόμβων ίσο με 5:

1. prune.OJ=prune.misclass(tree.OJ,best=5)

Ας δούμε και τη διαγραμματική απεικόνιση του νέου δέντρου

A diagram of a graph

AI-generated content may be incorrect.

Παρατηρούμε ότι τα φύλλα αδέρφια που είχαν τα ίδια labels (τα οποία σημειώσαμε στο [ερώτημα d](#_Γραφική_αναπαράσταση_δέντρου)) έχουν «κλαδευτεί».

## Διαφορά στο train

A screenshot of a computer code

AI-generated content may be incorrect.

## Διαφορά στο test

A screenshot of a computer code

AI-generated content may be incorrect.

# Άσκηση 4

## Training και test set

## Εκπαίδευση, απεικόνιση και διερμήνευση αποτελεσμάτων

## Cross validation

## Bagging

## Random forests