Αποθήκες Δεδομένων και Εξόρυξη Γνώσης

Δέντρα Αποφάσεων

Decision Trees

Κατηγοριοποίηση: Ορισμός

- Δοσμένης μιας συλλογής εγγραφών (σώμα εκπαίδευσης-training set)
  - Κάθε εγγραφή περιέχει ένα σύνολο ιδιοτήτων-attributes, μιας εκ των οποίων είναι η κλάση-class
- Εύρεση ενός μοντέλου για την ιδιότητα της κλάσης ως συνάρτηση των τιμών των υπόλοιπων μεταβλητών

- Στόχος
  - Οι προηγουμένως αθέατες εγγραφές θα πρέπει να χαρακτηρισθούν με μια κλάση όσο ακριβέστερα γίνεται
- Ένα σύνολο αξιολόγησης-test set χρησιμοποιείται για να εξακριβωθεί η ακρίβεια του μοντέλου. Συνήθως, χωρίζουμε το σύνολο δεδομένων σε εκπαίδευσης και αξιολόγησης.

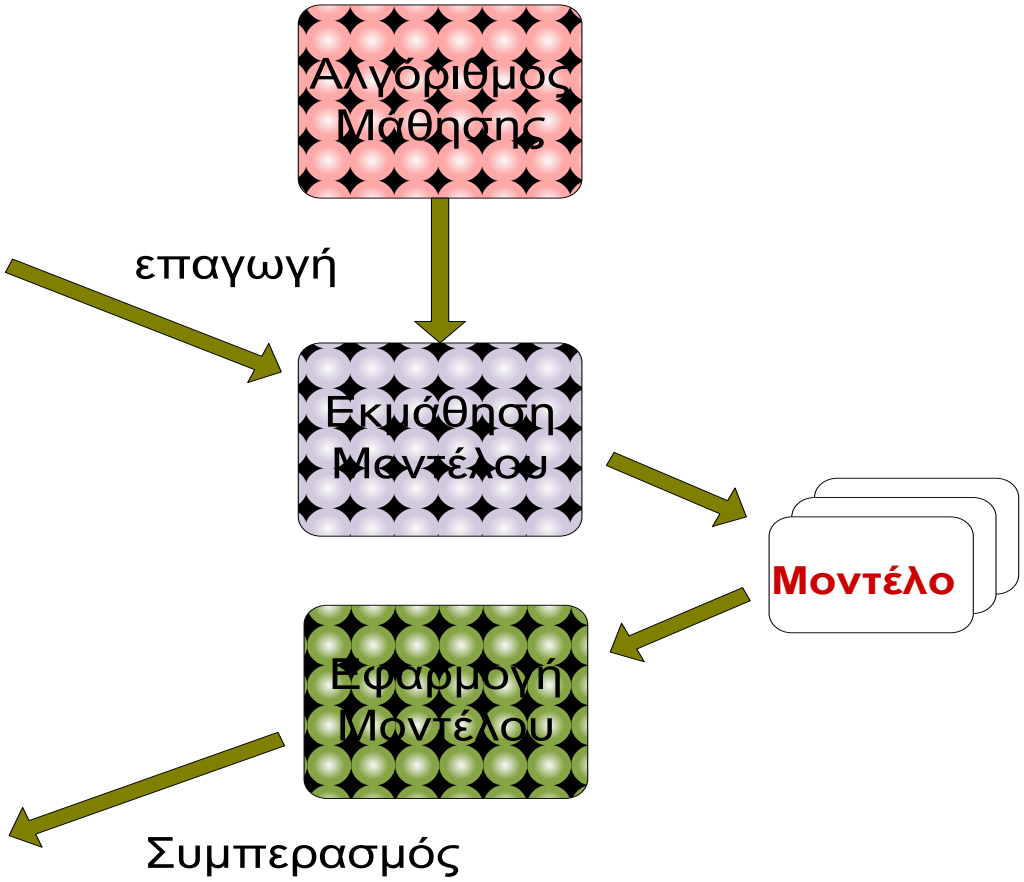
Γραφική απεικόνιση της κατηγοριοποίησης

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Σώμα Εκπαίδευσης

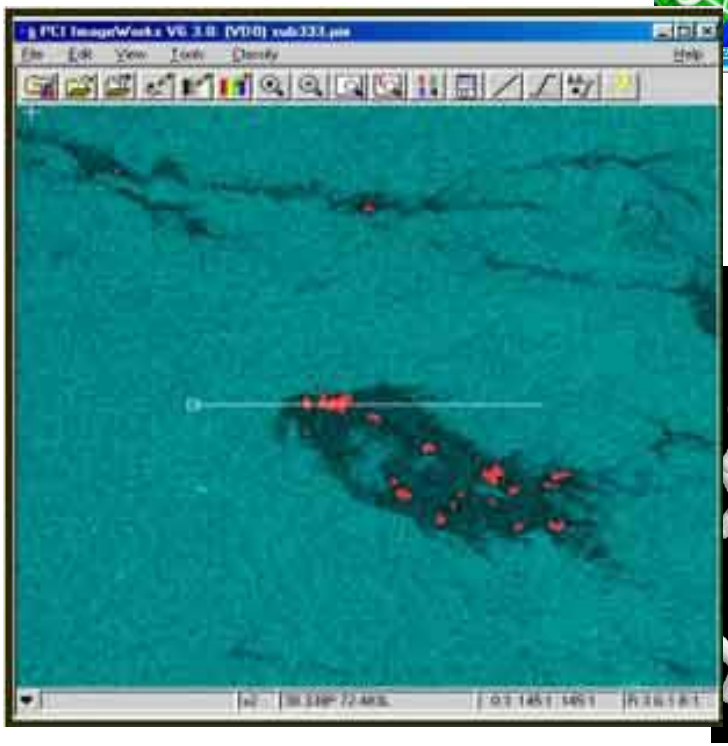
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

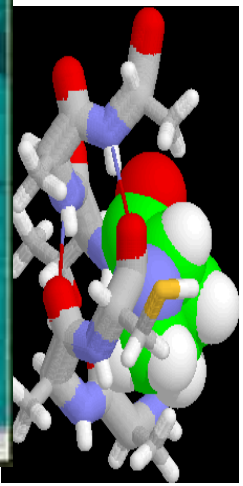
Σώμα Αξιολόγησης



Παραδείγματα εφαρμογών

- Πρόβλεψη καρκινικών κυττάρων για το αν είναι καλοήθη ή κακοήθη
- Κατηγοριοποίηση συναλλαγών με πιστωτική κάρτα για τον αν είναι νόμιμες ή μη
- Κατηγοριοποίηση των δομών πρωτεΐνης
- Κατηγοριοποίηση άρθρων εφημερίδων ως οικονομικά, αθλητικά, κοινωνικά, κτλ.
- Κατηγοριοποίηση δορυφορικών εικόνων θαλάσσης για το αν είναι πετρελαϊκή διαρροή ή φύκια



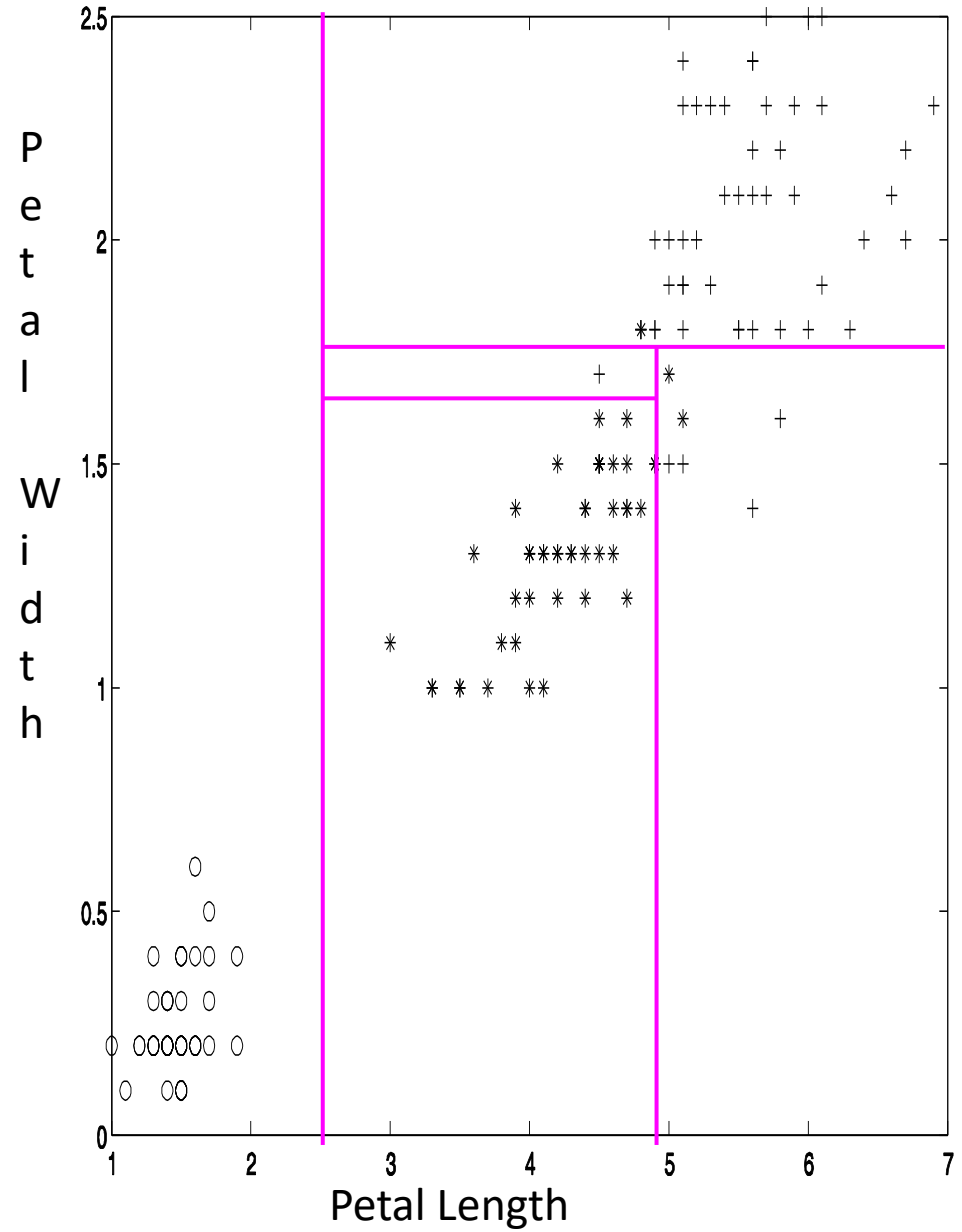


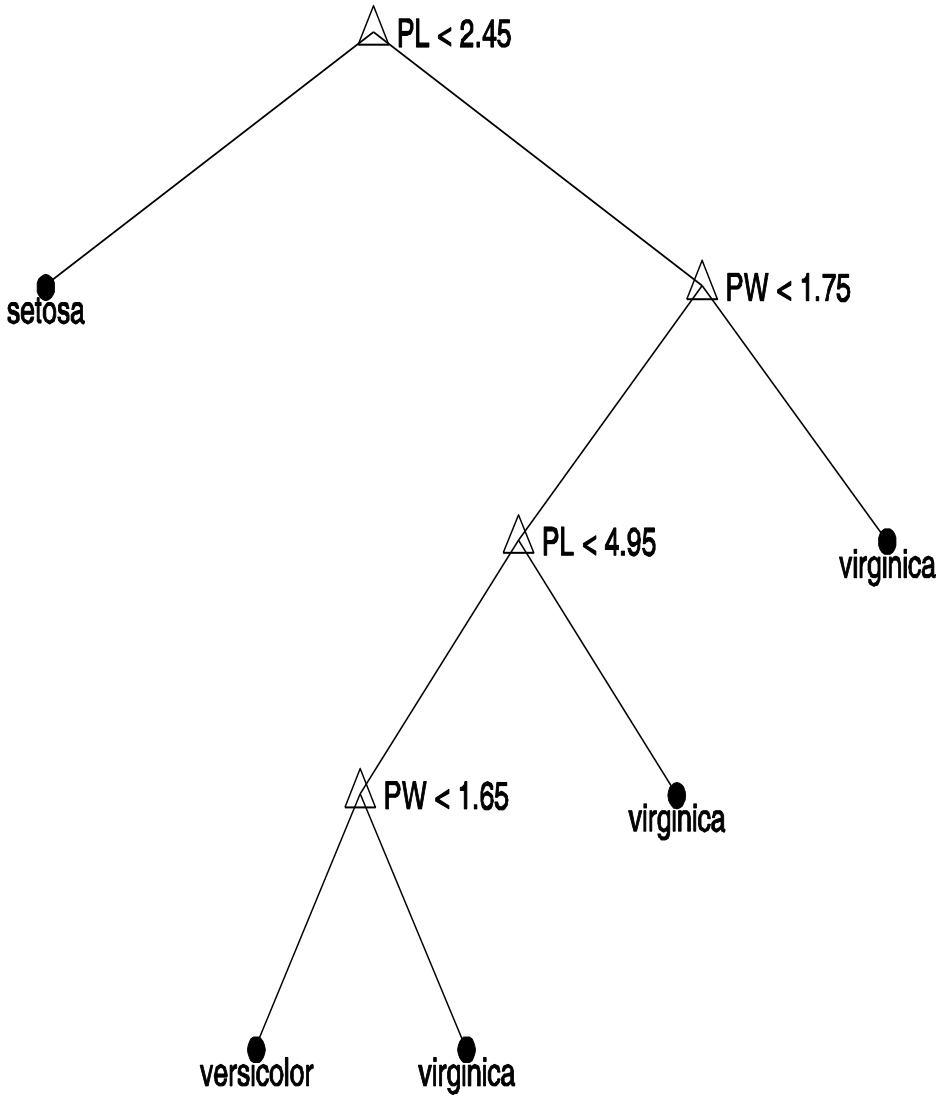
Τεχνικές

- Μέθοδοι με δέντρα αποφάσεων
- Μέθοδοι βασισμένοι σε κανόνες
- Μέθοδοι βασισμένοι στη μνήμη
- Νευρωνικά δίκτυα
- Δίκτυα Bayes και απλοϊκή μέθοδος Bayes
- Μηχανές διανυσμάτων υποστήριξης

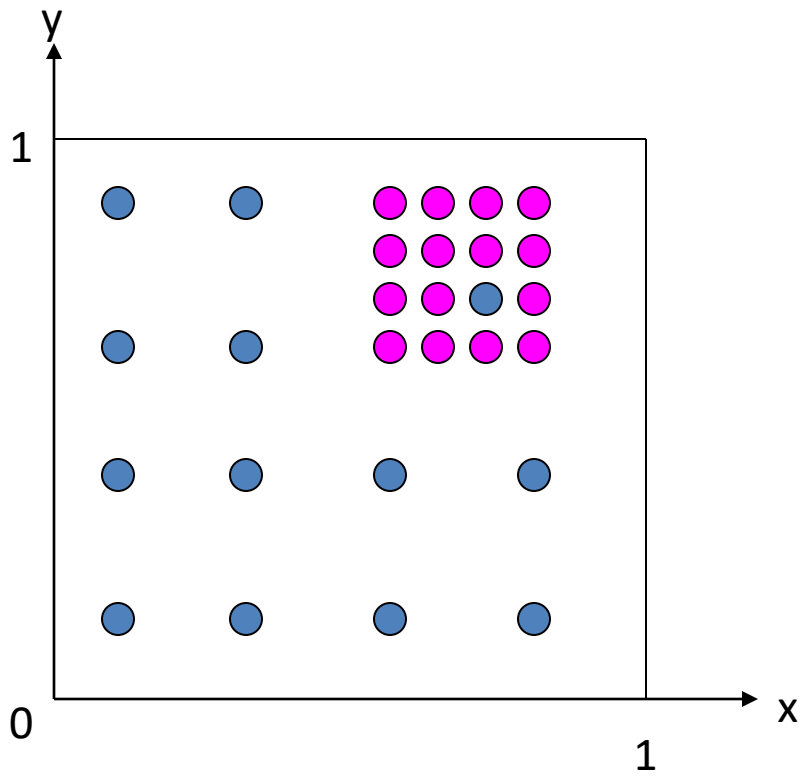
- Decision Tree based Methods
- Rule-based Methods
- Memory based reasoning
- Neural Networks
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines

Ένα δέντρο απόφασης





Παράδειγμα: δημιουργία ενός δέντρου απόφασης



13 15

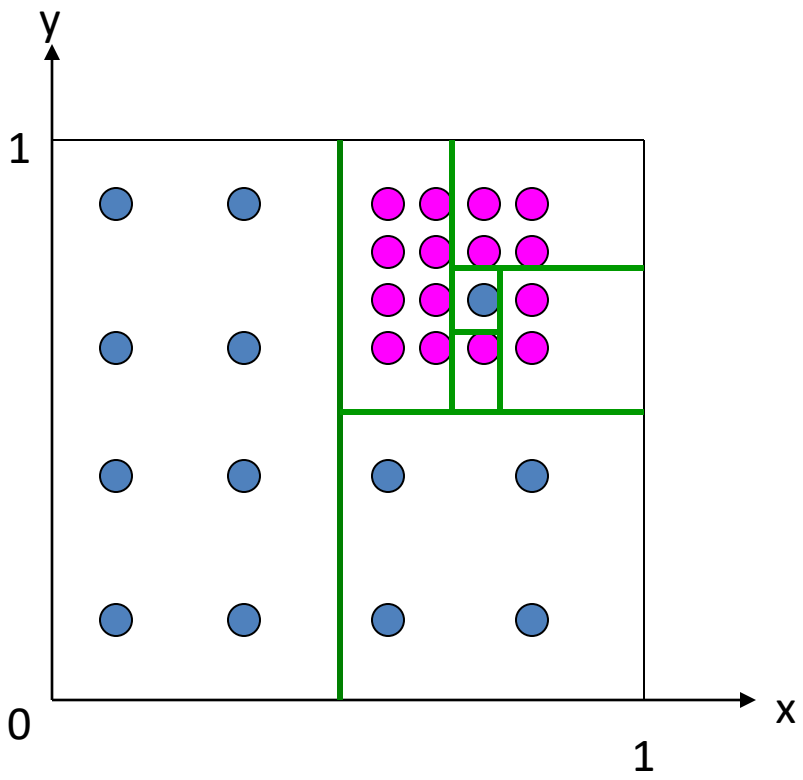
Αρχικά

13 15

Αν έπρεπε να διαλέξουμε, θα διαλέγαμε

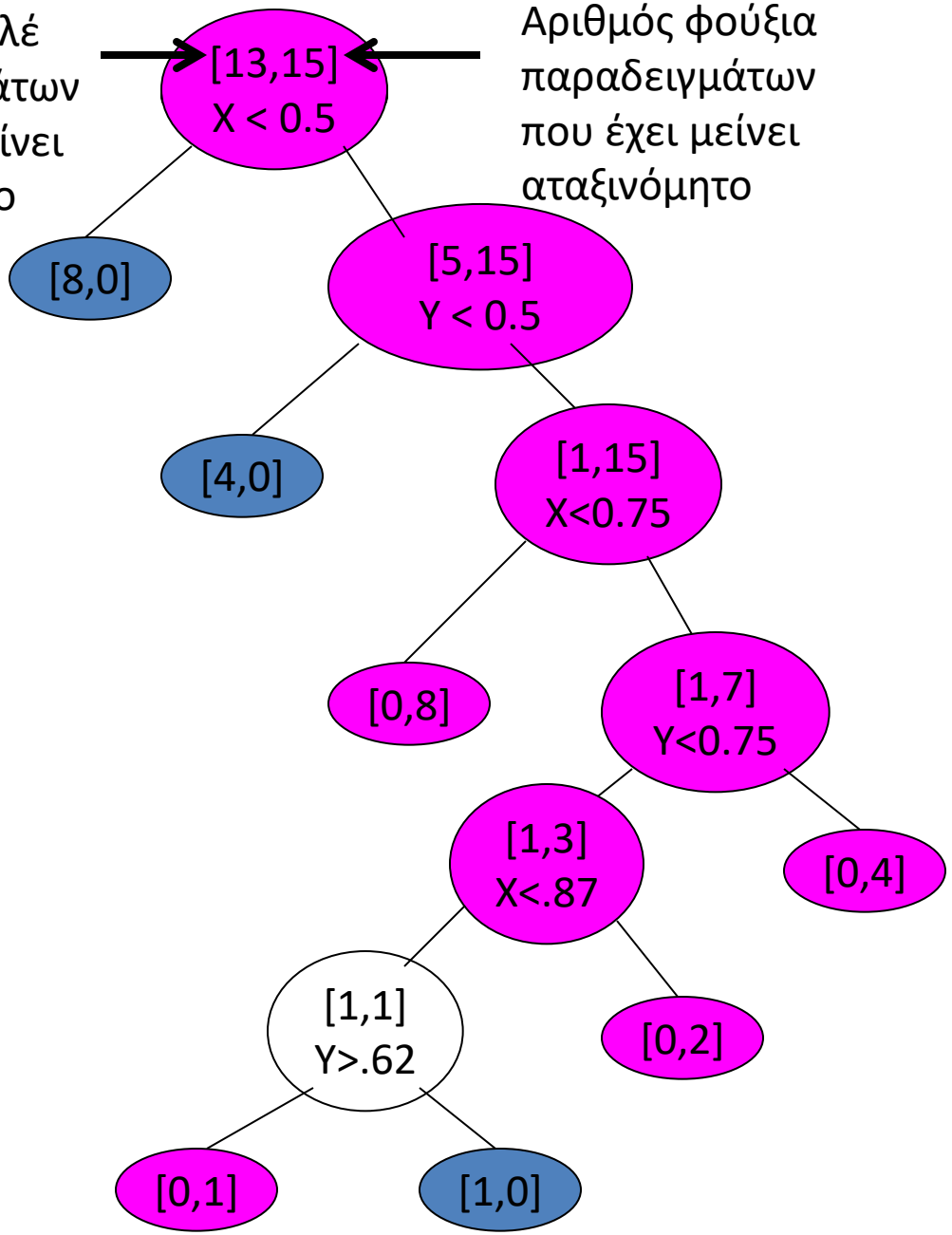
Αλλά είναι πολύ μεγάλη η αβεβαιότητα (uncertainty).

Θέλουμε καλύτερες πιθανότητες, οπότε διασπάμε τα δεδομένα.

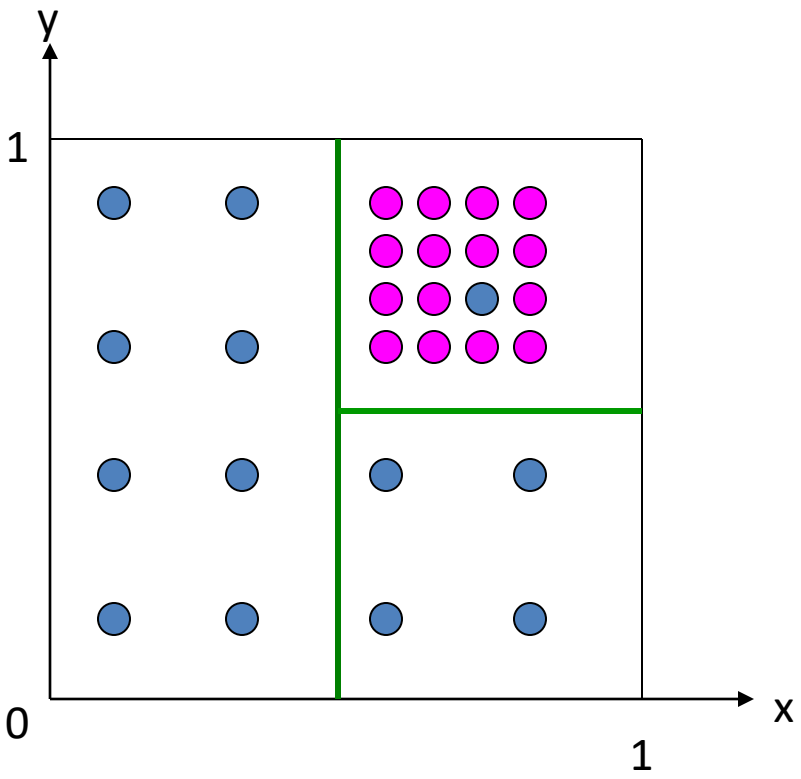


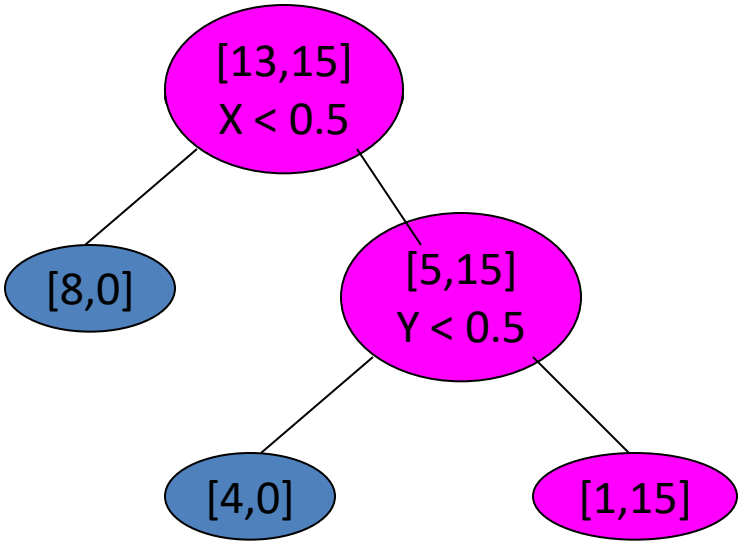
Αριθμός μπλέ παραδειγμάτων που έχει μείνει αταξινόμητο

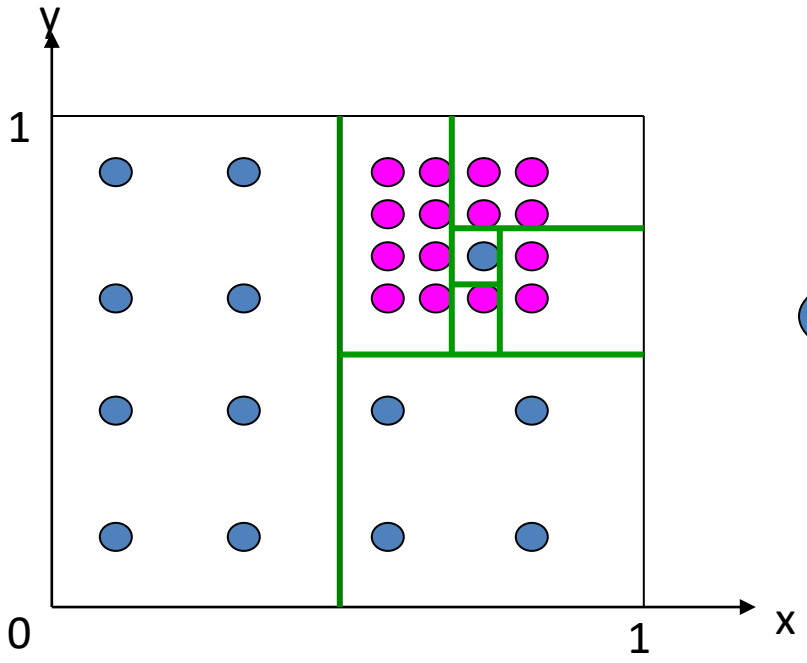
Αριθμός φούξια παραδειγμάτων που έχει μείνει αταξινόμητο

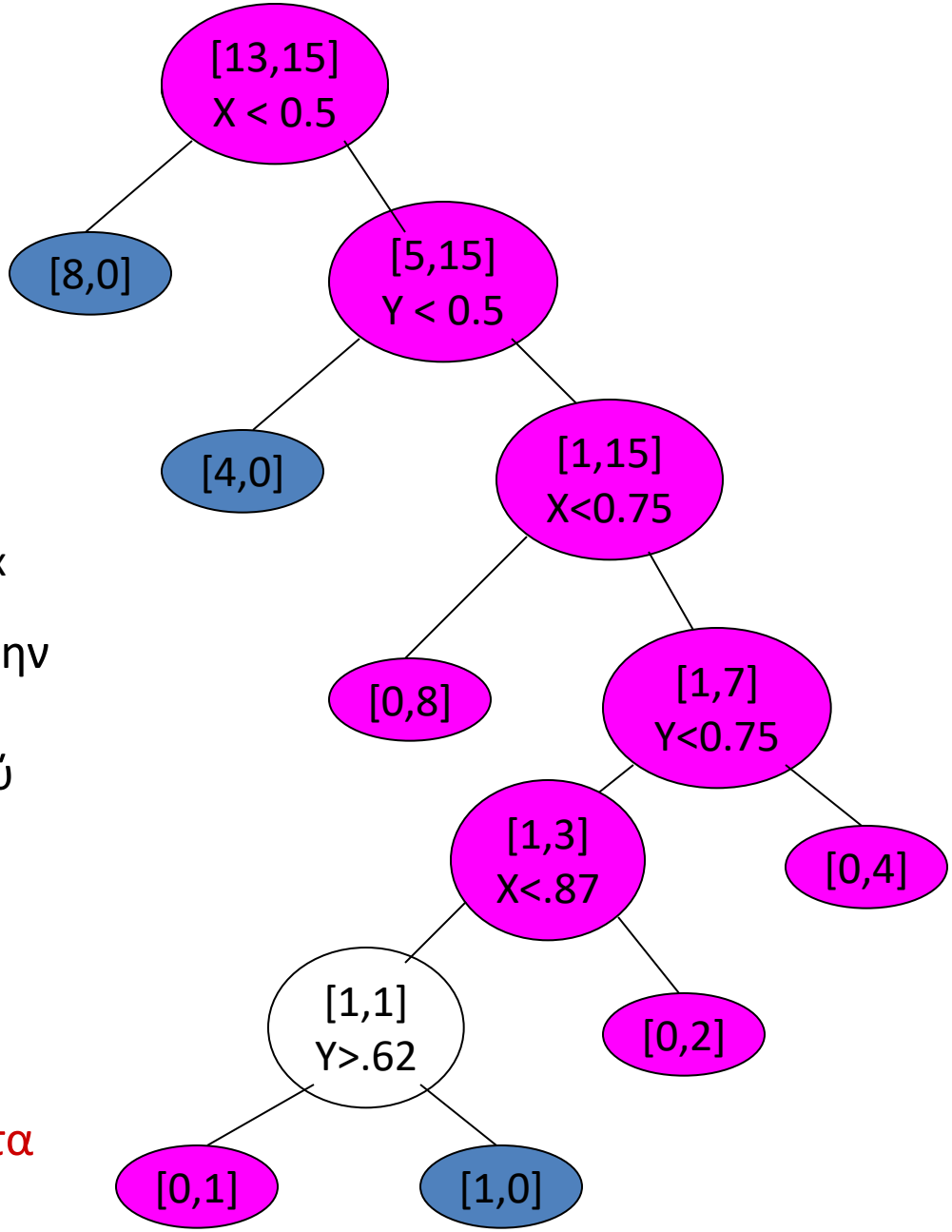


Μερικά βήματα πίσω…









Αυτό το δέντρο τα πάει καλύτερα στην ταξινόμηση των παραπάνω παραδειγμάτων. Είναι όμως πολύ πιο πολύπλοκο.

Αυτό το ένα μπλε πιθανότατα είναι outlier ούτως η άλλως.

Πιθανότατα έχουμε κάνει υπερπροσαρμογή (overfitting) στα δεδομένα εκπαίδευσης.

Χτίσιμο του δέντρου

Το δέντρο σχηματίζεται top-down (ξεκινώντας από τη ρίζα)

Σε κάθε φάση

- Κοίτα όλα τα παραδείγματα που μένουν ακόμα αταξινόμητα και όλα τα χαρακτηριστικά (όλα τα πιθανά splits)
- Διάλεξε εκείνο το split που μειώνει περισσότερο την αβεβαιότητα

Χρειαζόμαστε ένα μέτρο της αβεβαιότητας…

Εντροπία

Πόσο δύσκολο είναι (σε πόση αβεβαιότητα βρίσκομαι) αρχικά να απαντήσω στην ερώτηση σε ποιο χρώμα ανήκει ένα αντικείμενο – στο μπλε ή στο φούξια;

$$I_{\alpha\rho\chi} = -\frac{13}{28}\log\frac{13}{28} - \frac{15}{28}\log\frac{15}{28}$$

Πόση αβεβαιότητα έχω/πόση πληροφορία χρειάζομαι για να απαντήσω σε μια ερώτηση δυο απαντήσεων;

$$p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$$

Εντροπία μετά το split1:

$$I_{split1} = \sum_{i=1}^{p} \frac{p_i + n_i}{p + n}\left(-\frac{p_i}{p_i + n_i}\log\frac{p_i}{p_i + n_i} - \frac{n_i}{p_i + n_i}\log\frac{n_i}{p_i + n_i}\right)$$

Εντροπία αρχική ($I_{\alpha\rho\chi}$)

p_{split1} p_{split2} .... p_{splitN}

Εντροπία μετά το split1

Εντροπία μετά το split2

Εντροπία μετά το splitN

Πλήθος τιμών του χαρακτηριστικού που έχω επιλέξει για το split1

Πλήθος αρχικών μπλε παραδειγμάτων

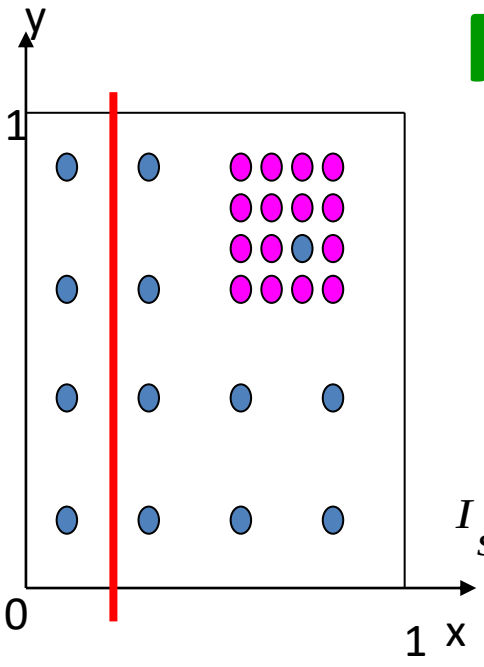
Πλήθος αρχικών φούξια παραδειγμάτων

Πλήθος φούξια παραδειγμάτων που έχουν μείνει αταξινόμητα μετά το split 1

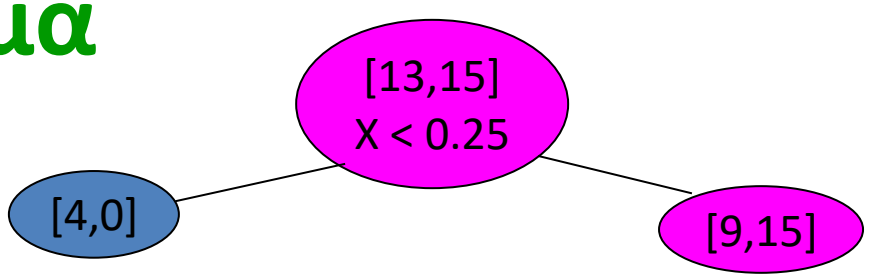
Πλήθος μπλε παραδειγμάτων που έχουν μείνει αταξινόμητα μετά το split 1

Information Gain Κέρδος Πληροφορίας (Μείωση Εντροπίας) $IG = I_{\alpha\rho\chi} - I_{split1}$

Παράδειγμα



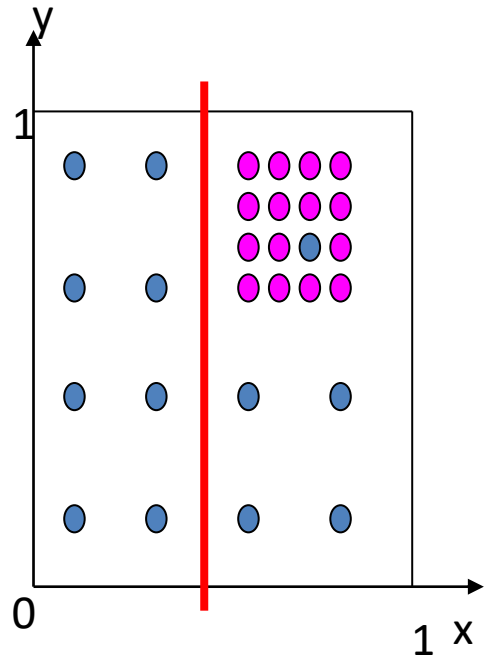
split1



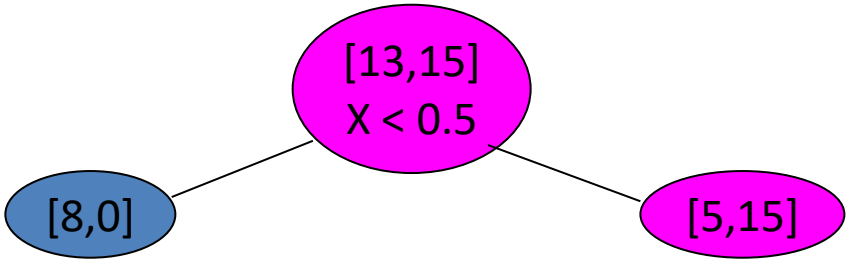
$$I_{split1} = \sum_{i=1}^{\upsilon} \frac{p_i + n_i}{p+n} \left(-\frac{p_i}{p_i+n_i}\log\frac{p_i}{p_i+n_i} - \frac{n_i}{p_i+n_i}\log\frac{n_i}{p_i+n_i}\right)$$

$$I_{split1} = \frac{4+0}{13+15}\left(-\frac{4}{4}\log\frac{4}{4} - \frac{0}{4}\log\frac{0}{4}\right) + \frac{9+15}{13+15}\left(-\frac{9}{24}\log\frac{9}{24} - \frac{15}{24}\log\frac{15}{24}\right)$$

$$IG_{split2} = I_{\alpha\rho\chi} - I_{split1} = -\frac{13}{28}\log\frac{13}{28} - \frac{15}{28}\log\frac{15}{28} - I_{split1}$$



split2

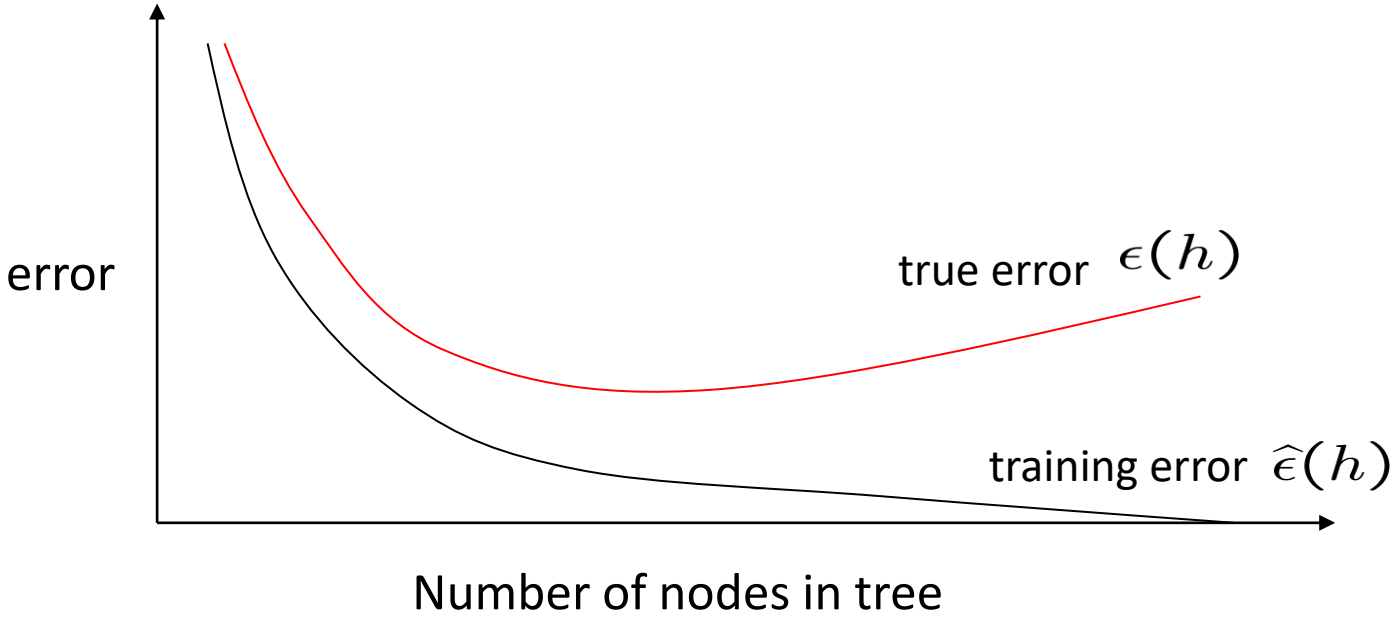


$$I_{split2} = \frac{8+0}{13+15}\left(-\frac{8}{8}\log\frac{8}{8} - \frac{0}{8}\log\frac{0}{8}\right) + \frac{5+15}{13+15}\left(-\frac{5}{20}\log\frac{5}{20} - \frac{15}{20}\log\frac{15}{20}\right)$$

$$IG_{split2} = I_{\alpha\rho\chi} - I_{split2} = -\frac{13}{28}\log\frac{13}{28} - \frac{15}{28}\log\frac{15}{28} - I_{split2}$$

Ο αλγόριθμος τερματίζει

- Όταν όλα τα φύλλα είναι «καθαρά» (χωρίς μπλεγμένα παραδείγματα)
- Όταν ξεμείνει από πιθανά splits
- Στην συνέχεια μειώνεται το μέγεθος του δέντρου με κλάδεμα για να αποφευχθεί το πρόβλημα της υπερπροσαρμογής.



Μετα-κλάδεμα (Post-pruning)
Reduced Error Pruning

[1] Χωρίζω τα δεδομένα εκπαίδευσης (S_{full}) σε δυο τμήματα:
Ένα μικρότερο σετ εκπαίδευσης S
Ένα σετ επικύρωσης (validation set) V

[2] Κατασκευάζω ένα δέντρο απόφασης T, χρησιμοποιώντας το S.

[3] Κλαδεύω χρησιμοποιώντας το V

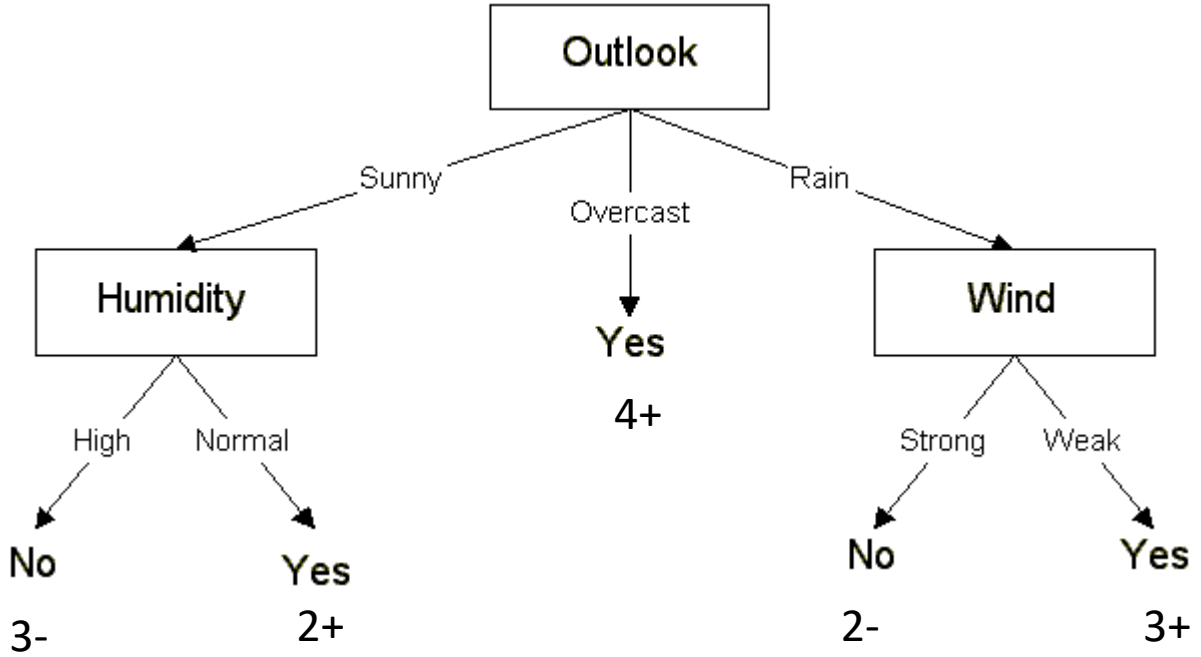
επανάληψη

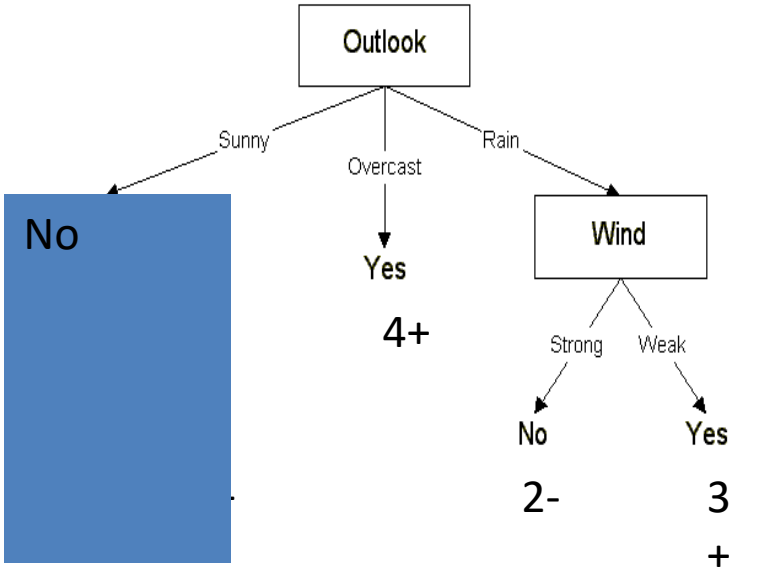
Για κάθε κόμβο u στο T μετράω στο υπο-δέντρο που τον έχει ρίζα πόσα παραδείγματα στα φύλλα του ταξινομούνται σε κάθε τιμή της κλάσης.

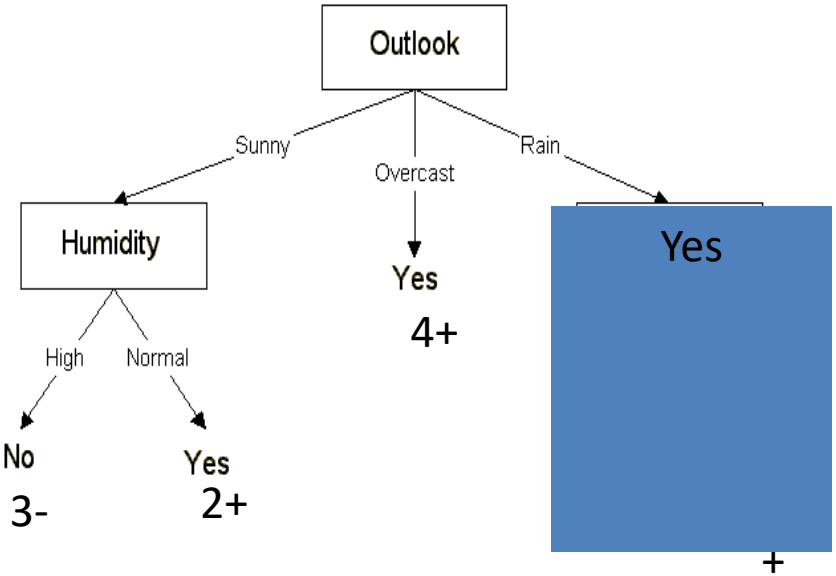
Υπολογίζω την κλάση που έχει πλειοψηφική τιμή (c) στο υπο-δέντρο.

Αφαιρώ το υπο-δέντρο, και στη θέση του τοποθετώ ένα φύλλο με τιμή c στην κλάση ταξινόμησης.

Αν το μικρότερο δέντρο T’ μειώνει το σφάλμα στο V, σε σχέση με το T, τότε T = T’


Outlook
Sunny
Overcast
Rain
Humidity
Yes
4+
Wind
High
Normal
Strong
Weak
No
Yes
No
Yes
3-
2+
2-
3+


Outlook
Sunny
Overcast
Rain
No
Yes
4+
Wind
Strong
Weak
No
Yes
2-
3
+


Outlook
Sunny
Overcast
Rain
Humidity
Yes
4+
Yes
High
Normal
No
Yes
3-
2+
+

Προ-κλάδεμα (Pre-pruning)

- Στο προ-κλάδεμα αποφασίζουμε κατά την διάρκεια σχηματισμού του δέντρου πότε πρέπει να σταματήσουμε να προσθέτουμε χαρακτηριστικά στο δέντρο.
- Αν πχ το κέρδος πληροφορίας όλων των πιθανών χαρακτηριστικών που μπορούμε να προσθέσουμε είναι κάτω από ένα συγκεκριμένο κατώφλι, σταματάει η κατασκευή του δέντρου.
- Αυτό μπορεί να είναι προβληματικό
  - Μερικές φορές μπορεί μεμονωμένα χαρακτηριστικά να μην συμβάλουν σε μια απόφαση, αλλά σε συνδυασμό με κάποιο/α άλλο/α μπορεί να έχουν σημαντική επίδραση στην απόφαση.