

# Üzleti Elemzések Módszertana

## 4. Gyakorlat: Döntési fák

Kuknyó Dániel  
Budapesti Gazdasági Egyetem

2023/24  
2.félév

1 Bevezetés

2 Tanítás

3 Döntési fák tulajdonságai

1 Bevezetés

2 Tanítás

3 Döntési fák tulajdonságai

# Döntési fák a gépi tanulásban

A döntési fák olyanok, mint a svájci bicska: nagyon sok mindenre jó, de szinte semmire sem a legalkalmasabb.

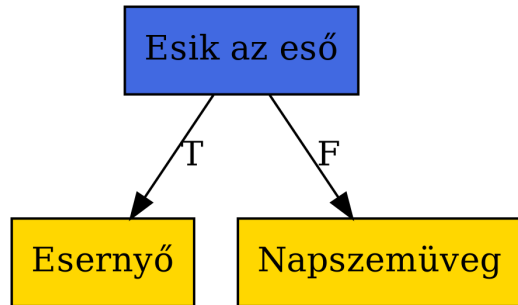
A döntési fák kifejezetten hasznosak gyors taníthatóságuk, jól értelmezhetőségük és pontosságuk miatt.



# Döntési fák alapjai

A döntési fák képesek mind regressziós és osztályozási problémákat is végrehajtani. Könnyen illeszthetők komplex adathalmazokra.

Az algoritmus alapja, hogy **mintaegyedeket osztályoz** változóikban felvett értékeik alapján.

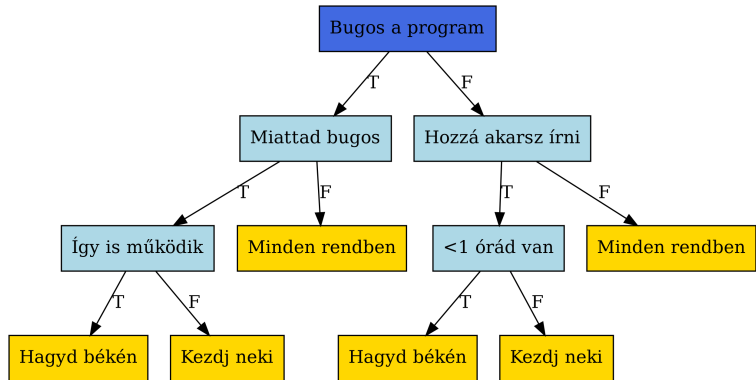


# Egy kezdeti döntési fa

A folyamat a fa gyökerénél kezdődik.

A mintaegyedek a **csomópontok kérdéseire válaszolnak** változóikban felvett értékeik alapján.

A végső osztály lehet **folytonos és diszkrét** változó is.



# A döntési fa komponensei

**Gyökér csomópont**

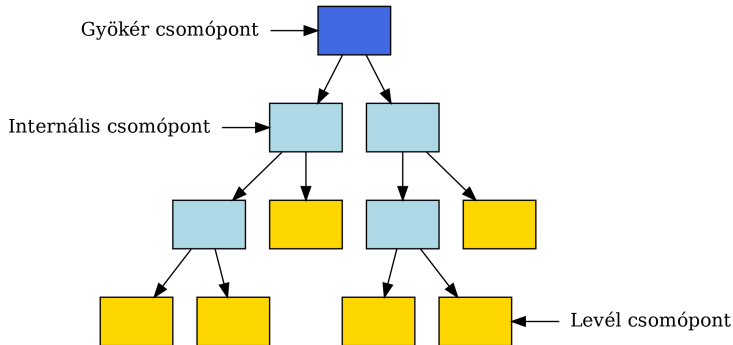
Csak outputja van.

**Internális csomópont**

Van inputja és outputja is.

**Levél csomópont**

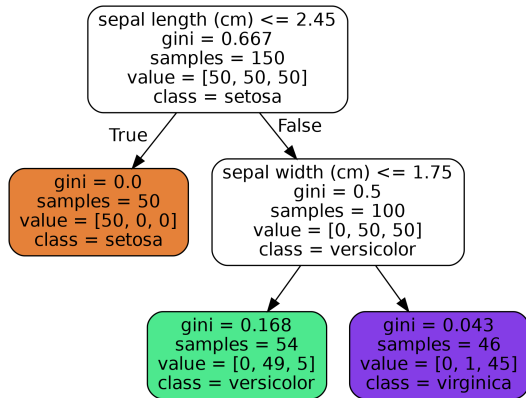
Csak inputja van.



# Döntési fa az Írisz adathalmazon

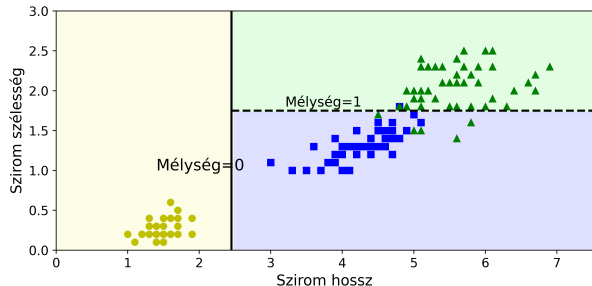
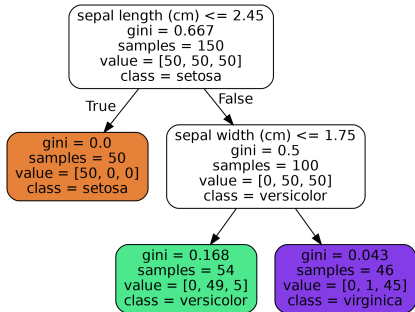
Az első szeparálási változó szirmok hossz, aminek a küszöbértéke 2.45 cm. Ha az adott virág szirmok hossza kevesebb mint ez az érték akkor a modell szerint a becsült osztály Setosa.

Ha viszont nagyobb akkor a következő szeparálási ponthoz ér az osztályozás, ami szerint a következő kérdés, hogy a szirmok szélesség kisebb-e mint 1.75 cm. Ha igen, a becsült osztály versicolor, egyébként pedig Virginica.





# A fa ábrázolása



A vastag vonal a gyökérből származó határ. Mivel a bal oldali halmaz teljesen tiszta, nem lehet tovább bontani. De a jobb oldali részhalmaz továbbra is kevert, ezért a jobb oldali első szintű belső nódus tovább bontja 1.75cm küszöbnél.

1 Bevezetés

2 Tanítás

3 Döntési fák tulajdonságai

# Tisztátalanság

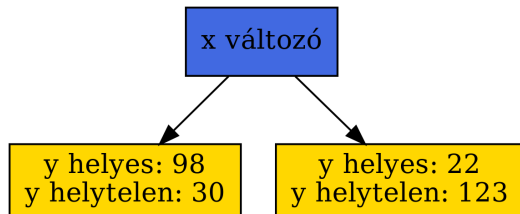
Azok a változók, amelyek nem képesek 1 : 0 arányban szeparálni az egyedeket tisztátalannak számítanak. Ennek egyik mutatószáma a Gini-index.

## Gini

$$G(x) = 1 - P(A)^2 - P(B)^2$$

- $P(\cdot)$ : adott levélbe kerülés valószínűsége

Egy változó Gini-indexe leveleinek Gini-indexeinek súlyozott átlaga.



$$G(A) = 1 - \left(\frac{98}{98 + 30}\right)^2 - \left(\frac{30}{98 + 30}\right)^2 = 0.35$$

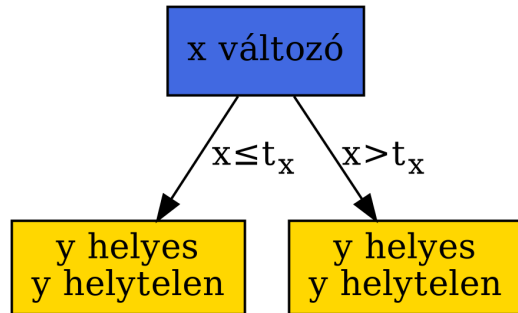
$$G(B) = 1 - \left(\frac{22}{22 + 123}\right)^2 - \left(\frac{123}{22 + 123}\right)^2 = 0.25$$

$$G(x) = \left(\frac{128}{128 + 145}\right) \cdot 0.35 + \left(\frac{145}{128 + 145}\right) \cdot 0.25 = 0.3$$

# Szeparáció folytonos változó esetén

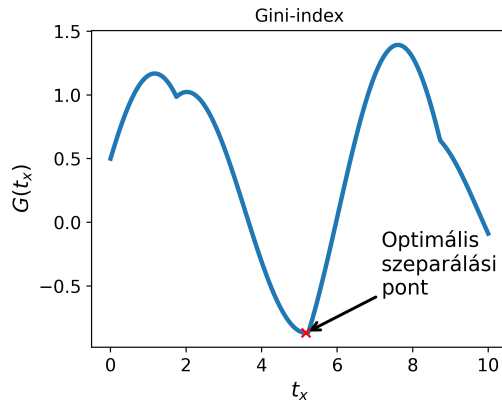
A folytonos változónak minden értékéhez tartozik egy Gini-index.

Egy adott  $x$  változóra a  $t_x$  küszöbérték menti szeparáció, hogy az egyik partícióba azon mintaegyedek kerülnek, amelyekre  $x \leq t_x$  a másikba pedig amelyekre  $x > t_x$ .



# Szeparáció folytonos változó esetén

Ennek megfelelően a szeparáció ott a legjobb, ahol a  $G(t_x)$  függvénynek minimuma van.

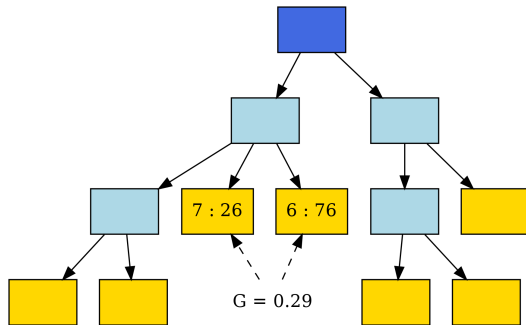


# Mikor érdemes szeparálni?

Amikor egy csomópontnak **magasabb a tisztátalansága tovább bontáskor**, felesleges a szeparáció és levélcsomópont válik belőle.

Gyökércsomópont abból a változóból válik, amelynek **a legalacsonyabb a tisztátalansága**.

Ebben az esetben szeparációval  $G = 0.29$

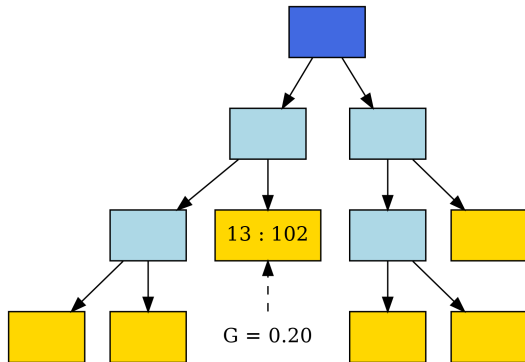


# Mikor érdemes szeparálni?

Amikor egy csomópontnak **magasabb a tisztátalansága tovább bontáskor**, felesleges a szeparáció és levélcsomópont válik belőle.

Gyökércsomópont abból a változóból válik, amelynek **a legalacsonyabb a tisztátalansága**.

Szeparáció nélkül  $G = 0.20$ , tehát a szeparáció felesleges.



# A CART tanító algoritmus

A **C**lassification **A**nd **R**egression **T**rees egy döntési fák tanítására használt algoritmus.

Az eljárás  $x$  változóra és  $t_x$  küszöbértékre olyan  $(x, t_x)$  párokat keres, amelyekre a létrejövő részhalmazoknak a lehető legalacsonyabb a tisztátalansága.

Ezt rekurzívan ismétli kilépésig.

## A CART költségfüggvénye

$$J(x, t_x) = \frac{m_A}{m} G_A + \frac{m_B}{m} G_B$$

Ahol:

- $G_A$ : Bal oldali nódus Gini-indexe
- $G_B$ : Jobb oldali nódus Gini-indexe
- $m_A$ : Bal oldali nódusba bekerült egyedek száma
- $m_B$ : Jobb oldali nódusba bekerült egyedek száma
- $m$ : Egyedek száma a teljes halmazban

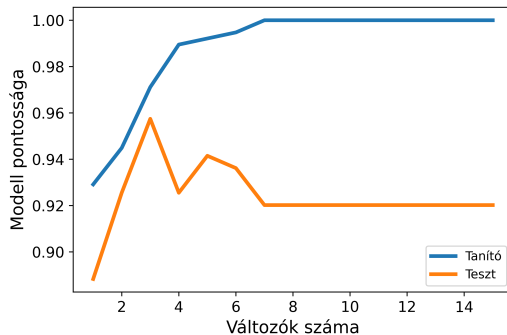


## Korai leállás döntési fák esetén

Túltanulás esetén a **tanító pontosság nagyon magas lesz, viszont a teszt pontosság alacsony.**

Döntési fák esetén annyi változót érdemes meghagyni a modellezés során, amennyivel a lehető legmagasabb a teszt pontosság.

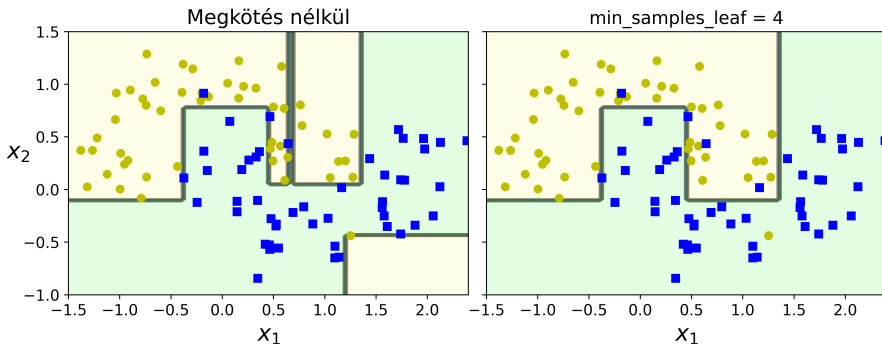
Korai leállás esetén a **modell kiszáll a tanításból, ha a validációs pontosság elkezd csökkenni.**



# Döntési fák regularizálása

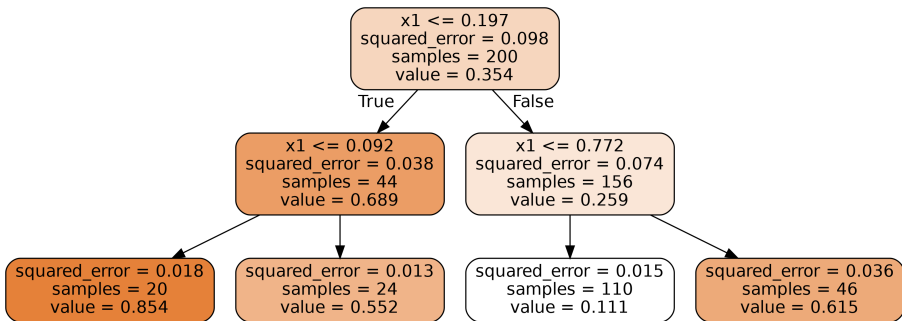
A döntési fák meglehetősen kevés előfeltételezéssel élnek az adathalmaz irányába. Ha megkötések nélkül van tanítva, **könnyen túltanulhat a modell**.

A bal oldali ábrán egy regularizáció nélküli, a jobb oldalon pedig egy `min_samples_leaf=4` paraméterrel tanított döntési fa látható.



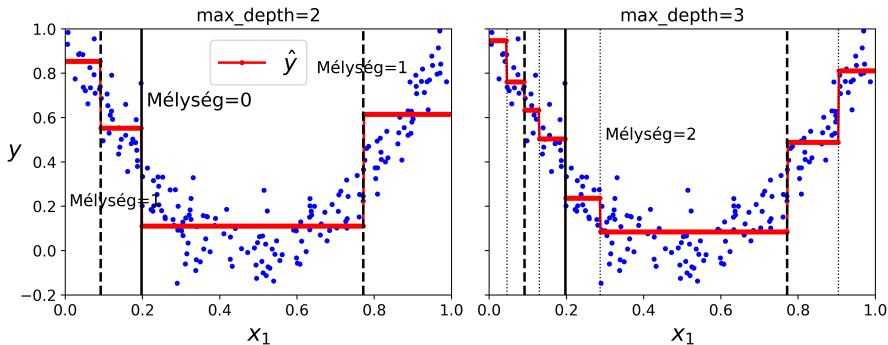
# Regresszió döntési fákkal

Regresszió esetén a döntési fák leveleikben folytonos változókhoz tartozó értékeket vesznek fel. Ebben az esetben a predikció a levelekbe bekerült mintaegyedek célváltozóikban felvett értékeinek az átlaga.



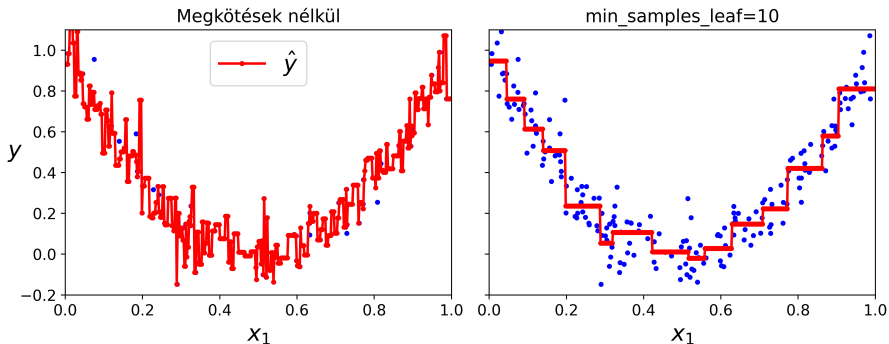
# Regresszió döntési fákkal

Regresszió esetén a döntési fák leveleikben folytonos változókhoz tartozó értékeket vesznek fel. Ebben az esetben a predikció a levelekbe bekerült mintaegyedek célváltozóikban felvett értékeinek az átlaga.



# Regularizáció regresszor fák esetén

Az osztályozó fákhoz hasonlóan a regresszor fák is hajlamosak a túltanulásra. A regularizáció olyan paraméterek állításával érhető el, mint a `min_samples_leaf`, `min_samples_split`, `max_leaf_nodes`, `max_depth`.



# CART tanító algoritmus regressziós fákra

A regresszor fák a tisztátalanság helyett az MSE mutatót minimalizálják.

Egy  $V$  nódus becsült értéke a bele került mintaegyedek célváltozóikban felvett értékeinek átlaga:

$$\hat{y}_V = \frac{1}{m_V} \sum_{i \in m_A} y_i$$

## A CART regresszor algoritmus költségfüggvénye

$$J(x, t_x) = \frac{m_A}{m} MSE_A + \frac{m_B}{m} MSE_B$$

Ahol:

- $MSE_A = \sum_{i \in m_A} (\hat{y} - y_i)^2$ :  $A$  csomópont átlagos négyzetes hibája
- $MSE_B = \sum_{i \in m_B} (\hat{y} - y_i)^2$ :  $B$  csomópont átlagos négyzetes hibája

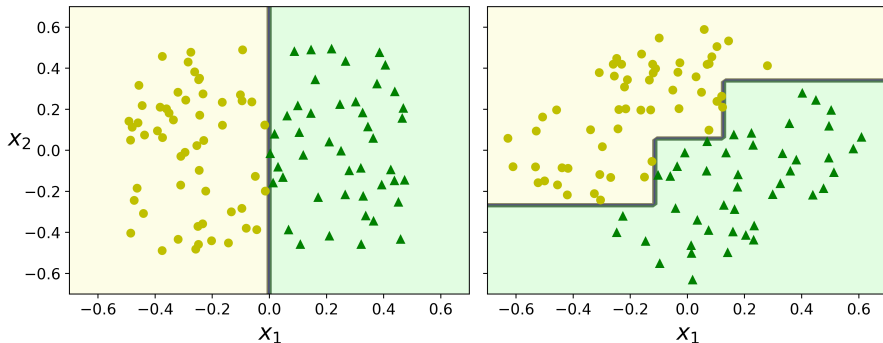
1 Bevezetés

2 Tanítás

3 Döntési fák tulajdonságai

# Instabilitás: rotáció az adathalmazon

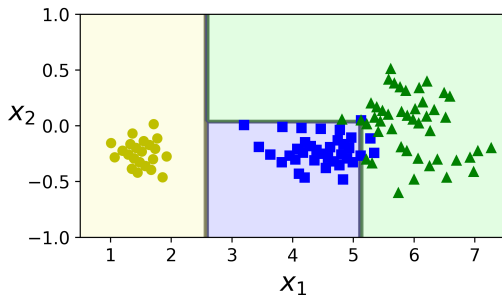
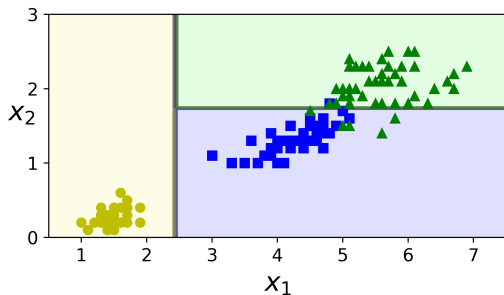
Az alábbi példában egy lineárisan szeparálható adathalmazon történt  $45^\circ$ -os forgatás után látható ugyanannak a modellnek a predikciója. A létrejövő döntési határ jóval komplexebb a transzformált adathalmaz esetén.





# Instabilitás: rotáció az adathalmazon

Az következő példában az Írisz adathalmazon egy  $180^\circ$ -os forgatás után láthatóak hasonló módon paraméterezett modellek döntési határai. Érdekes megfigyelni, mennyire különbözik a predikció a torzított adathalmazon.



# Instabilitás: variációk az adathalmazban

Ebben az esetben a legszélesebb Versicolor (kék) nem került bele a minta adathalmazba. Egyetlen minta adatpont változása is nagy torzítást képes bevinni a modellbe.

