Mesterséges Intelligencia I.
ID3 futás

Futtassuk az $ID3$ algoritmust az S adatbázison az

$$E(S) = 4p_+p_-$$

entrópia használatával.

Kezdeti hívás: $ID3\,(S, \{Outlook, Temperature, Humidity, Wind\})$

	Outlook	Temperature	Humidity	Wind	Play tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

1. ábra. S adatbázis (eredeti)

A következő *Gain* értékek maximumaként tudja az algoritmus kiválasztani az aktuális hívásban létrejövő csomópont címkéjét.

$$E(S) = 4\frac{9}{14}\frac{5}{14} = \frac{90}{98}$$

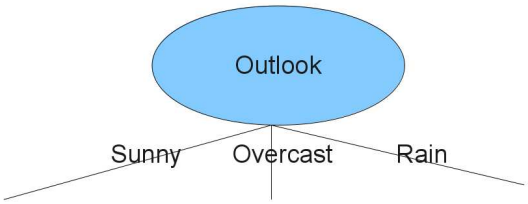
$$Gain(S, Outlook) = \frac{90}{98} - \left(\frac{5}{14}4\frac{2}{5}\frac{3}{5} + \frac{4}{14}4\frac{4}{4}\frac{0}{4} + \frac{5}{14}4\frac{3}{5}\frac{2}{5}\right) = \frac{90}{98} - \frac{24}{35} = 0.23$$

$$Gain(S, Temp) = \frac{90}{98} - \left(\frac{4}{14}4\frac{2}{4}\frac{2}{4} + \frac{6}{14}4\frac{4}{6}\frac{2}{6} + \frac{4}{14}4\frac{3}{4}\frac{1}{4}\right) = \frac{90}{98} - \frac{37}{42} = 0.03$$

$$Gain(S, Hum) = \frac{90}{98} - \left(\frac{7}{14}4\frac{3}{7}\frac{4}{7} + \frac{7}{14}4\frac{6}{7}\frac{1}{7}\right) = \frac{90}{98} - \frac{36}{49} = 0.18$$

$$Gain(S, Wind) = \frac{90}{98} - \left(\frac{8}{14}4\frac{6}{8}\frac{2}{8} + \frac{6}{14}4\frac{3}{6}\frac{3}{6}\right) = \frac{90}{98} - \frac{6}{7} = 0.06$$

Maximális a *Sunny* attributum $\rightarrow$ ő lesz a gyökér, 3 leszármazottal.



2. ábra. Első hívás által létrehozott csomópont

Minden egyes leszármazott egy-egy rekurzív hívással lesz létrehozva.
A *Sunny* él mentén történő hívás: $ID3\left(S_{Outlook=Sunny}, \{Temperature, Humidity, Wind\}\right)$

	Outlook	Temperature	Humidity	Wind	Play tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes

3. ábra. $S_{Outlook=Sunny}$ adatbázis

A következő *Gain* értékek maximumaként tudja az algoritmus kiválasztani az aktuális hívásban létrejövő csomópont címkéjét.

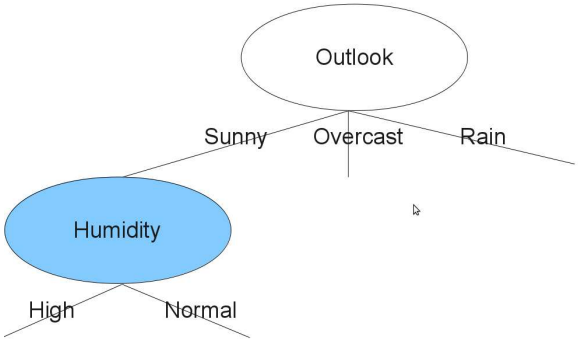
$$E(S) = 4\frac{2}{5}\frac{3}{5} = \frac{24}{25}$$

$$Gain(S_{Outlook=Sunny}, Temp) = \frac{24}{25} - \left(\frac{2}{5}4\frac{0}{2}\frac{2}{2} + \frac{2}{5}4\frac{1}{2}\frac{1}{2} + \frac{1}{5}4\frac{1}{1}\frac{0}{1}\right) = \frac{24}{25} - \frac{4}{10}$$

$$Gain(S_{Outlook=Sunny}, Hum) = \frac{24}{25} - \left(\frac{3}{5}4\frac{0}{3}\frac{3}{3} + \frac{2}{5}4\frac{2}{2}\frac{0}{2}\right) = \frac{24}{25}$$

$$Gain(S_{Outlook=Sunny}, Wind) = \frac{24}{25} - \left(\frac{3}{5}4\frac{1}{3}\frac{2}{3} + \frac{2}{5}4\frac{1}{2}\frac{1}{2}\right) = \frac{24}{25} - X > 0$$

Maximális a *Humidity* attributum $\rightarrow$ ő lesz az *Outlook* csúcs *Sunny* él menti leszármazottja.



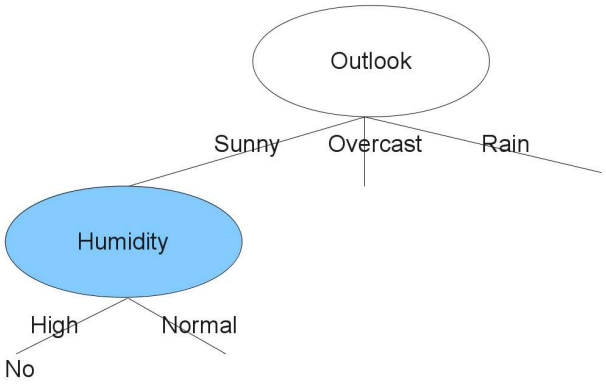
4. ábra. Második hívás által létrehozott csomópont

Minden egyes leszármazott egy-egy rekurzív hívással lesz létrehozva.
A $High$ él mentén történő hívás: $ID3\left(S_{Outlook=Sunny \wedge Humidity=High}, \{Temperature, Wind\}\right)$

	Outlook	Temperature	Humidity	Wind	Play tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D8	Sunny	Mild	High	Weak	No

5. ábra. $S_{Outlook=Sunny \wedge Humidity=High}$ adatbázis

Ahogy látható az adatbázis homogén (csak No címkével rendelkező példákat tartalmaz), ami az ID3 algoritmus megállási feltétele. Tehát megállunk és egy No címkével rendelkező levelet hozunk létre.



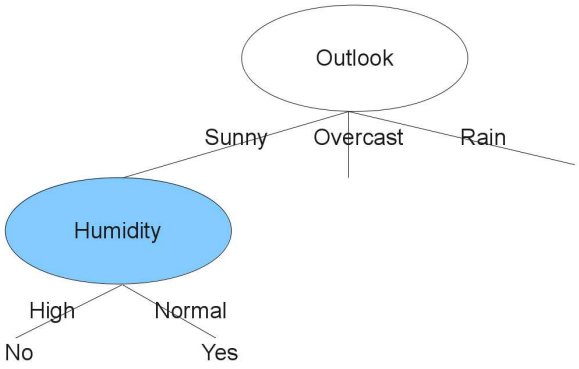
6. ábra. Második hívás által létrehozott csomópont

A $Normal$ él mentén történő hívás: $ID3\,(S_{Outlook=Sunny \wedge Humidity=Normal}, \{Temperature, Wind\})$

	Outlook	Temperature	Humidity	Wind	Play tennis
D9	Sunny	Cool	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes

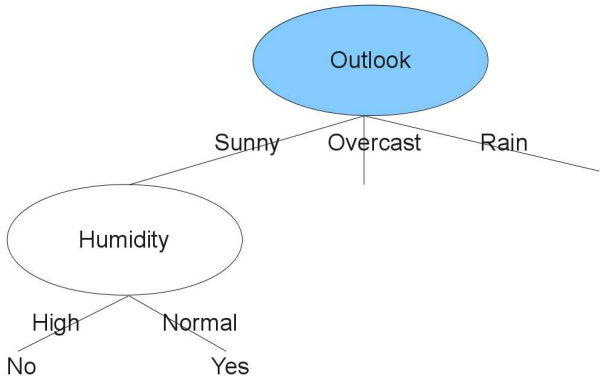
7. ábra. $S_{Outlook=Sunny \wedge Humidity=Normal}$ adatbázis

Ahogy látható az adatbázis homogén (csak Yes címkével rendelkező példákat tartalmaz), ami az ID3 algoritmus megállási feltétele. Tehát megállunk és egy Yes címkével rendelkező levelet hozunk létre.



8. ábra. Harmadik hívás által létrehozott csomópont

Mivel a $Humidity$ csomópont teljes részfáját létrehoztuk visszatér a rekurzió.



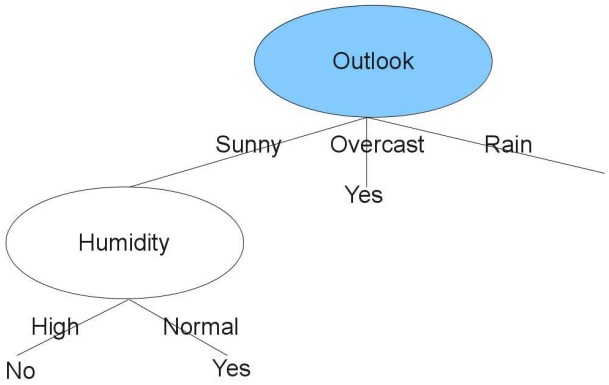
9. ábra. Visszatérünk a gyökér szintjére

A $Overcast$ él mentén történő hívás: $ID3\,(S_{Outlook=Overcast}, \{Temperature, Humidity, Wind\})$

	Outlook	Temperature	Humidity	Wind	Play tennis
D3	Overcast	Hot	High	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes

10. ábra. $S_{Outlook=Overcast}$ adatbázis

Ahogy látható az adatbázis homogén (csak Yes címkével rendelkező példákat tartalmaz), ami az ID3 algoritmus megállási feltétele. Tehát megállunk és egy Yes címkével rendelkező levelet hozunk létre.



11. ábra. A következő hívás által létrehozott csomópont

A *Rain* él mentén történő hívás: $ID3\left(S_{Outlook=Rain}, \{Temperature, Humidity, Wind\}\right)$

	Outlook	Temperature	Humidity	Wind	Play tennis
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D10	Rain	Mild	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

12. ábra. $S_{Outlook=Rain}$ adatbázis

A következő *Gain* értékek maximumaként tudja az algoritmus kiválasztani az aktuális hívásban létrejövő csomópont címkéjét.

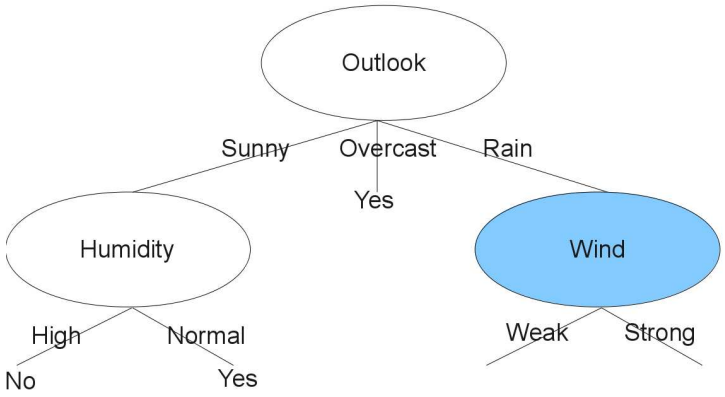
$$E(S) = 4\frac{3}{5}\frac{2}{5} = \frac{24}{25}$$

$$Gain(S_{Outlook=Rain}, Temp) = \frac{24}{25} - \left(\frac{3}{5}4\frac{2}{3}\frac{1}{3} + \frac{2}{5}4\frac{1}{2}\frac{1}{2}\right) = \frac{24}{25} - \frac{14}{15}$$

$$Gain(S_{Outlook=Rain}, Hum) = \frac{24}{25} - \left(\frac{2}{5}4\frac{1}{2}\frac{1}{2} + \frac{3}{5}4\frac{2}{3}\frac{1}{3}\right) = \frac{24}{25} - \frac{14}{15}$$

$$Gain(S_{Outlook=Rain}, Wind) = \frac{24}{25} - \left(\frac{3}{5}4\frac{3}{3}\frac{0}{3} + \frac{2}{5}4\frac{0}{2}\frac{2}{2}\right) = \frac{24}{25}$$

Maximális a $Wind$ attributum $\rightarrow$ ő lesz az *Outlook* csúcs *Rain* él menti leszármazottja.



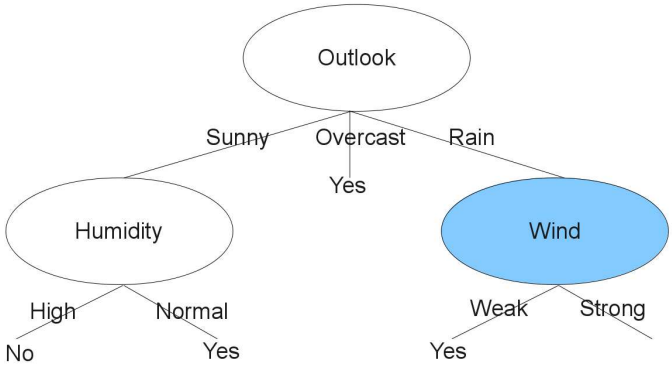
13. ábra. A következő hívás által létrehozott csomópont

Minden egyes leszármazott egy-egy rekurzív hívással lesz létrehozva.
A $Weak$ él mentén történő hívás: $ID3\left(S_{Outlook=Rain \wedge Wind=Weak}, \{Temperature, Humidity\}\right)$

	Outlook	Temperature	Humidity	Wind	Play tennis
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes

14. ábra. $S_{Outlook=Rain \wedge Wind=Weak}$ adatbázis

Ahogy látható az adatbázis homogén (csak Yes címkével rendelkező példákat tartalmaz), ami az ID3 algoritmus megállási feltétele. Tehát megállunk és egy Yes címkével rendelkező levelet hozunk létre.



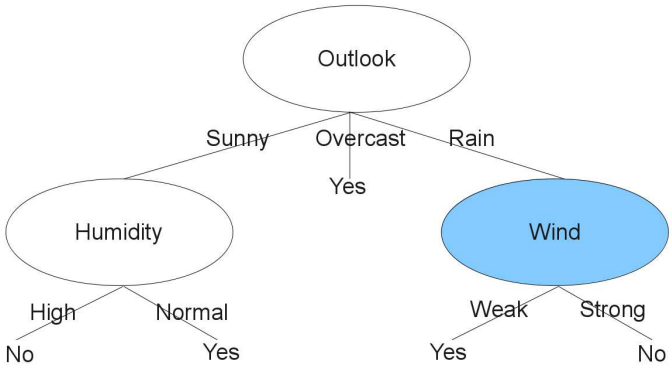
15. ábra. Következő hívás által létrehozott csomópont

A *Strong* él mentén történő hívás: $ID3\left(S_{Outlook=Rain \wedge Wind=Strong}, \{Temperature, Humidity\}\right)$

	Outlook	Temperature	Humidity	Wind	Play tennis
D6	Rain	Cool	Normal	Strong	No
D14	Rain	Mild	High	Strong	No

16. ábra. $S_{Outlook=Rain \wedge Wind=Strong}$ adatbázis

Ahogy látható az adatbázis homogén (csak No címkével rendelkező példákat tartalmaz), ami az ID3 algoritmus megállási feltétele. Tehát megállunk és egy No címkével rendelkező levelet hozunk létre.



17. ábra. Következő hívás által létrehozott csomópont

Mivel a $Wind$ csomópont teljes részfáját létrehoztuk visszatér a rekurzió.
Mivel a *Outlook* csomópont teljes részfáját létrehoztuk visszatér a rekurzió.
A futás véget ér.

Ugyan ez a feladat Shanon entropiával számolva:

$$E : -\sum p_i log_2 pi$$

$$E(S) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} = 0.94$$

$$Gain(S, Outlook) = 0.94 - \left[\frac{4}{14}(-\frac{4}{4}log_2\frac{4}{4} - \frac{0}{4}log_2\frac{0}{4}) + \frac{5}{14}(-\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5}) + \frac{5}{14}(-\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{3}{5})\right] = \underline{0.25}$$

$$Gain(S, Temp) = 0.94 - \left[\frac{4}{14}(-\frac{3}{4}log_2\frac{3}{4} - \frac{1}{4}log_2\frac{1}{4}) + \frac{4}{14}(-\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4}) + \frac{6}{14}(-\frac{4}{6}log_2\frac{4}{6} - \frac{2}{6}log_2\frac{2}{6})\right] = 0.029$$

$$Gain(S, Hum) = 0.94 - \left[\frac{7}{14}(-\frac{3}{7}log_2\frac{3}{7} - \frac{4}{7}log_2\frac{4}{7}) + \frac{7}{14}(-\frac{6}{7}log_2\frac{6}{7} - \frac{1}{7}log_2\frac{1}{7})\right] = 0.15$$

$$Gain(S, Wind) = 0.94 - \left[\frac{6}{14}(-\frac{3}{6}log_2\frac{3}{6} - \frac{3}{6}log_2\frac{3}{6}) + \frac{8}{14}(-\frac{6}{8}log_2\frac{6}{8} - \frac{2}{8}log_2\frac{2}{8})\right] = 0.048$$

$$E(S_{Outlook=Sunny}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.97$$

$$Gain(S_{Outlook=Sunny}, Temp) = 0.97 - \left[\frac{2}{5}(-\frac{0}{2}log_2\frac{0}{2} - \frac{2}{2}log_2\frac{2}{2}) + \frac{2}{5}(-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2}) + \frac{1}{5}(-\frac{1}{1}log_2\frac{1}{1} - \frac{0}{1}log_2\frac{0}{1})\right.$$

$$Gain(S_{Outlook=Sunny}, Hum) = 0.97 - \left[\frac{3}{5}(-\frac{0}{3}log_2\frac{0}{3} - \frac{3}{3}log_2\frac{3}{3}) + \frac{2}{5}(-\frac{2}{2}log_2\frac{2}{2} - \frac{0}{2}log_2\frac{0}{2})\right] = \underline{0.97}$$

$$Gain(S_{Outlook=Sunny}, Wind) = 0.97 - \left[\frac{2}{5}(-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2}) + \frac{3}{5}(-\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3})\right] = 0.02$$

$$E(S_{Outlook=Rain}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.97$$

$$Gain(S_{Outlook=Rain}, Temp) = 0.97 - \left[0 + \frac{3}{5}(-\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3}) + \frac{2}{5}(-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2})\right] = 0.02$$

$$Gain(S_{Outlook=Rain}, Hum) = 0.97 - \left[\frac{2}{5}(-\frac{1}{2}log_2\frac{1}{2} - \frac{1}{2}log_2\frac{1}{2}) + \frac{3}{5}(-\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3})\right] = 0.02$$

$$Gain(S_{Outlook=Rain}, Wind) = 0.97 - \left[\frac{2}{5}(-\frac{0}{2}log_2\frac{0}{2} - \frac{2}{2}log_2\frac{2}{2}) + \frac{3}{5}(-\frac{3}{3}log_2\frac{3}{3} - \frac{0}{3}log_2\frac{0}{3})\right] = \underline{0.97}$$