Aufgabe: Entropie und Informationsgewinn in einem Entscheidungsbaum

Aufgabenstellung:

Ein Datensatz enthält Informationen über die Abschlusswahrscheinlichkeit von Studierenden basierend auf den Merkmalen **Lernzeit**, **Kursart** und **Abschlussniveau**. Dein Ziel ist es, das Kriterium für den ersten Split im Entscheidungsbaum zu bestimmen. Berechne dazu die **Entropie des Gesamtdatensatzes** und den **Informationsgewinn** für jedes Merkmal, um das Attribut mit dem höchsten Informationsgewinn zu finden.

Gegebener Datensatz:

Lernzeit	Kursart	Abschlussniveau	Abschluss erreicht
Hoch	Theorie	Mittel	Ja
Mittel	Praxis	Hoch	Ja
Hoch	Theorie	Hoch	Ja
Niedrig	Theorie	Mittel	Nein
Niedrig	Praxis	Mittel	Nein
Mittel	Theorie	Mittel	Ja
Hoch	Theorie	Mittel	Ja
Mittel	Praxis	Niedrig	Nein
Hoch	Theorie	Hoch	Ja
Niedrig	Praxis	Niedrig	Nein
Mittel	Theorie	Mittel	Ja

Lernzeit	Kursart	Abschlussniveau	Abschluss erreicht
Mittel	Theorie	Hoch	Ja
Niedrig	Theorie	Niedrig	Nein
Hoch	Praxis	Hoch	Ja

- Zielvariable: **Abschluss erreicht (Ja oder Nein)**
- Merkmale: **Lernzeit (Hoch, Mittel, Niedrig)**, **Kursart (Theorie, Praxis)**, **Abschlussniveau (Hoch, Mittel, Niedrig)**

1. Berechnung der Gesamtentropie

Die Zielvariable *Abschluss erreicht* hat zwei mögliche Werte: *Ja* oder *Nein*.
Die Entropie wird berechnet mit:

$$H(S) = -p_+ \log_2(Ja) - p_- \log_2(Nein)$$

Zählen der Werte:

- **Ja** = 9
- **Nein** = 5
- **Gesamtanzahl** = 14

$$p_+ = \frac{9}{14}, \quad p_- = \frac{5}{14}$$

$$H(S) = -\left(\frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14}\right)$$

$$H(S) = 0.9402859586706311 \approx 0.94$$

2. Berechnung der Entropie für jedes Merkmal

(a) Split nach *Lernzeit*

Lernzeit	Anzahl	Ja	Nein	Entropie
Hoch	5	5	0	0.000
Mittel	5	4	1	0.722
Niedrig	4	0	4	0.000

$$H(S_H) = -p_+ \log_2(Ja) - p_- \log_2(Nein) = 0$$

$$H(S_M) = -p_+ \log_2(Ja) - p_- \log_2(Nein) = \left(-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}\right) \approx 0.722$$

$$H(S_N) = -p_+ \log_2(Ja) - p_- \log_2(Nein) = 0$$

Gewichtete Entropie:

$$H_{Lernzeit} = \frac{5}{14}\cdot 0 + \frac{5}{14}\cdot 0.722 + \frac{4}{14}\cdot 0 \approx 0.258$$

(b) Split nach *Kursart*

Kursart	Anzahl	Ja	Nein	Entropie
Theorie	9	7	2	0.764
Praxis	5	2	3	0.971

$$H(S_T) = -p_+ \log_2(Ja) - p_- \log_2(Nein) = \left(-\frac{7}{9}\log_2\frac{2}{9} - \frac{2}{9}\log_2\frac{3}{9}\right) \approx 0.764$$

$$H(S_P) = -p_+ \log_2(Ja) - p_- \log_2(Nein) = \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) \approx 0.971$$

Gewichtete Entropie:

$$H_{Kursart} = \frac{9}{14}\cdot 0.764 + \frac{5}{14}\cdot 0.971 \approx 0.838$$

© Split nach *Abschlussniveau*

Abschlussniveau	Anzahl	Ja	Nein	Entropie
Hoch	5	5	0	0.000
Mittel	6	4	2	0.918
Niedrig	3	0	3	0.000

$$H(S_H) = -p_+ \log_2(Ja) - p_- \log_2(Nein) = 0$$

$$H(S_M) = -p_+ \log_2(Ja) - p_- \log_2(Nein) = \left(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right) \approx 0.918$$

$$H(S_N) = -p_+ \log_2(Ja) - p_- \log_2(Nein) = 0$$

Gewichtete Entropie:

$$H_{Abschluss} = \frac{5}{14}\cdot 0 + \frac{6}{14}\cdot 0.918 + \frac{3}{14}\cdot 0 \approx 0.393$$

3. Berechnung des Informationsgewinns

$$IG(\text{Lernzeit}) = H(S) - H_{Lernzeit} = 0.940 - 0.258 = 0.682$$

$$IG(\text{Kursart}) = H(S) - H_{Kursart} = 0.940 - 0.838 = 0.102$$

$$IG(\text{Abschluss}) = H(S) - H_{Abschluss} = 0.940 - 0.393 = 0.547$$

Da **Lernzeit** den höchsten Informationsgewinn hat, wählen wir dies als erstes Split-Kriterium.