DTL01: Entscheidungsbäume mit CAL3 und ID3
Vollständige Handsimulation

Datensatz

Zielvariable: `Kandidat` in {O, M}. Attribute: `Alter` {<35, ≥35}, `Einkommen` {hoch, niedrig}, `Bildung` {Abitur, Bachelor, Master}.

Nr.	Alter	Einkommen	Bildung	Kandidat
1	≥ 35	hoch	Abitur	O
2	< 35	niedrig	Master	O
3	≥ 35	hoch	Bachelor	M
4	≥ 35	niedrig	Abitur	M
5	≥ 35	hoch	Master	O
6	< 35	hoch	Bachelor	O
7	< 35	niedrig	Abitur	M

CAL3 (S_1=4, S_2=0,7), Pfadreihenfolge: Alter → Einkommen → Bildung

Start: ein Blatt mit Stern $*$ (Unwissen). Bei jedem Blatt werden Klassen gezählt. Sobald $n \geq S_1$, wird geprüft: Dominanz $\geq S_2$ und eindeutig ⇒ Abschluss auf diese Klasse, sonst Differenzierung mit dem nächsten auf dem Pfad noch nicht verwendeten Attribut.

Durchlauf 1

1. Bsp 1: {≥ 35, hoch, Abitur, O} Blatt: Zähler $\{O:1\}$.
2. Bsp 2: {< 35, niedrig, Master, O} Blatt: $\{O:2\}$.
3. Bsp 3: {≥ 35, hoch, Bachelor, M} Blatt: $\{O:2, M:1\}$.
4. Bsp 4: {≥ 35, niedrig, Abitur, M} Blatt: $\{O:2, M:2\}$, jetzt $n=4$, keine Dominanz ⇒ **Differenzierung an der Wurzel auf** `Alter`. Das aktuelle Beispiel wird in Ast `Alter` =≥ 35 eingetragen ($\{M:1\}$).
5. Bsp 5: {≥ 35, hoch, Master, O} Blatt `Alter` =≥ 35: $\{M:1, O:1\}$.
6. Bsp 6: {< 35, hoch, Bachelor, O} Blatt `Alter` =< 35: $\{O:1\}$.
7. Bsp 7: {< 35, niedrig, Abitur, M} Blatt `Alter` =< 35: $\{O:1, M:1\}$.

Durchlauf 2

1. Bsp 1 $\rightarrow$ `Alter` $=\geq 35$: $\{M:1, O:2\}$.
2. Bsp 2 $\rightarrow$ `Alter` $=< 35$: $\{O:2, M:1\}$.
3. Bsp 3 $\rightarrow$ `Alter` $=\geq 35$: $\{M:2, O:2\}$ mit $n=4 \Rightarrow$ **Differenzierung auf** `Einkommen`. Aktuelles Beispiel in Ast `Einkommen` = hoch ($\{M:1\}$).
4. Bsp 4 $\rightarrow$ `Alter` $=\geq 35$, `Einkommen` = niedrig: $\{M:1\}$.
5. Bsp 5 $\rightarrow$ `Alter` $=\geq 35$, `Einkommen` = hoch: $\{M:1, O:1\}$.
6. Bsp 6 $\rightarrow$ `Alter` $=< 35$: $\{O:3, M:1\}$ mit $n=4$ und Anteil $3/4 = 0{,}75 \geq S_2 \Rightarrow$ **Abschluss:** `Alter` $=< 35 \Rightarrow O$.
7. Bsp 7 $\rightarrow$ festes Blatt O (keine Aktion).

Durchlauf 3

1. Bsp 1 $\rightarrow$ `Alter` $=\geq 35$, `Einkommen` = hoch: $\{M:1, O:2\}$.
2. Bsp 2 $\rightarrow$ festes Blatt O.
3. Bsp 3 $\rightarrow$ `Alter` $=\geq 35$, `Einkommen` = hoch: $\{M:2, O:2\}$ mit $n=4 \Rightarrow$ **Differenzierung auf** `Bildung`. Aktuelles Beispiel in Ast `Bildung` = Bachelor ($\{M:1\}$).
4. Bsp 4 $\rightarrow$ `Alter` $=\geq 35$, `Einkommen` = niedrig: $\{M:2\}$.
5. Bsp 5 $\rightarrow$ `Alter` $=\geq 35$, `Einkommen` = hoch, `Bildung` = Master: $\{O:1\}$.
6. Bsp 6 $\rightarrow$ festes Blatt O.
7. Bsp 7 $\rightarrow$ festes Blatt O.

Durchlauf 4

Es erfolgen nur noch Zählererhöhungen in den Blättern unter `Alter` $=\geq 35$. Keine Dominanz und keine weitere Differenzierung nötig. In der nächsten Runde tritt keine strukturelle Änderung mehr auf. Abbruch.

Ergebnisbaum (CAL3)

.

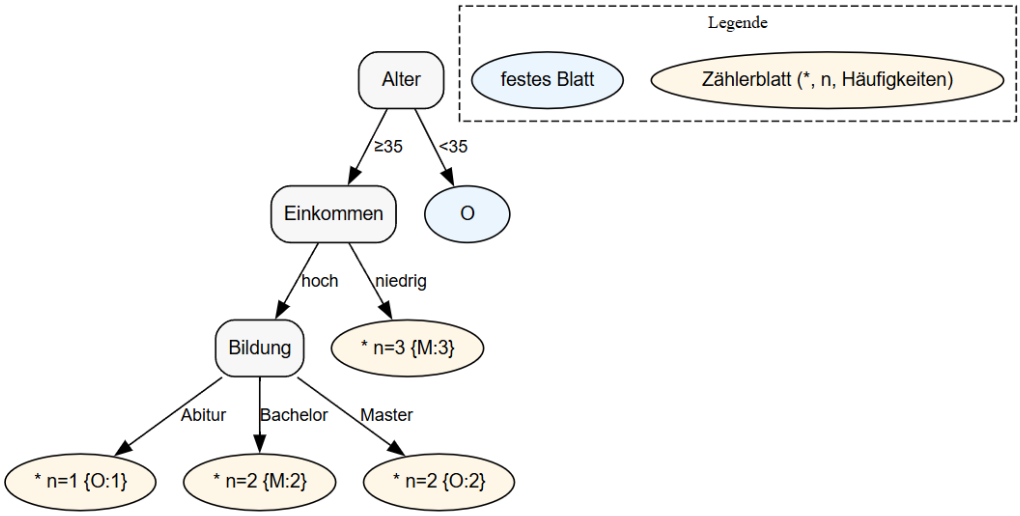

Abbildung 1: Ergebnisbaum nach CAL3 (S_1=4, S_2=0,7)

ID3: Information Gain und Baum

Gesamtentropie mit 4 mal O, 3 mal M:

$$H(S) = -\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7} \approx 0{,}985 \text{ Bit}.$$

Remainder und Gain

- **Bildung**:

$$\begin{aligned} S_{\text{Abitur}} &= \{O, M, M\} \Rightarrow H = -\tfrac{1}{3}\log_2\tfrac{1}{3} - \tfrac{2}{3}\log_2\tfrac{2}{3} \approx 0{,}9183, \\ S_{\text{Bachelor}} &= \{O, M\} \Rightarrow H = 1, \\ S_{\text{Master}} &= \{O, O\} \Rightarrow H = 0, \\ R(S, \text{Bildung}) &= \tfrac{3}{7}\cdot 0{,}9183 + \tfrac{2}{7}\cdot 1 + \tfrac{2}{7}\cdot 0 \approx 0{,}6793, \\ \text{Gain}(S, \text{Bildung}) &= 0{,}985 - 0{,}6793 \approx \mathbf{0{,}306}. \end{aligned}$$

- **Einkommen**:

$$\begin{aligned} S_{\text{hoch}} &= \{O, M, O, O\} \Rightarrow H = -\tfrac{3}{4}\log_2\tfrac{3}{4} - \tfrac{1}{4}\log_2\tfrac{1}{4} \approx 0{,}8113, \\ S_{\text{niedrig}} &= \{O, M, M\} \Rightarrow H \approx 0{,}9183, \\ R(S, \text{Einkommen}) &= \tfrac{4}{7}\cdot 0{,}8113 + \tfrac{3}{7}\cdot 0{,}9183 \approx 0{,}8572, \\ \text{Gain}(S, \text{Einkommen}) &\approx 0{,}985 - 0{,}8572 \approx 0{,}128. \end{aligned}$$

- **Alter**:

$$\begin{aligned} S_{\geq 35} &= \{O, M, M, O\} \Rightarrow H = 1, \\ S_{<35} &= \{O, O, M\} \Rightarrow H \approx 0{,}9183, \\ R(S, \text{Alter}) &= \tfrac{4}{7}\cdot 1 + \tfrac{3}{7}\cdot 0{,}9183 \approx 0{,}965, \\ \text{Gain}(S, \text{Alter}) &\approx 0{,}985 - 0{,}965 \approx 0{,}020. \end{aligned}$$

Größter Gain bei **Bildung** $\Rightarrow$ Wurzel ist `Bildung`.

Rekursiv in den Teilmengen

- `Bildung` = Master: $\{O, O\}$ ist rein $\Rightarrow O$.
- `Bildung` = Abitur: $\{O, M, M\}$. Bester Split ist `Einkommen`: hoch $\Rightarrow O$, niedrig $\Rightarrow M$ (beide rein).
- `Bildung` = Bachelor: $\{O, M\}$. Bester Split ist `Alter`: $\geq 35 \Rightarrow M$, $< 35 \Rightarrow O$ (beide rein).

Ergebnisbaum (ID3)

Dieser ID3-Baum trennt die Trainingsmenge perfekt (7 von 7 richtig).

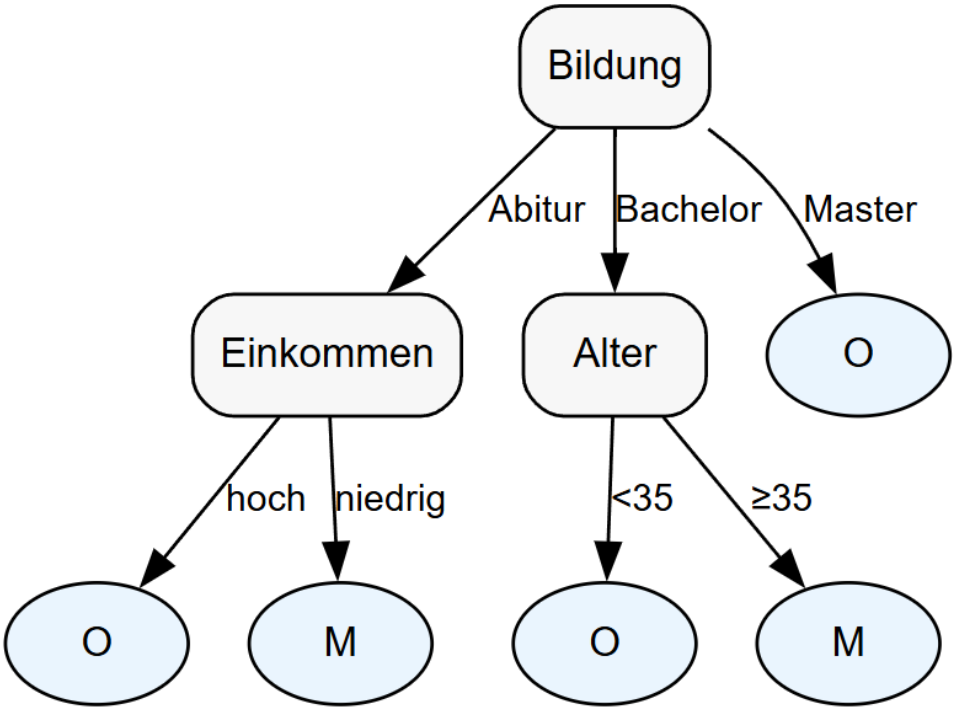

Abbildung 2: Ergebnisbaum nach ID3