Lösung: Entscheidungsbäume

DTL.01

Trainingsdaten

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

Teil A: CAL3 (S1=4, S2=0.7)

CAL3: Knoten darf Blatt werden, wenn

Support ≥ 4 oder Reinheit ≥ 0.7

Attributprüfung

Einkommen:

hoch (1,3,5,6): O=3, M=1 → Reinheit = 3/4 = 0.75 ≥ 0.7 → Blatt = O

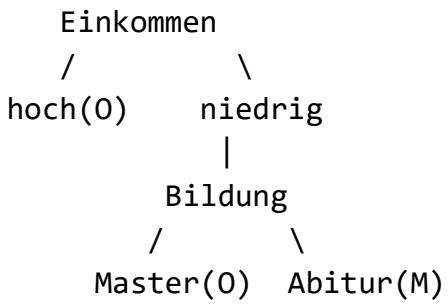
niedrig (2,4,7): O=1, M=2 → Reinheit = 1/3 = 0.33 → weiter splitten

niedrig ⇒ Split nach Bildung:

Master → (Nr.2) → O → vollständig rein

Abitur → (Nr. 4,7) → M → vollständig rein

CAL3-Entscheidungsbaum



Teil B: ID3 (Entropie / Information Gain)

Entropie gesamt

O = 4, M = 3

$$[H(S) = -4/7 \cdot \log2(4/7) - 3/7 \cdot \log2(3/7) \approx 0.9852]$$

Information Gain:

IG(Alter)

hoch: H ≈ 1.0

niedrig: H ≈ 0.918

$$H(\text{Alter}) = 37 * 0.918 + 47 * 1 \quad H(\text{Alter}) \approx 0.964$$

$$\text{Gain}(H, H(\text{Alter})) = 0.986 - 0.964 \quad \text{Gain}(H, H(\text{Alter})) \approx 0.022$$

IG(Einkommen)

hoch: H ≈ 0.8113

niedrig: H ≈ 0.9183

$$[H\_\{split\} = 4/7 \cdot 0.8113 + 3/7 \cdot 0.9183 \approx 0.8571]$$

$$[IG(\text{Einkommen}) = 0.9852 - 0.8571 = 0.1281]$$

IG(Bildung)

Abitur: H ≈ 0.9183

Master: H = 0

Bachelor: H = 1

$$[H\_\{split\} = 3/7 \cdot 0.9183 + 2/7 \cdot 0 + 2/7 \cdot 1 \approx 0.6793]$$

$$[IG(\text{Bildung}) = 0.9852 - 0.6793 = 0.3060]$$

→ Bildung bester Split!

Teilbäume

Master → O (rein)

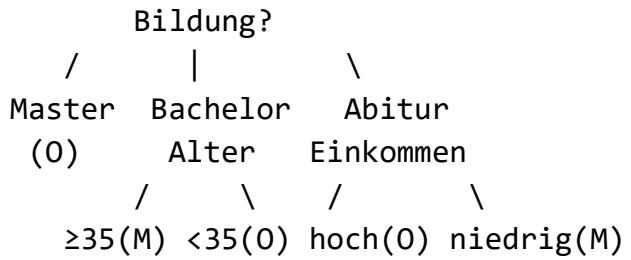
Bachelor (Nr.3,6):

- Alter trennt: ≥35 → M / <35 → O → 2 reine Blätter

Abitur (Nr.1,4,7):

- Einkommen trennt: hoch → O / niedrig → M

ID3-Entscheidungsbaum



DTL.02
Ausgang: x3(x2(x1(C,A), x1(B,A)), x1(x2(C,B), A))

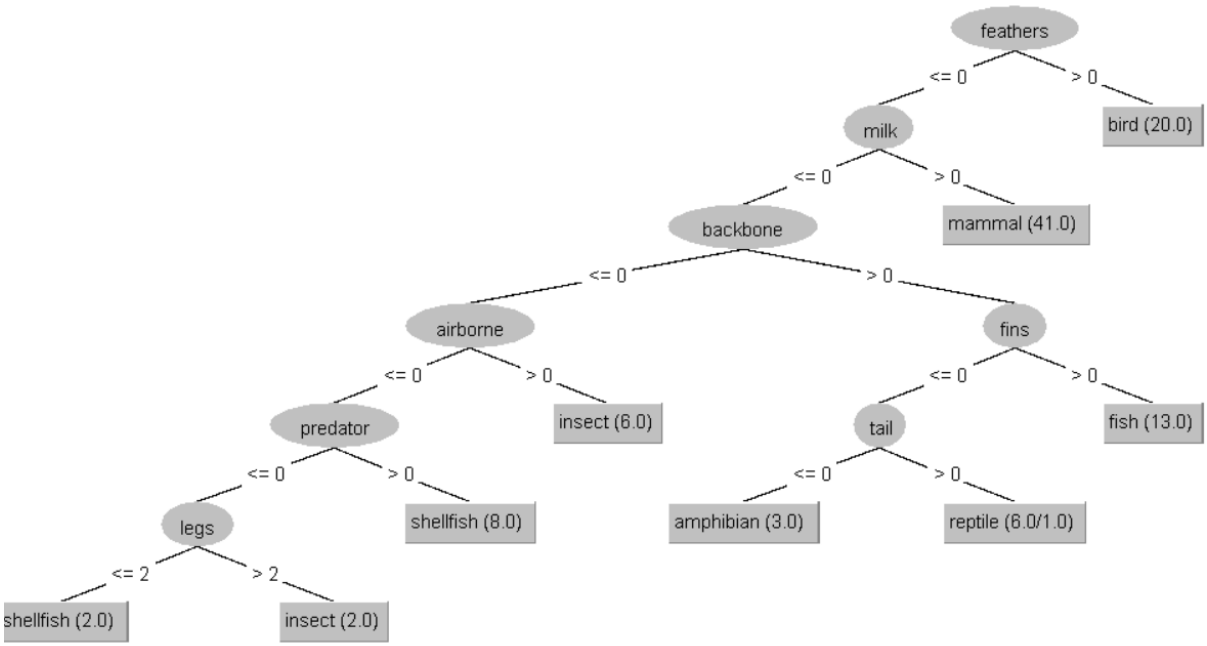
1) Faktorisierung (Distributivität):
x3(x1(x2(C,B), A), x1(x2(C,B), A))

2) Vereinfachung gleichartiger Unterbäume:
x1(x2(C,B), A)

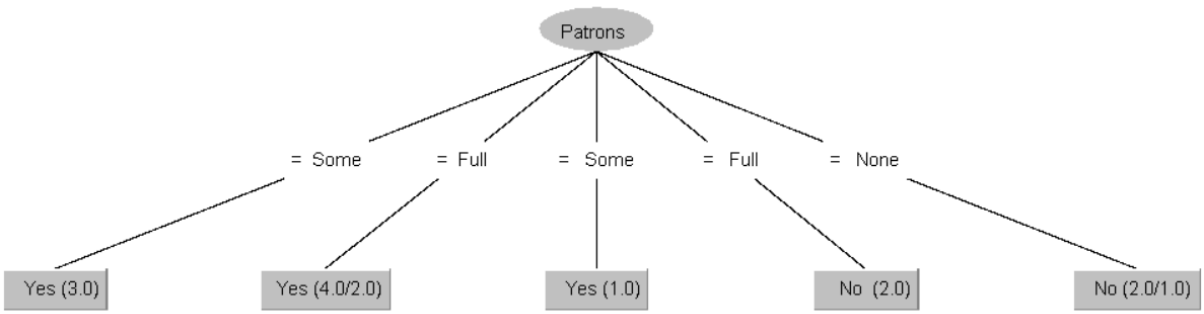
DTL.03

1)

Zoo



Restaurant



Zoo-Datensatz:

- Wichtige Merkmale: Federn, Milch, Rückenmark, Flossen, Schwanz, flugfähig

- Logische Klassifikation: Federn → Vogel, Milch → Säugetier
- Fehlerrate :0.9901 %
- **Confusion Matrix:** fast alle Tiere richtig, nur wenige Verwechslungen bei ähnlichen Klassen (Reptilien ↔ Amphibien, Insekten ↔ Schalentiere)

Restaurant-Datensatz:

- Wichtigstes Merkmal: Patrons (Anzahl Gäste)
    - None → No, Some/Full → Yes
- Fehlerrate :25 %
- **Confusion Matrix:** die meisten Entscheidungen richtig, wenige Yes fälschlicherweise als No

2)

Nominal: Kategorien ohne Reihenfolge

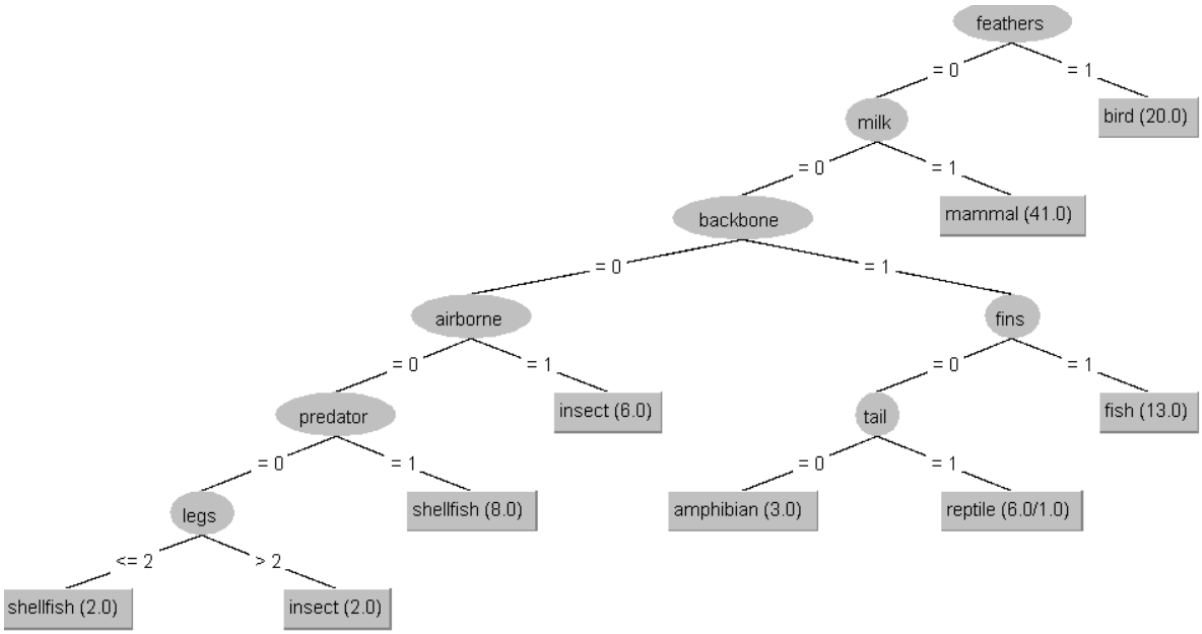
Ordinal/Numeric: Zahlen oder geordnete Kategorien

String: Text/Zeichen, nicht für Berechnung geeignet

3)

Training mit J48

Zoo



Restaurant