

DTL.04

CAL3

$A = \text{Alter}$, $E = \text{Einkommen}$, $B = \text{Bildung}$, $K = \text{Kandidat}$ Künftig die auch x_1, x_2, x_3 heißen

*

1 $|O_1|$

$$S_1 = 4 \quad S_2 = 0.7$$

2 $|O_2|$

3 $|O_2, M_1|$

4 $|O_2, M_2|$

$A(*, /M_1/)$

$$P(O) = \frac{2}{4} = 0.5$$

5 $A(*, /M_1, O_1/)$

$$P(M) = -1$$

6 $A(|O_1|, /M_1, O_1/)$

Bei der Lösung
hat es das
vertauscht.
dass sollte
trotzdem passen.

$$< 35 = 0$$

$$\geq 35 = 1$$

7 $A(|O_1, M_1|, /M_1, O_1/)$

x_{t+1} Teste nächste Attribut das noch nicht
verwendet wurde

1 $A(|O_1, M_1|, /M_1, O_2|)$

$$\text{niedrig} = 0$$

2 $A(|O_2, M_1|, /M_1, O_2|)$

$$\text{hoch} = 1$$

3 $A(|O_2, M_1|, /M_2, O_2|)$

Differenzierungen erzeugen keine Pfade

$A(|O_2, M_1|, E(*, /M_1/))$

4 $A(|O_2, M_1|, E(*, /M_1, /M_1/))$

5 $A(|O_2, M_1|, E(*, /M_1, /M_1, O_1/))$

6 $A(|O_2, M_1|, E(*, /M_1, /M_1, O_2/))$

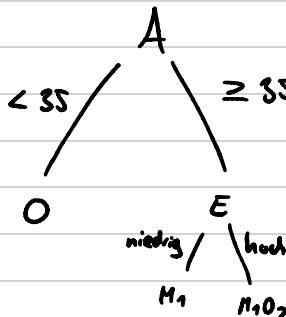
Ergebnis: $A(O, E(*, /M_1, /M_1, O_2/))$

(Nicht ganz fertig!)

Könnte in den Klassen

dann kommen aber viel
kleiner

(Müssen weiter entwickeln, bis
alle erschöpft wurden durch eine
Klasse)



Richtige Lösung anscheinend:

$$(43) 1 \times_1 (x_2(k_3(O, O, m), m), 0)$$

1D3 Klassiert die auch x_1, x_2, x_3 kennen

$A = \text{Alter}, E = \text{Einkommen}, B = \text{Bildung}, K = \text{Kandidat}$

$\{1, 2, 3, 4, 5, 6, 7\} \setminus \{A, E, B\} = \{O\}$

$$\text{Anzahl Klasse } O = 4 \quad P(O) = \frac{4}{7} = 0,57$$

$$\text{Anzahl Klasse } M = 3 \quad P(M) = \frac{3}{7} = 0,43$$

Gesamtzahl: 7

$$H(S) = -\left(\frac{4}{7} \cdot \log_2 \frac{4}{7} + \frac{3}{7} \cdot \log_2 \frac{3}{7}\right) = 0.99 \text{ Bit}$$

(Entropie für die Trainingsmenge / gesamten Datensatz)

| 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 |

$$\text{Gain}(S, A) = H(S) - R(S, A)$$

Alter (Berechnung Gain)

$$\begin{aligned} \geq 35 &= \{1, 3, 4, 5\} = \{2 \times O, 2 \times M\} \quad \frac{4}{7} \cdot \left(-\left(\frac{2}{4} \cdot \log_2 \left(\frac{2}{4}\right) + \frac{2}{4} \cdot \log_2 \left(\frac{2}{4}\right)\right)\right) = 0.371 \\ < 35 &= \{2, 6, 7\} = \{2 \times O, 1 \times M\} \quad \frac{3}{7} \cdot \left(-\left(\frac{2}{3} \cdot \log_2 \left(\frac{2}{3}\right) + \frac{1}{3} \cdot \log_2 \left(\frac{1}{3}\right)\right)\right) = 0.393 \end{aligned} \quad \left. \right\}_+$$

$$R(S, D) = 0.96$$

$$\text{Gain}(S, D) = 0.99 - 0.96 = 0.03 \text{ Bit}$$

Logarithmische Tabelle eventuell
in den Klausuren.

Einkommen

$$\begin{aligned} \text{hoch} &= \{1, 3, 5, 6\} = \{3 \times O, 1 \times M\} \quad \frac{4}{7} \cdot \left(-\left(\frac{3}{4} \cdot \log_2 \left(\frac{3}{4}\right) + \frac{1}{4} \cdot \log_2 \left(\frac{1}{4}\right)\right)\right) = 0.463 \\ \text{niedrig} &= \{2, 4, 7\} = \{1 \times O, 2 \times M\} \quad \frac{3}{7} \cdot \left(-\left(\frac{1}{3} \cdot \log_2 \left(\frac{1}{3}\right) + \frac{2}{3} \cdot \log_2 \left(\frac{2}{3}\right)\right)\right) = 0.393 \end{aligned} \quad \left. \right\}_+$$

$$R(S, E) = 0.856$$

$$\text{Gain}(S, E) = 0.99 - 0.856 = 0.13 \text{ Bit}$$

Bildung

$$\begin{aligned} \text{Abitur} &= \{1, 4, 7\} = \{1 \times O, 2 \times M\} \quad \frac{3}{7} \cdot \left(-\left(\frac{1}{3} \cdot \log_2 \left(\frac{1}{3}\right) + \frac{2}{3} \cdot \log_2 \left(\frac{2}{3}\right)\right)\right) = 0.393 \\ \text{Master} &= \{2, 5\} = \{2 \times O\} \quad \frac{2}{7} \cdot \left(-\left(\frac{2}{2} \cdot \log_2 \left(\frac{2}{2}\right)\right)\right) = 0 \\ \text{Bachelor} &= \{3, 6\} = \{1 \times M, 1 \times O\} \quad \frac{2}{7} \cdot \left(-\left(\frac{1}{2} \cdot \log_2 \left(\frac{1}{2}\right) + \frac{1}{2} \cdot \log_2 \left(\frac{1}{2}\right)\right)\right) = \frac{2}{7} \end{aligned} \quad \left. \right\}_+$$

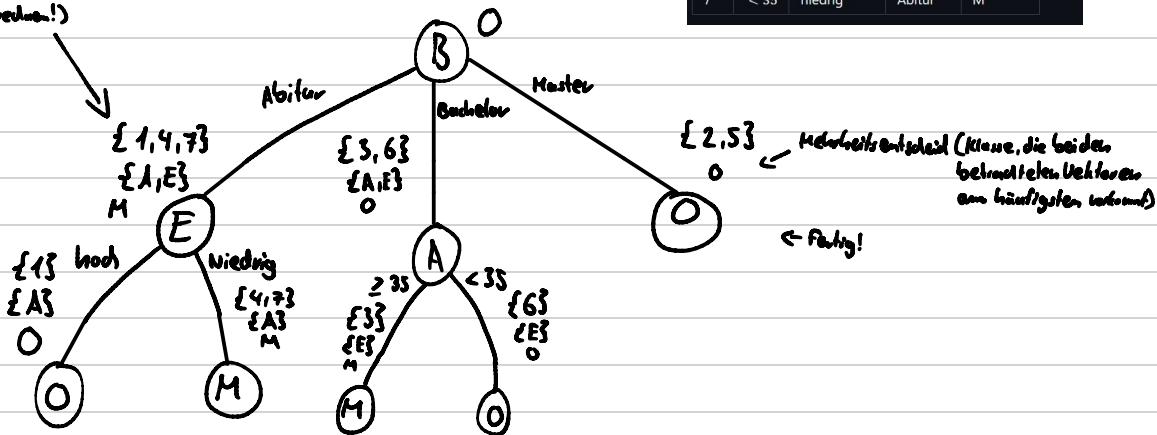
$$R(S, B) = 0.68$$

$$\text{Gain}(S, B) = 0.99 - 0.68 = 0.31 \text{ Bit}$$

$\{1, 2, 3, 4, 5, 6, 7\} \setminus \{A, E, B\} \setminus O$

Müssen Attribut
nicht höheren
Grau wählen
(Mit dem nicht
ausgewählten Attributen
rechnen!)

| 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 |



$$S = \{1, 4, 7\} \text{ Anzahl } O = 1 \text{ Anzahl } M = 2 \text{ Gesamtanzahl } = 3$$

$$H(S') = -\left(\frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right) + \frac{2}{3} \cdot \log_2\left(\frac{2}{3}\right)\right) = 0.92$$

Alter

$$\begin{aligned} \geq 35 &= \{1, 4, 7\} = \{1 \times O, 1 \times M\} & \frac{2}{3} \cdot \left(-\left(\frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right) + \frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right)\right)\right) = \frac{2}{3} \\ < 35 &= \{2, 5\} = \{1 \times M\} & \frac{1}{3} \cdot \left(-\left(1 \cdot \log_2(1)\right)\right) = 0 \end{aligned} \quad \left. \right\} +$$

$$R(S', A) = \frac{2}{3}$$

$$G(S', A) = 0.92 - \frac{2}{3} = 0.25 \text{ Bit}$$

Einkommen

$$\text{hoch} = \{1\} = \{1 \times O\} \quad \frac{1}{3} \cdot \left(-\left(1 \cdot \log_2(1)\right)\right) = 0 \quad \left. \right\} +$$

$$\text{niedrig} = \{4, 7\} = \{2 \times M\} \quad \frac{2}{3} \cdot \left(-\left(\frac{2}{2} \cdot \log_2\left(\frac{2}{2}\right)\right)\right) = 0 \quad \left. \right\}$$

$$R(S', E) = 0$$

$$Gain(S', E) = 0.92 - 0 = 0.92 \text{ Bit}$$

$$\{3, 6\} \text{ Gesamtanzahl } = 2 \quad \text{Anzahl } M=1 \quad \text{Anzahl } O=1$$

$$H(S'') = -\left(\frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right) + \frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right)\right) = 1$$

Alter

$$\geq 35 = \{3\} = \{1 \times M\} \quad \frac{1}{2} \cdot (-1 \cdot \log_2(1)) = 0$$

$$< 35 = \{6\} = \{1 \times O\} \quad \frac{1}{2} \cdot (-1 \cdot \log_2(1)) = 0$$

$$R(S', A) = 0$$

$$G(S', A) = 1 - 0 = 1 \text{ Bit}$$

| 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 |

Einkommen

$$\text{hoch} = \{3, 6\} = \{1 \times M, 1 \times O\} \quad \frac{1}{2} \cdot \left(-\left(\frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right) + \frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right)\right)\right) = 1 \quad \left.\right\} +$$

$$\text{niedrig} = \{3\} \quad = 0$$

$$R(S', E) = 1$$

$$G(S', E) = 1 - 1 = 0 \text{ Bit}$$

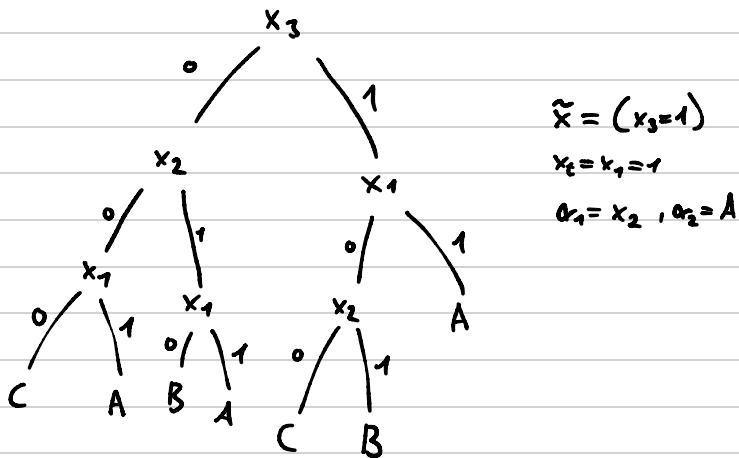
Klausur:

- Aufstellen und sagen welchen R wir nehmen oder welchen Gain
- Gegebenen Knoten
- $\{3, 6\} \{x_1, x_2\} \ O \ ?$ Welches R nehmen wir gewinnt x_1 oder x_2 ?
- Und dann wie ist die Aufteilung der Vektoren
- Sollen da unterschiedlich hältreden (Logiktaubes Tafel?)

DTL.02 Pruning

$$\alpha = x_3(x_2(x_1(C, A), x_1(B, A)), x_1(x_2(C, B), A))$$

\tilde{x} Weg zu Nichtendknoten x_t



$$\tilde{x} = (x_3=1)$$

$$x_t = x_1 = 1$$

$$\alpha_1 = x_2, \alpha_2 = A$$

$$\tilde{x} = (x_3=0, x_2=0)$$

$$x_t = x_1$$

$$\alpha_1 = C, \alpha_2 = A$$

$$\alpha_1 \neq \alpha_2$$

(Ersetzen x_t wenn alle gleich)

Alle * und ein Klassensymbol dann x_t = Klassensymbol (Bedingt redundante Attribute)

$$\alpha = x_3(x_2(x_1(C, A), x_4(B, A)), x_4(x_2(C, B), A)) \quad \begin{matrix} 1. Transformation \\ 2. Reduzierbarkeit \end{matrix}$$

$$x_3(x_1(x_2(C,B), x_2(A,A)), x_1(x_2(C,B), A))$$

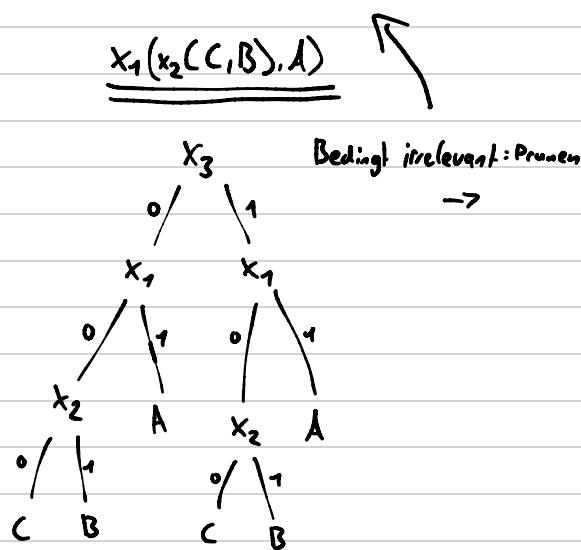
$$x_3(x_1(x_2(c, b), A)), x_1(x_2(c, b), A)) \in$$

(Entscheidungen müssen bei der Transformation zum selben Ergebnis zeigen)

(0,0,0 → C bei beiden Bspw.)

(Wie man zur Wurzel

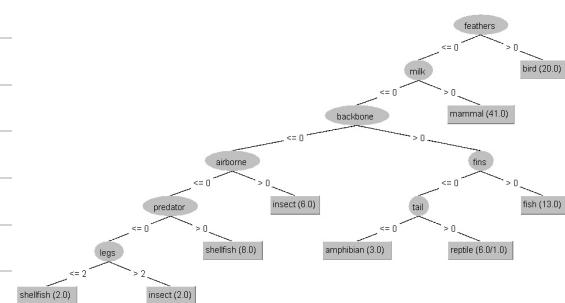
(kannst ist global irrelevant)



DTL.03 (Weka)

1. Training mit JS48

Aussehen Baum: (zoo.csv)



Fehler rate für den Trainingssatz (incorrectly classified instances):

1 (In Prozent: 0.9901%)

Interpretation Confusion Matrix:

(Visualisiert die Performance des Algorithmus)

(Zur Beurteilung der Qualität eines machine learning models)

(Spaltenüberschriften, welche Klasse das Modell wie oft vorhersagt hat)

(Zeilenüberschriften rechts, zeigen die wahre Klasse jeder Zeile)

| ==== Confusion Matrix === | | | | | | |
|---------------------------|----|----|----|---|---|---|
| a | b | c | d | e | f | g |
| 41 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 13 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 20 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 10 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 8 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 3 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 5 |

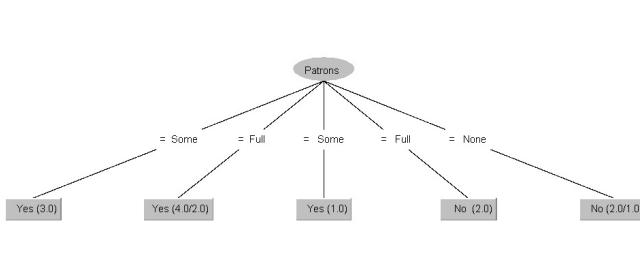
<- classified as
a = mammal
b = fish
c = bird
d = shellfish
e = insect
f = amphibian
g = reptile

Orientierung Diagonale
alle richtig klassifiziert)

0 0 0 0 3 1 (Zeile f)

Der Algorithmus hat 3 Amphibien korrekt klassifiziert und 1 Amphibie falsch klassifiziert.
Wir haben also 1 falsch klassifizierte Instanz.

Aussehen Baum: (restaurant.csv)



Fehlerrate für den Trainingsatz (incorrectly classified instances):

3 (In Prozent: 25%)

Interpretation Confusion Matrix:

(Visualisiert die Performance des Algorithmus)

(Zur Beurteilung der Qualität eines machine learning models)

| ==== Confusion Matrix === | | | | | | | |
|---------------------------|---|---|---|---|---|---|-------------------|
| a | b | c | d | e | f | g | <-- classified as |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | a = Yes |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | b = No |
| 0 | 0 | 2 | 0 | 0 | 0 | 0 | c = Yes |
| 0 | 0 | 0 | 2 | 0 | 0 | 0 | d = No |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | e = No |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | f = No |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | g = No |

(Will the customer wait
for a table?)

Es wurde die Klasse b, f und g falsch klassifiziert. Genauer gesagt, haben wir 3 falsch klassifizierte Instanzen.

2. Unterschied "nominal", "ordinal" und "string"

nominal:

Attribut für Instanzen das mögliche Werte aufliest die eine Instanz annehmen kann. Instanzen müssen genau einen dieser Werte annehmen. Die Liste wird auch "nominal specification" genannt.

ordinal (numeric):

Attribut für ganze oder reelle Zahlen. Ein Wert pro Instanz.

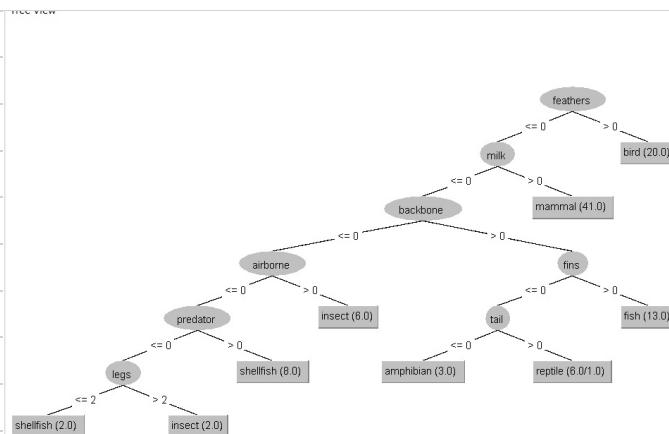
string:

Attribut in dem wir beliebige Textwerte hineinschreiben können für eine Instanz.

3. Training ARFF-Format Datensätze

J48

Aussehen Baum: (zoo.arff)



Fehler rate für den Trainingsatz (incorrectly classified instances):
1 (In Prozent: 0.9901 %)

Interpretation Confusion Matrix:

Dasselbe Ergebnis wie bei der zoo.csv mit J48.

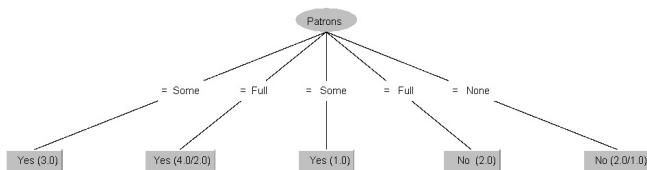
| ==== Confusion Matrix === | | | | | | |
|---------------------------|----|----|----|---|---|---|
| a | b | c | d | e | f | g |
| 41 | 0 | 0 | 0 | 0 | 0 | 1 |
| 0 | 13 | 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 20 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 10 | 0 | 0 | 1 |
| 0 | 0 | 0 | 0 | 8 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 3 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| <-- classified as | | | | | | |
| a = mammal | | | | | | |
| b = fish | | | | | | |
| c = bird | | | | | | |
| d = shellfish | | | | | | |
| e = insect | | | | | | |
| f = amphibian | | | | | | |
| g = reptile | | | | | | |

Fazit:

Kein Unterschied zu den Ergebnissen mit dem J48-Lauf mit .csv Dateien.

J48

Aussehen Baum: (restaurant.arff)



Fehlerrate für den Trainingssatz (incorrectly classified instances):

3 (In Prozent: 23%)

Interpretation Confusion Matrix:

Dasselbe Ergebnis wie bei der restaurant.csv mit J48.

| ==== Confusion Matrix === | | | | | | |
|---------------------------|---|---|---|---|---|---|
| a | b | c | d | e | f | g |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 |

(Will the customer wait for a table?)

Fazit:

Kein Unterschied zu den Ergebnissen mit dem J48-Loof mit .csv Dateien.

ID3 (Knoten mit Strings nicht angeben) (Spalte muss vorgeschrieben werden)

Aussehen Baum: (zoo.arff) (Muss gefüllt werden) (Direktive)

Remove Name Attribut Grund: Einstufigen Baum mit einer Kante beschriftet mit dem Tiersymbol und als Blatt die Klasse.

(Auszchnitt in Textform)

```
legs = '(-inf-0.8)'
| fins = '(-inf-0.1)'
| | toothed = '(-inf-0.1)': shellfish
| | toothed = '(0.1-0.2)': null
| | toothed = '(0.2-0.3)': null
| | toothed = '(0.3-0.4)': null
| | toothed = '(0.4-0.5)': null
| | toothed = '(0.5-0.6)': null
| | toothed = '(0.6-0.7)': null
| | toothed = '(0.7-0.8)': null
| | toothed = '(0.8-0.9)': null
| | toothed = '(0.9-inf)': reptile
fins = '(0.1-0.2)': null
| fins = '(0.2-0.3)': null
| fins = '(0.3-0.4)': null
| fins = '(0.4-0.5)': null
| fins = '(0.5-0.6)': null
| fins = '(0.6-0.7)': null
| fins = '(0.7-0.8)': null
| fins = '(0.8-0.9)': null
| fins = '(0.9-inf)'
```

Kennzeichnung Doppelpunkt für Blatt

Attribut = Ausprägung: Blatt

Fehler rate für den Trainingssatz (incorrectly classified instances):

0 (0%)

Interpretation Confusion Matrix:

Es gibt keine Klassifizierungsfelder mit dem ID3.

| ==== Confusion Matrix === | | | | | | |
|---------------------------|----|----|----|---|---|---|
| a | b | c | d | e | f | g |
| 41 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 13 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 20 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 10 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 8 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 5 |

<-- classified as

a = mammal
b = fish
c = bird
d = shellfish
e = insect
f = amphibian
g = reptile

ID3

Aussuchen Baum: (restaurant.arff)

```
WaitEstimate = 0-10
| Patrons = Some: Yes
| Patrons = Full: null
| Patrons = Some: Yes
| Patrons = Full: null
| Patrons = None: No
WaitEstimate = 30-60: No
WaitEstimate = 10-30: Yes
WaitEstimate = >60: No
WaitEstimate = 0-10: Yes
WaitEstimate = 0-10
| Hungry = Yes: Yes
| Hungry = No: No
| Hungry = Yes: null
| Hungry = No: null
WaitEstimate = 10-30: No
WaitEstimate = 30-60: Yes
```

Fehler rate für den Trainingsatz (incorrectly classified instances):
0 (0%)

Interpretation Confusion Matrix:

Es gibt keine Klassifizierungsfehler mit dem ID3.

==== Confusion Matrix ===

| a | b | c | d | e | f | g | <-- classified as |
|---|---|---|---|---|---|---|-------------------|
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | a = Yes |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | b = No |
| 0 | 0 | 2 | 0 | 0 | 0 | 0 | c = Yes |
| 0 | 0 | 0 | 2 | 0 | 0 | 0 | d = No |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | e = No |
| 0 | 0 | 0 | 0 | 0 | 1 | 0 | f = No |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | g = No |

Vergleich untereinander:

Bei der Anwendung des Entscheidungsbaum Lernalgorithmus ID3 für die Datensätze Zoo und Restaurant, gibt es keine Klassifizierungsfehler und eine Fehlerrate von 0%. Auf der anderen Seite macht der J48 Klassifizierungsfehler, pruned jedoch den Baum. Der Baum wird bei ID3 ohne Pruning wiedergegeben und ist sehr präzise.