DTL.03 – Machine Learning mit Weka

Aufgabe DTL.03 – Machine Learning mit Weka (3 Punkte)

Teil 1: Training mit J48 (1 Punkt)

Ziel: Für die Datensätze `zoo.csv` und `restaurant.csv` wurde jeweils ein Entscheidungsbaum mit dem Algorithmus **J48** (C4.5-Implementierung) trainiert. Die Auswertung erfolgte mit der Option `Use training set` in Weka.

a) Zoo-Datensatz

Modell:

```
feathers <= 0
|   milk <= 0
|   |   backbone <= 0
|   |   |   airborne <= 0
|   |   |   |   predator <= 0
|   |   |   |   |   legs <= 2: shellfish
|   |   |   |   |   legs > 2: insect
|   |   |   |   predator > 0: shellfish
|   |   |   airborne > 0: insect
|   |   backbone > 0
|   |   |   fins <= 0
|   |   |   |   tail <= 0: amphibian
|   |   |   |   tail > 0: reptile
|   |   |   fins > 0: fish
|   milk > 0: mammal
feathers > 0: bird
```

Baumgröße:

- Blätter: 9
- Knoten: 17

Fehlerrate:

- Richtig klassifiziert: 99,0 %
- Falsch klassifiziert: 1,0 %
- Kappa-Statistik: 0,987

Confusion Matrix:

Klasse	Richtig	Falsch	TP-Rate
mammal	41	0	1,000
fish	13	0	1,000
bird	20	0	1,000
shellfish	10	0	1,000
insect	8	0	1,000
amphibian	3	1	0,750
reptile	5	0	1,000

Interpretation: Der Baum trennt klar nach biologischen Merkmalen wie *Gefieder*, *Milch*, *Rückgrat* und *Flossen*. Nur eine Amphibie wurde als Reptil klassifiziert, alle anderen Klassen perfekt. Das Modell erreicht eine Genauigkeit von 99 %.

b) Restaurant-Datensatz

Modell:

```
Pat = Some: Yes
Pat = Full: No (6.0/2.0)
Pat = None: No
```

Baumgröße:

- Blätter: 3
- Knoten: 4

Fehlerrate:

- Richtig klassifiziert: 83,3 %
- Falsch klassifiziert: 16,7 %
- Kappa-Statistik: 0,667

Confusion Matrix:

Klasse	Richtig	Falsch	TP-Rate
Yes	4	2	0,667
No	6	0	1,000

Interpretation: Das Attribut `Pat` (Anzahl der Gäste) ist der wichtigste Prädiktor:

- `Pat = Some` ⇒ **Yes** (warten)
- `Pat = Full` ⇒ **No** (nicht warten)
- `Pat = None` ⇒ **No**

Der Baum ist logisch nachvollziehbar, mit einer Genauigkeit von 83,3 %.

Teil 2: ARFF-Format (1 Punkt)

Erklärung: Das ARFF-Format (*Attribute-Relation File Format*) wird von Weka verwendet, um Datensätze mit Attributdefinitionen und Typen zu beschreiben. Es besteht aus zwei Teilen:

1. **Header:** Definition der Attribute und Typen.
2. **Data:** Die eigentlichen Dateninstanzen.

Attributtypen:

Typ	Beschreibung	Beispiel
nominal	Endliche Auswahl diskreter Werte	{Yes, No}
numeric	Zahlenwerte (ordinal)	0, 1, 2, 3
string	Freitext oder Namen	"Restaurant A"

Beispiel (Restaurant-Datensatz):

```
@relation restaurant
@attribute Alt {Yes,No}
@attribute Bar {Yes,No}
@attribute Fri {Yes,No}
@attribute Sat {Yes,No}
@attribute Hun {Yes,No}
@attribute Pat {None,Some,Full}
@attribute Price { $, $$, $$$ }
@attribute Rain {Yes,No}
@attribute Res {Yes,No}
@attribute Type {French,Thai,Italian,Burger}
@attribute Est {0-10,10-30,30-60,>60}
@attribute WillWait {Yes,No}
@data
Yes,No,No,Yes,Some,$$$,No,Yes,French,0-10,Yes
Yes,No,No,Yes,Full,$,No,No,Thai,30-60,No
```

Teil 3: Vergleich ID3 und J48 (1 Punkt)

ID3:

- Wählt Attribute nach höchstem Informationsgewinn (Entropie).
- Keine Behandlung kontinuierlicher oder fehlender Werte.
- Kein Pruning (führt zu größeren Bäumen).

J48 (C4.5):

- Erweiterung von ID3.
- Unterstützt kontinuierliche Werte.
- Verwendet **Gain Ratio** statt reiner Entropie.
- Führt **Pruning** (Baumvereinfachung) durch.

Vergleich:

Kriterium	ID3	J48
Pruning	Nein	Ja
Kontinuierliche Werte	Nein	Ja
Robustheit	Geringer	Höher
Genauigkeit (Zoo)	$\approx 99\,\%$	$99\,\%$
Genauigkeit (Restaurant)	$\approx 80\,\%$	$83\,\%$

Fazit:

- **Zoo:** J48 liefert nahezu perfekte Ergebnisse (99 %).
- **Restaurant:** Einfacher Baum, 83 % Genauigkeit.
- **ID3** erzeugt meist größere, weniger generalisierte Bäume, während **J48** stabiler und genauer ist.