BW2 Praktikum

Aufgabe 6 – Gruppe 1

Adrian Helberg
Version 1.0

21.12.2018

Aufgabe 9 b)

Informationsgewinn:

Für jedes Attribut A:

$$Gain(S, A) = H(S) - \sum_{v \in A} \frac{|S_v|}{|S|} H(S_v)$$

S	:	Datensatz
$H(S)$	:	Entropie im Datensatz oder Untermenge davon
S_v	:	Untermenge von S für die A den Wert v hat
$\|S\|$	:	Mächtigkeit von S
$\|S_v\|$	:	Mächtigkeit von S_v

Entropie im Datensatz:

$$H(S) = -\sum_{c=1}^{|C|} p_c \log_2 p_c$$

S: Datensatz

$|C|$: Anzahl Kategorien

p_c: Anteil der Instanzen in S, die Kategorie k angehören

Gesamtentropie berechnen (DAMAGE):

$$52\ Instanzen, 20x\ HIGH, 32x\ LOW$$
$$H(S) =$$

$$-\frac{20}{52}\log_2\left(\frac{20}{52}\right) - \frac{32}{52}\log_2\left(\frac{32}{52}\right)$$

= 0.9612347…

Entropie für die Teildatensätze berechnen:

$$A: SEX = M: 24\ Instanzen, 20x\ \text{DAMAGE=HIGH}, 4x\ DAMAGE = LOW:$$
$$H(S_M) =$$

$$-\frac{20}{24}\log_2\left(\frac{20}{24}\right) - \frac{4}{24}\log_2\left(\frac{4}{24}\right)$$

= 0.650022…

$$A: SEX = F: 28\ Instanzen, 0x\ \text{DAMAGE=HIGH}, 28x\ DAMAGE = LOW:$$

$$H(S_M) = 0$$

Gesamtentropie zusammensetzen:

$$Gain(S, SEX) = H(S) - \frac{|S_M|}{|S|}H(S_M) - \frac{|S_F|}{|S|}H(S_F) =$$
$$= 0.961237 - \frac{24}{52} * 0.65 - 0$$
$$= 0.661237$$

Alle fehlenden Attribute berechnen:

$$G(S, SEX) = 0.661237 \sim 66.12\%$$
$$G(S, CARTYPE) = 0.0069 \sim 0.69\%$$
$$G(S, AGE^*) = 0.218813 \sim 21.88\%$$

* AGE wird in <30 und >=30 aufgeteilt

„SEX“ erzielt den größten Informationsgewinn, daher wird es als Baumwurzel gewählt.

Gesamtentropie berechnen (SEX):

$$52\ Instanzen, 24x\ M, 28x\ F$$
$$H(S) =$$

$$-\frac{24}{52}\log_2\left(\frac{24}{52}\right) - \frac{28}{52}\log_2\left(\frac{28}{52}\right)$$

= 0.995727…

Entropie für die Teildatensätze berechnen:

$$A: CARTYPE = COUPE: 28\ Instanzen, 12x\ \text{SEX} = \text{M}, 16x\ SEX = F:$$
$$H(S_{COUPE}) =$$

$$-\frac{12}{28}\log_2\left(\frac{12}{28}\right) - \frac{16}{28}\log_2\left(\frac{16}{28}\right)$$

= 0.985228…

$$A: CARTYPE = VAN: 24\ Instanzen, 12x\ \text{SEX} = \text{M}, 12x\ SEX = F:$$

$$H(S_{VAN}) = 1$$

Gesamtentropie zusammensetzen:

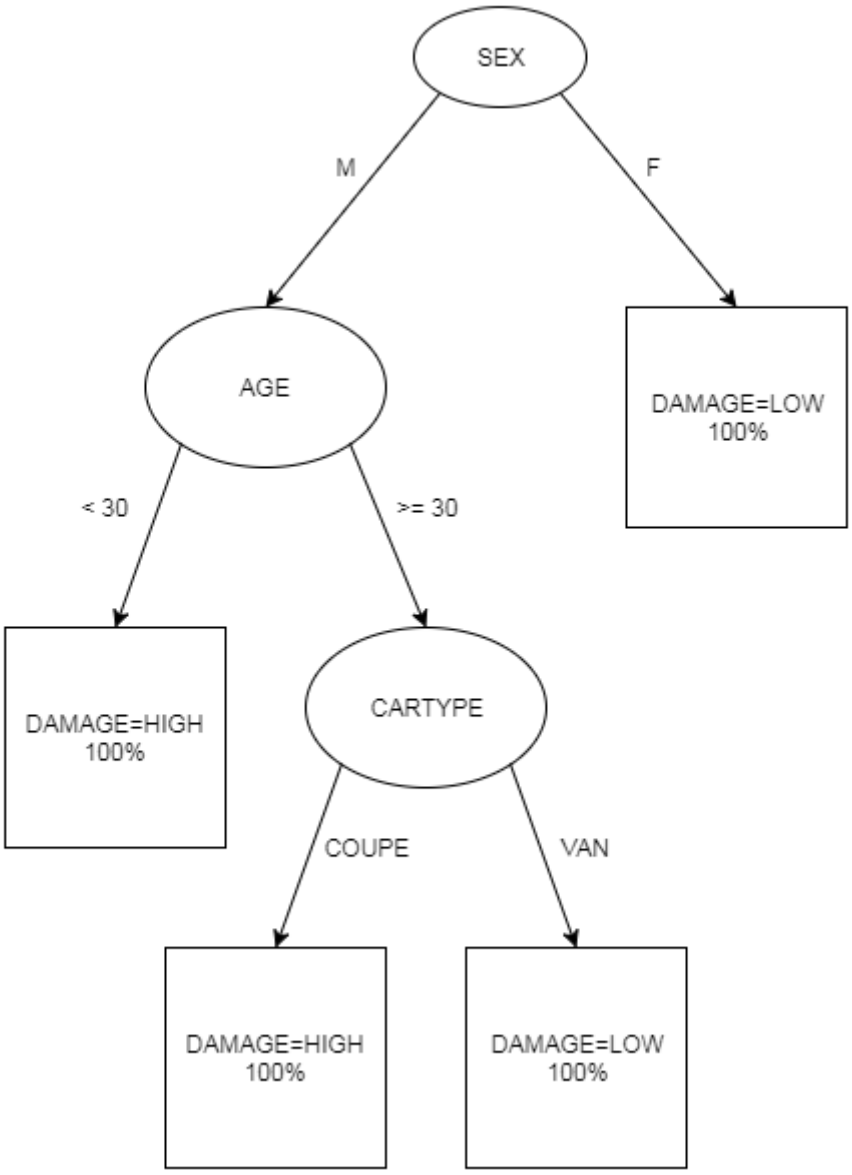
$$Gain(S, CARTYPE) = H(S) - \frac{|S_M|}{|S|}H(S_{COUPE}) - \frac{|S_F|}{|S|}H(S_{VAN}) =$$
$$= 0.995727 - \frac{28}{52} * 0.985228 - \frac{24}{52} * 1$$

$$= 0.003668$$

Alle fehlenden Attribute berechnen:

$$G(S, CARTYPE) = 0.003668 \sim 0.37\%$$
$$G(S, AGE^*) = 0.218812 \sim 21.88\%$$

„AGE“ erzielt den größten Informationsgewinn, daher wird es als weiterer Knoten gewählt.
Alle übrigen Informationen können ohne Berechnungen in den Baum eingefügt werden.



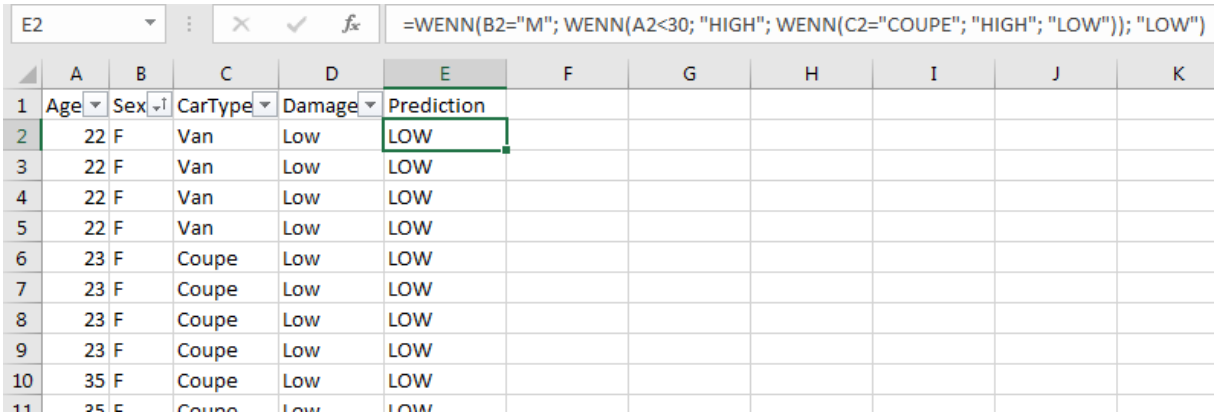
Der Entscheidungsbaum zeigt, dass das Geschlecht in den vorgegebenen Daten sehr aussagekräftig ist im Bezug auf das Attribute „DAMAGE“. Eine Versicherungsgesellschaft kann diese Datenanalyse nutzen, um teurere Versicherungsverträge für gewisse „Risikogruppen“ (Hier: Männer unter 30 Jahre und Männer, die ein Coupe fahren) zu erstellen, um eine Kostendeckung zu erreichen, sollte ein Schaden am Auto auftreten. Billigere Verträge gehen z.B. an die Gruppe „Frauen“.

Abgeleitete Regeln:

IF *SEX = M* ***THEN***
    IF *AGE < 30* ***THEN*** *DAMAGE* ***IS*** *HIGH*
    ELSE
        IF *CARTYPE = COUPE* ***THEN*** *DAMAGE* ***IS*** *HIGH*
        ELSE *DAMAGE* ***IS*** *LOW*
ELSE *DAMAGE* ***IS*** *LOW*

Error Matrix (52 Datensätze):

DAMAGE	HIGH	LOW
HIGH	20	0
LOW	0	32



	A	B	C	D	E	F	G	H	I	J	K
1	Age	Sex	CarType	Damage	Prediction						
2	22	F	Van	Low	LOW						
3	22	F	Van	Low	LOW						
4	22	F	Van	Low	LOW						
5	22	F	Van	Low	LOW						
6	23	F	Coupe	Low	LOW						
7	23	F	Coupe	Low	LOW						
8	23	F	Coupe	Low	LOW						
9	23	F	Coupe	Low	LOW						
10	35	F	Coupe	Low	LOW						

Der Algorithmus hält einen Vorhersagewert von 100%

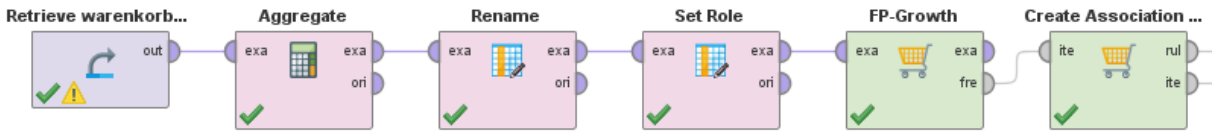
Aufgabe 9 c)

Wichtige Kenngrößen:

Support	Relative Häufigkeit der Beispiele, in denen die Regel anwendbar ist.
Konfidenz	Relative Häufigkeit der Beispiele, in denen die Regel richtig ist.
Lift	Angabe, wie hoch der Konfidenzwert für eine Regel den Erwartungswert übertriift, also die generelle Bedeutung einer Regel.

Beispiel 1:
Minimum Support: 0.2
Minimum Items pro Itemset: 3
Maximum Items pro Itemset: 8
Minimum Konfidenz: 0.8

Prozess:



1. Gruppiere nach Order-Nummer und aggregiere Produkt ID als Konkatenation
2. Setze die Order-nummern als ID
3. FP-Groth und Association Rules Konfiguration siehe oben

FP-Growth (Maximaler Support):

Size	Support ↓	Item 1	Item 2	Item 3
3	0.605	ORHT2000	DXRD1000	ORMN1000

Association Rules:

No.	Premises	Conclusion
77964	ORHT2000, ORHT1000, DXTR1000	DXRD1000, ORMN1000, PRTR2000, DXTR2000

Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
0.361	0.820	0.945	-0.519	0.138	1.617	2.738

AssociationRules

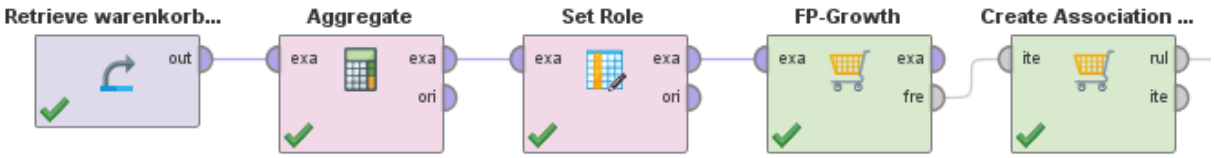
```
Association Rules
[ORHT2000, ORWN1000, DXTR3000] --> [DXTR2000, DXTR1000] (confidence: 0.800)
[ORMN1000, DXTR2000, DXTR1000] --> [DXRD1000, PRTR1000] (confidence: 0.800)
[ORHT1000, PRRD1000, DXRD2000] --> [DXRD1000, ORWN1000] (confidence: 0.800)
[DXRD1000, PRTR1000, DXRD2000] --> [DXTR2000, DXTR1000] (confidence: 0.800)
[DXRD1000, PRTR1000, DXTR1000, PRRD2000] --> [PRRD3000] (confidence: 0.800)
[ORMN1000, ORHT1000, DXTR3000] --> [PRRD1000, PRTR1000] (confidence: 0.800)
[ORMN1000, ORHT1000, DXTR3000] --> [DXTR2000, PRTR3000] (confidence: 0.800)
[ORMN1000, PRTR1000, DXRD2000] --> [DXTR2000, DXTR1000] (confidence: 0.800)
[ORHT1000, DXTR2000, DXTR1000] --> [PRTR2000, PRTR1000] (confidence: 0.800)
[PRRD1000, DXRD2000, DXTR3000] --> [DXTR2000, DXTR1000] (confidence: 0.800)
[DXTR2000, ORWN1000, PRTR1000, PRRD3000] --> [PRRD2000] (confidence: 0.800)
[DXRD1000, ORMN1000, ORHT1000, PRTR3000] --> [ORHT2000, DXRD2000] (confidence: 0.800)
[DXRD1000, ORMN1000, ORHT1000, PRTR3000] --> [ORHT2000, DXTR1000] (confidence: 0.800)
[ORHT2000, DXRD1000, ORMN1000, DXTR1000] --> [ORHT1000, PRTR3000] (confidence: 0.800)
[ORHT2000, DXRD1000, DXTR2000, PRTR1000] --> [ORMN1000, DXTR3000] (confidence: 0.800)
[ORMN1000, ORWN1000, DXTR1000] --> [ORHT2000, DXRD1000, PRTR1000] (confidence: 0.800)
[DXRD1000, PRTR3000, DXTR1000] --> [ORHT2000, ORMN1000, DXTR3000] (confidence: 0.800)
[DXRD1000, PRTR2000, ORHT1000, ORWN1000] --> [ORHT2000, PRTR3000] (confidence: 0.800)
[ORHT2000, PRTR2000, ORHT1000, ORWN1000] --> [DXRD1000, DXTR1000] (confidence: 0.800)
[DXRD1000, PRTR2000, PRRD1000, ORWN1000] --> [ORHT2000, PRTR3000] (confidence: 0.800)
```

Erläuterung

Die Regeln beschreiben das Kaufverhalten der Kunden.
Z.B. wird, wenn die Produktkombination {ORHT2000, ORWN1000, DXTR3000} gekauft wird, mit einem Konfidenzwert von 80% auch die Kombination {DXTR200, DXTR1000} gekauft.

Aus diesen Assoziationen kann Dirt Bikes gezielt Werbung einsetzen, Produktkombinationen zusammenstellen und Über das Kaufverhalten der Kunden lernen, um so ein besserer Anbieter zu werden.

Aufgabe 9 d)



1. Gruppiere nach „Time“ und CUSTOMER_ID und aggregiere die PRODUCT_ID als Konkatenation
2. Setze „CUSTOMER_ID“ als Label und „TIME“ als ID
3. FP-Groth und Association Rules wie folgt:
    - Minimum Support: 0.8
    - Minimum Items pro Itemset: 3
    - Maximum Items pro Itemset: 8
    - Minimum Konfidenz: 0.9

No.	Premises	Conclusion	Support	Confidence
6	PUMP1000, ORHT2000, FAID1000	DXRD1000	0.545	0.912

AssociationRules

```
Association Rules
[PUMP1000, CAGE1000, ORMN1000] --> [ORHT2000] (confidence: 0.906)
[PUMP1000, CAGE1000, DXRD1000] --> [ORHT2000] (confidence: 0.907)
[PUMP1000, CAGE1000, ORHT1000] --> [ORHT2000] (confidence: 0.909)
[PUMP1000, ORHT2000, FAID1000] --> [CAGE1000] (confidence: 0.909)
[PUMP1000, DXRD1000, ORMN1000] --> [ORHT2000] (confidence: 0.909)
[PUMP1000, ORHT2000, FAID1000] --> [DXRD1000] (confidence: 0.91)
[PUMP1000, DXRD1000, PRTR2000] --> [ORHT2000] (confidence: 0.913)
[PUMP1000, DXRD1000, ORHT1000] --> [ORHT2000] (confidence: 0.913)
[PUMP1000, CAGE1000, PRTR2000] --> [ORHT2000] (confidence: 0.915)
[PUMP1000, ORHT2000, DXRD1000] --> [CAGE1000] (confidence: 0.915)
[PUMP1000, DXRD1000, FAID1000] --> [ORHT2000] (confidence: 0.916)
[CAGE1000, DXRD1000, ORMN1000] --> [ORHT2000] (confidence: 0.917)
```

Erläuterung

Mit einem Konfidenzwert von 90,6% wird zu dem Produkt „Enduro 550 (silver)“ die Zubehörteile „LargeAir Pump“, „Water Bottle Cage“ und „Men’s Off Road Bike Fully“ gekauft.
Um den Kaufwunsch des Kunden zu stärken, könnte Dirt Bikes die Kombination als Produkt-Paket verkaufen mit einem geringen Preisnachlass o.ä.