Beslutningstræ

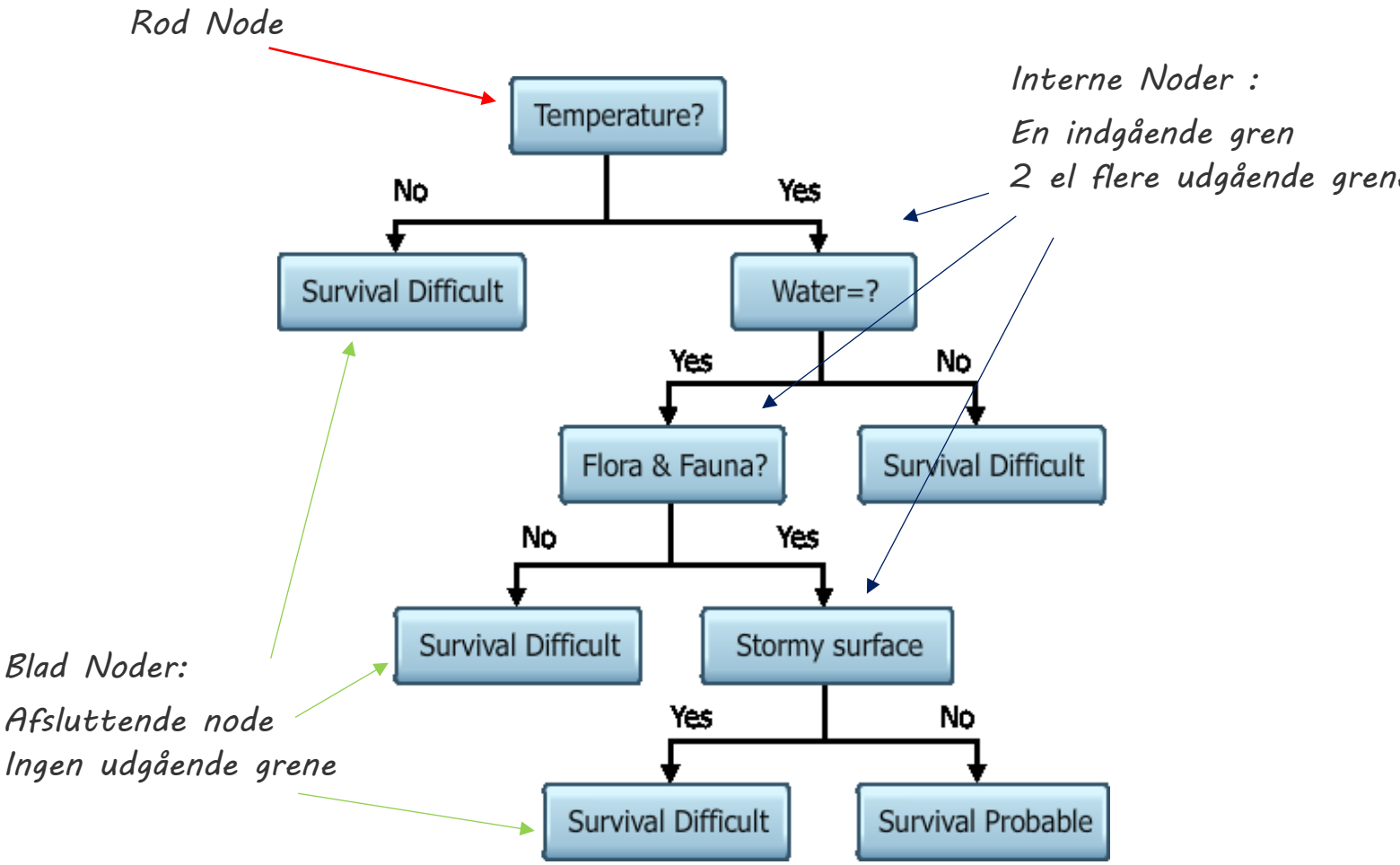
Struktur og metode til beslutningstræ.

Efter denne indledning skal vi se på struktur og metode til at bygge et beslutningstræ.
Lad os starte med at bygge et simpelt eksempel, *overlevelse på en ny planet*, den kategoriske variabel vi derfor ønsker at forudsiger er *overlevelse mulig/overlevelse svær*.

Overlevelsechancerne på en ny planet afgøres i prioriteret rækkefølge af

- Temperaturen
- Vandet
- Flora & Fauna
- Vejrforholdene

Beslutningstræet om vi skal befolke planeten kan derfor se ud som herunder



Entropi, hvilken node skal bruges som rod?

Der findes flere metoder. Som tidligere nævnt kan χ^2- test for uafhængighed anvendes til at beslutte rækkefølgen, der findes også grådige og tilgivende metoder. Her vil vi dog bruge ideen om *informations tillvækst* baseret på *entropi*.

Entropi.: I et beslutningstræ betyder begrebet homogenitet. Hvis data er komplet blandet (50%-50%) vil entropien være 1, mens entropien vil gå mod 0 jo renere datasættet er – se eksempel her til højre.

Informationstilvækst.: Er her defineret som tilvæksten/faldet i entropi (renheden) når en node splittes op efter en variabel.

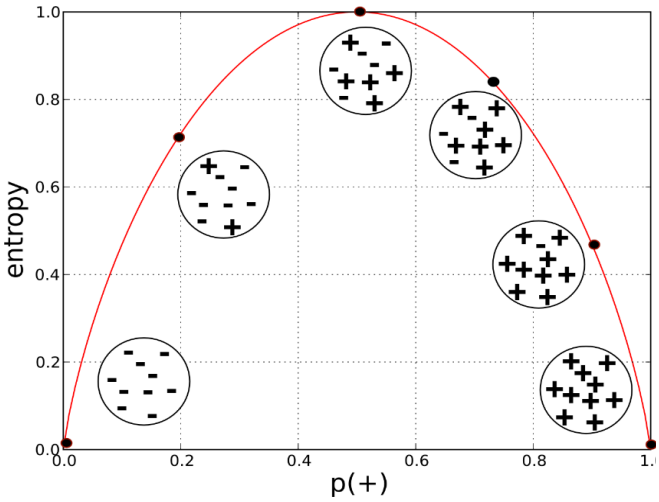


Fig.: F Provost & T Fawcett, *Data Science for Business*, O'Reilly 2013

Metoden er simpel;

1. Ved hvert niveau (node) beregnes informationstilvæksten ved opslit efter de enkelte variable.
2. Noden splittes i nye noder efter de mulige udfald (interne noder)
3. Ovenstående gentages indtil alle muligheder er udtømt og vi kun har bladnoder tilbage.

Det er vigtigt at være opmærksom på muligheden for at overfitte (overtilpasse) modellen så den bliver alt for deltaljeret. Det er derfor vigtigt at holde sig for øje at alle de afsluttende bladnoder skal repræsentere en væsentlig del af datasættet, eksempelvis mere end 5%.

Matematikken bag?

Beregning af entropien i rodnoden;

$$info(\text{Parent}) = \sum_{i=1}^{n} -p_i \cdot \log_2(p_i) = -p_1 \cdot \log_2(p_1) - p_2 \cdot \log_2(p_2) - \ldots - p_n \cdot \log_2(p_n)$$

Beregning af entropien i en intern node;

$$info\,(Parent, Child) = \sum_{i=1}^{l} -p_{C_i} \cdot info(C_i)$$

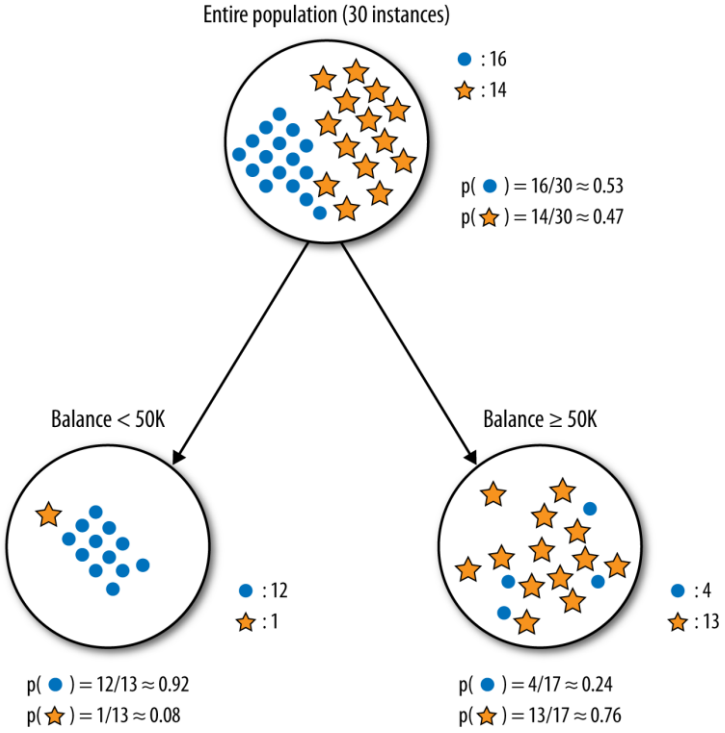
Beregning af informationstilvæksten ved et split;

$$Gain = info(Parent) - info(Parent, Child)$$

Et simpelt beregnet eksempel?

Inden vi går i gang med det fulde datasæte og Excel i et worked example er det godt lige at få overblikket over hvordan informationstilvæks og entropi algoritmen fungerer.

Vi tager udgangspunkt i en stikprøve af størrelse 30 delt i 16 kvinder og 14 mænd. Dette datasæt kan splittes efter enten indestående i banken eller bopælstype. Vi ønsker at forudsige personens køn.



Entropi rodnode.:

$$info(rod) = \frac{16}{30} \cdot \log_2\left(\frac{16}{30}\right) + \frac{14}{30} \cdot \log_2\left(\frac{14}{30}\right)$$
$$= 0{,}997$$

Entropi split.:

$$Info(< 50K) = \frac{12}{13} \cdot \log_2\left(\frac{12}{13}\right) + \frac{1}{13} \cdot \log_2\left(\frac{1}{13}\right)$$
$$= 0{,}391$$
$$Info(\geq 50K) = \frac{4}{17} \cdot \log_2\left(\frac{4}{17}\right) + \frac{13}{17} \cdot \log_2\left(\frac{13}{17}\right)$$
$$= 0{,}787$$
$$Info(split) = \frac{13}{30} \cdot 0{,}391 + \frac{17}{30} \cdot 0{,}787$$
$$= 0{,}615$$

Informationtilvækst 50K split er da

$$Gain = 0{,}997 - 0{,}615 = 0{,}382$$

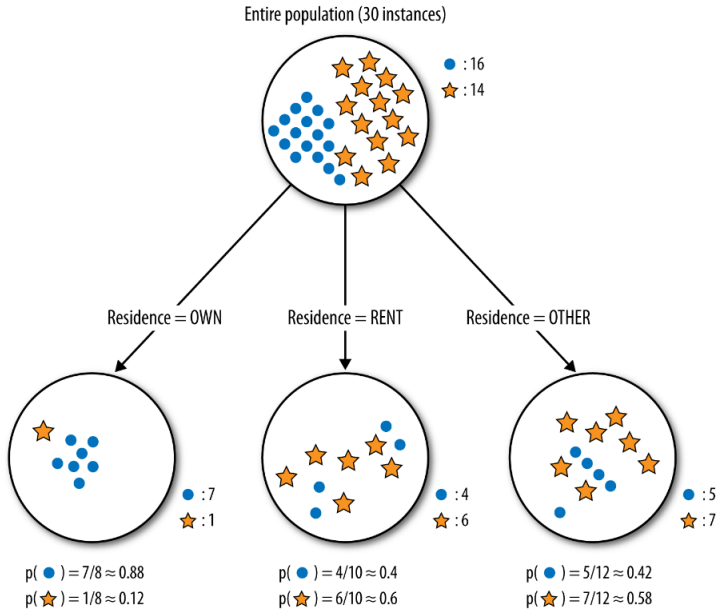


Fig og eksempel.: F Provost & T Fawcett, *Data Science for Business*, O'Reilly 2013

Entropi rodnode, som før.: 0,997

Entropi split.:

$$info(OW) = \frac{7}{8} \cdot \log_2\left(\frac{7}{8}\right) + \frac{1}{8} \cdot \log_2\left(\frac{1}{8}\right)$$
$$= 0{,}570$$
$$info(RE) = \frac{4}{10} \cdot \log_2\left(\frac{4}{10}\right) + \frac{6}{10} \cdot \log_2\left(\frac{6}{10}\right)$$
$$= 0{,}970$$
$$info(OW) = \frac{5}{12} \cdot \log_2\left(\frac{5}{12}\right) + \frac{7}{12} \cdot \log_2\left(\frac{7}{12}\right)$$
$$= 0{,}980$$

$$Info(split) = \frac{8}{30} \cdot 0{,}57 + \frac{10}{30} \cdot 0{,}97 + \frac{12}{30} \cdot 0{,}98$$
$$= 0{,}867$$

Informationtilvækst boligform split er da

$$Gain = 0{,}997 - 0{,}867 = 0{,}130$$

Da informationstilvæksten er størst ved $50K$ splittet vælges denne variabel. De to splits ($> \ og \ \leq$) bliver nu de nye interne noder og processen gentager sig.

Worked Example.:

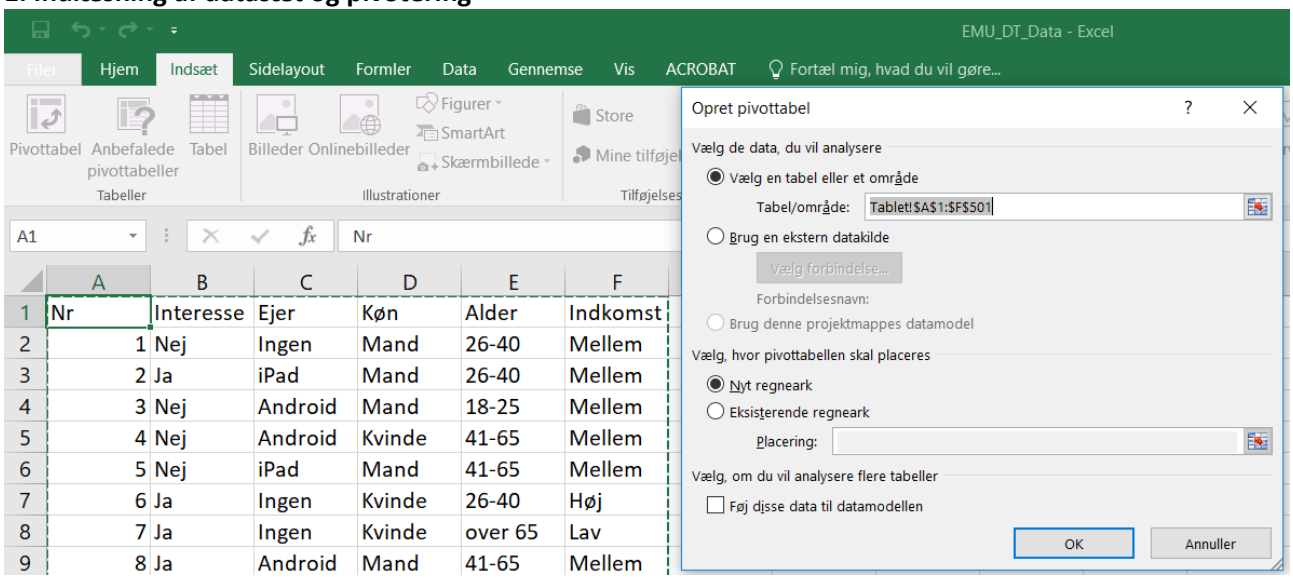
Det lokale pengeinstitut har besluttet at undersøge muligheden for at tilbyde homebanking til deres kunder. Til det formål har de udsendt et spørgeskema med 5 spørgsmål beskrevet her

a) Er du interesseeret i homebanking? [Ja, Nej]
b) Ejer du en tablet? [Android, Ipad, Ingen]
c) Køn? [Kvinde, Mand]
d) Alder? [18-25, 26-40, 41-65, over 65]
e) Indkomst? [Lav, Mellem, Høj]
Ialt 500 kunder har svaret.

Vi skal nu have lavet et beslutningstræ, der hurtigt kan vejlede bankrådgiverne om de skal tilbyde nye kunder muligheden for homebanking.

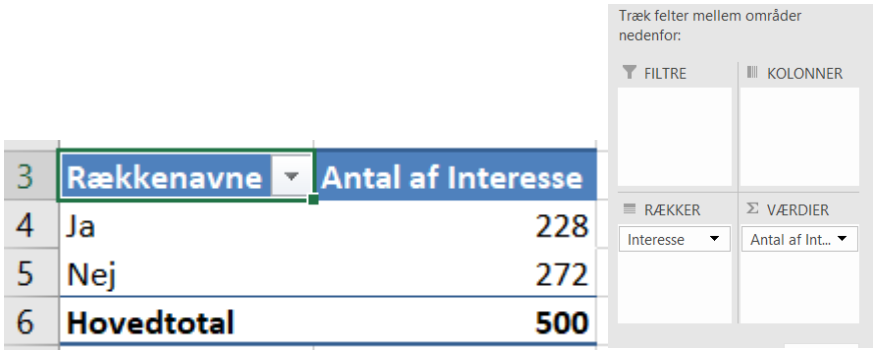
Metodemæssigt indlæser vi først datasættet og strukturerer det ved hjælp af *pivot*-tabeller.
Dernæst undersøger vi datasættets overordnede *interesse* entropi, hvorefter vi ser på informations tilvæksten ved at dele datasættet op efter de enkelte variable *tablet, køn, alder* og *indkomst.*
Denne proces gentages i dybden indtil det ikke længere giver mening at gå dybere ned i strukturen.

1. Indlæsning af datasæt og pivotering



	A	B	C	D	E	F
1	Nr	Interesse	Ejer	Køn	Alder	Indkomst
2	1	Nej	Ingen	Mand	26-40	Mellem
3	2	Ja	iPad	Mand	26-40	Mellem
4	3	Nej	Android	Mand	18-25	Mellem
5	4	Nej	Android	Kvinde	41-65	Mellem
6	5	Nej	iPad	Mand	41-65	Mellem
7	6	Ja	Ingen	Kvinde	26-40	Høj
8	7	Ja	Ingen	Kvinde	over 65	Lav
9	8	Ja	Android	Mand	41-65	Mellem

Vi pivoterer i henhold til interesse og ser en fordeling med 228 interesserede og 272 ikke interesserede bland de 500 forespurgte kunder.
Vi ser dermed et udelte data med en næsten 50/50 fordeling og forventer derfor en entropi på omkring 1.

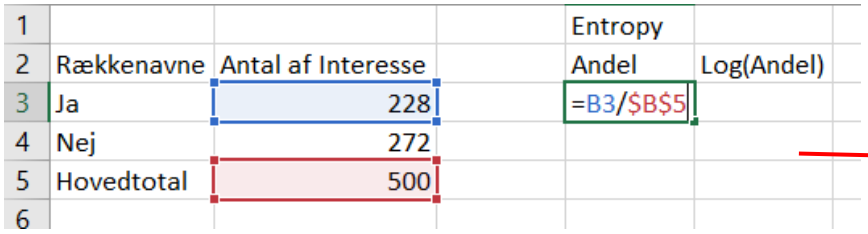


2. Informationsniveau i rod-noden.

$$info(Udelt) = \sum_{i=1}^{n} p_i \cdot \log_2(p_i)$$

Vi kopierer tallene fra Excel *pivot* fanebladet ind i nyt faneblad, som vi kalder [0. Info]

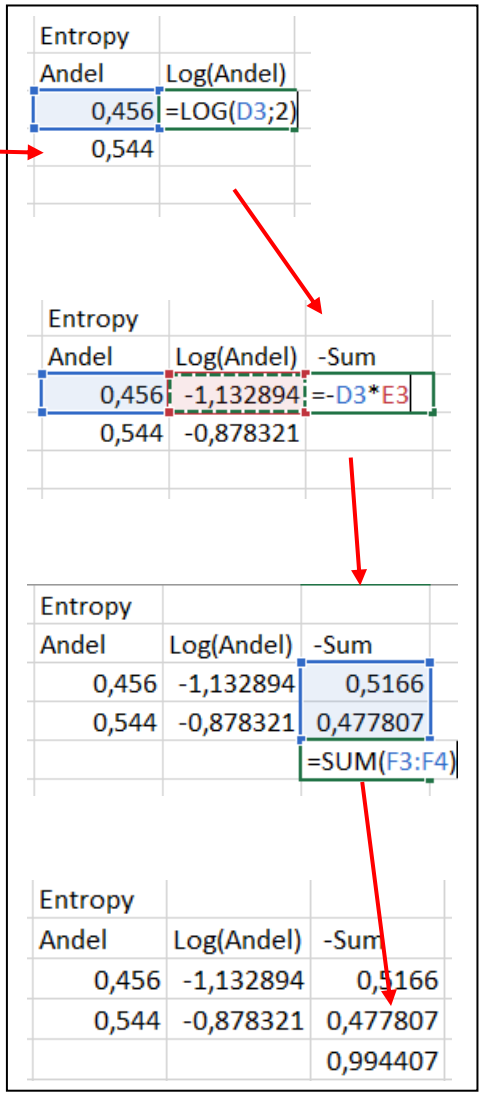
a) Først finder vi andelene (p_i)



b) Så tager vi totals logaritmen af andelene $(\log_2(p_i))$

c) Multiplicerer andelene og logaritmen af samme $(p_i \cdot \log_2(p_i))$

d) Summerer og har dermed vores basisinformations niveau.



Som forventet ser vi den meget høj informationsværdi 0,994 i vores rodnode.

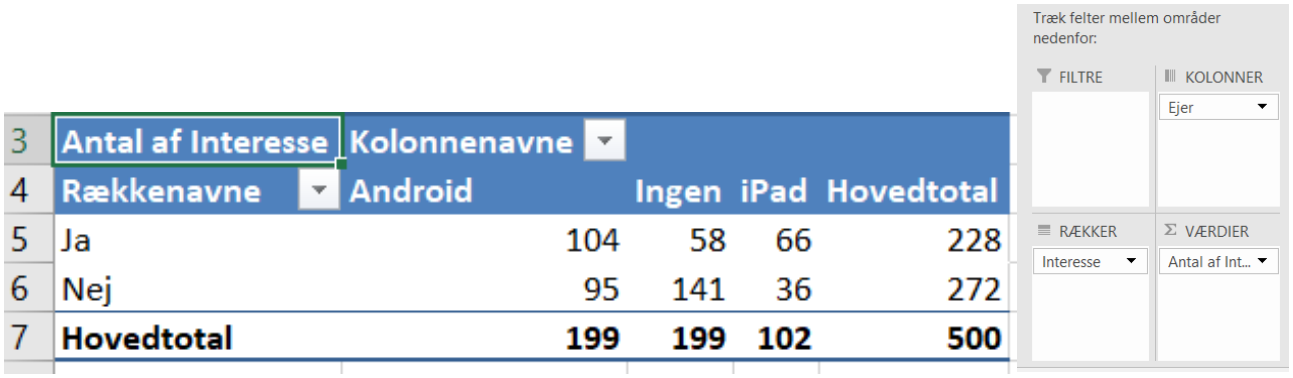
Næste skridt er at tage de fire variable én for én for at undersøge Hvilken, der giver den højeste informationstilvækst.

Rækkefølgen bliver

i) Tabletejer
ii) Køn
iii) Alder
iv) Indkomst

Tilbage i *pivot* fanebladet deler vi informationerne op efter

i) Tabletejer



3	Antal af Interesse	Kolonnenavne			
4	Rækkenavne	Android	Ingen	iPad	Hovedtotal
5	Ja	104	58	66	228
6	Nej	95	141	36	272
7	**Hovedtotal**	**199**	**199**	**102**	**500**

3. Informationsniveau i første Child variabel (Tabletejer)

Vi kopierer tallene fra Excel *pivot* fanebladet ind i nyt faneblad, som vi kalder [1. Tabletejer] og for at få beregnet informationsniveauet for data opdelt i henhold til variablen

	A	B	C	D	E
1	Rækkenavne	Android	Ingen	iPad	
2	Ja	104	58	66	
3	Nej	95	141	36	
4	Hovedtotal	199	199	102	500
5					
6	Andel JA				
7	log(Andel JA)				
8	-Produkt				
9					
10	Andel NEJ				
11	log(Andel NEJ)				
12	Produkt				
13					
14	Entropi				
15	Vægtet Entropi				

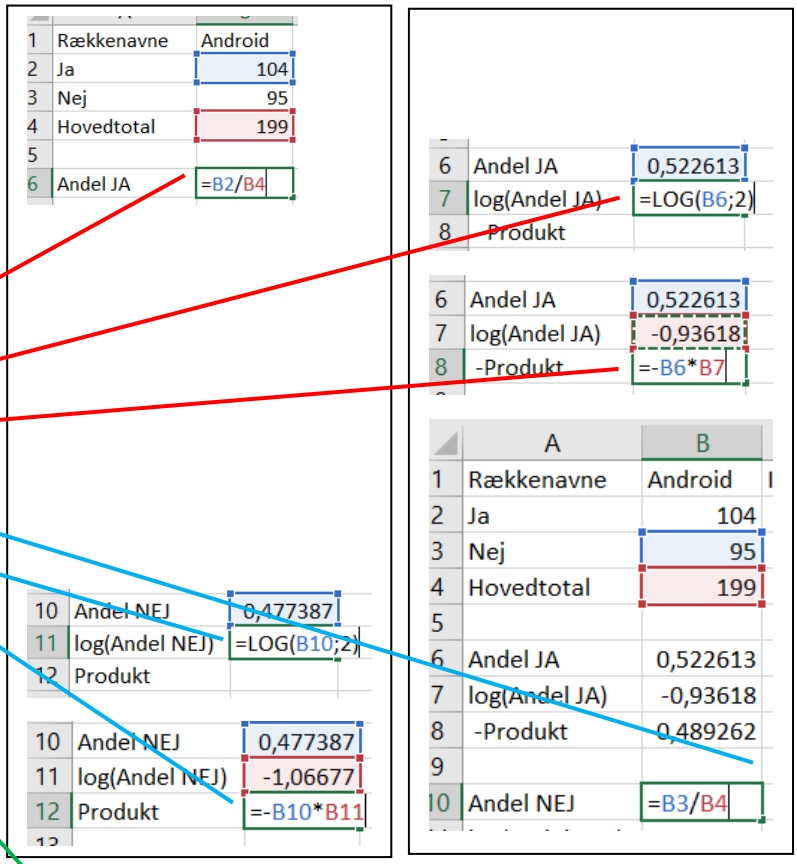
$$info(Var, Child) = \sum_{i=1}^{l} p_{C_i} \cdot info(p_{C_i})$$

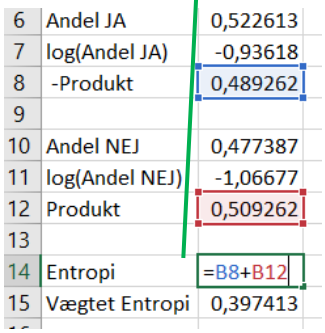
bygger arket op med rækker som vist her til venstre.

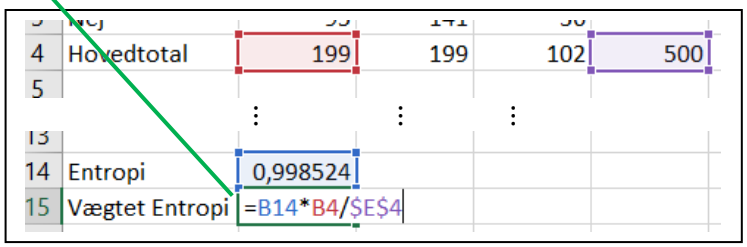
Dermed kan vi finde $info()$ for de enkelte $Child$ i variablen inden vi laver den vægtede sammenlægning.

Lad os se på beregningerne

	A	B
1	Rækkenavne	Android
2	Ja	104
3	Nej	95
4	Hovedtotal	199
5		
6	Andel JA	0,522613
7	log(Andel JA)	-0,93618
8	-Produkt	0,489262
9		
10	Andel NEJ	0,477387
11	log(Andel NEJ)	-1,06677
12	Produkt	0,509262
13		
14	Entropi	0,998524
15	Vægtet Entropi	0,397413







Vi markerer kolonnen og kopierer den ud i nødvendig bredde.

	Rækkenavne	Android
2	Ja	104
3	Nej	95
4	Hovedtotal	199
5		
6	Andel JA	0,522613
7	log(Andel JA)	-0,93618
8	-Produkt	0,489262
9		
10	Andel NEJ	0,477387
11	log(Andel NEJ)	-1,06677
12	Produkt	0,509262
13		
14	Entropi	0,998524
15	Vægtet Entropi	0,397413

	Rækkenavne	Android	Ingen	iPad	
1					
2	Ja	104	58	66	
3	Nej	95	141	36	
4	Hovedtotal	199	199	102	500
5					
6	Andel JA	0,522613	0,291457	0,647059	
7	log(Andel JA)	-0,93618	-1,77864	-0,62803	
8	-Produkt	0,489262	0,518399	0,406373	
9					
10	Andel NEJ	0,477387	0,708543	0,352941	
11	log(Andel NEJ)	-1,06677	-0,49707	-1,5025	
12	Produkt	0,509262	0,352198	0,530294	
13					
14	Entropi	0,998524	0,870596	0,936667	
15	Vægtet Entropi	0,397413	0,346497	0,19108	=SUM(B15:D15)
16					

SUM(**tal1**; [tal2];

Afslutningsvis tages summen af *Vægtet Entropi* kolonnerne og vi er klar til at beregne informationstilvæksten.

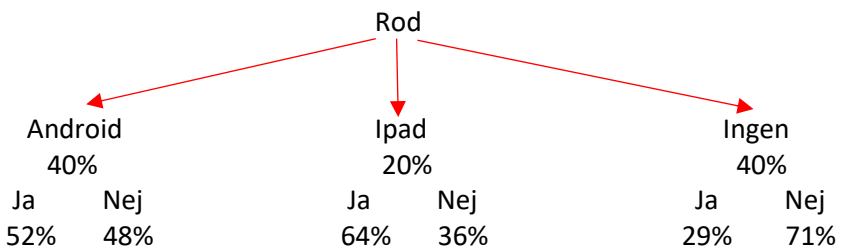
15	Vægtet Entropi	0,397413	0,346497	0,19108	0,93499

$$Tilvækst = Rod\ Information - Child\ Information$$
$$= 0{,}994 - 0{,}935 = 0{,}059$$

Processen gentages for de 3 andre variable, hvormed vi har

Variabel	*Child information*	*Tilvækst*	*Faneblad*
Rod noden	0,994		0. Rodinfo
Tabletejer	0,935	0,059	1. Tabletejer
Køn	0,964	0,031	1. Køn
Alder	0,988	0,006	1. Alder
Indkomst	0,975	0,019	1. Indkomst

Da splittet på *Tabletejer* giver den største informationstilvækst deles rodnoten her i tre interne blad noder.



Og processen gentager for om bladnoderne skal være interne noder.

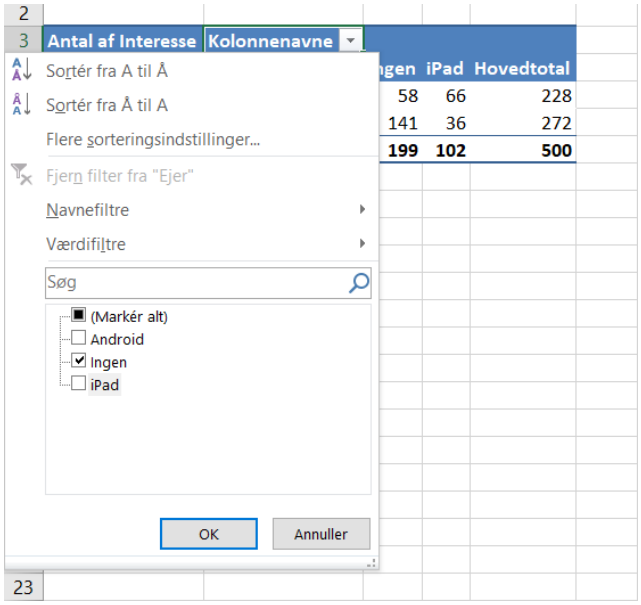
4. Informationsniveau i andet niveau

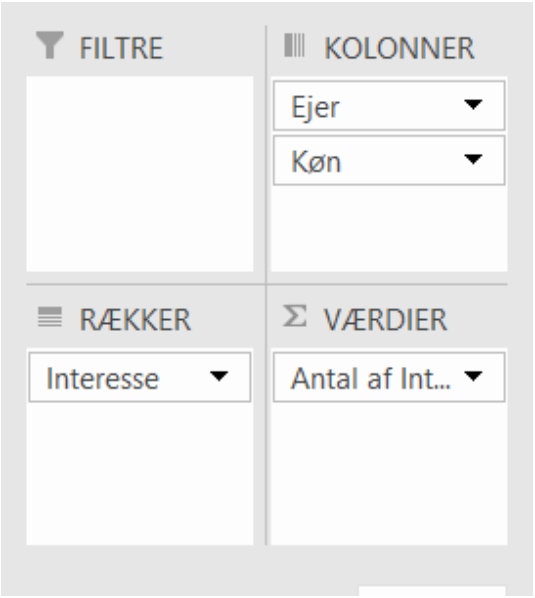
Disse beregninger skal foretages tre gange, det vil sig vi skal gennemgå beregningerne for såvel *Android, Ipad* som *Ingen.*

Først skal vi have data så fanebladet *pivot* besøges igen og vi bruger indstillingerne fra *Tabletejer* igen

Antal af Interesse	Kolonnenavne			
Rækkenavne	Android	Ingen	iPad	Hovedtotal
Ja	104	58	66	228
Nej	95	141	36	272
Hovedtotal	**199**	**199**	**102**	**500**

Da vi begynder med bladnoden *Ingen* skal vi have sorteret de andre to værdier væk og så skal vi søge at splitte op i *Køn, Alder* og *Indkomst* i den rækkefølge.
Altså tryk på pilen ved *Kolonnenavne* og afmarker *Android* og *Ipad*, træk derefter køn ned i valgfeltet kolonner.





Resultatet ses herunder.

Antal af Interesse	Kolonnenavne			
	Ingen		Ingen Total	Hovedtotal
Rækkenavne	Kvinde	Mand		
Ja	45	13	**58**	58
Nej	51	90	**141**	141
Hovedtotal	**96**	**103**	**199**	**199**

Vi splitter nu op i *Ingen* baldnoden så dybt så muligt. Metoden fra før benyttes.
De tre beregningskørsler giver følgende

	A	B	C	D
1	Rækkenavne	Kvinde	Mand	
2	Ja	45	13	
3	Nej	51	90	
4	Hovedtotal	96	103	199
5				
6	Andel JA	0,46875	0,126214	
7	log(Andel JA)	-1,09311	-2,98606	
8	-Produkt	0,512395	0,376881	
9				
10	Andel NEJ	0,53125	0,873786	
11	log(Andel NEJ)	-0,91254	-0,19465	
12	Produkt	0,484785	0,17008	
13				
14	Entropi	0,99718	0,546962	
15	Vægtet Entropi	0,481052	0,283101	0,764153

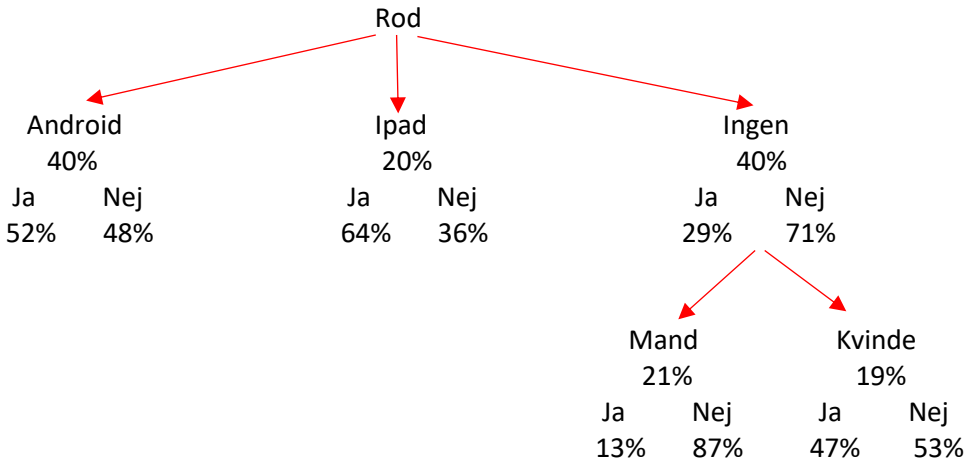
	A	B	C	D	E	F
1	Rækkenavne	18-25	26-40	41-65	over 65	
2	Ja	9	12	12	25	
3	Nej	23	23	32	63	
4	Hovedtotal	32	35	44	88	199
5						
6	Andel JA	0,28125	0,342857	0,272727	0,284091	
7	log(Andel JA)	-1,83007	-1,54432	-1,87447	-1,81558	
8	-Produkt	0,514709	0,529481	0,511219	0,515788	
9						
10	Andel NEJ	0,71875	0,657143	0,727273	0,715909	
11	log(Andel NEJ)	-0,47644	-0,60572	-0,45943	-0,48215	
12	Produkt	0,34244	0,398045	0,334132	0,345177	
13						
14	Entropi	0,857148	0,927527	0,845351	0,860965	
15	Vægtet Entropi	0,137833	0,163133	0,186912	0,380728	0,868606

	A	B	C	D	E
1	Rækkenavne	Høj	Lav	Mellem	
2	Ja	14	37	7	
3	Nej	22	92	27	
4	Hovedtotal	36	129	34	199
5					
6	Andel JA	0,388889	0,286822	0,205882	
7	log(Andel JA)	-1,36257	-1,80177	-2,28011	
8	-Produkt	0,529888	0,516788	0,469434	
9					
10	Andel NEJ	0,611111	0,713178	0,794118	
11	log(Andel NEJ	-0,71049	-0,48767	-0,33258	
12	Produkt	0,43419	0,347792	0,264104	
13					
14	Entropi	0,964079	0,86458	0,733538	
15	Vægtet Entro	0,174406	0,560456	0,125328	0,860191

Fra første split resultat ved vi at *info(Ingen)=0,871* hvorfor

Variabel	*Child inf.*	*Tilvækst*	*Faneblad*
Blad Ingen	0,871		
Køn	0,764	0,107	2a.Køn
Alder	0,869	0,002	2a.Alder
Indkomst	0,860	0,011	2a.Indkomst

Anden split bliver derfor på variablen *Køn*



Vi bemærker her at datasættet er meget rent for mænd, der ikke ejer en tablet. Her er kun 13% interesseret i homebanking mens 87% er ikke interesserede. Vi betragter derfor *Ingen->Mænd* som en *bladnode*. Billedet er dog noget mere mudret for kvindernes vedkommende, hvor informationerne stadig er splittet 50/50.
Kvinder uden tablet udgør 20% af datasættet, hvorfor det kan hjælpe at atomisere yderligere. Det sker på næste side, hvor vi søger at gå i niveau tre og betragter *Ingen->Kvinder* som en *intern node*.

5. Information på tredje niveau

Antal af Interesse	Kolonnenavne							
	Ingen							
	Kvinde					Kvinde Total	Mand	Ingen Total	Hovedtotal
Rækkenavne	18-25	26-40	41-65	over 65				
Ja	7	9	9	20		45	13	**58**	58
Nej	10	5	12	24		51	90	**141**	141
Hovedtotal	**17**	**14**	**21**	**44**		**96**	**103**	**199**	**199**

Antal af Interesse	Kolonnenavne					
	Ingen					
	Kvinde				Kvinde Total	Mand	Ingen Total	Hovedtotal
Rækkenavne	Høj	Lav	Mellem				
Ja	11	28	6		45	13	**58**	58
Nej	9	31	11		51	90	**141**	141
Hovedtotal	**20**	**59**	**17**		**96**	**103**	**199**	**199**

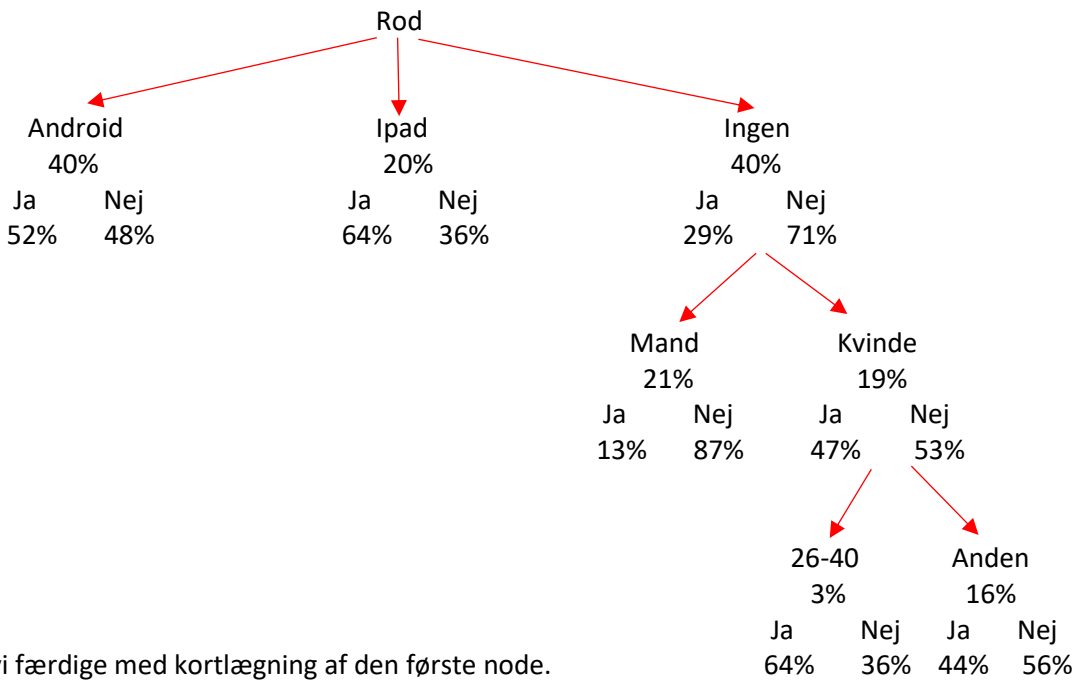
	A	B	C	D
1	Rækkenavne	26-40	Andet	
2	Ja	9	36	
3	Nej	5	46	
4	Hovedtotal	14	82	96
5				
6	Andel JA	0,642857	0,439024	
7	log(Andel JA)	-0,63743	-1,18763	
8	-Produkt	0,409776	0,521397	
9				
10	Andel NEJ	0,357143	0,560976	
11	log(Andel NEJ	-1,48543	-0,83399	
12	Produkt	0,53051	0,467848	
13				
14	Entropi	0,940286	0,989245	
15	Vægtet Entro	0,137125	0,84498	0,982105

	A	B	C	D
1	Rækkenavne	Høj	Andet	
2	Ja	11	34	
3	Nej	9	42	
4	Hovedtotal	20	76	96
5				
6	Andel JA	0,55	0,447368	
7	log(Andel JA)	-0,8625	-1,16046	
8	-Produkt	0,474373	0,519155	
9				
10	Andel NEJ	0,45	0,552632	
11	log(Andel NEJ	-1,152	-0,85561	
12	Produkt	0,518401	0,472837	
13				
14	Entropi	0,992774	0,991992	
15	Vægtet Entro	0,206828	0,785327	0,992155

Vi lægger mærke til at ved opsplitning efter såvel alder som indkomst er der kun én gruppe der stikker ud – ved alder er det gruppen kvinder 26-40 år der har en overvægt til ja og, ved indkomst er det kvinder med høj indkomst der har en overvægt til ja – alle andre afslår tilbuddet. Derfor giver det mening at dele såvel alder som indkomst op i ja/nej grupper.

Fra forrige split resultat ved vi at *info(Kvinder|Ingen)*=0,997 hvorfor tredje split bliver på *Alder*.

Variabel	*Child inf.*	*Tilvækst*	*Faneblad*
Blad Kvinder	0,997		
Alder	0,982	0,015	3a.Alder
Indkomst	0,992	0,005	3a.Indkomst



Dermed er vi færdige med kortlægning af den første node.

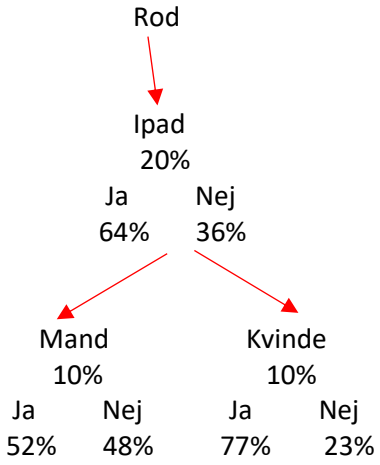
6. De andre splits

Andet split anden node (Ipad)

Fra første split resultat ved vi at *info(Ipad)=0,937* hvorfor

Variabel	*Child inf.*	*Tilvækst*	*Faneblad*
Blad Ipad	0,937		
Køn	0,887	0,050	2b.Køn
Alder	0,918	0,019	2b.Alder
Indkomst	0,932	0,005	2b.Indkomst

Anden split bliver derfor på også variablen *Køn*.
Efter dette split indeholder blad noderne kun 10% af det oprindelige datasæt, hvorfor det ikke giver meget mening at fortsætte – risikoen for overfitting er simpelthen for stor.

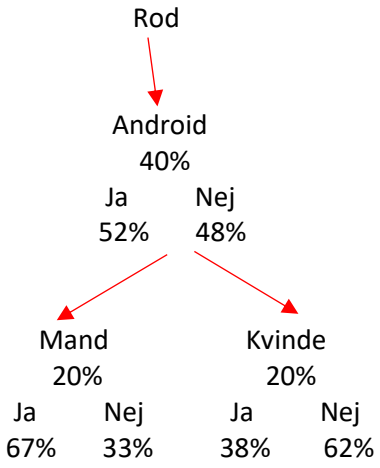


Andet split tredje node (Android)

Fra første split resultat ved vi at *info(Android)=0,999* hvorfor næste alle yderligere spilt kun kan blive renere

Variabel	*Child inf.*	*Tilvækst*	*Faneblad*
Blad Android	0,999		
Køn	0,938	0,061	2c.Køn
Alder	0,991	0,008	2c.Alder
Indkomst	0,998	0,001	2c.Indkomst

Anden split bliver derfor på også variablen *Køn*. Der er stadig 20% tilbage af datasættet i begge bladnoder.



Andet split tredje node (Android)

Mænd

Fra andet split resultat ved vi at *info(Mænd)=0,958*

Variabel	*Child inf.*	*Tilvækst*	*Faneblad*
Blad Mænd	0,958		
Alder	0,932	0,026	3cM.Alder
Indkomst	0,943	0,015	3cM.Indk.

Tredje split efter mand er variablen *Alder*.
Bemærk at grupperne over *41* er lagt sammen da andelen af datasættet ellers bliver for lille og overfitting sandsynlig.

Kvinder

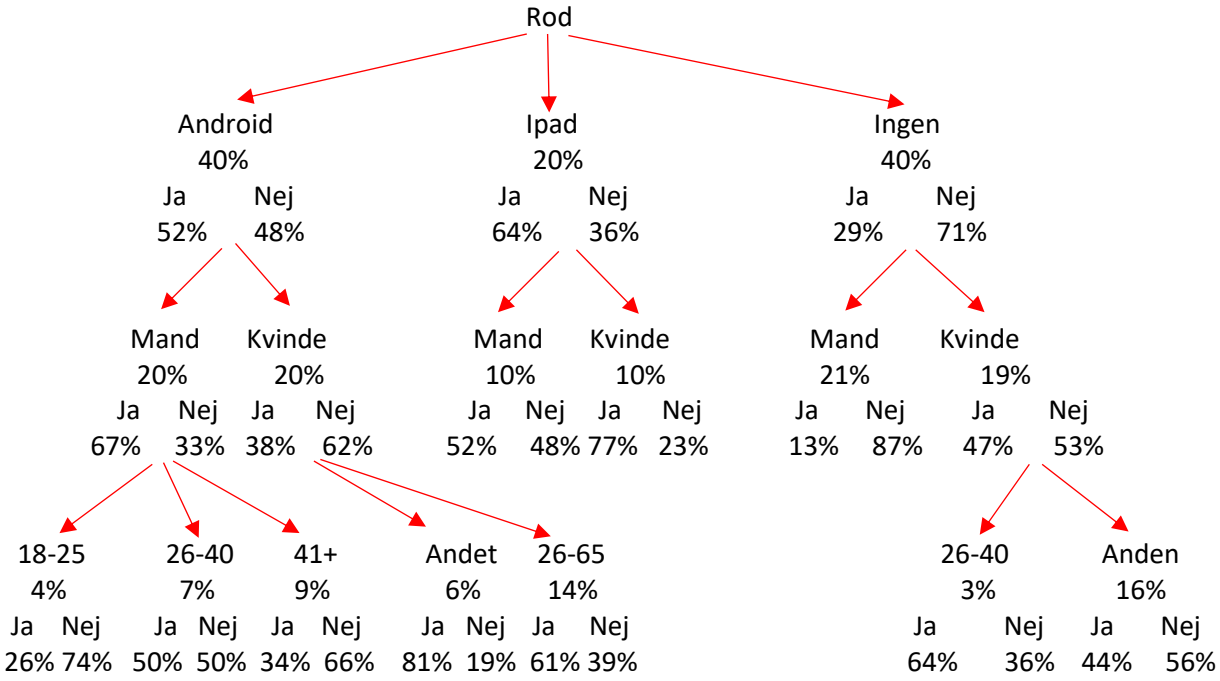
Fra andet split resultat ved vi at *info(Kvinder)=0,918*

Variabel	*Child inf.*	*Tilvækst*	*Faneblad*
Blad Kvinder	0,		
Alder	0,909	0,046	3cK.Alder
Indkomst	0,890	0,055	3cK.Indk.

Tredje split efter kvinde er variablen Alder.
Bemærk at grupperne over 26-65 samt Andet (18-25 og 65+) er lagt sammen således andelene i datasættet ikke bliver for små og overfitting sandsynlig.

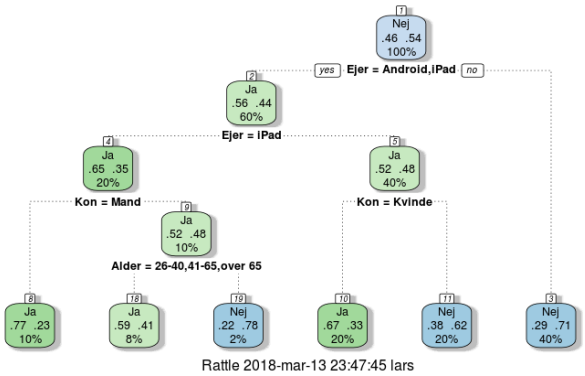
Det færdige beslutningstræ ses på næste side.

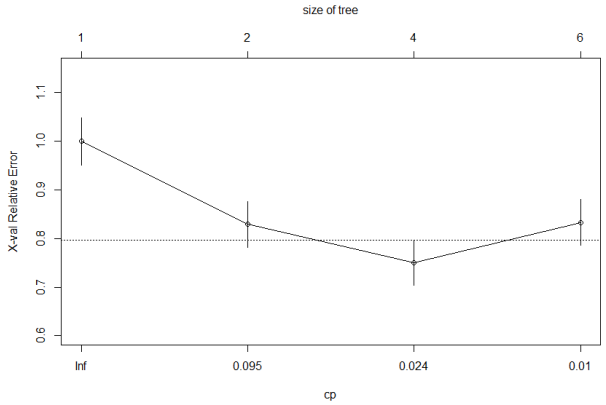
6. Beslutningstræet



7. Det bliver en del lettere med R.

```
setwd("~./EMU")
library(rpart)
library(FSelector)
library(rattle)
df<-as.data.frame(read.table("EMU_DT_Data.CSV",sep=";",header=T))
colnames(df)<-c("Id","Interesse","Ejer","Kon","Alder","Indkomst")
fit <- rpart(Interesse ~ Ejer + Køn + Alder + Indkomst, method="class", data=df)
plocp(fit)
fancyRpartPlot(fit)
```





Og informationstilvæksten fås simpelthen ved:

```
information.gain(Interesse~.,data=df,unit="log2")
```

	attr_importance
Id	0.000000000
Ejer	0.059416612
Kon	0.029945983
Alder	0.006416491
Indkomst	0.019827010

LT 2019.08.01 ■