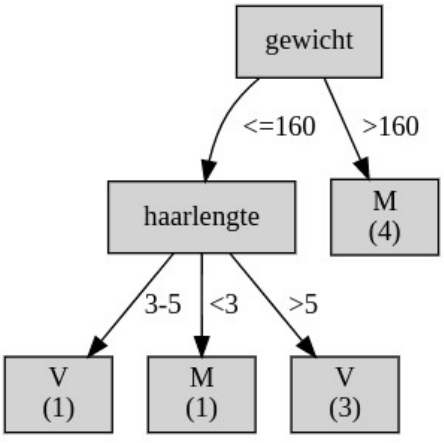
Beslissingsbomen
Oplossingen

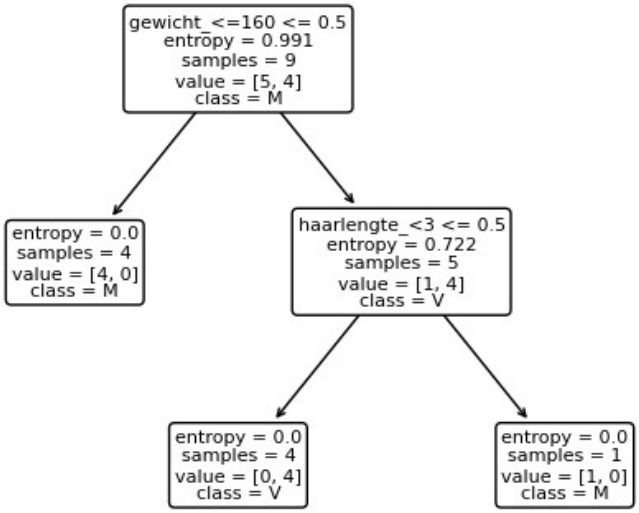
1 The Simpsons

1.
2. -4/9 * log2(4/9) - 5/9 * log2(5/9) = 0,991076
3. resultaten:
   - haarlengte:
     E(tabel) - 3/9*E(haarlengte<3) - 1/9*E(haarlengte tussen 3 en 5) - 5/9*E(haarlengte>5)
     = 0,991076 - 3/9*0 - 1/9*0 - 5/9* 0,9709506
     = 0,4516591
   - gewicht:
     E(tabel) - 5/9*E(gewicht<=160) - 4/9*E(gewicht>160)
     = 0,991076 - 5/9*0,7219281 - 4/9*0
     = 0,5900049
   - leeftijd:
     E(tabel) - 3/9*E(leeftijd<30) - 3/9*E(leeftijd tussen 30 en 40) - 3/9*E(leeftijd>40)
     = 0,991076 - 3/9*0,9182958 - 3/9*0,9182958 - 3/9*0,9182958
     = 0,07278023
   - geslacht:
     E(tabel) - 5/9*E(geslacht=M) - 4/9*E(geslacht=V)
     = 0,991076 - 5/9*0 - 4/9*0
     = 0,991076
   - de entropie van de laatste kolom is het hoogst en gelijk aan de entropie van de tabel. Maar deze kolom komt niet in aanmerking voor het ID3 algoritme omdat we deze willen voorspellen.
4.

	haarlengte	gewicht	leeftijd	geslacht
1	>5	<=160	30-40	V
2	<3	<=160	<30	M
3	>5	<=160	<30	V
4	3-5	<=160	<30	V
6	>5	<=160	>40	V

5. 0,7219281
6. haarlengte: 0,7219281
   gewicht: 0
   leeftijd: 0,1709506
   we kiezen dus haarlengte
7.
```
[gewicht]
 - <=160: [haarlengte]
 - <3: M
 - 3-5: V
 - >5: V
 - >160: M
```
8. 

9.

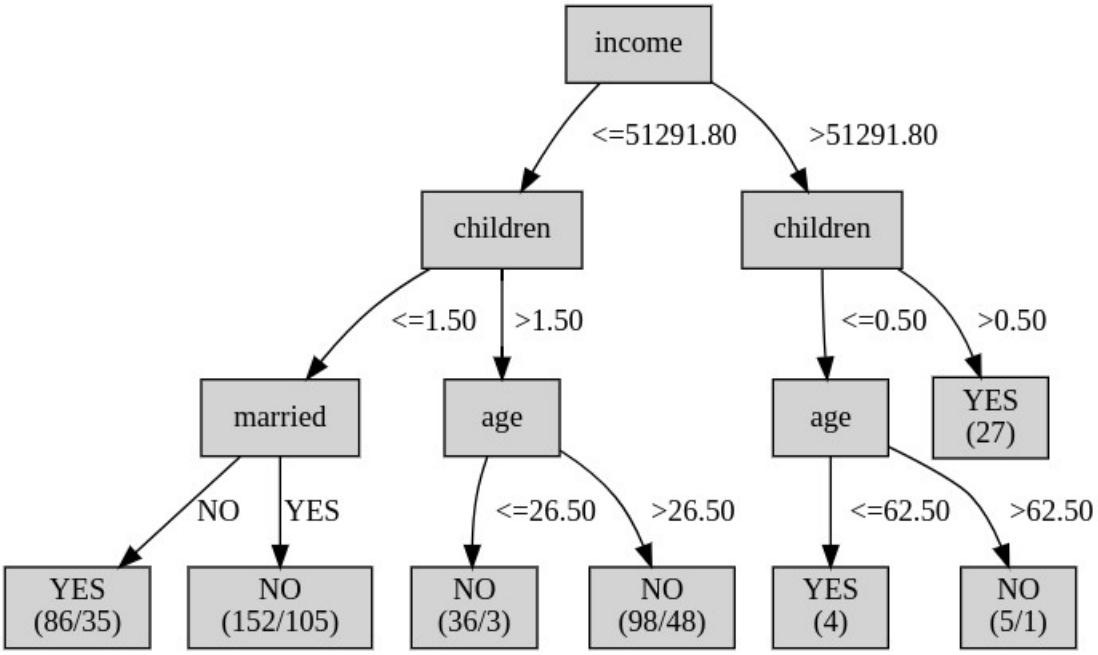


Het resultaat is hetzelfde als de voorgaande boom.

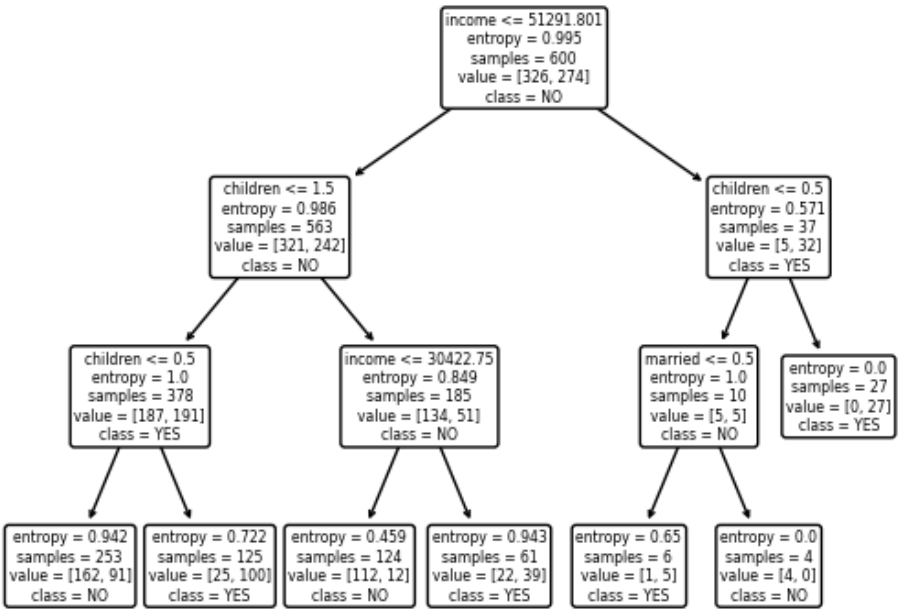
10. het gewicht is groter dan 160, dus M

2 Bank

1.
2.



3. income
4.

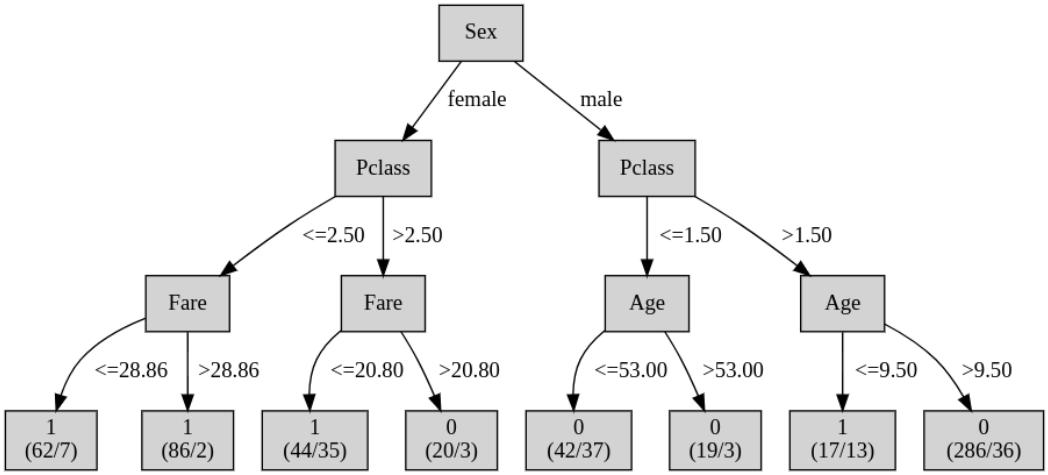


De boom is heel vergelijkbaar met de vorige, maar niet helemaal gelijk.

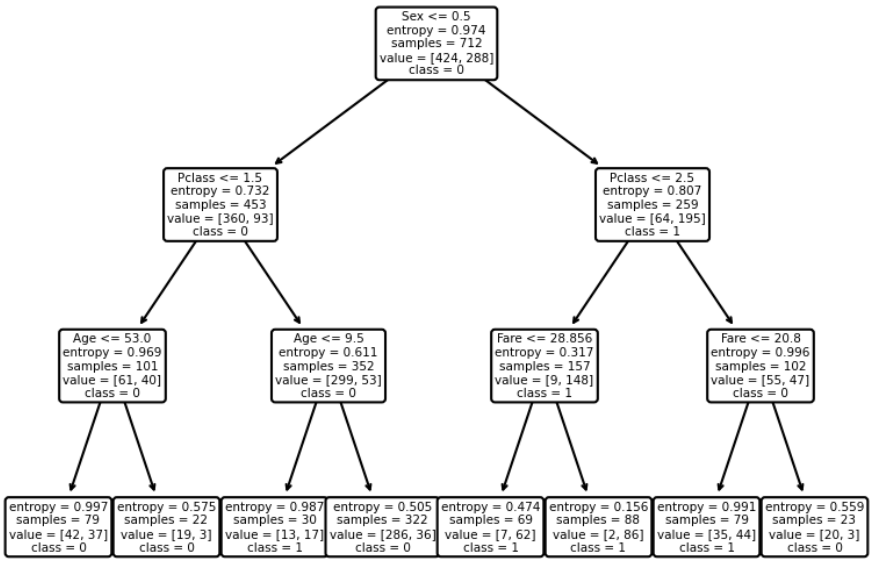
5. De getallen tussen vierkante haken geven de frequenties weer van de 2 categorieën (NO en YES) die nog overbleven in iedere situatie. Bij de Id3Estimator worden deze frequenties enkel bij de bladeren weergegeven en staan ze gesorteerd zodat de hoogste frequentie steeds eerst komt. Als je die getallen deelt door de som van de twee krijg je een soort kans dat het antwoord die waarde is.
6. het inkomen en het aantal kinderen
7. YES (tweede blad-node van links)
8. De frequenties in de node zijn 25 en 100. Er is dus 25/(25+100)=20% kans dat de persoon haar lening niet zal kunnen afbetalen en 100/(25+100)=80% kans dat ze die wel zal kunnen afbetalen

3 Titanic

1. Cabin bevat de meeste onbekende waarden
2.
3. 712
4.



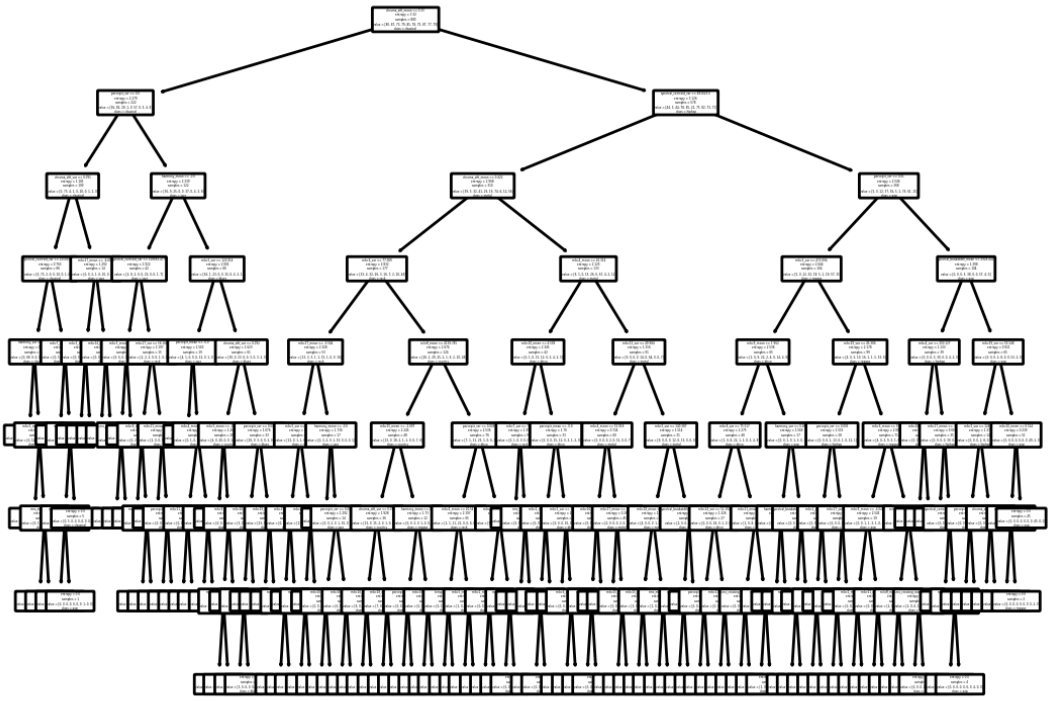
   De kolom ”Sex” heeft de hoogste information gain.
5. Een man van 20 jaar in klasse 2, komt terecht in de uiterst rechtse node. Daar heb je 286/(286+36)=88,8% kans om niet te overleven.
6. Als de man in klasse 1 had gereisd, zouden zijn overlevingskansen (volgens deze boomstructuur) gestegen zijn tot 46,8% (37/(42+37))
7.
8.



   De boomstructuur is helemaal dezelfde.
9. x
10. Beide algoritmes gebruiken dezelfde boomstructuur en de resultaten zijn dan ook helemaal gelijk.
11. De bomen voorspellen dat 134 van de 331 test-passagiers overleefd zouden hebben. Dat is 40,5%

4 Music Genre Classification

1.
2.
3. music_train, music_test = train_test_split(music, test_size=0.2, random_state=42)
4. De boom is nu wat groot om duidelijk te zijn:



5. De boom kan maar 51% van de genres juist voorspellen. Dat is niet heel hoog maar er zijn veel categorieën (10). Als je dus volledig random zou voorspellen, zou je slechts 1 op de tien juist raden. De boom kan dus wel een pak beter dan willekeurig het muziekgenre herkennen. Maar het is niet goed genoeg om met enige zekerheid het genre te bepalen.