3.2 Mišmaš

Kot vemo, pri Mišmašu že od nekdaj pečejo kruh miši. Včasih je bilo enostavno: črne miši so pekle črn kruh, bele belega, sive polbelega in rumene koruznega. Kriza tudi Miševu ni prizanesla. Polbeli in koruzni kruh so začasno prenehali peči, sive in rumene miši pa je Mišmaš skrivnostno prerazporedil na beli in črni kruh. S pomočjo spodnje tabele ugotovi, kako se je odločal. Zgradi odločitveno drevo z uporabo informacijskega prispevka (*angl. Information Gain*).

BARVA	REP	KLOBUK	KRUH
črna	kratek	nima	črn
črna	kratek	nima	črn
črna	kratek	nima	črn
črna	kratek	nima	črn
siva	dolg	ima	črn
siva	dolg	nima	črn
rumena	dolg	nima	črn
rumena	dolg	nima	črn
siva	dolg	ima	črn
siva	dolg	nima	črn
rumena	dolg	ima	bel
rumena	kratek	ima	bel
črna	dolg	ima	bel
siva	kratek	nima	bel
bela	dolg	ima	bel
bela	dolg	nima	bel
bela	dolg	ima	bel
bela	kratek	ima	bel
bela	kratek	ima	bel
bela	kratek	ima	bel

Rešitev:
Entropija razreda je:

$$H(\text{KRUH}) = -p(\text{bel})\log_2(\text{bel}) - p(\text{črn})\log_2(\text{črn}) = 1\text{bit},$$

kar sicer lahko izračunamo tudi na pamet, saj je razred enakomerno porazdeljen.

V koren drevesa bomo postavili tisti atribut, ki najbolj zmanjša nedoločenost H. Za vsak atribut izračunamo zmanjšano entropijo:

$$\text{IG(X)} = H(\text{KRUH}) - H_{\text{res}}(\text{X}).$$

$H_{\text{res}}(X)$ je preostala nedoločenost, če podatke razdelimo glede na atribut X. Izračunamo jo kot uteženo vsoto entropij posameznih vrednosti atributa X:

$H_{\text{res}}(X) = \sum_v p(v)H(v)$, kjer je: $H(v) = -\sum_r p(r|X = v)\log_2 p(r|X = v)$ prek vseh razredov r.
Imamo 5 črnih, 4 rumene, 6 belih in 5 sivih miši, torej:
$H_{\text{res}}(\text{BARVA}) = \frac{5}{20}H(\text{črna}) + \frac{4}{20}H(\text{rumena}) + \frac{6}{20}H(\text{bela}) + \frac{5}{20}H(\text{siva})$
$H(\text{črna}) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.722$
$H(\text{rumena}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$
$H(\text{bela}) = -\frac{6}{6}\log_2\frac{6}{6} - \frac{0}{6}\log_2\frac{0}{6} = 0$
$H(\text{siva}) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.722$
$H_{\text{res}}(\text{BARVA}) = 0.561$ in
$\text{IG}(\text{BARVA}) = H(\text{KRUH}) - H_{\text{res}}(\text{BARVA}) = 1 - 0.561 = 0.439.$

Podobno izračunamo IG(REP) in IG(KLOBUK):
$H_{\text{res}}(\text{REP}) = \frac{11}{20}H(\text{dolg}) + \frac{9}{20}H(\text{kratek})$
$H(\text{dolg}) = -\frac{5}{11}\log_2\frac{5}{11} - \frac{6}{11}\log_2\frac{6}{11} = 0.994$
$H(\text{kratek}) = -\frac{5}{9}\log_2\frac{5}{9} - \frac{4}{9}\log_2\frac{4}{9} = 0.991$
$H_{\text{res}}(\text{REP}) = 0.993$ in
$\text{IG}(\text{REP}) = H(\text{KRUH}) - H_{\text{res}}(\text{REP}) = 1 - 0.993 = 0.007.$

$H_{\text{res}}(\text{KLOBUK}) = \frac{10}{20}H(\text{ima}) + \frac{10}{20}H(\text{nima})$
$H(\text{ima}) = -\frac{8}{10}\log_2\frac{8}{10} - \frac{2}{10}\log_2\frac{2}{10} = 0.722$
$H(\text{nima}) = -\frac{2}{10}\log_2\frac{2}{10} - \frac{8}{10}\log_2\frac{8}{10} = 0.722$
$H_{\text{res}}(\text{KLOBUK}) = 0.722$ in
$\text{IG}(\text{KLOBUK}) = H(\text{KRUH}) - H_{\text{res}}(\text{KLOBUK}) = 1 - 0.722 = 0.278.$

Največji informacijski prispevek (IG) ima BARVA, zato ta atribut izberemo za koren drevesa, ki ga gradimo. Trenutno drevo je globine 1 in ima eno notranje vozlišče (BARVA) in štiri liste, v katerih se nahajajo učni primeri z ustreznimi vrednostmi atributa BARVA. Postopek računanja informacijskih prispevkov in izbire atributa ponovimo v vsakem od listov. Pri tem upoštevamo samo učne primere v vsakem listu. Ker smo barvo že določili, bomo izbirali samo med atributoma REP in KLOBUK. Za vsako barvo si naredimo ustrezno tabelo primerov in izračunajmo informacijska prispevka preostalih atributov:

BARVA = črna:

BARVA	REP	KLOBUK	KRUH
črna	kratek	nima	črn
črna	kratek	nima	črn
črna	kratek	nima	črn
črna	kratek	nima	črn
črna	dolg	ima	bel

Porazdelitev razreda je 4 : 1 (4 črne miši pečejo črn kruh, 1 pa belega). Entropija razreda je:
$H(\text{KRUH}) = -p(\text{bel})\log_2(\text{bel}) - p(\text{črn})\log_2(\text{črn}) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} =$ 0.722bit

Informacijska prispevka atributov REP in KLOBUK za črne miši sta:

$H_{\text{res}}(\text{REP}) = \frac{1}{5}H(\text{dolg}) + \frac{4}{5}H(\text{kratek})$
$H(\text{dolg}) = -\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0$
$H(\text{kratek}) = -\frac{0}{4}\log_2\frac{0}{4} - \frac{4}{4}\log_2\frac{4}{4} = 0$
$H_{\text{res}}(\text{REP}) = 0$ in
$\text{IG}(\text{REP}) = H(\text{KRUH}) - H_{\text{res}}(\text{REP}) = 0.722 - 0 = 0.722.$

Ker je porazdelitev vrednosti pri obeh atributih enaka, je račun za KLOBUK isti: $\text{IG}(\text{KLOBUK}) = H(\text{KRUH}) - H_{\text{res}}(\text{KLOBUK}) = 0.722 - 0 = 0.722$.

Atributa REP in KLOBUK sta torej enakovredna in vseeno je, katerega izberemo za koren tega poddrevesa. To bi pravzaprav lahko ugotovili že brez računanja, ker sta atributa praktično identična.

BARVA = rumena:

BARVA	REP	KLOBUK	KRUH
rumena	dolg	nima	črn
rumena	dolg	nima	črn
rumena	dolg	ima	bel
rumena	kratek	ima	bel

Tudi pri rumenih miših zadostuje že metoda ostrega pogleda: rumene miši s klobukom pečejo bel kruh, tiste brez pa črnega. Glede na rep očitno ne moremo natančno določiti vrste kruha. Opaženo potrdimo z računanjem:

Porazdelitev razreda je 2 : 2 – tega še ne bomo računali, vemo že, da je njegova entropija 1bit.
$H_{\text{res}}(\text{REP}) = \frac{3}{4}H(\text{dolg}) + \frac{1}{4}H(\text{kratek})$

$H(\text{dolg}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918$
$H(\text{kratek}) = -\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0$
$H_{\text{res}}(\text{REP}) = 0.689$ in
$\text{IG}(\text{REP}) = H(\text{KRUH}) - H_{\text{res}}(\text{REP}) = 1 - 0.689 = 0.311.$

$H_{\text{res}}(\text{KLOBUK}) = \frac{2}{4}H(\text{ima}) + \frac{2}{4}H(\text{nima})$
$H(\text{ima}) = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0$
$H(\text{nima}) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} = 0$
$H_{\text{res}}(\text{KLOBUK}) = 0$ in
$\text{IG}(\text{KLOBUK}) = H(\text{KRUH}) - H_{\text{res}}(\text{REP}) = 1 - 0 = 1.$

BARVA = bela:

BARVA	REP	KLOBUK	KRUH
bela	dolg	ima	bel
bela	dolg	nima	bel
bela	dolg	ima	bel
bela	kratek	ima	bel
bela	kratek	ima	bel
bela	kratek	ima	bel

Bele miši očitno pečejo samo bel kruh. Iz tega sklepamo, da je entropija oz. nedoločenost enaka 0. Z drugimi besedami, nesmiselno bi bilo nadalje vejiti to vozlišče drevesa. Naše sklepanje potrdi tudi račun:

$$H(\text{KRUH}) = -p(\text{bel})\log_2(\text{bel}) - p(\text{črn})\log_2(\text{črn}) = -\frac{6}{6}\log_2\frac{6}{6} - \frac{0}{6}\log_2\frac{0}{6} = 0\text{bit}.$$

BARVA = siva:

BARVA	REP	KLOBUK	KRUH
siva	dolg	ima	črn
siva	dolg	nima	črn
siva	dolg	ima	črn
siva	dolg	nima	črn
siva	kratek	nima	bel

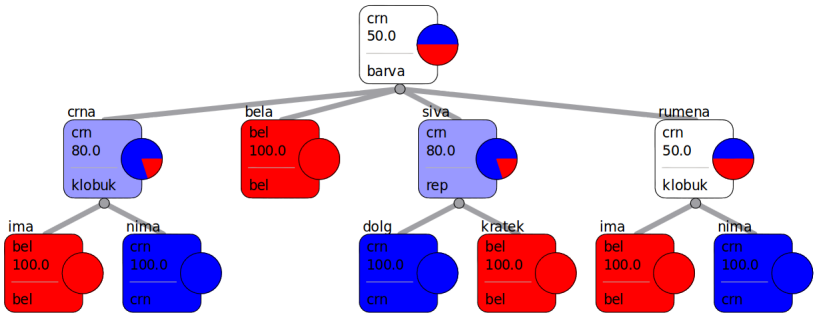
Ostale so še sive miši. Nedoločenost pri njih je $H(\text{KRUH}) = -p(\text{bel})\log_2(\text{bel}) - p(\text{črn})\log_2(\text{črn}) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.722\text{bit}.$

$H_{\text{res}}(\text{REP}) = \frac{4}{5}H(\text{dolg}) + \frac{1}{5}H(\text{kratek})$
$H(\text{dolg}) = -\frac{0}{4}\log_2\frac{0}{4} - \frac{4}{4}\log_2\frac{4}{4} = 0$

$H(\text{kratek}) = -\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0$
$H_{\text{res}}(\text{REP}) = 0$ in
$\text{IG}(\text{REP}) = H(\text{KRUH}) - H_{\text{res}}(\text{REP}) = 0.722 - 0 = 0.722.$

$H_{\text{res}}(\text{KLOBUK}) = \frac{2}{5}H(\text{ima}) + \frac{3}{5}H(\text{nima})$
$H(\text{ima}) = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0$
$H(\text{nima}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918$
$H_{\text{res}}(\text{KLOBUK}) = 0.551$ in
$\text{IG}(\text{KLOBUK}) = H(\text{KRUH}) - H_{\text{res}}(\text{KLOBUK}) = 0.722 - 0.551 = 0.171.$

Atribut REP ima višji informacijski prispevek, zato ga damo v koren poddrevesa. V naslednjem koraku spet razdelimo množico učnih primerov glede na trenutno drevo(slika 3.1). Ugotovimo, da so listi tega drevesa čisti in da nadaljna vejitev drevesa ni smiselna. Zato je drevo na sliki 3.1 že končno drevo. Še enkrat poudarimo, da sta REP in KLOBUK pri črnih miših enakovredna in da je pravilno tudi drevo, ki ima v korenu poddrevesa pri črnih miših atribut REP.



Slika 3.1: Mišmaš info gain in gini

3.3 Mišmaš z razmerjem inf. prispevka

Zgradite odločitveno drevo za 2. nalogo z uporabo razmerja informacijskega prispevka (*angl. Information Gain Ratio*).

Rešitev:

Večino dela za izračun razmerja informacijskega prispevka smo opravili že zgoraj. Izračunati moramo le še entropije atributov, s katerimi bomo delili njihove informacijske prispevke. Entropija atributa X je $H(X) = \sum_v p(X = v) \log_2 p(X = v)$, kjer je v vrednost atributa X.

$H(\text{BARVA}) = -p(\text{BARVA} = \text{črna}) \log_2 p(\text{BARVA} = \text{črna}) - p(\text{BARVA} = \text{rumena}) \log_2 p(\text{BARVA} = \text{rumena}) - p(\text{BARVA} = \text{bela}) \log_2 p(\text{BARVA} = \text{bela}) - p(\text{BARVA} = \text{siva}) \log_2 p(\text{BARVA} = \text{siva}) =$
$= -\frac{5}{20} \log_2 \frac{5}{20} - \frac{4}{20} \log_2 \frac{4}{20} - \frac{6}{20} \log_2 \frac{6}{20} - \frac{5}{20} \log_2 \frac{5}{20} = 1.985\text{bit}$
$H(\text{REP}) = -p(\text{REP} = \text{dolg}) \log_2 p(\text{REP} = \text{dolg}) - p(\text{REP} = \text{kratek}) \log_2 p(\text{REP} = \text{kratek}) = -\frac{11}{20} \log_2 \frac{11}{20} - \frac{9}{20} \log_2 \frac{9}{20} = 0.993\text{bit}$
$H(\text{KLOBUK}) = -p(\text{KLOBUK} = \text{ima}) \log_2 p(\text{KLOBUK} = \text{ima}) - p(\text{KLOBUK} = \text{nima}) \log_2 p(\text{KLOBUK} = \text{nima}) = -\frac{10}{20} \log_2 \frac{10}{20} - \frac{10}{20} \log_2 \frac{10}{20} = 1\text{bit}$
Razmerje informacijskega prispevka atributa X, $IGR(X)$, je

$$IGR(X) = IG(X)/H(X).$$

Na celotni množici podatkov je:
$IGR(\text{BARVA}) = IG(\text{BARVA})/H(\text{BARVA}) = 0.439/1.985 = 0.221$
$IGR(\text{REP}) = IG(\text{REP})/H(\text{REP}) = 0.007/0.993 = 0.007$
$IGR(\text{KLOBUK}) = IG(\text{KLOBUK})/H(\text{KLOBUK}) = 0.278/1 = 0.278$

Glede na IGR je za koren najbolj primeren atribut KLOBUK. V nadaljevanju postopamo kot prej. Začetno množico podatkov razdelimo glede na vrednosti atributa KLOBUK. Tako za KLOBUK=ima kot za KLOBUK=nima moramo izračunati informacijska prispevka atributov BARVA in REP.

KLOBUK = ima:

BARVA	REP	KLOBUK	KRUH
siva	dolg	ima	črn
siva	dolg	ima	črn
rumena	dolg	ima	bel
rumena	kratek	ima	bel
črna	dolg	ima	bel
bela	dolg	ima	bel
bela	dolg	ima	bel
bela	kratek	ima	bel
bela	kratek	ima	bel
bela	kratek	ima	bel

$H(\text{KRUH}) = -p(\text{bel})\log_2(\text{bel}) - p(\text{črn})\log_2(\text{črn}) = -\frac{8}{10}\log_2\frac{8}{10} - \frac{2}{10}\log_2\frac{2}{10} = 0.722\text{bit}$.

$H_{\text{res}}(\text{BARVA}) = \frac{1}{10}H(\text{črna}) + \frac{2}{10}H(\text{rumena}) + \frac{5}{10}H(\text{bela}) + \frac{2}{10}H(\text{siva})$
$H(\text{črna}) = -\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0$
$H(\text{rumena}) = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0$
$H(\text{bela}) = -\frac{5}{5}\log_2\frac{5}{5} - \frac{0}{5}\log_2\frac{0}{5} = 0$
$H(\text{siva}) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} = 0$
$H_{\text{res}}(\text{BARVA}) = 0$ in
$\text{IG}(\text{BARVA}) = H(\text{KRUH}) - H_{\text{res}}(\text{BARVA}) = 0.722 - 0 = 0.722$.

$H(\text{BARVA}) = -\frac{2}{10}\log_2\frac{2}{10} - \frac{2}{10}\log_2\frac{2}{10} - \frac{1}{10}\log_2\frac{1}{10} - \frac{5}{10}\log_2\frac{5}{10} = 1.761\text{bit}$
$IGR(\text{BARVA}) = IG(\text{BARVA})/H(\text{BARVA}) = 0.722/1.761 = 0.410$

$H_{\text{res}}(\text{REP}) = \frac{6}{10}H(\text{dolg}) + \frac{4}{10}H(\text{kratek})$
$H(\text{dolg}) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.918$
$H(\text{kratek}) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$
$H_{\text{res}}(\text{REP}) = 0.551$ in
$\text{IG}(\text{REP}) = H(\text{KRUH}) - H_{\text{res}}(\text{REP}) = 0.722 - 0.551 = 0.171$.

$H(\text{REP}) = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10} = 0.971\text{bit}$
$IGR(\text{REP}) = IG(\text{REP})/H(\text{REP}) = 0.171/0.971 = 0.176$

Izračun pokaže, da je po kriteriju razmerja informacijskega prispevka BARVA boljši atribut, zato ga damo v koren poddrevesa.

KLOBUK=nima:

BARVA	REP	KLOBUK	KRUH
črna	kratek	nima	črn
črna	kratek	nima	črn
črna	kratek	nima	črn
črna	kratek	nima	črn
siva	dolg	nima	črn
rumena	dolg	nima	črn
rumena	dolg	nima	črn
siva	dolg	nima	črn
siva	kratek	nima	bel
bela	dolg	nima	bel

$H(\text{KRUH}) = -p(\text{bel})\log_2(\text{bel}) - p(\text{črn})\log_2(\text{črn}) = -\frac{2}{10}\log_2\frac{2}{10} - \frac{8}{10}\log_2\frac{8}{10} = 0.722\text{bit}$.

$H_{\text{res}}(\text{BARVA}) = \frac{4}{10}H(\text{črna}) + \frac{2}{10}H(\text{rumena}) + \frac{1}{10}H(\text{bela}) + \frac{3}{10}H(\text{siva})$
$H(\text{črna}) = -\frac{0}{4}\log_2\frac{0}{4} - \frac{4}{4}\log_2\frac{4}{4} = 0$
$H(\text{rumena}) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} = 0$
$H(\text{bela}) = -\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0$
$H(\text{siva}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.918$
$H_{\text{res}}(\text{BARVA}) = 0.276$ in
$\text{IG}(\text{BARVA}) = H(\text{KRUH}) - H_{\text{res}}(\text{BARVA}) = 0.722 - 0.276 = 0.446$.

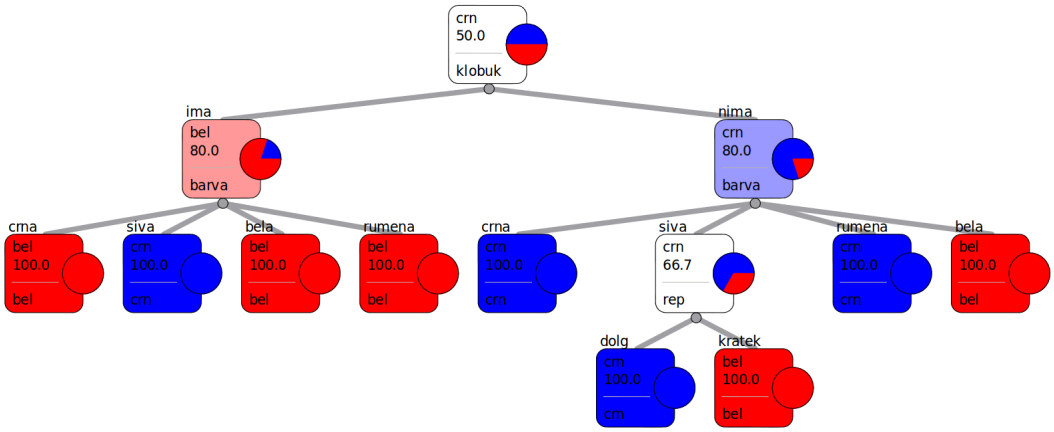
$H(\text{BARVA}) = -\frac{4}{10}\log_2\frac{4}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{2}{10}\log_2\frac{2}{10} - \frac{1}{10}\log_2\frac{1}{10} = 1.846\text{bit}$
$IGR(\text{BARVA}) = IG(\text{BARVA})/H(\text{BARVA}) = 0.446/1.846 = 0.242$

$H_{\text{res}}(\text{REP}) = \frac{5}{10}H(\text{dolg}) + \frac{5}{10}H(\text{kratek})$
$H(\text{dolg}) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.722$
$H(\text{kratek}) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.722$
$H_{\text{res}}(\text{REP}) = 0.722$ in
$\text{IG}(\text{REP}) = H(\text{KRUH}) - H_{\text{res}}(\text{REP}) = 0.722 - 0.722 = 0$.

$H(\text{REP}) = -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10} = 1\text{bit}$
$IGR(\text{REP}) = IG(\text{REP})/H(\text{REP}) = 0/1 = 0$

Atribut BARVA je spet bolje ocenjen od atributa REP, vendar sive miši brez klobuka še vedno pečejo različen kruh, zato gradnja našega drevesa še ni končana. Ker pa smo drevo vejili že po dveh atributih, nam ostane samo še eden, REP. Zato računanje ni potrebno. Vidimo, da siva miš brez klobuka in

s kratkim repom peče bel kruh, oni dve z dolgim repom pa črnega. Končno drevo za naš primer je na sliki 3.2. Prav lahko bi se zgodilo, da bi v listih ne dobili čistih porazdelitev – če zmanjka atributov za nadaljne vejitve, se gradnja drevesa ne glede na to ustavi.



Slika 3.2: Mišmaš gain ratio

Drevesi 3.1 in 3.2 sta različni zato, ker informacijski prispevek in razmerje informacijskega prispevka različno rangirata atribute. Razlog za to je, da informacijski prispevek daje prednost večvrednostnim atributom, torej atributu BARVA pred atributom REP.

3.4 Mišmaš z Gini indeksom

Rešite 2. nalogo še z uporabo Gini indeksa.
Rešitev:
Gini razreda je: $Gini(\text{KRUH}) = 1-p(\text{bel})^2-p(\text{črn})^2 = 1-0.5^2-0.5^2 = 0.5$.

V koren drevesa bomo postavili tisti atribut, ki najbolj zmanjša *Gini* razreda. Za vsak atribut izračunamo njegov *Gini* prispevek:
$Gini(\text{BARVA}) = Gini(\text{KRUH}) - Gini_{\text{res}}(\text{BARVA})$
Če podatke razdelimo glede na atribut X, je $Gini_{\text{res}}(X) = \sum_{v \in X} p(v)Gini(v)$, kjer je: $Gini(v) = 1 - \sum_r p(r|X = v)^2$ prek vseh razredov r.

Imamo 5 črnih, 4 rumene, 6 belih in 5 sivih miši, torej:
$Gini_{\text{res}}(\text{BARVA}) = \frac{5}{20}Gini(\text{črna})+\frac{4}{20}Gini(\text{rumena})+\frac{6}{20}Gini(\text{bela})+\frac{5}{20}Gini(\text{siva})$
$Gini(\text{črna}) = 1 - (\frac{1}{5})^2 - (\frac{4}{5})^2 = 0.320$
$Gini(\text{rumena}) = 1 - (\frac{2}{4})^2 - (\frac{2}{4})^2 = 0.5$
$Gini(\text{bela}) = 1 - (\frac{6}{6})^2 - (\frac{0}{6})^2 = 0$
$Gini(\text{siva}) = 1 - (\frac{1}{5})^2 - (\frac{4}{5})^2 = 0.320$
$Gini_{\text{res}}(\text{BARVA}) = 0.260$ in
$GiniGain(\text{BARVA}) = Gini(\text{KRUH}) - Gini_{\text{res}}(\text{BARVA}) = 0.5 - 0.260 = 0.240$.

$Gini_{\text{res}}(\text{REP}) = \frac{11}{20}Gini(\text{dolg}) + \frac{9}{20}Gini(\text{kratek})$
$Gini(\text{dolg}) = 1 - (\frac{5}{11})^2 - (\frac{6}{11})^2 = 0.496$
$Gini(\text{kratek}) = 1 - (\frac{5}{9})^2 - (\frac{4}{9})^2 = 0.494$
$Gini_{\text{res}}(\text{REP}) = 0.495$ in
$GiniGain(\text{REP}) = Gini(\text{KRUH}) - Gini_{\text{res}}(\text{REP}) = 0.5 - 0.495 = 0.005$.

$Gini_{\text{res}}(\text{KLOBUK}) = \frac{10}{20}Gini(\text{ima}) + \frac{10}{20}Gini(\text{nima})$
$Gini(\text{ima}) = 1 - (\frac{8}{10})^2 - (\frac{2}{10})^2 = 0.320$
$Gini(\text{nima}) = 1 - (\frac{2}{10})^2 - (\frac{8}{10})^2 = 0.320$
$Gini_{\text{res}}(\text{KLOBUK}) = 0.320$ in
$GiniGain(\text{KLOBUK}) = Gini(\text{KRUH}) - Gini_{\text{res}}(\text{KLOBUK}) = 0.5 - 0.320 = 0.180$.

Največji Gini prispevek ima BARVA, zato jo damo v koren drevesa. Kot v prejšnjih dveh nalogah postopek ponovimo na listih trenutnega drevesa.

BARVA = črna:

BARVA	REP	KLOBUK	KRUH
črna	kratek	nima	črn
črna	kratek	nima	črn
črna	kratek	nima	črn
črna	kratek	nima	črn
črna	dolg	ima	bel

$$Gini(\text{KRUH}) = 1 - (p(\text{bel}))^2 - (p(\text{črn}))^2 = 1 - (\frac{1}{5})^2 - (\frac{4}{5})^2 = 0.320$$

$Gini_{\text{res}}(\text{REP}) = \frac{1}{5}Gini(\text{dolg}) + \frac{4}{5}Gini(\text{kratek})$
$Gini(\text{dolg}) = 1 - (\frac{1}{1})^2 - (\frac{0}{1})^2 = 0$
$Gini(\text{kratek}) = 1 - (\frac{0}{4})^2 - (\frac{4}{4})^2 = 0$
$Gini_{\text{res}}(\text{REP}) = 0$ in
$GiniGain(\text{REP}) = Gini(\text{KRUH}) - Gini_{\text{res}}(\text{REP}) = 0.320 - 0 = 0.320.$

Ker je porazdelitev vrednosti pri obeh atributih enaka, je račun za KLOBUK isti: $GiniGain(\text{KLOBUK}) = Gini(\text{KRUH}) - Gini_{\text{res}}(\text{KLOBUK}) = 0.320 - 0 = 0.320$.

Za atributa REP in KLOBUK velja enako kot prej – vseeno je, katerega izberemo za koren tega poddrevesa.

BARVA = rumena:

BARVA	REP	KLOBUK	KRUH
rumena	dolg	nima	črn
rumena	dolg	nima	črn
rumena	dolg	ima	bel
rumena	kratek	ima	bel

$$Gini(\text{KRUH}) = 1 - (p(\text{bel}))^2 - (p(\text{črn}))^2 = 1 - (\frac{2}{4})^2 - (\frac{2}{4})^2 = 0.5$$

$Gini_{\text{res}}(\text{REP}) = \frac{3}{4}Gini(\text{dolg}) + \frac{1}{4}Gini(\text{kratek})$
$Gini(\text{dolg}) = 1 - (\frac{1}{3})^2 - (\frac{2}{3})^2 = 0.444$
$Gini(\text{kratek}) = 1 - (\frac{1}{1})^2 - (\frac{0}{1})^2 = 0$
$Gini_{\text{res}}(\text{REP}) = 0.333$ in
$GiniGain(\text{REP}) = Gini(\text{KRUH}) - Gini_{\text{res}}(\text{REP}) = 0.5 - 0.333 = 0.167.$

$Gini_{\text{res}}(\text{KLOBUK}) = \frac{2}{4}Gini(\text{ima}) + \frac{2}{4}Gini(\text{nima})$
$Gini(\text{ima}) = 1 - (\frac{2}{2})^2 - (\frac{0}{2})^2 = 0$
$Gini(\text{nima}) = 1 - (\frac{0}{2})^2 - (\frac{2}{2})^2 = 0$
$Gini_{\text{res}}(\text{KLOBUK}) = 0$ in
$GiniGain(\text{KLOBUK}) = Gini(\text{KRUH}) - Gini_{\text{res}}(\text{KLOBUK}) = 0.5 - 0 = 0.5.$

BARVA = bela:

BARVA	REP	KLOBUK	KRUH
bela	dolg	ima	bel
bela	dolg	nima	bel
bela	dolg	ima	bel
bela	kratek	ima	bel
bela	kratek	ima	bel
bela	kratek	ima	bel

Sklepamo tako kot v nalogi 3.2.

BARVA = siva:

BARVA	REP	KLOBUK	KRUH
siva	dolg	ima	črn
siva	dolg	nima	črn
siva	dolg	ima	črn
siva	dolg	nima	črn
siva	kratek	nima	bel

$$Gini(\text{KRUH}) = 1 - p(\text{bel})^2 - p(\text{črn})^2 = 1 - (\frac{1}{5})^2 - (\frac{4}{5})^2 = 0.320$$

$Gini_{\text{res}}(\text{REP}) = \frac{4}{5}Gini(\text{dolg}) + \frac{1}{5}Gini(\text{kratek})$
$Gini(\text{dolg}) = 1 - (\frac{0}{4})^2 - (\frac{4}{4})^2 = 0$
$Gini(\text{kratek}) = 1 - (\frac{1}{1})^2 - (\frac{0}{1})^2 = 0$
$Gini_{\text{res}}(\text{REP}) = 0$ in
$GiniGain(\text{REP}) = Gini(\text{KRUH}) - Gini_{\text{res}}(\text{REP}) = 0.320 - 0 = 0.320.$

$Gini_{\text{res}}(\text{KLOBUK}) = \frac{2}{5}Gini(\text{ima}) + \frac{3}{5}Gini(\text{nima})$
$Gini(\text{ima}) = 1 - (\frac{2}{2})^2 - (\frac{0}{2})^2 = 0$
$Gini(\text{nima}) = 1 - (\frac{1}{3})^2 - (\frac{2}{3})^2 = 0.444$
$Gini_{\text{res}}(\text{KLOBUK}) = 0.267$ in

$GiniGain(\text{KLOBUK}) = Gini(\text{KRUH}) - Gini_{\text{res}}(\text{KLOBUK}) = 0.053.$

Atribut REP ima višji gini prispevek, zato ga damo v koren poddrevesa.

Tako kot v nalogi 3.2 tudi na tem mestu končamo gradnjo drevesa in drevo je natančno tako kot tisto, ki smo ga dobili z informacijskim prispevkom (slika 3.1). Ugotovimo, da sta gini in informacijski prispevek enako razvrstila atribute po pomembnosti.