Umetna inteligenca 3

Seminarska naloga- 1. sklop

Slavko Žitnik, 63060254
Fakulteta za računalništvo in informatiko
Ljubljana, 4.4.2011

1 Opis naloge

V tej seminarski nalogi generiramo umeten odločitveni problem tako, da za dva izbrana odločitvena modela velja, da ga eden reši dobro, drugi pa slabo. Poleg tega si izberemo eno realno bazo in modela testiramo še nad njo. Za obe domeni si izberemo 3 primere in za vsak model vizualiziramo razlago predikcij zanje. Za obe domeni vizualiziramo tudi razlago obeh modelov.
Pri reševanju seminarske naloge uporabljamo statistični paket R.

2 Algoritmi za predikcijo

Za prvi model si izberemo Naivni Bayesov klasifikator, ki predpostavlja pogojno neodvisnost med atributi.
Za drugi model si izberemo Odločitveno drevo.

3 Opis podatkov

3.1 Umetna učna množica

V umetni množici smo generirali XOR problem. Množica vsebuje 4 atribute (C1, C2, C3, C4) in razred (C). Vsi atributi imajo diskretne vrednosti 0 in 1. Vrednost razreda C = C1 XOR C2. Ostali trije atributi so generirani naključno. Množica vsebuje skupaj 100.000 primerov in smo jo z naključnim semenom 42 razdelili na učni (70%) in testni del (30%).

Predvidevamo, da bo Naivni Bayesov klasifikator slabo rešil problem za razliko od Odločitvenega drevesa. V tabeli 2 predstavimo klasifikacijske točnosti posameznih modelov. Opazimo, da se Naivni Bayes izkaže najslabši. Če bi obrnili njegovo napoved, bi lahko dobili natančnosti 0,56. Ker je delež obeh razredov v učnih podatkih enak, večinski klasifikator napove pravilno vsako drugo vrednost. Najbolje pa problem reši odločitveno drevo, ki pravilno napove vse testne primere.

	Klasifikacijska točnost
Naivni Bayes	0,44
Odločitveno drevo	1,0
Večinski klasifikator	0,5

Tabela 1: Uspešnost modelov na umetni množici

3.2 Realna učna množica

Realno učno množico izberemo iz UCI repozitorija. Odločili smo se za učno množico o vinih (angl. Wine Data Set).

Učna množica vsebuje 178 primerkov s trinajstimi zveznimi atributi. Atributi pomenijo – angl.: *Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline*. Ciljni je napovedati razred s tremi diskretnimi vrednostmi, katerih zastopanost je prikazana v tabeli 2.

Razred	Zastopanost v množici (#)	Delež
1	59	33%
2	71	40%
3	48	27%

Tabela 2: Priorna porazdelitev razredov na javni učni množici

Podatke pred obdelavo normaliziramo na interval [0,1] in množico razdelimo na učni (70%) in testni del (30%).

V tabeli 3 predstavimo klasifikacijske točnosti napovednih modelov nad izbrano učno množico. Na tej množici se najbolje izkaže naivni Bayesov klasifikator.

	Klasifikacijska točnost
Naivni Bayes	0,96
Odločitveno drevo	0,89
Večinski klasifikator	0,4

Tabela 3: Uspešnost modelov na realni učni množici

4 Razlaga testnih primerov

4.1 Umetna učna množica

Primere, ki jih razlagamo v tem razdelku, predstavimo v tabeli 4.

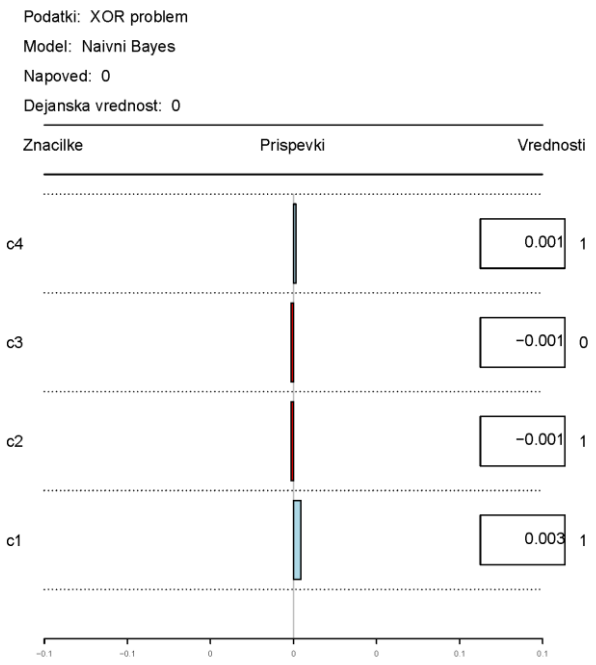
C1	C2	C3	C4	C
1	1	0	1	0
1	0	0	0	1
0	0	1	1	0

Tabela 4: Izbrani primeri za razlago nad umetno učno množico

*Vsi prispevki, ki jih omenjamo v primerih za učno množico, so prispevki atributov k klasifikacijski vrednosti 1.

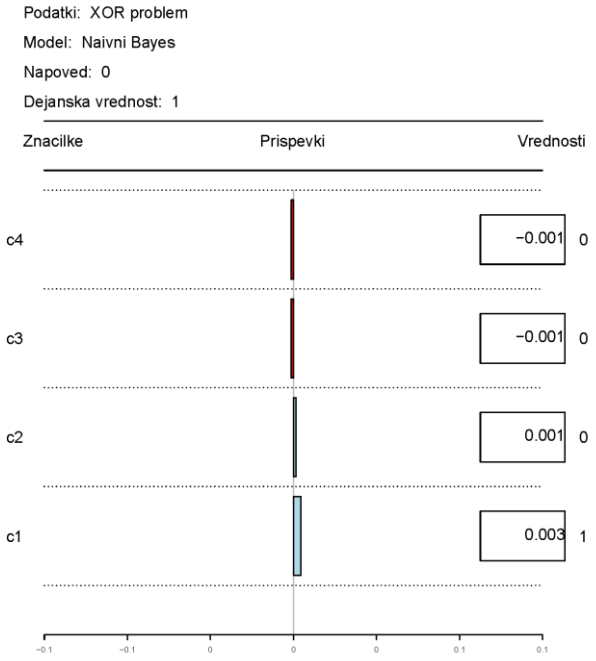
4.1.1 Naivni Bayes

4.1.1.1 Primer 1



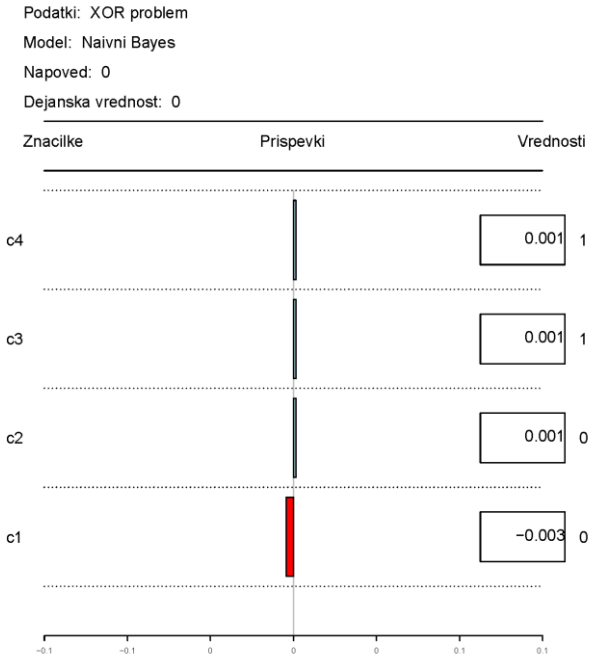
V tem primeru je model pravilno klasificiral primer. Tu in v naslednjih dveh primerih vidimo, da model ni ugotovil pomembnosti atributov C1 in C2, ki edina vplivata na napoved. Nekaj večji prispevek atributa C1 se ugotovi, vendar ni signifikanten.

4.1.1.2 Primer 2



V tem primeru je model napovedal napačen razred. Ugotovi se, da nekaj malega k razredu 1 prispevata atributa C1 in C2, ostala dva pa k razredu 0.

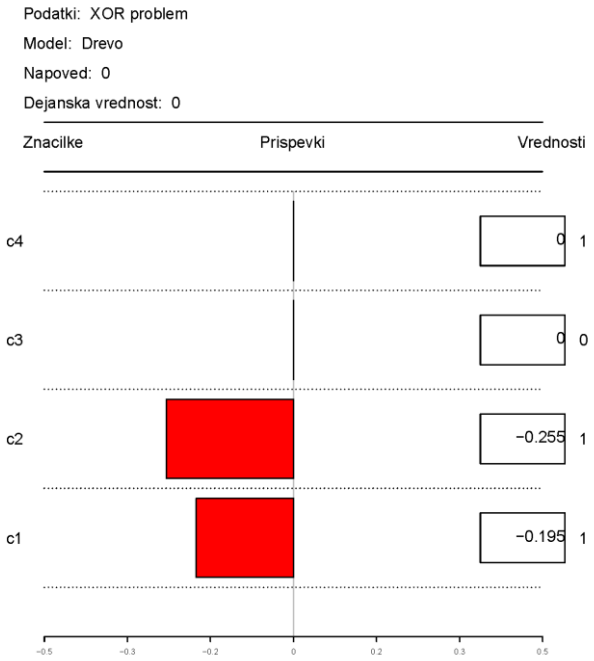
4.1.1.3 Primer 3



Enako kot v zgornjih dveh primerih model ni ugotovil bistvenega prispevka atributov C1 in C2. Primer je napovedal pravilno in k temu rezultatu je največ prispevala vrednost atributa C1.

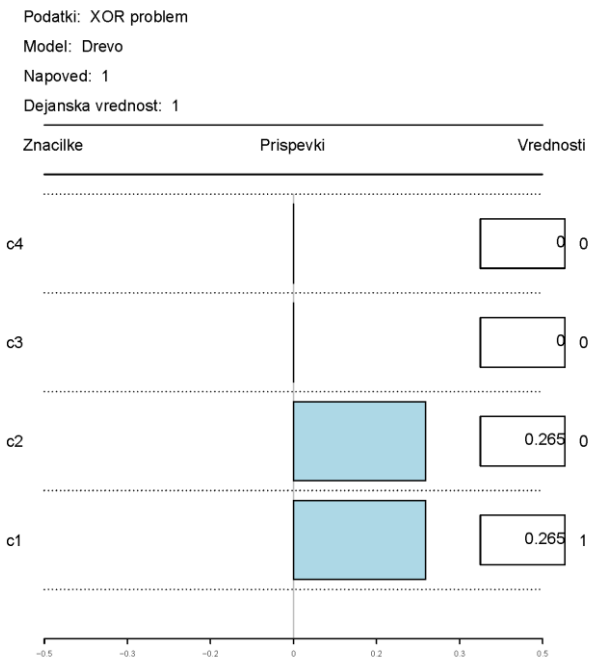
4.1.2 Odločitveno drevo

4.1.2.1 Primer 1

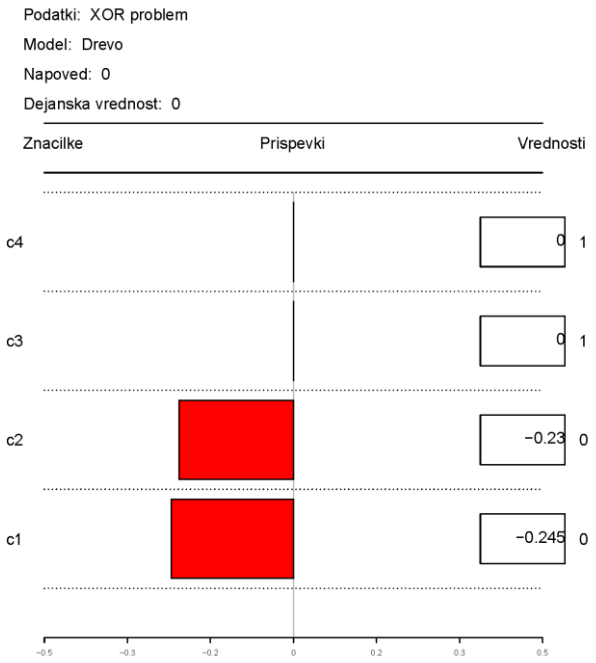


Pri modelu odločitvenega modela se očitno kaže, da sta vrednosti atributov C1 in C2 najbolj vplivali, da je bila napovedana vrednost 0, ki je pravilna. Za ostala dva atributa je bilo ugotovljeno, da imata minimalen prispevek, kar je tudi pravilno, saj sta naključna. Analogne ugotovitve se pokažejo tudi v naslednjih dveh primerih.

4.1.2.2 Primer 2



Vrednosti C1 in C2 sta odločilno prispevali k pravilni končni klasifikaciji v razred 1. Pravilno je bil ugotovljen tudi prispevek atributov C3 in C4, ki je skoraj ničeln.



Ciljna vrednost razreda v tem primeru je 0. Tudi v tem primeru smo ugotovili, da sta k tem najbolj prispevali vrednosti atributov C1 in C2.

4.1.2.3 Primer 3

4.2 Realna učna množica

Primere, ki jih razlagamo v tem razdelku, predstavimo v tabeli 4.

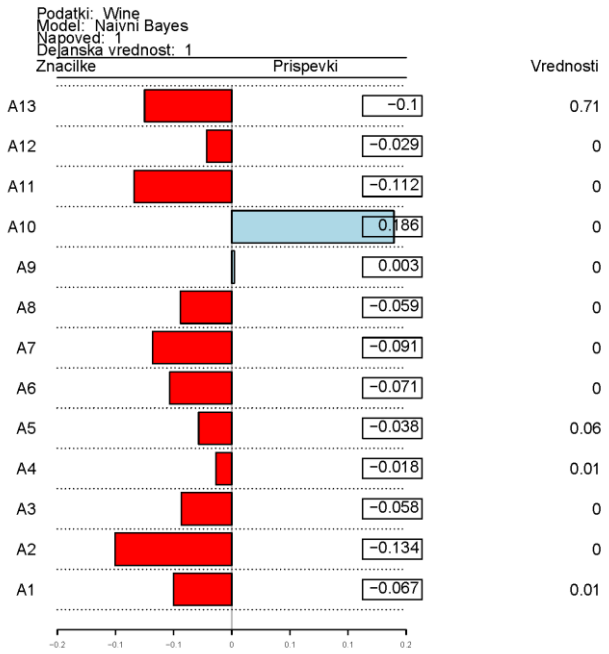
A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	C
.0077	.0013	.0015	.0109	.0600	.0015	.0018	.0001	.0015	.0033	.0005	.0018	.7053	1
.0071	.0010	.0013	.0010	.0481	.0008	.0008	.0002	.0008	.0013	.0005	.0012	.02856	2
.0072	.0022	.0013	.0124	.0523	.0012	.0004	.0002	.0005	.0044	.0002	.0008	.3094	3

Tabela 5: Izbrani primeri za razlago nad realno učno množico

*Prispevki, ki jih razlagamo v primerih nad realno učno množico so prispevki k vrednosti razreda 2. Torej če bo ciljna vrednost razreda 1 ali 3, bi morali biti prispevki negativni.

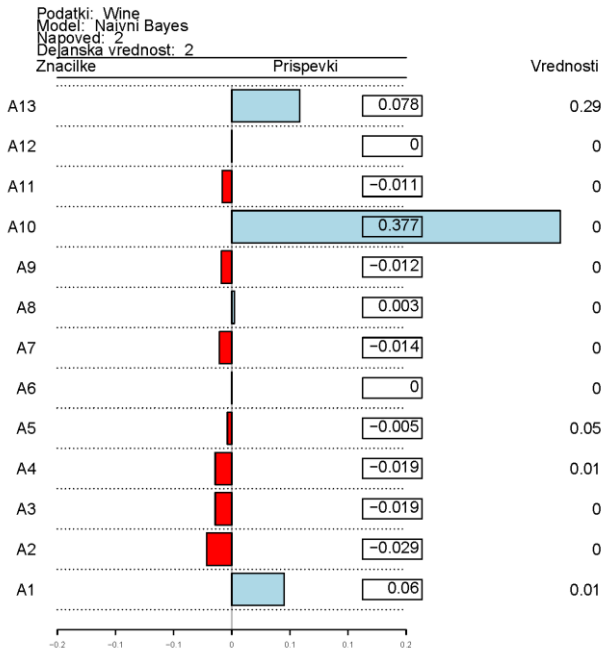
4.2.1 Naivni Bayes

4.2.1.1 Primer 1



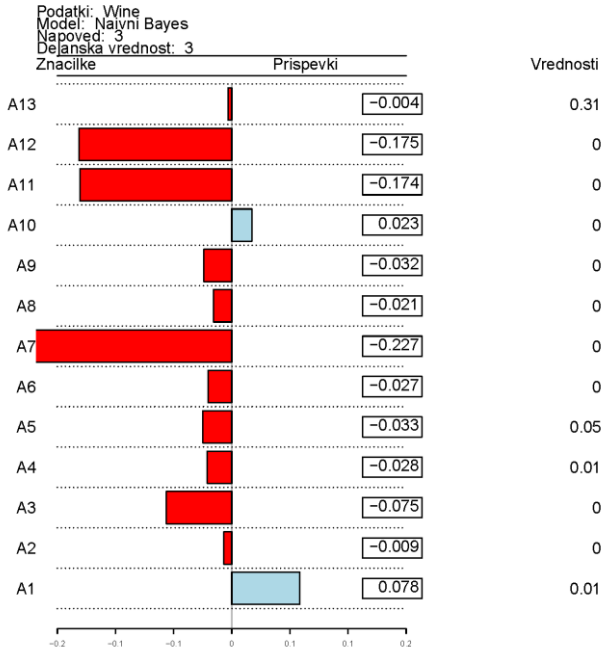
K klasifikaciji v razred 2 vpliva le vrednost spremenljivke A10 in malenkostno A9. Vrednosti ostalih atributov govorijo proti razredu 2 in najbrž največ razredu 1, kamor je model tudi pravilno klasificiral primer (to bi lahko pogledali tako, da bi izračunali še prispevke primerov za vrednosti 1 in 3 razredne spremenljivke).

4.2.1.2 Primer 2



Model je pravilno klasificiral primer v razred 2. K pravilni klasifikaciji so največ prispevali atributi A10, A13 in A1, ostali so bolj prispevali k ostalim vrednostim.

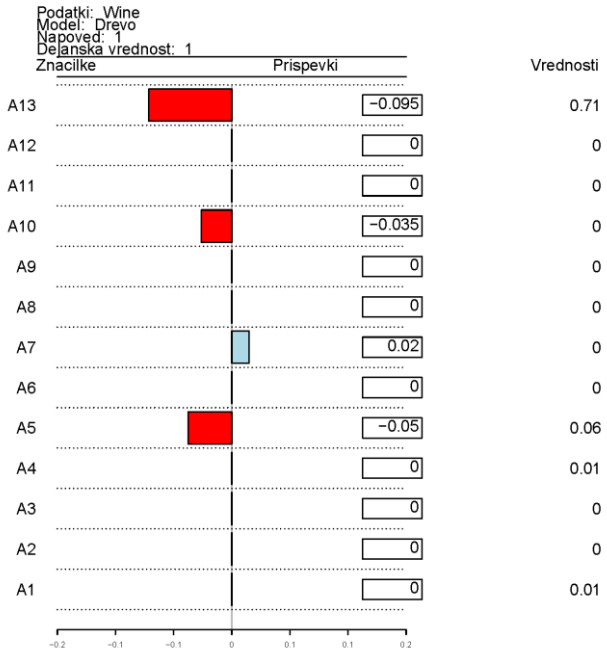
4.2.1.3 Primer 3



K razredu 2 sta v tem primeru prispevali vrednosti atributov A1 in A10. Ostali so prispevali bolj razredoma 1 in 3. Glede na to, da je model pravilno klasificiral primer v razred 3, so vrednosti k temu razredu največ prispevali.

4.2.2 Odločitveno drevo

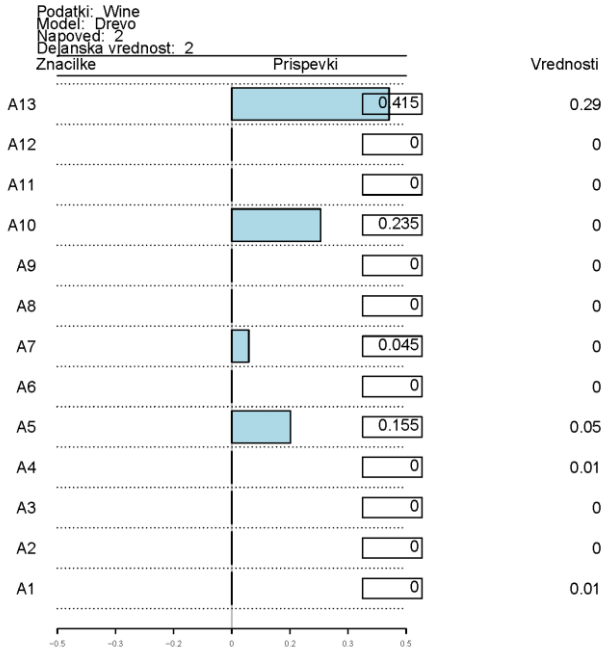
4.2.2.1 Primer 1



Odločitveno drevo je pravilno klasificiralo primer v razred 1. K razredu 2 prispeva samo vrednost atributa A7, ostali k razredu 1 in 3 in očitno najbolj k 1.

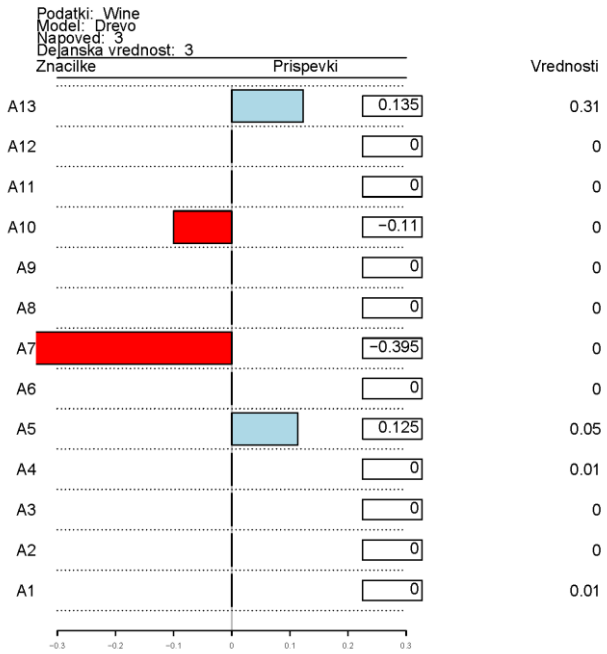
Glede na to, da opazimo, da imajo od nič različne prispevke le atributi A5, A7, A10, A13, lahko sklepamo, da so le ti atributi prisotni pri odločanju v drevesu.

4.2.2.2 Primer 2



V tem primeru so vsi atributi (A5, A7, A10, A13), ki so vsebovani v odločitvenem drevesu pozitivno prispevajo k razredu 2, ki je tudi pravilen razred, kamor se klasificira. Najbolj k razredu 2 prispeva vrednost atributa A13.

4.2.2.3 Primer 3



V tem primeru k razredu 2 prispevata vrednosti atributov A13 in A5. Vrednosti atributov A10 in A7 pa prispevata proti razredu 2, torej najbrž najbolj k razredu 3, kamor odločitveno drevo tudi pravilno klasificira primer.

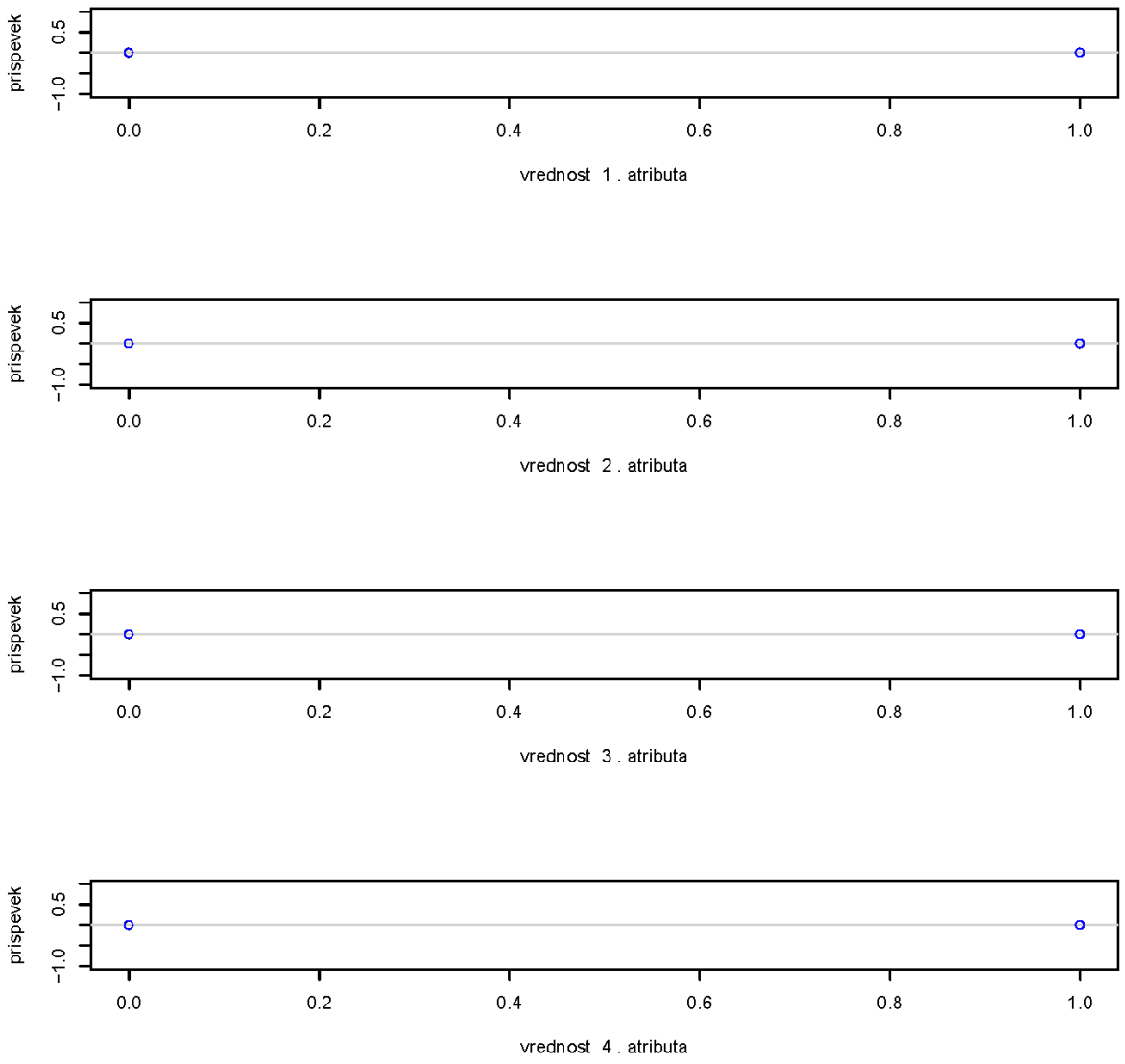
5 Razlaga modelov

5.1 Umetna učna množica

Na spodnjih dveh slikah razlagamo modela naivni Bayes in odločitveno drevo pri klasifikaciji nad umetno učno množico. Z sivimi krogci so označeni vrednosti posameznih atributov, z modro pa standardni odkloni, ki povejo pomembnost vrednosti določenega atributa.

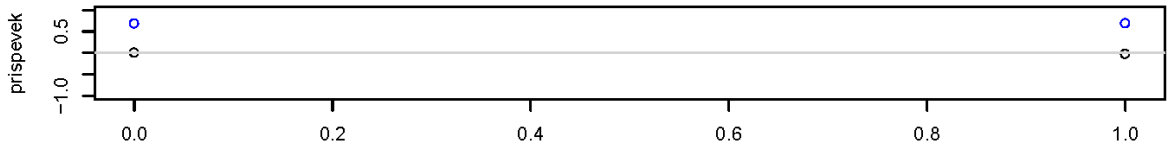
5.1.1 Naivni Bayesov model

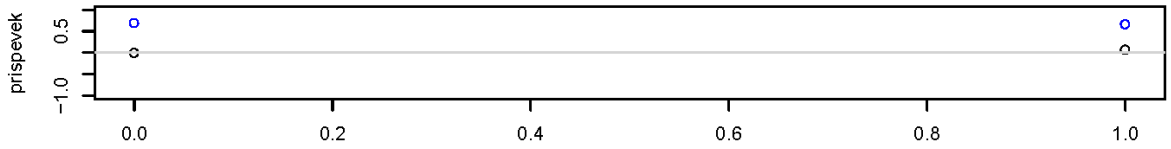
Na spodnji sliki predstavljamo razlago naivnega Bayesovega klasifikatorja na umetni množici. Kot pričakovano lahko iz slike ugotovimo, da razlaga pove, da noben atribut ne vpliva bistveno na klasifikacijsko vrednost. To vidimo iz tega, ker so standardni odkloni v bližini vrednosti 0.

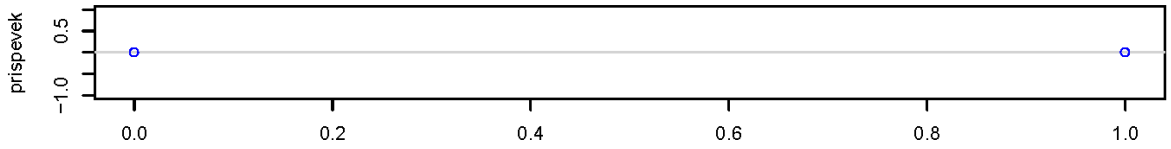


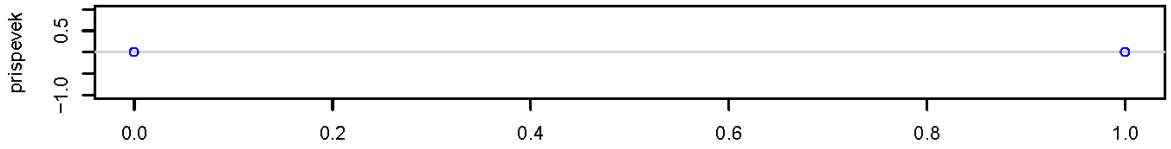
5.1.2 Model odločitvenega drevesa

Na spodnji sliki predstavljamo razlago odločitvenega drevesa na umetni množici. Iz vrednosti prispevkov lahko ugotovimo, da sta atributa C1 in C2 pomembna pri klasifikaciji, ostala dva pa ne. To je pravilno, saj sta atributa C3 in C4 naključna.







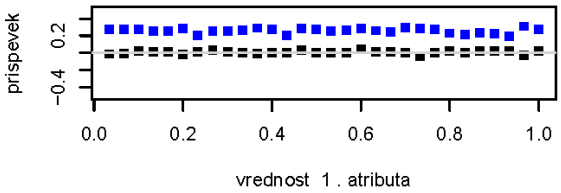


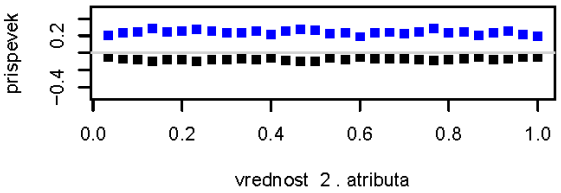
5.2 Realna učna množica

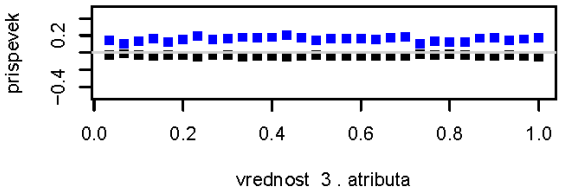
Na spodnjih slikah predstavimo prispevke posameznih atributov k razredu 2 nad realno učno množico vin.

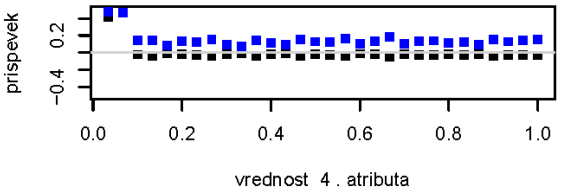
5.2.1 Naivni Bayesov model

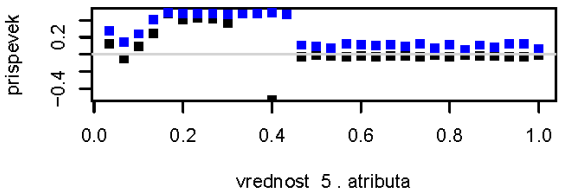
Iz spodnje slike lahko ugotovimo, da k razredu 2 najbolj prispevajo vrednosti atributov A10, vrednosti <0,5 atributa A13 in vrednosti <0,4 atributa A5. Negativno k razredu 2 oz. drugima dvema razredoma prispevajo atributi A2, A7, A11 ter vrednosti >0,5 atributa A13. Nekaj malega k razredu 2 prispeva tudi atribut A6. Pri atributu A1 vidimo, da nekatere vrednosti bolj prispevajo k razredu 2, nekatere proti. Za ostale atribute ne moremo tako očitno reči kako in kam prispevajo.

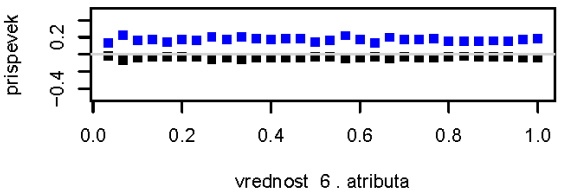

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 1 . atributa

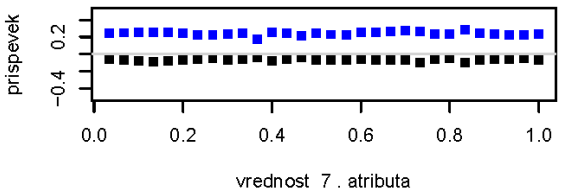

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 2 . atributa

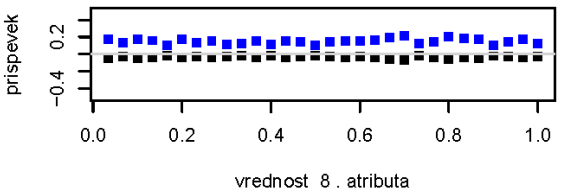

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 3 . atributa

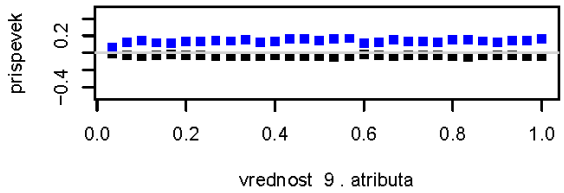

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 4 . atributa

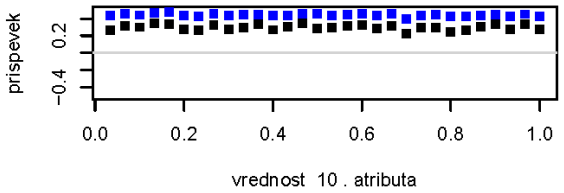

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 5 . atributa

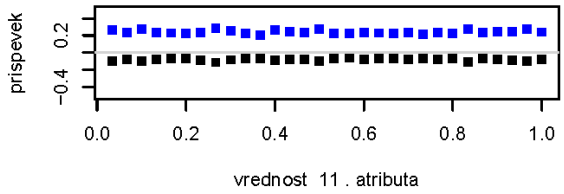

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 6 . atributa

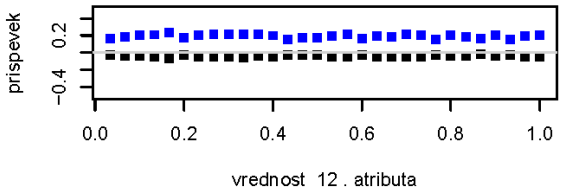

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 7 . atributa

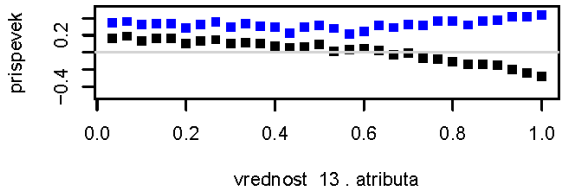

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 8 . atributa







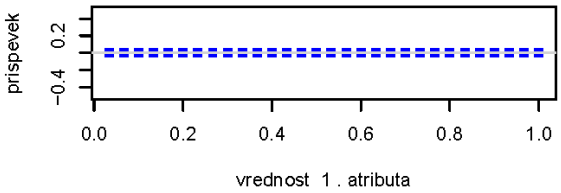


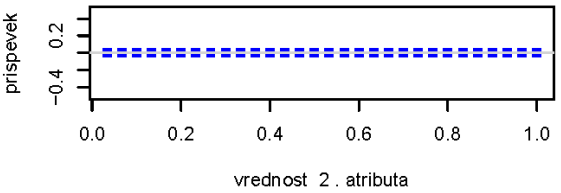


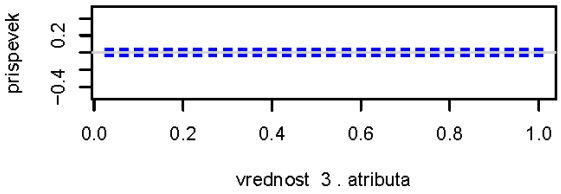
5.2.2 Model odločitvenega drevesa

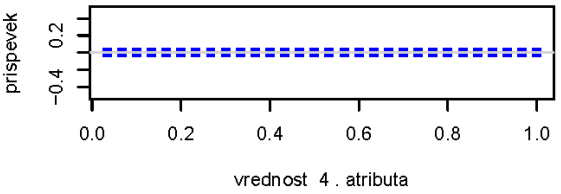
Na spodnjih dveh slikah razlagamo model odločitvenega drevesa na realni množici podatkov s področja vin. Kot smo v zgornjih primerih ugotovili, lahko opazimo le prispevke pri atributih A13, A10, A7 in A5. K razredu 2 najbolj prispevajo vrednosti <0,6 atributa A13 in atribut A7. Proti razredu 2 prispevajo atribut A13 >0,6, A5 in malo manj A10.

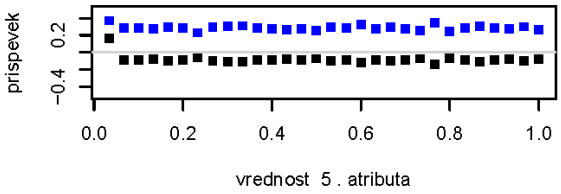
Ostali atributi na klasifikacijo ne vplivajo, ker tudi niso vsebovani v modelu klasifikatorja drevesa.

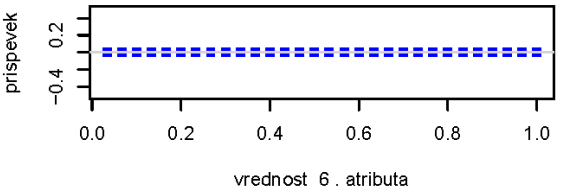

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 1 . atributa

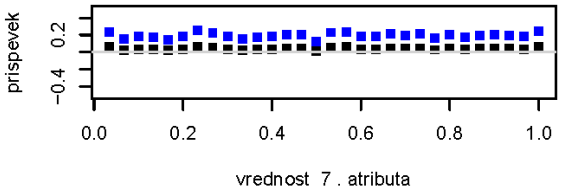

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 2 . atributa

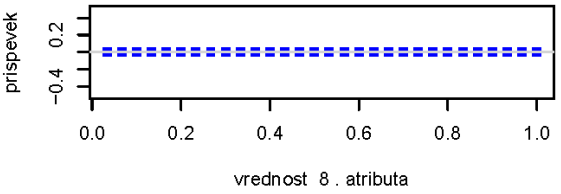

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 3 . atributa

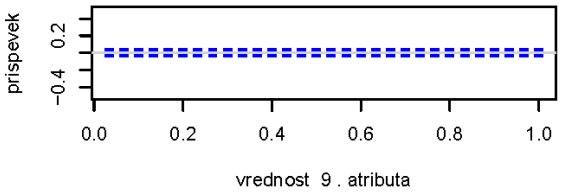

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 4 . atributa

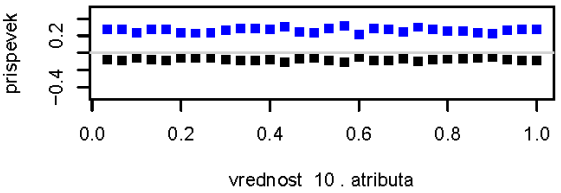

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 5 . atributa

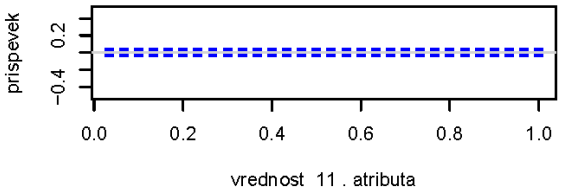

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 6 . atributa

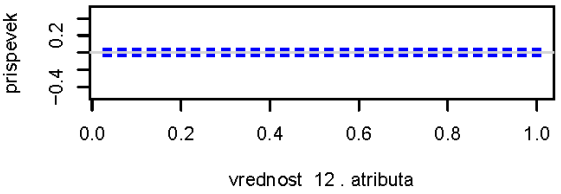

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 7 . atributa

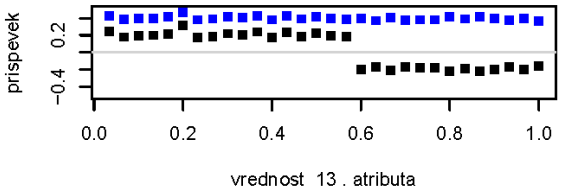

prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 8 . atributa


prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 9 . atributa


prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 10 . atributa


prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 11 . atributa


prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 12 . atributa


prispevek
−0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
vrednost 13 . atributa