Naïve Bayes

MAS Informatike – Nauka o podacima

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Iz prve tabele možemo prebrojati koliko puta svaki par atribut –vrednost pojavljuje za svaku od klasa.

U donjem delu tabele, iste informacije su napisane u obliku verovatnoća.

	Outlook			Temperature			Humidity			Windy		Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

Outlook			Temperature			Humidity			Windy		Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

Neka je korišćenjem informacija iz gornje tabele potrebno klasifikovati novi primer:

Outlook	Temperature	Humidity	Windy
sunny	cool	high	true

Verovatnoća da novi primer bude klasifikovan kao:

Yes = $\frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{9}{14} = 0.0053$

No= $\frac{3}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5} \cdot \frac{5}{14} = 0.0206$

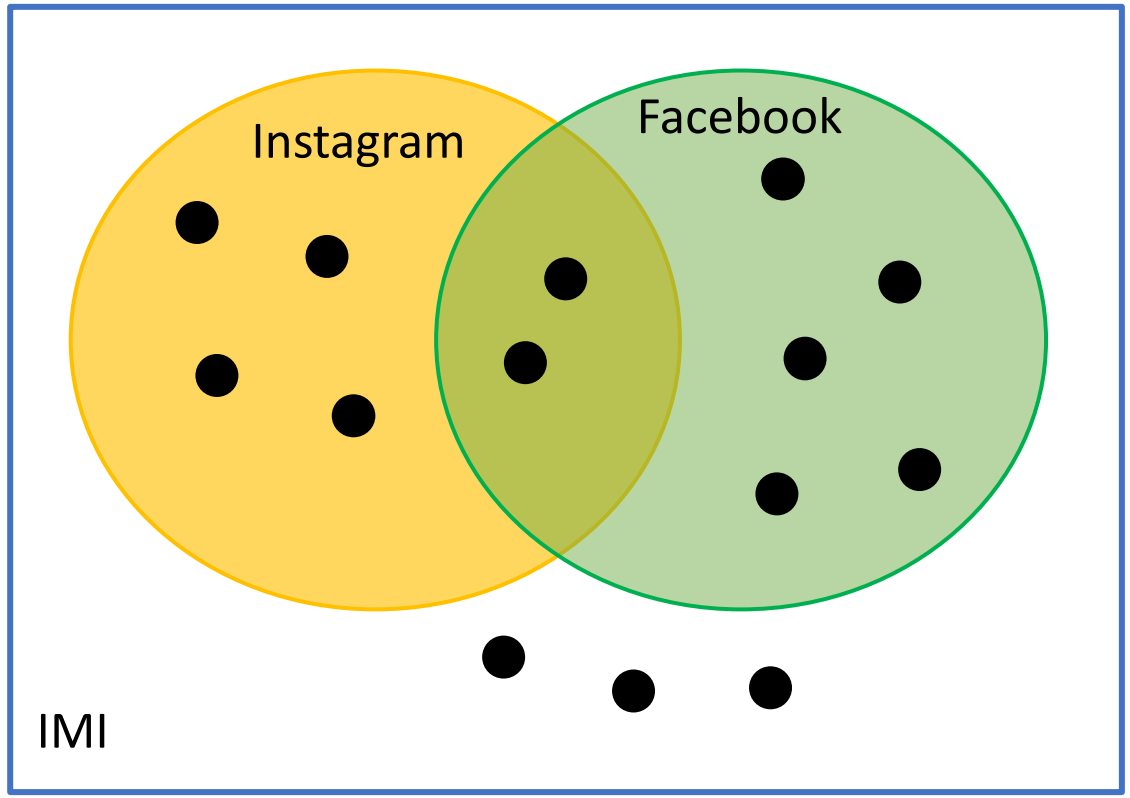
Ako prethodne vrednosti normalizujemo tako da u zbiru daju 1:

$$P(yes) = \frac{0.0053}{0.0053 + 0.0206} = 0.2046$$

$$P(no) = \frac{0.0206}{0.0053 + 0.0206} = 0.7953$$

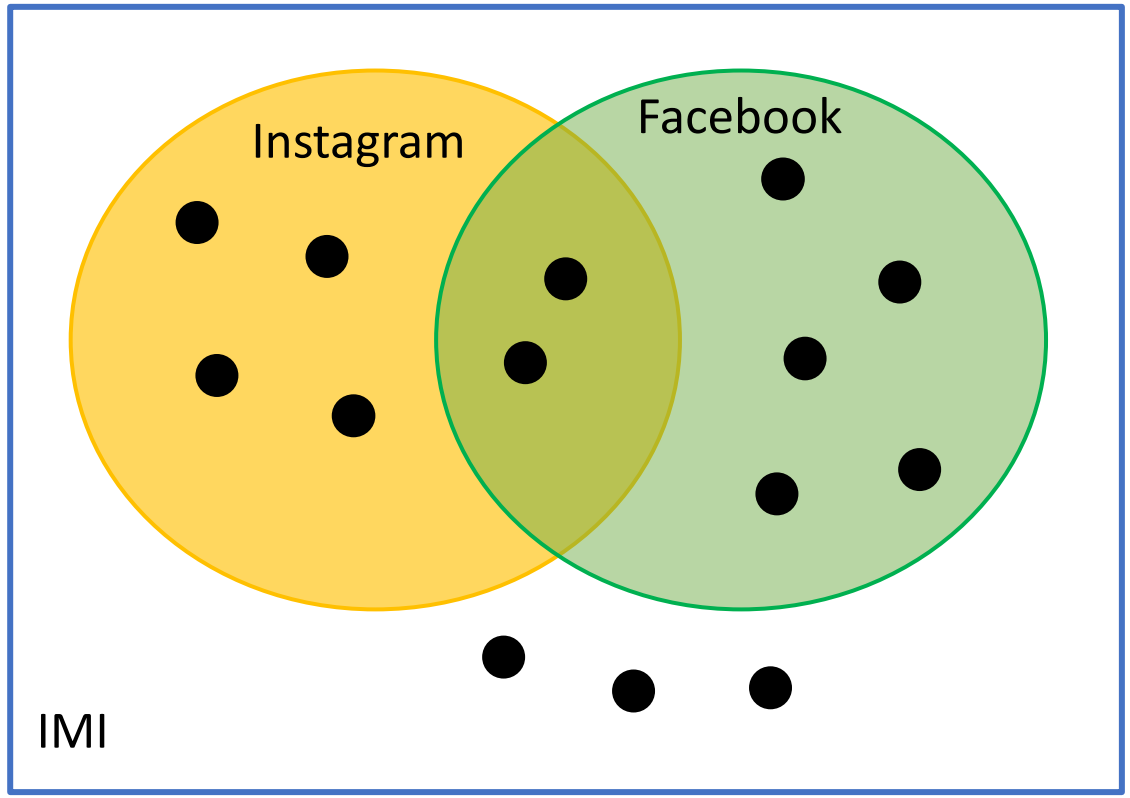
Veća je verovatnoća da novi primer bude klasifikovan kao no

Objašnjenje uslovne verovatnoće



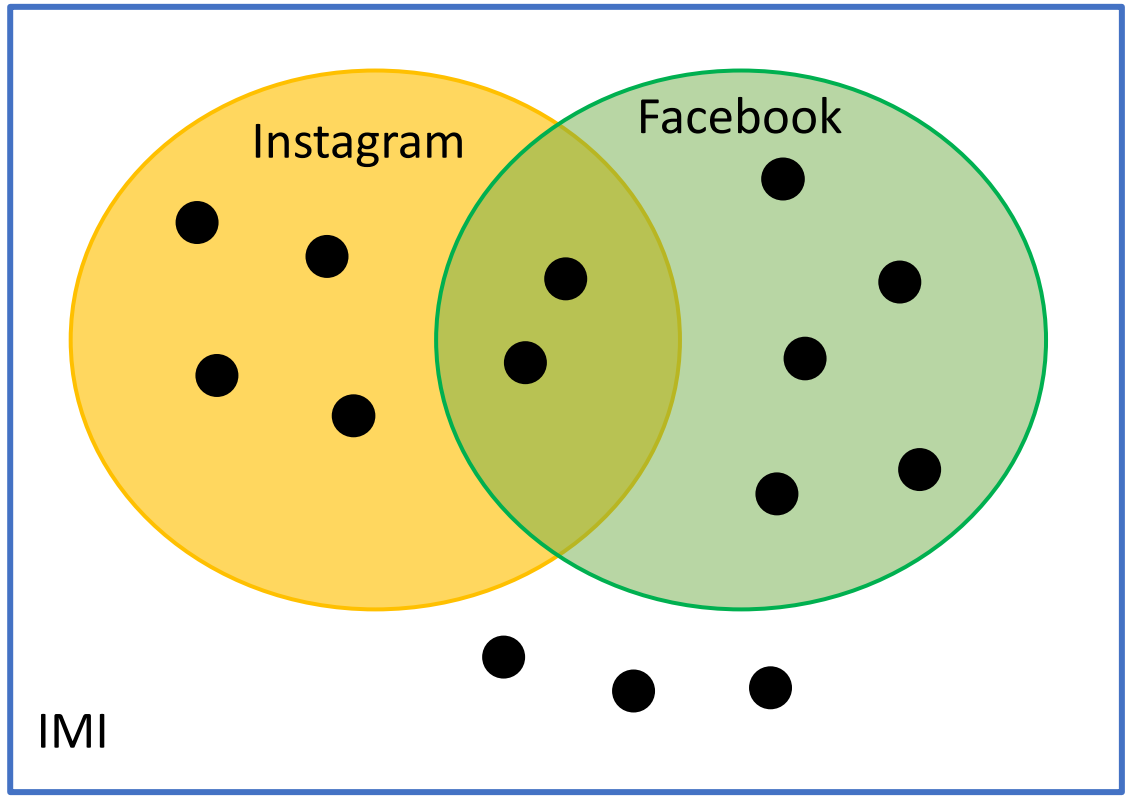
	Koristi Facebook	Ne koristi Facebook
Koristi Instagram		
Ne koristi Instagram		

Objašnjenje uslovne verovatnoće



	Koristi Facebook	Ne koristi Facebook
Koristi Instagram	2	5
Ne koristi Instagram	4	3

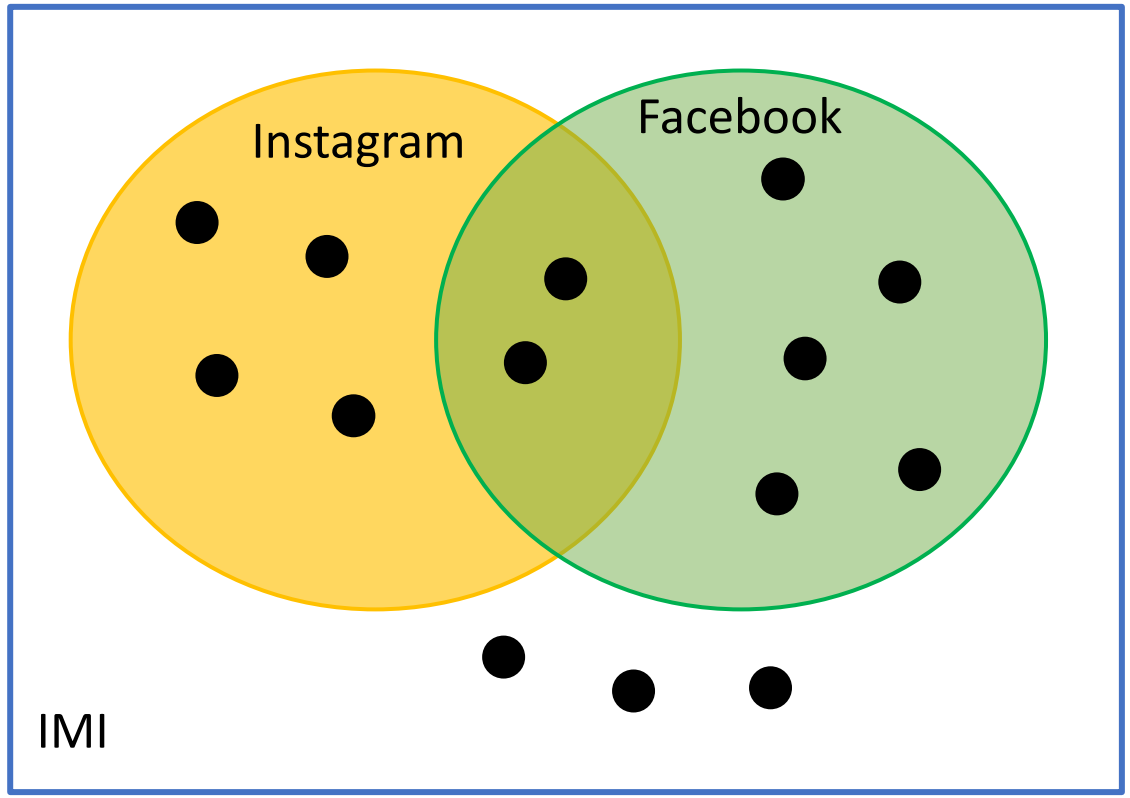
Objašnjenje uslovne verovatnoće



	Koristi Facebook	Ne koristi Facebook
Koristi Instagram	2	5
Ne koristi Instagram	4	3

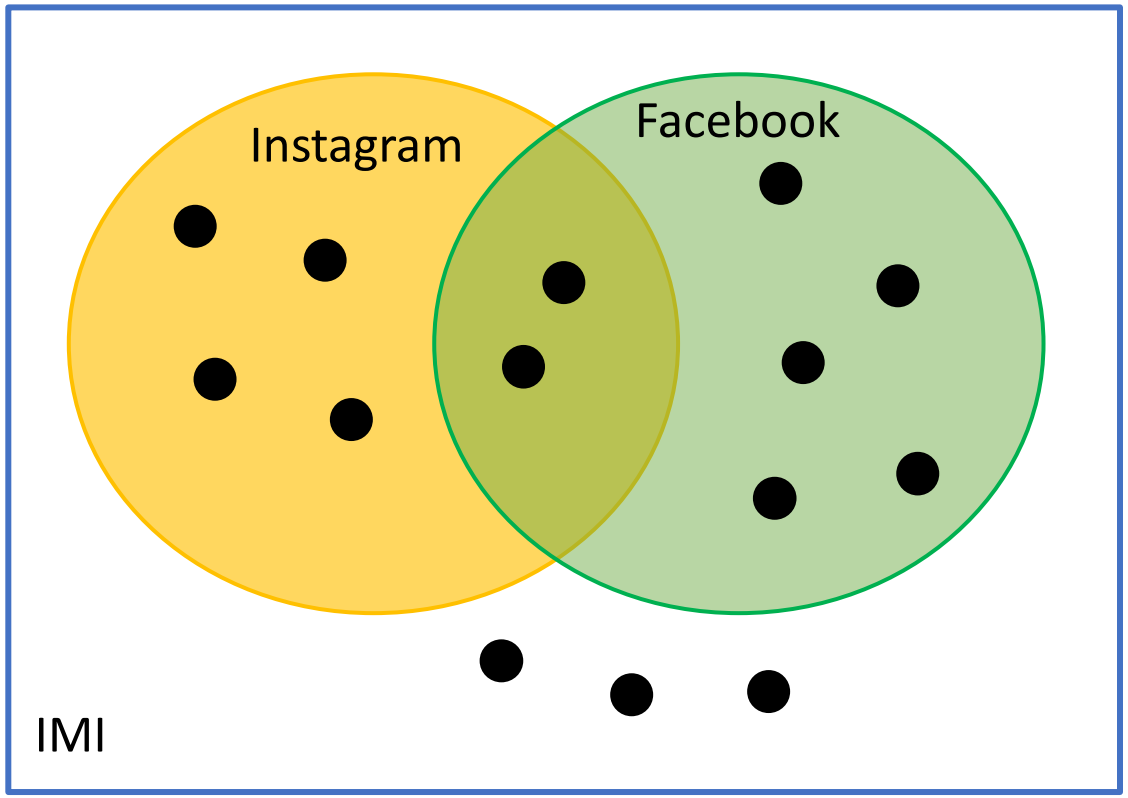
Kolika je šansa da profesor koji nam drži sledeće predavanje koristi obe društvene mreže?

Objašnjenje uslovne verovatnoće



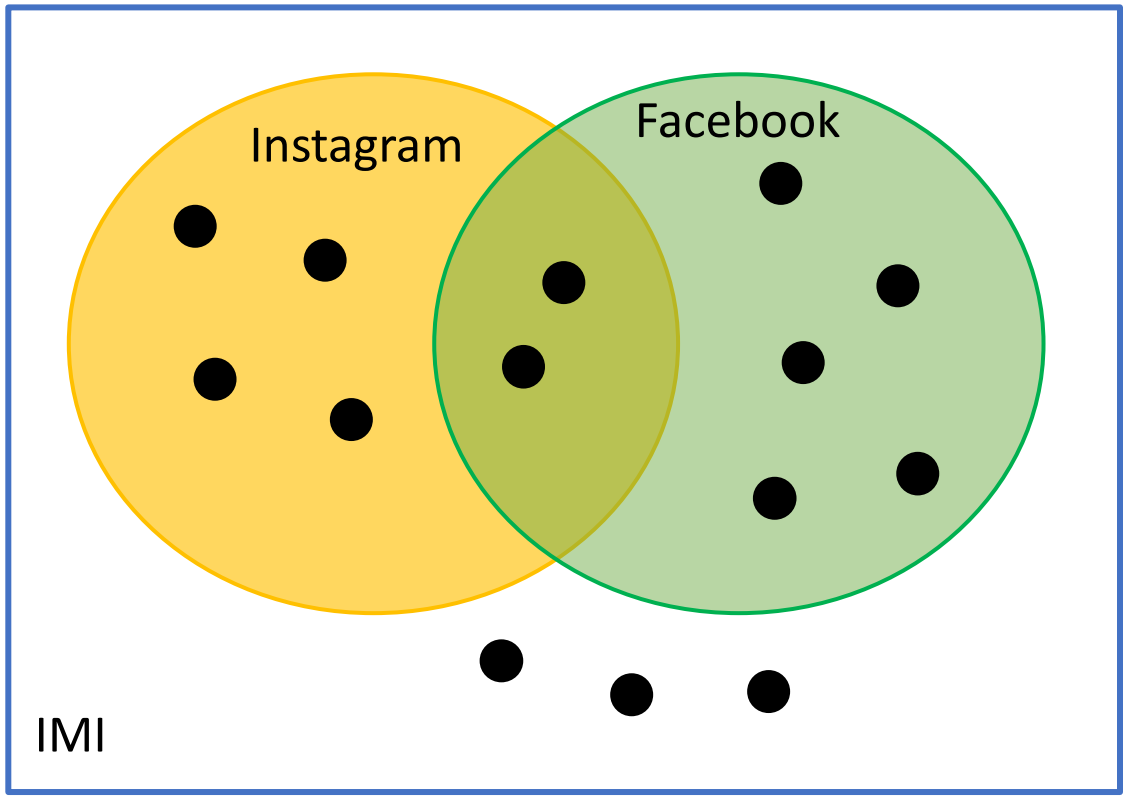
	Koristi Facebook	Ne koristi Facebook
Koristi Instagram	2 P=2/14	5 P=5/14
Ne koristi Instagram	4 P=4/14	3 P=3/14

Objašnjenje uslovne verovatnoće



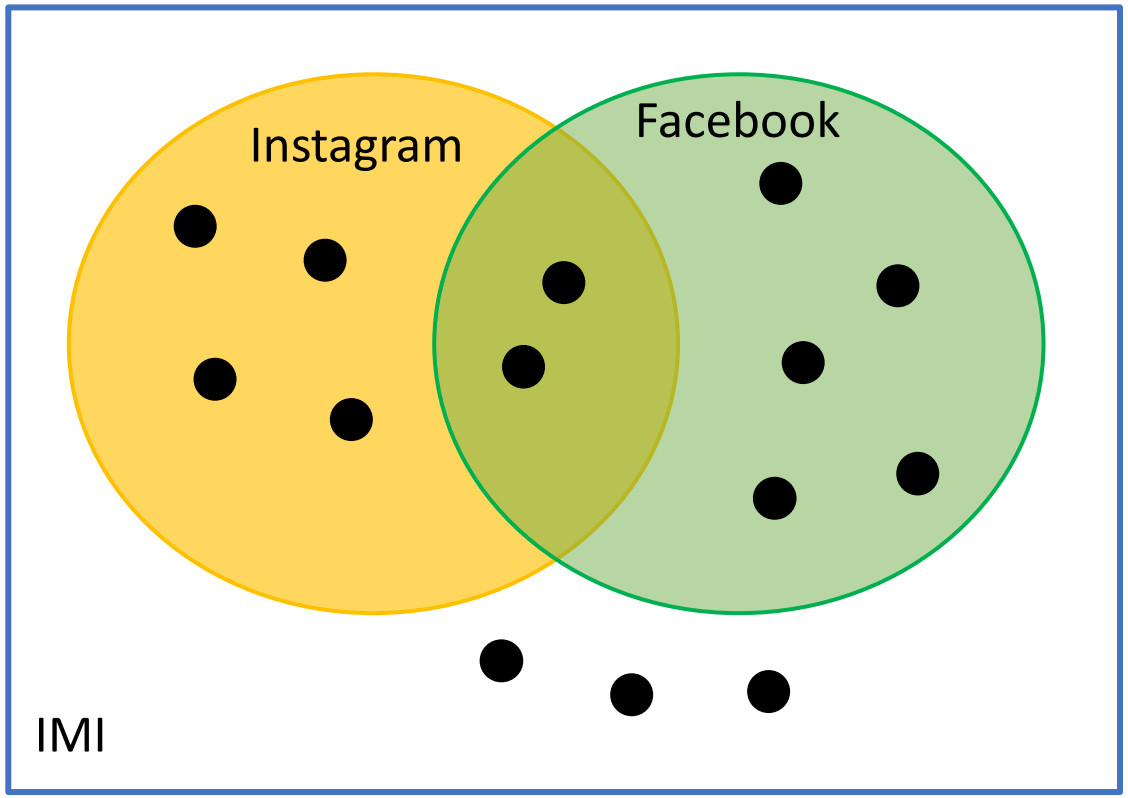
	Koristi Instagram	Ne koristi Instagram	Ukupno
Koristi Facebook	2 P=2/14	5 P=5/14	2+5=7
Ne koristi Facebook	4 P=4/14	3 P=3/14	4+3=7
	2+4=6	5+3=8	

Objašnjenje uslovne verovatnoće



	Koristi Instagram	Ne koristi Instagram	Ukupno
Koristi Facebook	2 P=2/14	5 P=5/14	2+5=7 P=7/14
Ne koristi Facebook	4 P=4/14	3 P=3/14	4+3=7 P=7/14
	2+4=6 P=6/14	5+3=8 P=8/14	

Objašnjenje uslovne verovatnoće



	Koristi Instagram	Ne koristi Instagram	Ukupno
Koristi Facebook	2 P=2/14	5 P=5/14	2+5=7 P=7/14
Ne koristi Facebook	4 P=4/14	3 P=3/14	4+3=7 P=7/14
	2+4=6 P=6/14	5+3=8 P=8/14	

P(Koristi Instagram | Koristi fejs)

Uslovna verovatnoća

- Uslovna verovatnoća. Predstavlja verovatnoću događaja ***A*** pod uslovom da se desio događaj ***B***

Conditional Probability

- $P(a \mid b)$ = fraction of possible worlds in which b is true in which a is also true
- $P(a \mid b) = P(a, b)/p(b)$
- thus, $P(a, b) = P(b)P(a \mid b)$

Bajesovo pravilo

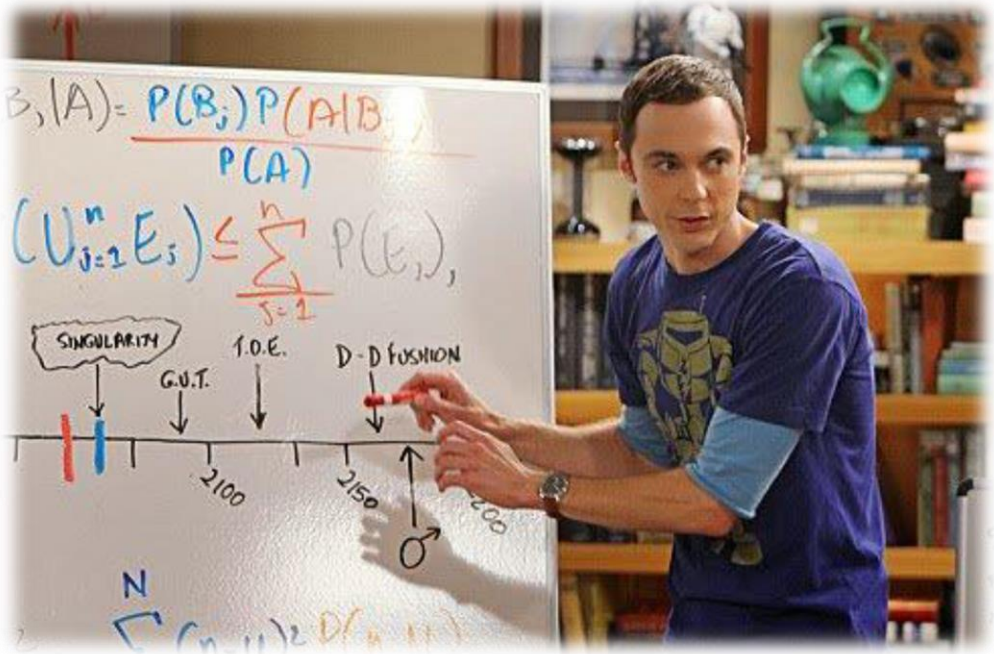
$$P(A|B) = \frac{P(A,B)}{P(B)}$$

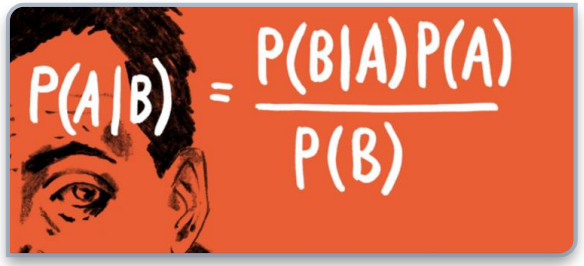
$$P(A|B) \cdot P(B) = P(A,B) = P(B|A) \cdot P(A)$$

$$\Rightarrow$$

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$



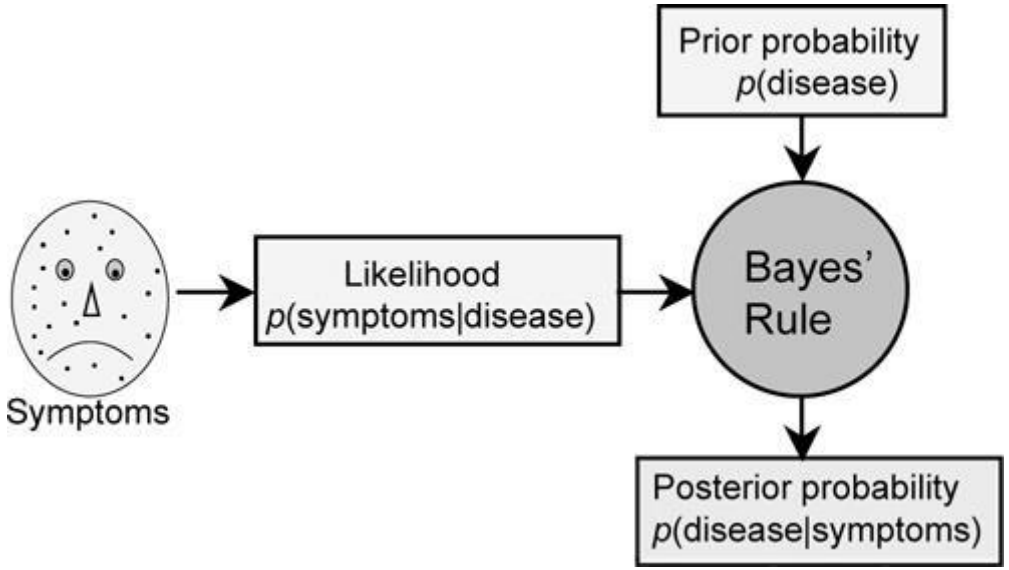


Primena u medicini

$$P(bolest|simptom) = \frac{P(simptom|bolest) \cdot P(bolest)}{P(simptom)}$$

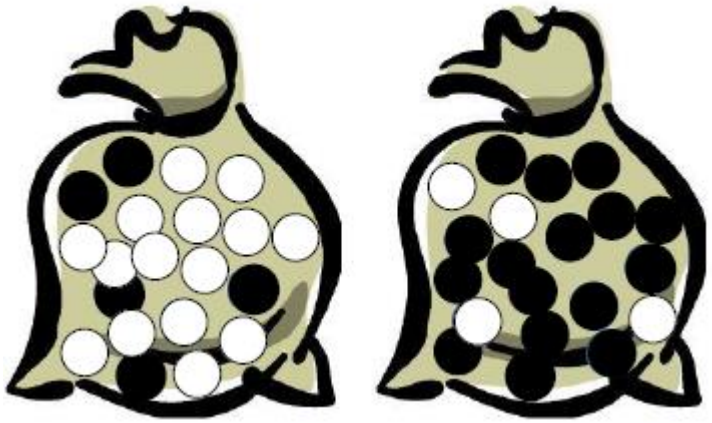
- Mnogo je lakše utvrditi verovatnoću simptoma, pod uslovom da nam je poznata bolest.

- Ako nam je poznata verovatnoća pojave nekog simptoma kod bolesti čije postojanje dokazujemo, i ako poznajemo apriori verovatnoću da se ta bolest javi, onda lako određujemo vrednost brojioca. Vrednost imenioca je ista za sev pretpostavljene bolesti.



Primer

- Dve vrste vrećica sa klikerima:
  - 4 vrećice tipa A (5 crnih, 15 belih kuglica)
  - 1 vrećica tipa B (16 crnih, 4 bele kuglice)



- Jedna vrećica je oštećena i kroz rupicu se vidi crni kliker. Kolika je verovatnoća da je u pitanju vreća tipa B?

$$P(B|C) = \frac{P(B) \cdot P(C|B)}{P(C)}$$

Upotreba Bajesove formule u problemima klasifikacije

- Simptomi -> Atributi
- Hipoteza -> Klasa

- Zanima nas kolika je verovatnoća klase C, pri vrednostima atributa $A_1, A_2, \dots, A_n$

$$P(C|A_1A_2 \dots A_n) = \frac{P(C) \cdot P(A_1A_2 \dots A_n|C)}{P(A_1A_2 \dots A_n)}$$

Upotreba Bajesove formule u problemima klasifikacije

- Naivni Bajesovski klasifikator se zasniva na pretpostavci da su vrednosti atributa međusobno nezavisne za određenu klasu.

- Za datu klasu, verovatnoća konjunkcije $A_1 A_2 \ldots A_n$ je proizvod verovatnoća pojedinačnih atributa

$$P(A_1 A_2 \ldots A_n | C) = P(A_1 | C) \cdot P(A_2 | C) \cdot \ldots \cdot P(A_n | C)$$

Upotreba Bajesove formule u problemima klasifikacije

- Verovatnoća klase C, pri vrednostima atributa $A_1, A_2, \dots, A_n$

$$P(C|A_1A_2 \dots A_n) = \frac{P(C) \cdot P(A_1|C) \cdot P(A_2|C) \cdot \dots \cdot P(A_n|C)}{P(A_1A_2 \dots A_n)}$$

- Imenilac je isti za sve pretpostavljene klase, tako da ne utiče na izbor one sa najvećom verovatnoćom. Zato ga možemo izostaviti.
- Napomena: Pojednostavljivanjem formule i izostavljanjem imenioca verovatnoće klasa više nemaju zbir 1. Ovo se može rešiti normalizacijom rezultata

Naivni Bajesovski klasifikator

- Primerak svrstavamo u najverovatniju klasu:

$$h(A_1 A_2 \dots A_n) = \text{argmax}_k \, P(C_k) \cdot \prod_{i=1}^{n} P(A_i|C_k)$$

- **Učenje:** oceniti vrednosti $P(C_k)$ i $P(A_i|C_k)$ za sve klase C_k i sve vrednosti atributa A_i
- **Predviđanje:** Koristimo gornju jednačinu za predviđanje klase novog primera

Primer

- Skup podataka za obuku sadrži 1200 primera voćki, pri čemu svaki primer pripada jednoj od tri moguće klase:
  - mango,
  - banana,
  - drugo voće.

- Atributi su sledeći:
  - da li je voće žuto ili ne,
  - da li je plod duguljast ili ne,
  - da li je slatko ili ne.

- Neka je potrebno klasifikovati voće koje je žuto, slatko i duguljasto.

Korak 1
(kreiranje tabele frekfencija)

- Šta možemo da zaključimo iz ove tabele?
  - Od 1200 plodova, 650 su mango, 400 su banane, a 150 je ostalo voće
  - 350 od ukupno 650 manga su žute boje
  - 800 plodova je žuto, 850 slatko i 400 duguljasto od ukupno 1200 plodova
  - …

Ime	Žuto	Slatko	Duguljasto	Ukupno
Mango	350	440	10	650
Banana	400	300	350	400
Ostalo	50	100	50	150
Ukupno	**800**	**850**	**400**	**1200**

Klasifikovanje nove voćke

- Da bismo klasifikovali primerak voća žute boje, slatkog ukusa i duguljastog ploda, potrebno je da utvrimo da li banana, mango ili ostalo voće ima najveću verovatnoću da ima ovakve osobine.

- Tražimo maksimalnu uslovnu verovatnoću:
  - P(Banana|Žuto Slatko Duguljasto)
  - P(Mango|Žuto Slatko Duguljasto)
  - P(Ostalo|Žuto Slatko Duguljasto)

Klasifikovanje nove voćke

$$h(\text{Žuto Slatko Duguljasto}) = \max\{$$

$$P(Banana) \cdot P(\text{Žuto}|Banana) \cdot P(Slatko|Banana) \cdot P(Duguljasto|Banana);$$
$$P(Mango) \cdot P(\text{Žuto}|Mango) \cdot P(Slatko|Mango) \cdot P(Duguljasto|Mango);$$
$$P(Ostalo) \cdot P(\text{Žuto}|Ostalo) \cdot P(Slatko|Ostalo) \cdot P(Duguljasto|Ostalo);$$
$$\}$$

Korak 2 (Tabela verovatnoća)

- Iz tabele frekfencija možemo preračunati i kreirati novu tabelu koja će sadržati potrebne verovatnoće.

Ime	Žuto	Slatko	Duguljasto	Ukupno
Mango	$P(\check{Z}uto\|Mango) = \frac{350}{650} = 0.54$	$P(Slatko\|Mango) = \frac{440}{650} = 0.67$	$P(Duguljasto\|Mango) = \frac{10}{650} = 0.015$	$P(Mango) = \frac{650}{1200} = 0.54$
Banana	$P(\check{Z}uto\|Banana) = \frac{400}{400} = 1$	$P(Slatko\|Banana) = \frac{300}{400} = 0.75$	$P(Duguljasto\|Banana) = \frac{350}{400} = 0.875$	$P(Banana) = \frac{400}{1200} = 0.33$
Ostalo	$P(\check{Z}uto\|Ostalo) = \frac{50}{150} = 0.33$	$P(Slatko\|Ostalo) = \frac{100}{150} = 0.66$	$P(Duguljasto\|Ostalo) = \frac{50}{150} = 0.33$	$P(Ostalo) = \frac{150}{1200} = 0.125$
Ukupno	**800**	**850**	**400**	**1200**

Klasifikovanje nove voćke

$$h(\text{Žuto Slatko Duguljasto}) = \max\{$$

$$P(Banana) \cdot P(\text{Žuto}|Banana) \cdot P(Slatko|Banana) \cdot P(Duguljasto|Banana);$$

$$P(Mango) \cdot P(\text{Žuto}|Mango) \cdot P(Slatko|Mango) \cdot P(Duguljasto|Mango);$$

$$P(Ostalo) \cdot P(\text{Žuto}|Ostalo) \cdot P(Slatko|Ostalo) \cdot P(Duguljasto|Ostalo);$$

$$\}$$

Ime	Žuto	Slatko	Duguljasto	Ukupno
Mango	$P(\text{Žuto}\|Mango) = \frac{350}{650} = 0.54$	$P(Slatko\|Mango) = \frac{440}{650} = 0.67$	$P(Duguljasto\|Mango) = \frac{10}{650} = 0.015$	$P(Mango) = \frac{650}{1200} = 0.54$
Banana	$P(\text{Žuto}\|Banana) = \frac{400}{400} = 1$	$P(Slatko\|Banana) = \frac{300}{400} = 0.75$	$P(Duguljasto\|Banana) = \frac{350}{400} = 0.875$	$P(Banana) = \frac{400}{1200} = 0.33$
Ostalo	$P(\text{Žuto}\|Ostalo) = \frac{50}{150} = 0.33$	$P(Slatko\|Ostalo) = \frac{100}{150} = 0.66$	$P(Duguljasto\|Ostalo) = \frac{50}{150} = 0.33$	$P(Ostalo) = \frac{150}{1200} = 0.125$
Ukupno	**800**	**850**	**400**	**1200**

Klasifikovanje nove voćke

$$h(\text{Žuto Slatko Duguljasto}) = \max\{$$

$0{,}33 \cdot 1 \cdot 0{,}75 \cdot \ 0{,}875;$

$P(Mango) \cdot P(\text{Žuto}|Mango) \cdot \ P(Slatko|Mango) \ \cdot \ P(Duguljasto|Mango);$

$P(Ostalo) \cdot P(\text{Žuto}|Ostalo) \cdot \ P(Slatko|Ostalo) \ \cdot \ P(Duguljasto|Ostalo);$

$\}$

Ime	Žuto	Slatko	Duguljasto	Ukupno
Mango	$P(\text{Žuto}\|Mango) = \frac{350}{650} = 0.54$	$P(Slatko\|Mango) = \frac{440}{650} = 0.67$	$P(Duguljasto\|Mango) = \frac{10}{650} = 0.015$	$P(Mango) = \frac{650}{1200} = 0.54$
Banana	$P(\text{Žuto}\|Banana) = \frac{400}{400} = 1$	$P(Slatko\|Banana) = \frac{300}{400} = 0.75$	$P(Duguljasto\|Banana) = \frac{350}{400} = 0.875$	$P(Banana) = \frac{400}{1200} = 0.33$
Ostalo	$P(\text{Žuto}\|Ostalo) = \frac{50}{150} = 0.33$	$P(Slatko\|Ostalo) = \frac{100}{150} = 0.66$	$P(Duguljasto\|Ostalo) = \frac{50}{150} = 0.33$	$P(Ostalo) = \frac{150}{1200} = 0.125$
Ukupno	**800**	**850**	**400**	**1200**

Klasifikovanje nove voćke

$$h(\text{Žuto Slatko Duguljasto}) = \max\{$$

$0{,}33 \cdot 1 \cdot 0{,}75 \cdot\ 0{,}875;$

$0{,}54 \cdot 0{,}67 \cdot\ 0{,}015 \cdot 0{,}54;$

$P(Ostalo) \cdot P(\text{Žuto}|Ostalo) \cdot\ P(Slatko|Ostalo)\ \cdot\ P(Duguljasto|Ostalo);$

$\}$

Ime	Žuto	Slatko	Duguljasto	Ukupno
Mango	$P(\text{Žuto}\|Mango) = \frac{350}{650} = 0.54$	$P(Slatko\|Mango) = \frac{440}{650} = 0.67$	$P(Duguljasto\|Mango) = \frac{10}{650} = 0.015$	$P(Mango) = \frac{650}{1200} = 0.54$
Banana	$P(\text{Žuto}\|Banana) = \frac{400}{400} = 1$	$P(Slatko\|Banana) = \frac{300}{400} = 0.75$	$P(Duguljasto\|Banana) = \frac{350}{400} = 0.875$	$P(Banana) = \frac{400}{1200} = 0.33$
Ostalo	$P(\text{Žuto}\|Ostalo) = \frac{50}{150} = 0.33$	$P(Slatko\|Ostalo) = \frac{100}{150} = 0.66$	$P(Duguljasto\|Ostalo) = \frac{50}{150} = 0.33$	$P(Ostalo) = \frac{150}{1200} = 0.125$
Ukupno	**800**	**850**	**400**	**1200**

Klasifikovanje nove voćke

$$h\left(\check{Z}uto\ Slatko\ Duguljasto\right) = \max\{$$

$0{,}33 \cdot 1 \cdot 0{,}75 \cdot\ 0{,}875;$

$0{,}54 \cdot 0{,}67 \cdot\ 0{,}015 \cdot 0{,}54;$

$0{,}125 \cdot 0{,}33 \cdot 0{,}66\ \cdot 0{,}33;$

$\}$

Ime	Žuto	Slatko	Duguljasto	Ukupno
Mango	$P(\check{Z}uto\|Mango) = \frac{350}{650} = 0.54$	$P(Slatko\|Mango) = \frac{440}{650} = 0.67$	$P(Duguljasto\|Mango) = \frac{10}{650} = 0.015$	$P(Mango) = \frac{650}{1200} = 0.54$
Banana	$P(\check{Z}uto\|Banana) = \frac{400}{400} = 1$	$P(Slatko\|Banana) = \frac{300}{400} = 0.75$	$P(Duguljasto\|Banana) = \frac{350}{400} = 0.875$	$P(Banana) = \frac{400}{1200} = 0.33$
Ostalo	$P(\check{Z}uto\|Ostalo) = \frac{50}{150} = 0.33$	$P(Slatko\|Ostalo) = \frac{100}{150} = 0.66$	$P(Duguljasto\|Ostalo) = \frac{50}{150} = 0.33$	$P(Ostalo) = \frac{150}{1200} = 0.125$
Ukupno	**800**	**850**	**400**	**1200**

Klasifikovanje nove voćke

$$h\left(\check{Z}uto\ Slatko\ Duguljasto\right) = \max\{$$

0,22;

0,003;

0,009;

}

Najveća je verovatnoća da je novo voće banana

Ime	Žuto	Slatko	Duguljasto	Ukupno
Mango	$P(\check{Z}uto\|Mango) = \frac{350}{650} = 0.54$	$P(Slatko\|Mango) = \frac{440}{650} = 0.67$	$P(Duguljasto\|Mango) = \frac{10}{650} = 0.015$	$P(Mango) = \frac{650}{1200} = 0.54$
Banana	$P(\check{Z}uto\|Banana) = \frac{400}{400} = 1$	$P(Slatko\|Banana) = \frac{300}{400} = 0.75$	$P(Duguljasto\|Banana) = \frac{350}{400} = 0.875$	$P(Banana) = \frac{400}{1200} = 0.33$
Ostalo	$P(\check{Z}uto\|Ostalo) = \frac{50}{150} = 0.33$	$P(Slatko\|Ostalo) = \frac{100}{150} = 0.66$	$P(Duguljasto\|Ostalo) = \frac{50}{150} = 0.33$	$P(Ostalo) = \frac{150}{1200} = 0.125$
Ukupno	**800**	**850**	**400**	**1200**

Klasifikovanje nove voćke

Zbir ovih verovatnoća nije 1, jer smo izostavili imenilac.

$$(\check{Z}uto\ Slatko\ Duguljasto) = \max\{$$

0,22;

0,003;

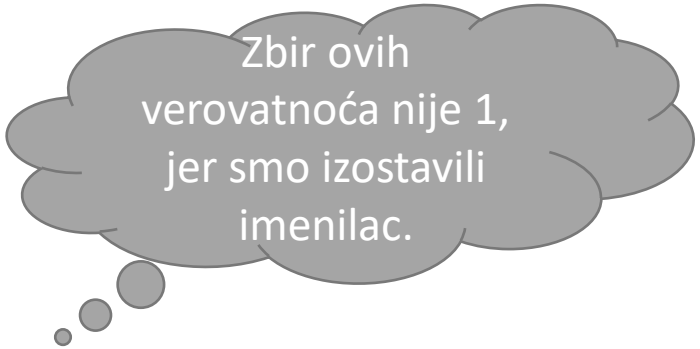
0,009;

}

Najveća je verovatnoća da je novo voće banana

Ime	Žuto	Slatko	Duguljasto	Ukupno
Mango	$P(\check{Z}uto\|Mango) = \frac{350}{650} = 0.54$	$P(Slatko\|Mango) = \frac{440}{650} = 0.67$	$P(Duguljasto\|Mango) = \frac{10}{650} = 0.015$	$P(Mango) = \frac{650}{1200} = 0.54$
Banana	$P(\check{Z}uto\|Banana) = \frac{400}{400} = 1$	$P(Slatko\|Banana) = \frac{300}{400} = 0.75$	$P(Duguljasto\|Banana) = \frac{350}{400} = 0.875$	$P(Banana) = \frac{400}{1200} = 0.33$
Ostalo	$P(\check{Z}uto\|Ostalo) = \frac{50}{150} = 0.33$	$P(Slatko\|Ostalo) = \frac{100}{150} = 0.66$	$P(Duguljasto\|Ostalo) = \frac{50}{150} = 0.33$	$P(Ostalo) = \frac{150}{1200} = 0.125$
Ukupno	**800**	**850**	**400**	**1200**

Klasifikovanje nove voćke



$0{,}22;$

$0{,}003;$

$0{,}009;$

Zato vršimo normalizaciju:

$$\frac{0{,}22}{0{,}22 + 0{,}003 + 0{,}009} = 0{,}95$$

$$\frac{0{,}003}{0{,}22 + 0{,}003 + 0{,}009} = 0{,}01$$

$$\frac{0{,}009}{0{,}22 + 0{,}003 + 0{,}009} = 0{,}04$$

- Dakle, verovatnoća da je voće banana je 0,95.