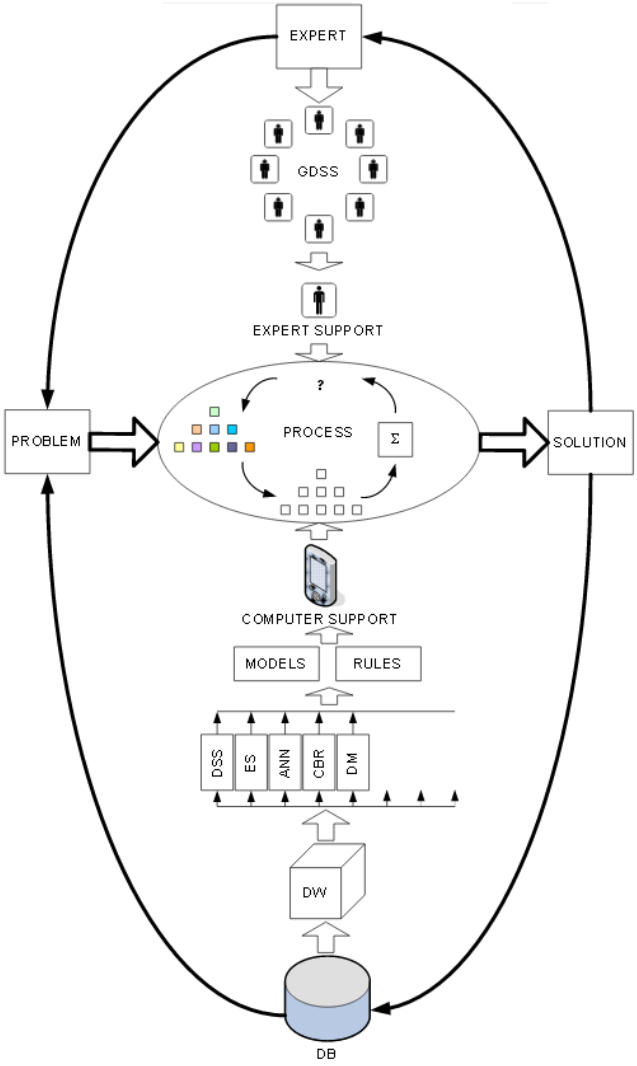
Odlučivanje naučeno iz podataka: Klasifikacija modelom Naivnog Bajesa

Tokom semestra koristili smo znanje eksperta ili grupe eksperata (npr. preferencije ili procene) kako bi modelovali proces donošenja odluka. Drugim rečima, u procesu rešavanja problema DO ima na raspolaganju podršku eksperata. Svaki put kada grupa ili ekspert reše određeni problem oni akumuliraju iskustvo i unapređuju svoje znanje. Ovakav pristup se koristi kada donosilac odluke (DO) ima dovoljno znanja i sposobnosti da reši problem koji se javlja. Međutim, neretko proces donošenja odluka u poslovnom procesu organizacije može brže i bolje da se sprovede ukoliko koristimo računarsku podršku, odnosno koristimo istorijske podatke. Za potrebe ovog predmeta računarska podrška u procesu odlučivanja se može zamisliti na sledeći način. Svako preduzeće vodi evidenciju poslovanja u bazama podataka. U bazi podataka se čuva iskustvo, odnosno znanje koje se može iskoristiti za bolje i brže donošenje odluka.



Slika 1. Model poslovne inteligencije

Problem koji je potrebno rešiti na osnovu iskustva može da se predstavi tabelom slučajeva, odnosno na sledeći način:

RB	x_1	x_2	x_3	y
1	x_{11}	x_{12}	x_{13}	y_1
2	x_{21}	x_{22}	x_{23}	y_2
3	x_{31}	x_{32}	x_{33}	y_3
4	x_{41}	x_{42}	x_{43}	y_4
5	x_{51}	x_{52}	x_{53}	y_5

Redovi predstavljaju slučajeve sa kojima se preduzeće već susrelo u prošlosti i ono predstavlja iskustvo. Kolone predstavljaju (ulazne) atribute i analogni su kriterijumima koje smo koristili. U tabeli su obeleženi sa x_1, x_2 i x_3. Dodatak je kolona y koja predstavlja ciljni atribut, odnosno atribut za koji treba uočiti zakonitosti na osnovu x_1, x_2 i x_3. Vrednosti u tabeli slučajeva predstavljaju stanje uzorka (red) za atribut (kolona). Tako x_{12} predstavlja vrednost prvog uzorka za drugi atribut.

Pravljenje modela odlučivanja naučenog iz podataka ima širok dijapazon primene. Uzmimo za primer bankarsko poslovanje i problem odlučivanja je da li nekome dati kredit ili ne. Svaka banka prilikom odobravanja kredita uzima podatke o klijentima, poput stanja na računu, mesečni prihodi, broj podignutih kredita, broj isplaćenih kredita, vrednost na štednom računu i slično. Ovi podaci predstavljaju atribute (x_1, …, x_5), a kako je reč o istorijskim podacima znamo da li je klijent vratio kredit ili ne (y). Zatim, se umesto eksperta uočavaju veze između ulaznih atributa i izlaznog atributa koje su se javljale u prošlosti (tabeli slučajeva) i na osnovu toga se pravi model odlučivanja. Na kraju, kada se pojavi novi klijent u banci, se uzmu podaci o stanju na računu, mesečnim prihodima, broju podignutih kredita, broju isplaćenih kredita, vrednosti na štednom računu i ostalim atributima, primeni model odlučivanja i dobije se predviđanje da li će klijent vratiti kredit ili ne. Ukoliko će klijent vratiti kredit, njemu se odobrava kredit. U suprotnom klijentu se ne odobrava kredit.

Sledeći primer pravljenja modela odlučivanja jeste prepoznavanje bankrota preduzeća. Svako preduzeće na kraju godine daje izveštaj o svom poslovanju Ministarstvu finansija. Iz izveštaja se mogu uzeti informacije npr. o kapitalu preduzeća (x_1), prodaji (x_2), ukupnim troškovima (x_3) i broju zaposlenih (x_4). Ciljni atribut (y) predstavlja da li je preduzeće bankrotiralo naredne godine. Podaci se uzimaju iz prethodnih godina i time se popunjava tabela slučajeva. Pravi se model odlučivanja i kada preduzeća predaju svoje izveštaje za tekuću godinu propuštaju se kroz model i dobija se predviđanje da li će preduzeće bankrotirati ili ne. Ta informacija može da se koristi kao signal da preduzeće možda nije pogodno u procesu javnih nabavki ili da im je možda potrebno pomoći.

Modeli odlučivanja naučeni iz podataka se primenjuju i u medicini kao pomoć u dijagnostikovanju bolesti. Pacijent dolazi kod doktora sa određenim simptomima. Za primer dijagnostikovanja upale pluća to mogu biti temperatura (x_1), glavobolja (x_2), bol u grudima (x_3), kašalj (x_4) i otežano disanje (x_5), a izlazni atribut je da li pacijent ima upalu pluća ili ne (y). Tokom godina sakuplja se tabela slučajeva, odnosno iskustvo na osnovu kojeg se pravi model odlučivanja. Zatim, kada dođe novi pacijent pogleda se koje simptome od gore pomenutih ima i daje se „predviđanje" da li je pacijent ima upalu pluća ili ne.

Razlog zašto se primenjuju modeli odlučivanja naučeni iz podataka, a ne ekspertsko znanje, može biti u prevelikom broju odluka koje je potrebno doneti pa se odluka može automatizovati. Dakle, proces donošenja odluke primenom modela naučenih iz podataka je brži, a odluka može biti bolja. Zatim, modeli odlučivanja naučeni iz podataka lakše mogu da uoče zavisnosti ciljnog atributa (y) i ulaznih atributa (x_1, x_2, …, x_n), naročito ukoliko postoji veći broj ulaznih atributa. Često ovakvim modelima nije potrebno da vide sva moguća stanja, odnosno kombinacije atributa da bi model mogao da donese odluku. Takođe, ne mora značiti da se pri istim stanjima (vrednostima za ulazne atribute) uvek desiti isti ishod (vrednost ciljnog atributa). Bitno je napomenuti da modeli odlučivanja naučeni iz podataka ne moraju nužno predviđati ishod tačno. Drugim rečima, ovakvi modeli često imaju grešku. Međutim, analiza modela odlučivanja naučenih iz podataka se radi na predmetima u narednim semestrima. Na kraju, modeli odlučivanja naučeni iz podataka nikako ne mogu da zamene DO i da ga oslobode odgovornosti, tj. DO je uvek odgovoran za krajnju odluku.

Modeli odlučivanja naučeni iz podataka predviđaju ishod (y) na osnovu ulaznih atributa (x_1, x_2, …, x_n), a primenjuju se na slučajevima (klijent, pacijent i slično) koje model nije video, odnosno primenjuje se na onim slučajevima za koje se ne zna ishod (y).

Bitno je napomenuti da kvalitet zaključaka zavisi od uzorka u tabeli slučajeva. Odnosno što je veći uzorak (ima više slučajeva) zaključak koji će se dobiti će biti bolji. Takođe, ciljni atribut (y) ne mora nužno da ima samo dva stanja, već može da ima i veći broj stanja. Na primer, rezultat fudbalske utakmice može biti pobeda domaćina, pobeda gosta ili nerešen rezultat.

Klasifikacija modelom Naivnog Bajesa

Banka želi da napravi model odlučivanja za davanje kredita klijentima. Gledajući u bazu podataka dostavili su sledeću tabelu slučajeva. Ona sadrži 10 klijenata koji su opisani preko atributa da li je klijent vlasnik stana, kakav mu je bračni status i koliki su mu prihodi. Sve vrednosti su kategoričke (tj. vrednosti pripadaju nekoj grupi) i za te klijente znamo da li su vratili kredit ili nisu (ciljni atribut).

	x_1	x_2	x_3	y
RB	Vlasnik stana	Bračni status	Prihodi	Vratio?
1	Da	Neoženjen	100+	Da
2	Ne	Oženjen	100+	Da
3	Ne	Neoženjen	50-70	Da
4	Da	Oženjen	100+	Da
5	Ne	Razveden	70-100	Ne
6	Ne	Oženjen	50-70	Da
7	Da	Razveden	100+	Da
8	Ne	Neoženjen	70-100	Ne
9	Ne	Oženjen	50-70	Da
10	Ne	Neoženjen	70-100	Ne

Cilj je napraviti model odlučivanja koji će za svakog novog klijenta dati predviđanje da li će novi klijent vratiti kredit ili ne i na osnovu toga doneti odluku da li mu dati kredit ili ne. Jedan od načina kreiranja modela odlučivanja na osnovu verovatnoća da li će klijent vratiti kredit u zavisnosti od njegovih osobina (ulaznih atributa). Drugim rečima, želimo da modelujemo $p(y|x_1, x_2, x_3)$. Prema Bajesovoj teoremi imamo sledeće:

$$p(y|x_1, x_2, x_3) = \frac{p(y) * \ p(x_1, x_2, x_3|y)}{p(x_1, x_2, x_3)}$$

U ovoj formuli $p(x_1, x_2, x_3|y)$ je problematično za računanje, tj. preko lančanog pravila dobijamo:

$$p(x_1, x_2, x_3|y) = \ p(x_1|y, x_2, x_3) * p(x_2|y, x_3) * p(x_3|y)$$

odnosno

$$p(y|x_1, x_2, x_3) = \frac{p(y) * \ p(x_1|y, x_2, x_3) * p(x_2|y, x_3) * p(x_3|y)}{p(x_1, x_2, x_3)}$$

Zbog toga uvodimo pretpostavku o uslovnoj nezavisnosti ulaznih promenljivih (stoga se algoritam naziva naivni). Nakon uvođenja pretpostavke dobijamo:

$$p(x_1, x_2, x_3|y) = \ p(x_1|y) * p(x_2|y) * p(x_3|y)$$

odnosno

$$p(y|x_1, x_2, x_3) = \frac{p(y) * \ p(x_1|y) * p(x_2|y) * p(x_3|y)}{p(x_1, x_2, x_3)}$$

Lepši zapis dobijamo na sledeći način:

$$p(y|x_1, x_2, x_3) = \frac{p(y) * \prod_{i=1}^{3} p(x_i|y)}{p(x_1, x_2, x_3)}$$

Na kraju, kako $p(x_1, x_2, x_3)$ ne zavisi od y možemo taj deo da eliminišemo jer neće uticati na rezultat. Nakon toga dobijamo:

$$p(y|x_1, x_2, x_3) \sim p(y) * \prod_{i=1}^{3} p(x_i|y)$$

U opštem slučaju može da bude *n* ulaznih atributa umesto tri koliko je u ovom primeru.

Za vraćanje kredita imamo dva moguća ishoda (y). Klijent je vratio kredit i klijent nije vratio kredit. Stoga potrebno je izračunati:

$$p(y = Da|x_1, x_2, x_3) \sim p(y = Da) * p(x_1|y = Da) * p(x_2|y = Da) * p(x_3|y = Da)$$

i

$$p(y = Ne|x_1, x_2, x_3) \sim p(y = Ne) * p(x_1|y = Ne) * p(x_2|y = Ne) * p(x_3|y = Ne)$$

Predviđanje će biti onaj gde je verovatnoća ishoda veća. Odnosno:

$$\hat{y} = argmax_{c \in y}\, p(y = c) * \prod_{i=1}^{n} p(x_i|y = c)$$

gde $\hat{y}$ predstavlja predviđanje, a *c* moguća stanja izlaznog atributa (ishoda).

Dakle, za model odlučivanja potrebno je da izračunamo $p(y)$ i $p(x_i|y)$.

Prvo ćemo izračunati apriori verovatnoće $p(y)$. To ćemo uraditi tako što ćemo u tabeli slučajeva videti koliko se puta javilo stanje $y = Da$ i koliko puta se javilo stanje $y = Ne$.

$$p(y = Da) = \frac{7}{10}$$

$$p(y = Ne) = \frac{3}{10}$$

Zatim računamo $p(x_i|y)$. Dakle, treba da vidimo koliko puta se javljalo stanje atributa x_i po ciljnom atributu y. Za atribut x_1 odnosno Vlasnik stana, imamo sledeću situaciju. Kada je klijent bio vlasnik stana onda je 3 puta vratio kredit, a kako je 7 puta ukupno vraćen kredit dobijamo da je $p(x_1 = Da|y = Da) = \frac{3}{7}$. Analogno tome popunjavamo ostatak tabele.

		Vlasnik stana	
		Da	Ne
Vratio?	Da	$\frac{3}{7}$	$\frac{4}{7}$
	Ne	$\frac{0}{3}$	$\frac{3}{3}$

Na isti način popunjavamo za atribut Bračni status.

		Bračni status		
		Oženjen	Neoženjen	Razveden
Vratio?	Da	$\frac{4}{7}$	$\frac{2}{7}$	$\frac{1}{7}$
	Ne	$\frac{0}{3}$	$\frac{2}{3}$	$\frac{1}{3}$

Na kraju, popunjavamo za atribut Prihodi.

		Prihodi		
		50-70	70-100	100+
Vratio?	Da	$\frac{3}{7}$	$\frac{0}{7}$	$\frac{4}{7}$
	Ne	$\frac{0}{3}$	$\frac{3}{3}$	$\frac{0}{3}$

Ovim postupkom dobili smo model odlučivanja. Kada se pojavi novi klijent možemo da predvidimo da li će taj klijent vratiti kredit ili ne. Recimo, pojavio se klijent. Za njega ne znamo da li će vratiti kredit ili ne (to treba da odlučimo). Ukoliko pogledamo tabelu slučajeva po x_1 i x_2 liči na peti slučaj koji nije vratio kredit, ali liči na sedmi slučaj (x_2 i x_3 su isti) koji je vratio kredit.

	x_1	x_2	x_3	y
RB	Vlasnik stana	Bračni status	Prihodi	Vratio?
Novi	Ne	Razveden	100+	?

Da bi doneli odluku treba da izračunamo:

$$p(y = Da|x_1 = Ne, x_2 = Razveden, x_3 = 100+) \sim p(y = Da) * p(x_1 = Ne|y = Da) * p(x_2 = Razveden|y = Da) * p(x_3 = 100+|y = Da)$$

i

$$p(y = Ne|x_1 = Ne, x_2 = Razveden, x_3 = 100 +) \sim p(y = Ne) * p(x_1 = Ne|y = Ne) * p(x_2 = Razveden|y = Ne) * p(x_3 = 100 + |y = Ne)$$

Ubacivanjem izračunatih vrednosti dobijamo:

$$p(y = Da|x_1 = Ne, x_2 = Razveden, x_3 = 100 +) = \frac{7}{10} * \frac{4}{7} * \frac{1}{7} * \frac{4}{7} = \frac{112}{3430} = 0.0327$$

i

$$p(y = Ne|x_1 = Ne, x_2 = Razveden, x_3 = 100 +) = \frac{3}{10} * \frac{3}{3} * \frac{1}{3} * \frac{0}{3} = \frac{0}{270} = 0$$

Kako je verovatnoća da klijent vratiti kredit veća od verovatnoće da klijent neće vratiti kredit zaključujemo da će klijent vratiti kredit, te njemu možemo dati kredit.

	x_1	x_2	x_3	y
RB	Vlasnik stana	Bračni status	Prihodi	Vratio?
Novi	Ne	Razveden	100+	**Da**

Razlog zašto se dobila verovatnoća 0 za ishod da klijent neće vratiti kredit je taj što nije viđen nijedan slučaj da je klijent sa prihodima preko 100 hiljada dinara nije vratio kredit. Stoga se takav slučaj smatra nemogućim.

Drugi primer

Policija želi da reši problem prevoza ukradenih vozila preko granice. Tokom vremena sakupili su sledeće podatke.

	x_1	x_2	x_3	y
RB	Boja	Tip	Poreklo	Ukraden?
1	Crvena	Sportska	Domaće	Da
2	Crvena	Sportska	Domaće	Ne
3	Crvena	Sportska	Domaće	Da
4	Žuta	Sportska	Domaće	Ne
5	Žuta	Sportska	Uvezeno	Da
6	Žuta	SUV	Uvezeno	Ne
7	Žuta	SUV	Uvezeno	Da
8	Žuta	SUV	Domaće	Ne
9	Crvena	SUV	Uvezeno	Ne
10	Crvena	Sportska	Uvezeno	Da

Pojavilo se novo vozilo za koje su sumnja da je ukradeno. Policija želi da utvrdi da li je vozilo kradeno ili nije u cilju sprovođenja istrage. Novo vozilo je prikazano ispod.

	x_1	x_2	x_3	y
RB	Boja	Tip	Poreklo	Ukraden?
Novi	Crvena	SUV	Domaće	?

Prvi korak je pravljenje modela odlučivanja. Dakle, treba da se izračunaju apriori i uslovne verovatnoće. Dobijamo sledeće apriori verovatnoće:

$$p(y = Da) = \frac{5}{10}$$

$$p(y = Ne) = \frac{5}{10}$$

Zatim, dobijamo sledeće uslovne verovatnoće:

		Boja	
		Crvena	Žuta
Ukraden?	Da	$\frac{3}{5}$	$\frac{2}{5}$
	Ne	$\frac{2}{5}$	$\frac{3}{5}$

		Tip	
		Sportska	SUV
Ukraden?	Da	$\frac{4}{5}$	$\frac{1}{5}$
	Ne	$\frac{2}{5}$	$\frac{3}{5}$

		Poreklo	
		Domaće	Uvezeno
Ukraden?	Da	$\frac{2}{5}$	$\frac{3}{5}$
	Ne	$\frac{3}{5}$	$\frac{2}{5}$

Da bi doneli odluku treba da izračunamo sledeće:

$$p(y = Da|x_1 = Crvena, x_2 = SUV, x_3 = Domaće) \sim p(y = Da) * p(x_1 = Crvena|y = Da) * p(x_2 = SUV|y = Da) * p(x_3 = Domaće|y = Da)$$

i

$$p(y = Ne|x_1 = Crvena, x_2 = SUV, x_3 = Domaće) \sim p(y = Ne) * p(x_1 = Crvena|y = Ne) * p(x_2 = SUV|y = Ne) * p(x_3 = Domaće|y = Ne)$$

Ubacivanjem izračunatih vrednosti dobijamo:

$$p(y = Da|x_1 = Crvena, x_2 = SUV, x_3 = Domaće) = \frac{5}{10} * \frac{3}{5} * \frac{1}{5} * \frac{2}{5} = \frac{30}{1250} = 0.024$$

I

$$p(y = Ne|x_1 = Crvena, x_2 = SUV, x_3 = Domaće) = \frac{5}{10} * \frac{2}{5} * \frac{3}{5} * \frac{3}{5} = \frac{90}{1250} = 0.072$$

Kako je verovatnoća da je automobile nije ukraden veća od verovatnoće da je automobil ukraden zaključujemo da automobil nije ukraden i ne pokrećemo dalju istragu.

	x_1	x_2	x_3	y
RB	Boja	Tip	Poreklo	Ukraden?
Novi	Crvena	SUV	Domaće	**Ne**