

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 4821

Primjena Bayesovog klasifikatora u analizi sportskih rezultata

Bruno Blažeka

Zagreb, svibanj 2017.

Zagreb, 9. ožujka 2017.

ZAVRŠNI ZADATAK br. 4821

Pristupnik: **Bruno Blažeka (0036478296)**
Studij: Računarstvo
Modul: Programsko inženjerstvo i informacijski sustavi

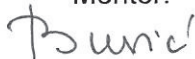
Zadatak: **Primjena Bayesovog klasifikatora u analizi sportskih rezultata**

Opis zadatka:

Tema rada je naivni Bayesov klasifikator. Potrebno je objasniti i implementirati primjenu Bayesovog klasifikatora na predviđanje ishoda utakmica na temelju zadane baze. Korisnik treba pristupati aplikaciji pomoću interaktivnog sučelja.

Zadatak uručen pristupniku: 10. ožujka 2017.
Rok za predaju rada: 9. lipnja 2017.

Mentor:



Doc. dr. sc. Tomislav Burić

Djelovođa:



Doc. dr. sc. Mirjana Domazet-Lošo

Predsjednik odbora za
završni rad modula:



Doc. dr. sc. Ivica Botički

SADRŽAJ

1. Uvod	1
2. Osnovni elementi teorije vjerojatnosti	4
2.1. Uvjetna vjerojatnost i Bayesova formula	4
2.2. Normalna razdioba	8
3. Primjena Bayesovog klasifikatora	13
3.1. Naivni Bayesov klasifikator	14
3.2. Primjena na diskretnom slučaju	15
3.3. Primjena na kontinuiranom slučaju	17
4. Program za analizu sportskih rezultata Bayesovim klasifikatorom	20
4.1. Korištene tehnologije i arhitektura aplikacije	20
4.2. Analiza diskretnih vrijednosti	22
4.3. Analiza kontinuiranih podataka	24
4.4. Testiranje točnosti predviđanja	28
5. Zaključak	29
Literatura	30

1. Uvod

Za temu završnog rada odabrao sam „Primjenu Bayesovog klasifikatora u analizi sportskih rezultata“. Taj problem i temu sam odabrao jer sam odlučio povezati naučena znanja u oblikovanju programske potpore sa svojim interesom za sportsku statistiku i analitiku baza podataka. Zanimalo me ako nad nekim velikim skupom podataka sportske statistike mogu uočiti nekakvu pravilnost i donekle pokušati automatizirati proces zaključivanja na temelju tog uzorka. Nakon konzultacija s mentorom, odlučio sam se za metodu Bayesova klasifikatora jer povezuje interes za sport s velikim bazama podataka i intrigantnom temom strojnog učenja.

Kada je došlo vrijeme za odabrati uzorak pokušao sam odabrati sportsko natjecanje gdje je najlakše primijeniti velike baze i koje će omogućiti najveći skup ulaznih podataka. Odabrao sam sjevernoameričku hokejsku ligu (engl. *National Hockey League*). Radi se o natjecanju s dosta velikim brojem utakmica (30 ekipa igra 82 utakmice redovnog dijela, utakmice doigravanja za prvaka nisam uzeo u obzir zbog njihove nepredvidivosti i dominantne ljudske komponente) tako da vjerujem kako mogu imati dosta veliku bazu utakmica tijekom kojih su sastavi ostali približno jednaki. Zato sam kao bazu odabrao susrete momčadi tijekom posljednje dvije sezone s tim da sam momčadi odabrao one koje su među svima prosječno najbližnjih rezultata. Iako se radi o natjecanju s kompleksnim sustavom pravila kojim se određuje upravo kako bi sve ekipe bile podjednako jake, ekipe mogu varirati u snazi. Uzevši sve to u obzir, pretpostavka kod uzimanja podskupa ekipa omogućava da u svaku utakmicu ekipe, prije analize ulaze s podjednakim šansama za pobjedu.

Među ekipama koje sam uključio u ovaj test je prosječan broj bodova u zadnje dvije godine 184.14, uz standardnu devijaciju od 12.3, a detaljnije možemo vidjeti u tablici 1.1.

Sjevernoamerička hokejska liga je natjecanje koje prema 1.2 ima prosječnu vjerojatnost iznenađujućeg pobjednika među najpopularnijim profesionalnim ligama u Sjevernoj Americi prema Ben-Naim et al. (2006). Prema tom izvoru, natjecanje koje je najmanje predvidio je engleska nogometna liga (engl. *English Premier League*), a

Tablica 1.1: Broj bodova odabranih ekipa u zadnje dvije godine

Ekipa	2015/2016	2016/2017	Ukupno
STL	107	99	206
NSH	96	94	190
BOS	93	95	188
MTL	82	103	185
OTT	85	98	183
EDM	70	103	173
TOR	69	95	164

Tablica 1.2: Vjerojatnost iznenađenja prema Ben-Naim et al. (2006)

Sport	Natjecanje	Vjerojatnost iznenađenja
1. Američki nogomet	NFL	30.9%
2. Košarka	NBA	31.6%
3. Hokej na ledu	NHL	38.3%
4. Bejzbol	MLB	41.3%
5. Nogomet	EPL	45.9%

najpredvidljivije je sjeverno američka liga u američkom nogometu (engl. *National Football League*).

U ovom radu istražujem radi li se o natjecanju u kojem je moguće uočiti pravilnosti koja odudara od standardnih algoritama kojim ljubitelji sporta "traže" pobjednika pa i tako povećati vjerojatnost ako se koncentriramo na druge aspekte. Ako zanemarimo faktore koje ljudi najčešće uzimaju u obzir kao što je nastup u prethodnoj utakmici i pogledamo brojeve koji prikazuju dugoročne tendencije možemo stvoriti model koji će imati možda imati veću točnost. Svoju tvrdnju mogu objasniti na primjeru jedne od varijabli na temelju kojih se vrši strojno učenje, trenutnog nizu pobjeda ili poraza. U ovoj ligi nema velikih nizova pobjeda i poraza što ima smisla upravo zbog okvirne jednake snage ekipa i čak se može reći da veliki niz pobjeda ili poraza obično znači da se tome nizu bliži kraj. Dakle, za sve varijable, broj pobjeda u sezoni i niz pobjeda ili poraza sam pretpostavio da ima normalnu razdiobu.

Prvi dio završnog rada je primjena Bayesova klasifikatora na diskretnom skupu, konkretno, u duhu generalne teme završnog rada, napravio sam programsku potporu koja omogućava korisniku da uz pomoć računala odredi ako su vani odgovarajući uvjeti

za bavljenje ledenim sportovima. Varijable koje se za proračun uzimaju u obzir su: Temperatura, Padaline, Visina snijega i Vjetar. Ispunio sam bazu s mogućim varijacijama navedenih parametara i sukladno njihovoj štetnosti za led sam dodijelio status ako je led siguran ili nije ili možda jednostavno vremenski uvjeti nisu odgovarajući za ledene sportove.

Tablica 1.3: Moguće vrijednosti varijabli

Varijabla	Moguće vrijednosti
Temperatura	negativna, oko nule, pozitivna
Padaline	blagi snijeg, jaki snijeg ili mećava, kiša, nema
Visina snijega	nema, niska, srednja, visoka
Vjetar	jak, slab, umjeren

U drugom dijelu ostvarena je funkcionalnost predviđanja rezultata. Vrijednosti su u ovom slučaju kontinuirane i uzeo sam u obzir da vrijednosti poprimaju normalnu razdiobu. Od korisnika se traži da odabere momčadi koje igraju, unese broj pobjeda u sezoni svake momčadi i niz zadnjih pobjeda i poraza, na temelju toga program će izbaci vjerojatnog pobjednika susreta. Kao bazu podataka bivših susreta program koristi popis međusobnih utakmica iz zadnje dvije godine, između ekipa koje su približno jednako kvalitetne, odabrane za potrebe ove aplikacije.

2. Osnovni elementi teorije vjerojatnosti

U ovom poglavlju objasniti ću matematičku pozadinu ovog završnog rada. Počet ćemo s objašnjavanjem same vjerojatnosti, uvjetne vjerojatnosti pa preko Bayesove formule i Gaussove razdiobe doći objašnjenja Bayesova klasifikatora. Bitno je objasniti neke od osnovnih pojmova kako bi kasnije razumijevanje čitatelju bilo lakše. U ovom poglavlju bit će navedeni osnovni pojmovi koji će biti korišteni u ovom radu.

Tablica 2.1: Objašnjenje najvažnijih pojmova

Izraz	Značenje
$P(h_1)$	apriorna vjerojatnost hipoteze
$P(h_1 A)$	aposteriorna vjerojatnost hipoteze
$P(A h_1)$	izglednost vjerojatnost hipoteze
MAP	maksimalni aposteriori
h_{MAP}	MAP hipoteza

2.1. Uvjetna vjerojatnost i Bayesova formula

Vjerojatnost događaja je mjera izglednosti da će se neki događaj dogoditi. Vjerojatnost događaja A zapisujemo kao $P(A)$. Prema Elezović (2016.a), vjerojatnost je preslikavanje $P : \mathcal{F} \rightarrow [0, 1]$ definirano na algebri događaja \mathcal{F} i ima svojstva:

- normiranost - $P(\Omega) = 1, P(\emptyset) = 0$,
- monotonost - ako je $A \subset B$, onda vrijedi $P(A) \leq P(B)$,
- aditivnost - ako su A i B disjunktni događaji, onda je $P(A \cup B) = P(A) + P(B)$.

Konačni vjerojatnosni prostor Ω je onaj vjerojatnosni prostor koji posjeduje konačno mnogo elementarnih događaja. Označimo njegove elemente, $\Omega = \{\omega_1, \omega_2, \dots, \omega_N\}$. Događaj u ovakvu prostoru je svaki podskup od Ω . Vjerojatnost bilo kojeg događaja možemo odrediti ako znamo vjerojatnost elementarnih događaja, što znači da kad želimo otkriti vjerojatnost skupa elementarnih događaja, jedino što trebamo je pobrojiti sve elementarne događaje i dobijemo vjerojatnost događaja kojeg tražimo.

Za skup elementarnih događaja koji svi imaju jednaku vjerojatnost i za koje vrijedi

$$P(\{\omega_i\}) = \frac{1}{N} \quad (2.1)$$

gdje je N broj elementarnih događaja, kažemo da čine klasični vjerojatnosni prostor. Taj naziv je dobio jer se, prema Elezović (2016.a), problemi iz kojih je iznikla teorija vjerojatnosti mogu opisati ovim modelom. Kada imamo događaj A koji se sastoji od M elementarnih događaja koji su unutar skupa Ω , vjerojatnost tog događaja računamo kao:

$$P(A) = M \cdot \frac{1}{N} = \frac{M}{N} \quad (2.2)$$

U klasičnom vjerojatnosnom prostoru vjerojatnost događaja možemo zapisivati kao broj povoljnih ishoda podijeljen s brojem skupa svih mogućih ishoda, na primjer, kolika je vjerojatnost da na kocki padne specifična brojka ili ista strana novčića. Navedena vjerojatnost računa se izrazom 2.3

$$P(A) = \frac{n(A)}{n(\Omega)} \quad (2.3)$$

Ako skup elementarnih događaja Ω je beskonačan, govorimo o beskonačnom vjerojatnosnom prostoru. U tom slučaju, ako se vodimo logikom koju smo koristili u dosadašnjim razmatranjima, možemo imati problema. Kao primjeri beskonačnog skupa elementarnog događaja su, u Elezović (2016.a), navedeni:

- Bacamo novčić dok se ne pojavi pismo,
- Biramo na sreću realan broj unutar intervala $[0, 1]$. Kolika je vjerojatnost da odaberemo jednu trećinu?

Da bi probleme otklonili trebamo precizno definirati svojstva algebre događaja i pripadne vjerojatnosti.

Ako si postavimo pitanje: "Kolika je vjerojatnost događaja A , ako nam je poznato da se realizirao događaj B ?" Tu vjerojatnost nazivamo *uvjetna vjerojatnost*, a zapisuje se u slučaju iz početnog pitanja kao $P_B(A)$. U Elezović (2016.a) uvjetna vjerojatnost je definirana na način:

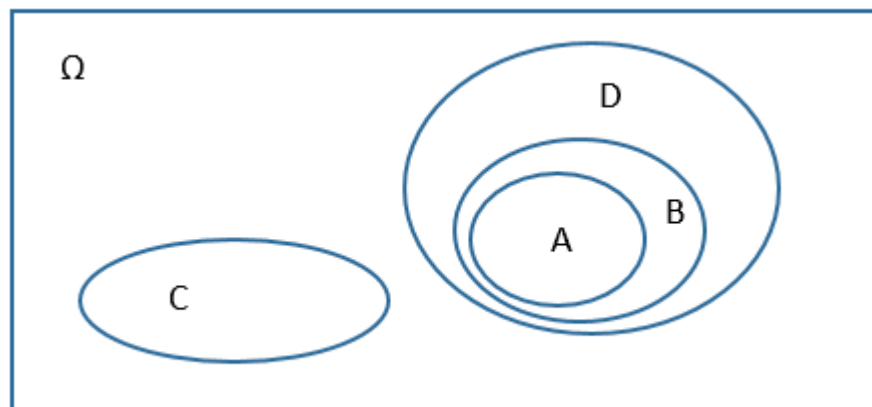
Neka je $B \in \mathcal{F}$ događaj pozitivne vjerojatnosti: $P(B) > 0$. Uvjetna vjerojatnost uz uvjet B je funkcija $P_B : \mathcal{F} \rightarrow [0, 1]$ i definirana je formulom

$$P_B(A) = \frac{P(AB)}{P(B)}, \forall A \in \mathcal{F}. \quad (2.4)$$

Kada bi htjeli znati kolika je vjerojatnost da je dvaput zaredom pala ista strana novčića, na primjer glava, i s obzirom na to da su rezultati dva bacanja nezavisni, tada bi trebali koristiti *Uvjetnu vjerojatnost nezavisnih događaja*. Tražena uvjetna vjerojatnost dana je izrazom 2.5:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (2.5)$$

Tu jednadžbu ću objasniti prema primjeru u Pauše (1988), Ako zamislimo da imamo stanje kao što je prikazano slikom 2.1 i želimo utvrditi uvjetne vjerojatnosti za $P(A|B)$, $P(C|B)$ i $P(D|B)$. Za ishode čiji skupovi nemaju presjek kao što su skupovi C i B je uvjetna vjerojatnost nula. Za skupove ishoda koji su nadskup nekog drugog, kao što je odnos skupa D i skupa B, vjerojatnost je jednaka jedan jer čim se dogodio događaj B se dogodio i D. Ostale vjerojatnosti računamo kao umnožak vjerojatnosti tih ishoda jer taj ishod koji želimo proračunati može i ne mora nastupiti. Te događaje nazivamo *stohastički nezavisnima*. Ovim proračunom smo na skupovima vidjeli primjenu formule 2.5.

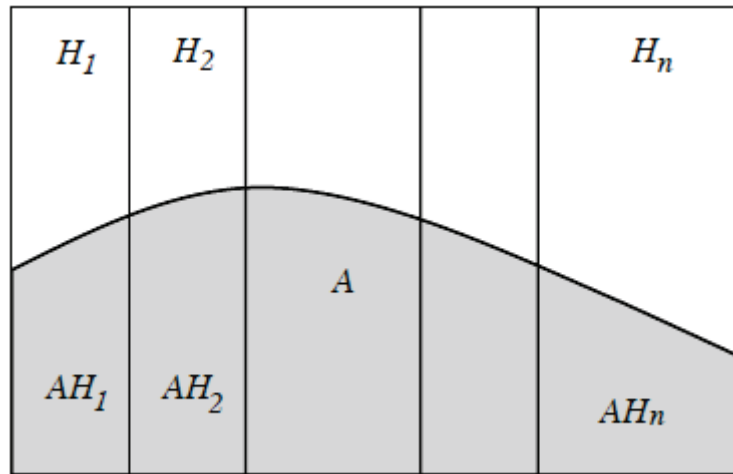


Slika 2.1: Primjer skupova

Ako pretpostavimo da skup elementarnih događaja rastavimo na n međusobno disjunktih događaja, kao u 2.2, za koje vrijedi:

$$\Omega = H_1 \cup H_2 \cup \dots \cup H_n \quad (2.6)$$

pri čemu su događaji disjunktne, taj rastav nazivamo particija vjerojatnosnog prostora. kažemo još da skup događaja čini potpun sustav događaja.



Slika 2.2: Rastav skupa Ω na disjunktne skupove

Prema Pauše (1988), ako cijeli vjerojatnosni prostor podijelimo na n dijelova koji se međusobno isključuju te za koje vrijede sljedeće jednakosti

$$\bigcup_{i=1}^n B_i = \Omega \text{ i } \sum_{i=1}^n P(B_i) = 1,$$

te uzmemo proizvoljni događaj A i uzmemo u obzir međusobno isključivanje, možemo reći da vrijedi:

$$\sum_{i=1}^n P(A \cap B_i) = P(A) \quad (2.7)$$

Primijenimo li na navedenu formulu izvedeni oblik formule 2.5 oblika $P(A \cap B) = P(A|B) \cdot P(B)$ dobijemo formulu koju zovemo *formula totalne vjerojatnosti*.

$$\sum_{i=1}^n P(A|H_i)P(H_i) = P(A) \quad (2.8)$$

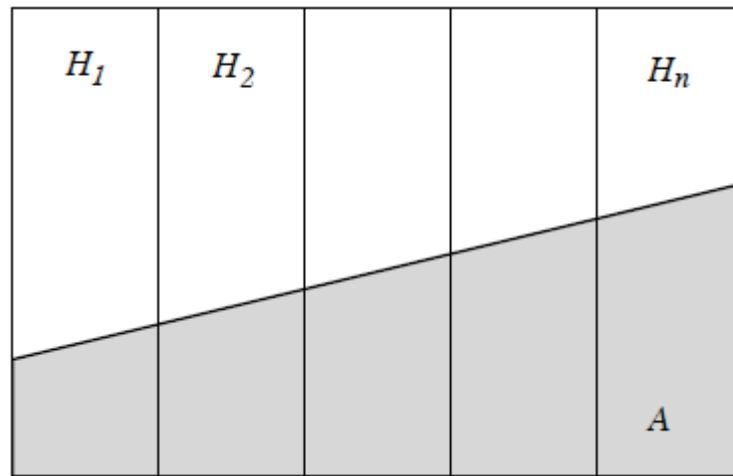
Koristeći sve navedene formule iz formule uvjetne vjerojatnosti dobivamo Bayesovu formulu koja je dobila ime po engleskom matematičaru Thomasu Bayesu.

$$P(H_i|A) = \frac{P(A|H_i)P(H_i)}{\sum_{j=1}^n P(A|H_j)P(H_j)} \quad (2.9)$$

"Relacija 2.9 se naziva Bayesova formula i ona se primjenjuje za računanje uvjetne vjerojatnosti hipoteze H_i kada se zna da je nastupio događaj A koji inače nastupa uvijek s jednom od hipoteza $P(H_i)$ gdje je $j = 1, \dots, n$." Pauše (1988)

Bayesovu formulu koristimo pri računanju aposteriornih vjerojatnosti pojedinih hipoteza. Prije početka pokusa svaka hipoteza ima svoju vjerojatnost realizacije $P(H_i)$.

Nakon realizacije pokusa, ako znamo koji se elementarni događaj ostvario, tad je nestala neizvjesnost: ostvarila se samo jedna od mogućih hipoteza H_1, \dots, H_n , dok za sve ostale znamo sa sigurnošću da se nisu ostvarile. Pretpostavimo međutim da nam nije poznato koji se elementarni događaj ostvario, već umjesto toga znamo da se ostvario događaj $A \subset \Omega$. U tom slučaju ne znamo točno koja je od hipoteza H_1, \dots, H_n nastupila, ali dodatna informacija o realizaciji događaja A mijenja apriorne vjerojatnosti pojedinih hipoteza. Pomoću Bayesove formule računamo uvjetne vjerojatnosti $P(H_1|A), \dots, P(H_n|A)$, koje nazivamo aposteriornim vjerojatnostima pojedinih hipoteza.



Slika 2.3: Interpretacija situacije kada su apriorne vjerojatnosti svih hipoteza jednake i nakon realizacije A (označeno sivim područjem) se vjerojatnosti pojedinih hipoteza mijenjaju

2.2. Normalna razdioba

Kad se bavimo s kontinuiranim podacima, pretpostavljamo da se radi o normalnoj razdiobi poznatoj po nazivu i Gaussova razdioba. Prema Elezović (2016.b), normalna razdioba je najvažnija neprekinuta razdioba ima li se u vidu učestalost i važnost modela u kojima se ona pojavljuje. Razlog tome je što se ta razdioba javlja kao granična u svim situacijama kad je slučajna varijabla dobivena kao zbroj velikog broja međusobno nezavisnih pribrojnika. Pauše (1988) kaže: "Gaussova razdioba je najvažniji primjer kontinuirane distribucije vjerojatnosti. Jednadžba po kojoj se računa vrijednost Gaussove razdiobe je:

$$p(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(t - \mu)^2}{2\sigma^2}\right) \quad (2.10)$$

Tu su μ i $\sigma > 0$ zadane konstante, tako da se govori o normalnoj distribuciji $\mathcal{N}(\mu, \sigma^2)$ s parametrima μ i σ ." Spomenute varijable predstavljaju za skup brojeva:

- μ medijan skupa,
- σ standardnu devijaciju skupa,
- σ^2 varijancu skupa.

Ako imamo X neprekinutu slučajnu varijablu čija je gustoća dana 2.10, možemo pisati $X \sim \mathcal{N}(\mu, \sigma^2)$

Analizom grafova prikazanih na slici 2.4 vidimo da se promjenom medijana danog simbolom μ pomiče tjeme funkcije lijevo smanjivanjem i desno povećavanjem. S druge strane, na slici 2.5 vidimo da povećanjem parametra σ graf postaje širi, a smanjenjem postaje uži. Prema tome zaključujemo da je strmina proporcionalna sa σ .

Iz Elezović (2016.b) izvući ću nekoliko svojstava normalne razdiobe, na primjer, izaberemo li parametre $a = 0$ i $\sigma = 1$ dobivamo slučajnu varijablu $\mathcal{N}(0, 1)$ i nju nazivamo *jedinična normalna razdioba*. Njezinu jednadžbu razdiobe možemo zapisati izrazom 2.11.

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right) \quad (2.11)$$

Ako imamo X koji je jedinična normalna razdioba, a bilo koji realni i σ bilo koji pozitivni broj. Definiramo slučajnu varijablu

$$Y = a + \sigma X$$

Za nju možemo izvesti jednadžbu normalne razdiobe identičnu kao 2.10. Možemo reći da se jedinična i opća normalna razdioba mogu dobiti jedna iz druge *linearnom transformacijom*.

Još jedno od svojstava normalne razdiobe je pravilo 3σ , to pravilo kaže da je normalna varijabla unutar intervala $a + 3\sigma$ i $a - 3\sigma$ s vjerojatnošću od 99.73

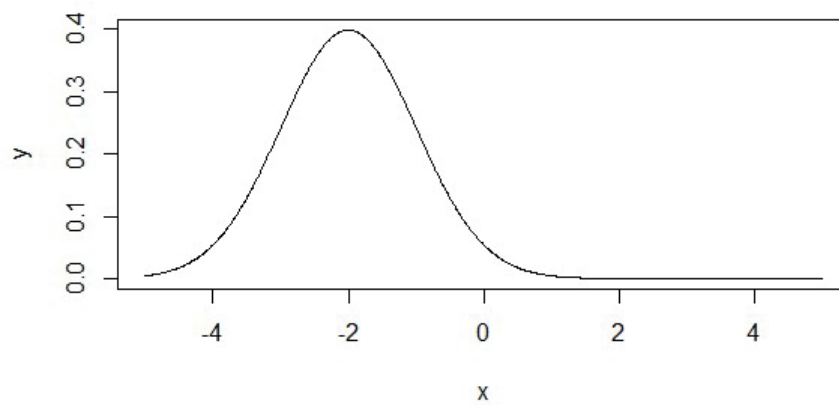
$$X \sim \mathcal{N}(0, 1) \implies a + \sigma X \sim \mathcal{N}(\mu, \sigma^2)$$

$$X \sim \mathcal{N}(\mu, \sigma^2) \implies \frac{X - a}{\sigma} \sim \mathcal{N}(0, 1).$$

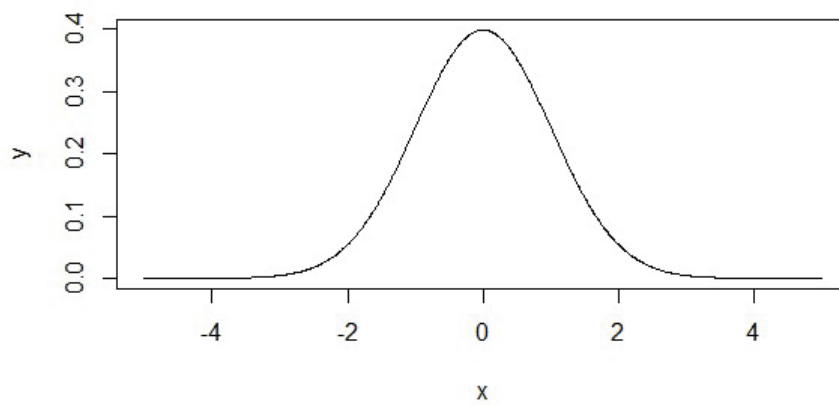
Stabilnost je svojstvo razdioba kad zbroj nezavisnih slučajnih varijabli ima razdiobu istog tipa, normalna razdioba je jedina koja ima pojačano svojstvo stabilnosti.

Neka imamo X_1 i X_2 i neka su nezavisne slučajne varijable s normalnim razdiobama

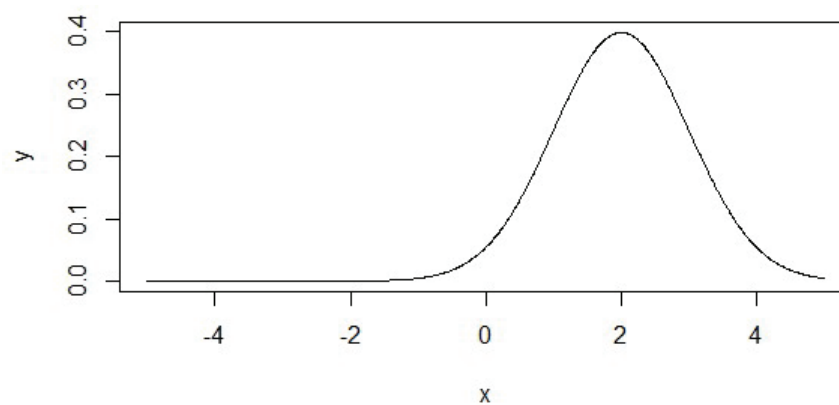
$$X_1 \sim \mathcal{N}(a_1, \sigma_1^2), X_2 \sim \mathcal{N}(a_2, \sigma_2^2)$$



(a) Graf Gaussove razdiobe s $\mu = -2$ i $\sigma = 1$

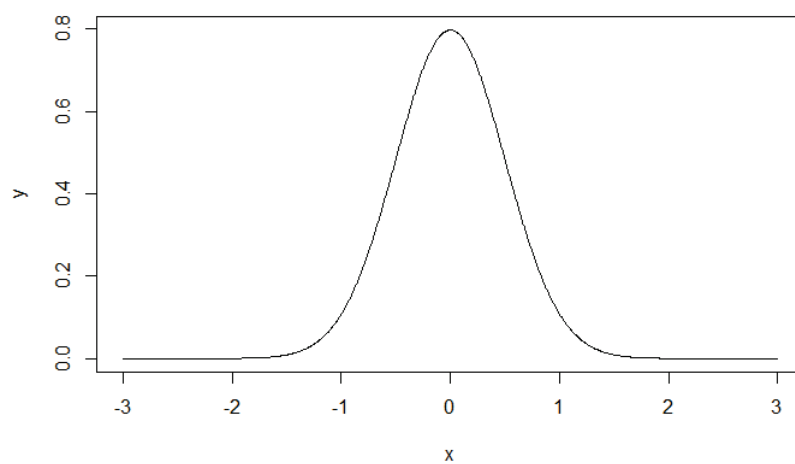


(b) Graf Gaussove razdiobe s $\mu = 0$ i $\sigma = 1$

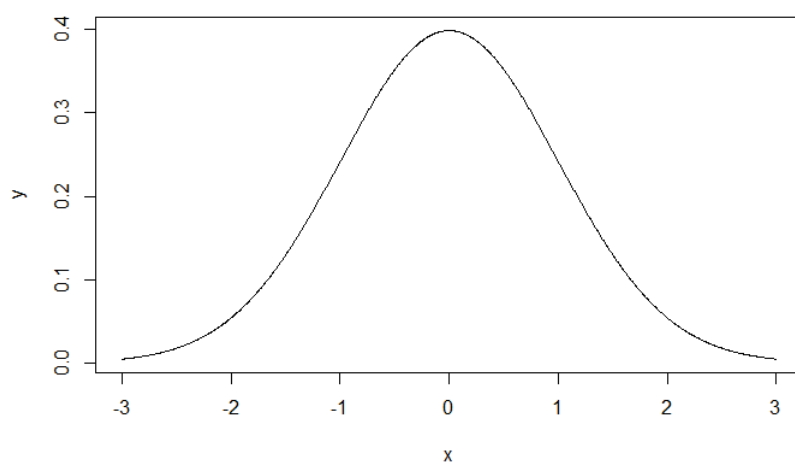


(c) Graf Gaussove razdiobe s $\mu = 2$ i $\sigma = 1$

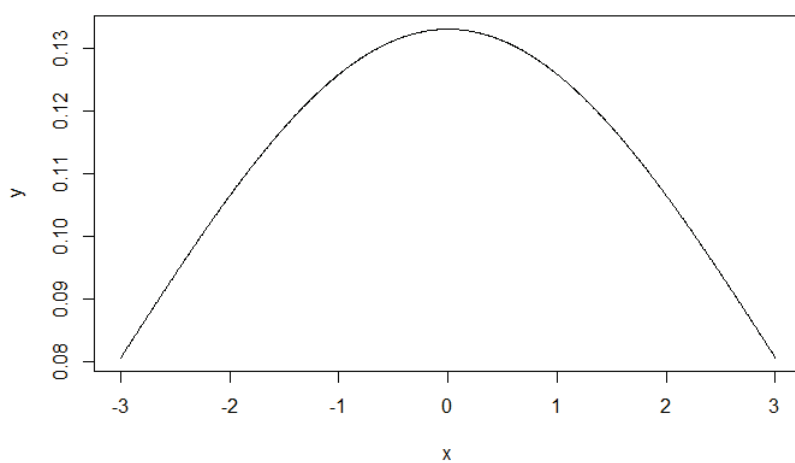
Slika 2.4: Grafovi razdioba s različitim parametrom μ



(a) Graf Gaussove razdiobe s $\mu = 0$ i $\sigma = 0.5$

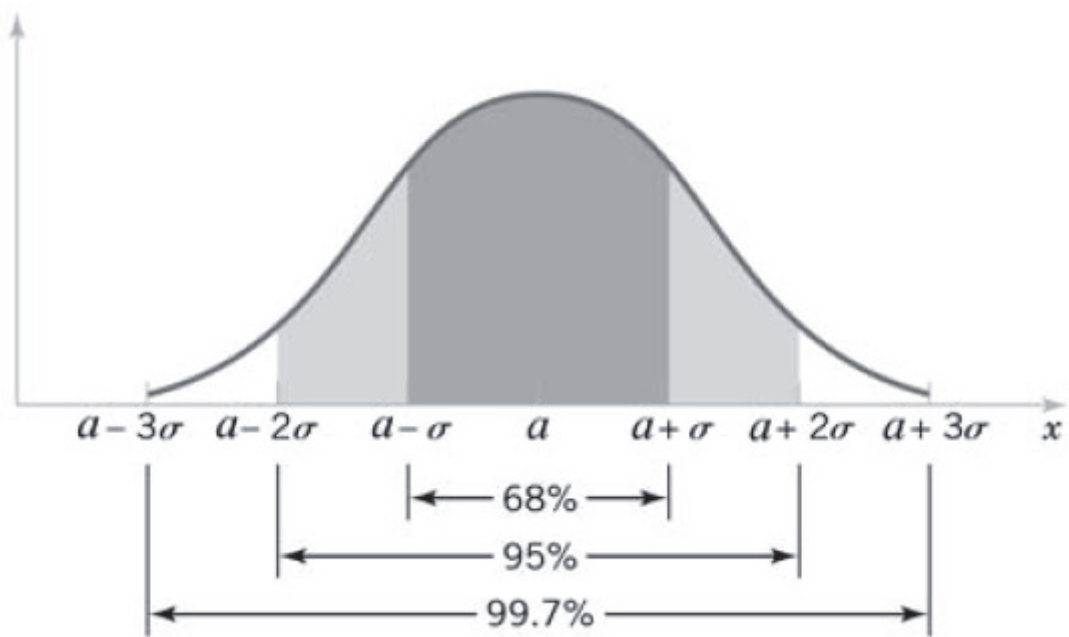


(b) Graf Gaussove razdiobe s $\mu = 0$ i $\sigma = 1$



(c) Graf Gaussove razdiobe s $\mu = 0$ i $\sigma = 3$

Slika 2.5: Grafovi razdioba s različitim parametrom σ



Slika 2.6: Prikaz pravila 3σ iz Elezović (2016.b)

i za s_1 i s_2 koji su bilo koji realni brojevi tada vrijedi

$$s_1X_1 + s_2X_2 \sim \mathcal{N}(s_1a_1 + s_2a_2, s_1^2\sigma_1^2 + s_2^2\sigma_2^2).$$

Čitatelj zainteresiran za ostale elemente teorije vjerojatnosti, bilo da su u poglavlju 2.1 ili 2.2 može pročitati u navedenoj literaturi

3. Primjena Bayesovog klasifikatora

Metoda naivnog Bayesova klasifikatora (engl. *Naive Bayes classifier*) je jedna od najjednostavnijih metoda među metodama strojnog učenja te je klasifikator zbog svoje "naivne" pretpostavke upravo dobio taj naziv. Temeljena je na Bayesovoj formuli obrađenoj u prethodnom poglavlju i ima naziv klasifikator jer temeljem obilježja uzorka pridružuje razred.

U ovom poglavlju ću objasniti primjenu Bayesova klasifikatora na dva tipa slučajeva kojim sam se susretao tijekom izrade ovog rada. Primjeri su jedan koji ima diskretne vrijednosti, a drugi koji ima kontinuirane. Neke od primjena metode naivnog Bayesova klasifikatora u stvarnom svijetu su:

- medicinske dijagnoze i proučavanje karakteristika genoma,
- filtriranje neželjene elektroničke pošte,
- predlaganje upita u tražilicama i zanimljivih tekstova,
- klasifikacija članaka po tipu (politika, sport, glazba, ...),
- prepoznavanje lica i emocija,
- prepoznavanje teksta,
- vremenska prognoza.

Diskretan slučaj će biti zaključivanje koji su vremenski uvjeti pogodni za vožnju bicikla u prirodi, a konkretan slučaj će biti klasifikacija ljudi po spolu uzevši u obzir njihovu visinu, težinu i veličinu stopala.

3.1. Naivni Bayesov klasifikator

Pretpostavka naivnog Bayesova klasifikatora je da su sva obilježja uvjetno nezavisna u odnosu na razred.

$$p(x|y = c) = \prod_{i=1}^D p(x_i|y = c) \quad (3.1)$$

Prema Murphy (2006) ta pretpostavka da su sva obilježja nezavisna je obično pogrešna, ali unatoč tome Bayesov klasifikator se uspijeva natjecati s kompleksnijim klasifikatorima. U slučaju normalne razdiobe, slučaj koji je korišten u ovom radu, imamo:

$$p(x|y = c) = \prod_{i=1}^D \mathcal{N}(x_i|\mu_{ic}, \sigma_{ic}) \quad (3.2)$$

Od metoda klasifikacije najjednostavnija je *diskriminativna funkcija* koja je prikazana u izrazu 3.3 i kada iskoristimo Bayesovu formulu 2.9 dobijemo 3.4.

$$h_{MAP} = \operatorname{argmax}_{h_i \in H} P(h_i|D) \quad (3.3)$$

$$h_{MAP} = \operatorname{argmax}_{h_i \in H} \frac{P(D|h_i)P(h_i)}{P(D)} \quad (3.4)$$

Sada uzmemo u obzir da je $P(D)$ konstanta, možemo izostaviti nazivnik i imamo:

$$h_{MAP} = \operatorname{argmax}_{h_i \in H} P(D|h_i)P(h_i) \quad (3.5)$$

Naime, prema Rish (2001), unatoč nerealističnim pretpostavkama nezavisnosti, naivni Bayesov klasifikator je iznenađujuće efikasan u praksi jer njegova klasifikacijska odluka može ponekad biti točna iako su procjene vjerojatnosti netočne. Pokazano je kako naivni Bayes najbolje djeluje u dva slučaja, kada su varijable potpuno nezavisne što je i očekivano i kod funkcionalno zavisnih obilježja što nije očekivano, dok najgore rezultate ostvaruje između ta dva ekstrema.

Leung (2007) kaže da su studije prikazale da Bayesov klasifikator ima usporedive sposobnosti klasifikacije s klasifikatorima stabla odluke (engl. *Decision Tree Classifier*) i odabrane neuronske mreže (engl. *Neural Network Classifier*) te da je klasifikator prikazao veliku točnost i brzinu kod velikih baza podataka.

Sve u svemu, metoda naivnog Bayesova klasifikatora može biti primijenjena na različite probleme. Primjena na problemima s diskretnim i kontinuiranim vrijednostima pokazala je da je efikasan za primjere koji jasno pripadaju nekom skupu kao što je na primjer kišan i vjetrovit dan za ostati kod kuće i isto tako je spreman predvidjeti spol za osobe koji imaju predvidljive proporcije. S druge strane vidimo da kada bi

imali osobu koja je viša ili manja od većine pripadnika populacije svog spola, klasifikator jer bi je svrstao samo po fizičkim parametrima, a ne po nekim drugim koji su bitniji. Ukratko, zaključujem da je za efikasnu metodu potrebno izraditi dobar model da se ne zapostave bitni parametri i napuniti bazu podataka s čim više adekvatnih podataka.

3.2. Primjena na diskretnom slučaju

Primjenu Bayesova klasifikatora na diskretnom slučaju ću objasniti na primjeru određivanja jesu li vremenski uvjeti pogodni za vožnju bicikla. Razmatranjem mogućih atmosferskih uvjeta koji utječu na vožnju biciklom dobio sam popis uvjeta koji je naveden u tablici 3.1. Različite kombinacije uvjeta znače da su uvjeti pogodni ili nisu pogodni za vožnju biciklom.

Tablica 3.1: Moguće vrijednosti varijabli za vožnju biciklom

Varijabla	Moguće vrijednosti
Temperatura	visoka, ugodna, niska
Vrijeme	sunčano, oblačno, kišno
Vjetar	jak, umjeren, slab

Od podataka navedenih u tablici 3.1, visoka i niska temperatura negativno utječu na mogućnost bicikliranja, dok ugodna temperatura povećava mogućnost bicikliranja, sunčano i oblačno vrijeme pozitivno utječu na uvjete bicikliranja, a kišno negativno i tako dalje. Treba napomenuti da recimo visoka temperatura i oblačno vrijeme daju rezultat da je vrijeme za bicikliranje u redu kao i kombinacija sunčano i ugodna temperatura. Vjetar odgovara ako je umjeren ili slab. Kako bi mogli izraditi model, uzimamo skup podataka iz tablice i na temelju njih možemo proanalizirati nekoliko primjera i stvoriti sustav za računanje prihvatljivosti vremenskih uvjeta. Za računanje posterior vrijednosti u diskretnom slučaju koristimo sljedeće formule:

$$h_{MAP} = \operatorname{argmax}_{h_i \in H} P(h_i | D)$$

$$h_{NB} = \operatorname{argmax}_{h \in [da, ne]} P(h) P(x | h)$$

$$h_{NB} = \operatorname{argmax}_{h \in [da, ne]} P(h) \prod_t P(a_t | h)$$

Tablica 3.2: Tablica podataka za bicikl

Dan	Temperatura	Vrijeme	Jačina vjetra	Voziti bicikl?
1	visoka	sunčano	slab	NE
2	ugodna	sunčano	umjeren	DA
3	visoka	oblačno	umjeren	DA
4	niska	sunčano	jak	NE
5	ugodna	kišno	umjeren	NE
6	ugodna	sunčano	slab	DA
7	visoka	oblačno	jak	NE
8	niska	kišno	slab	NE
9	visoka	sunčano	jak	NE
10	ugodna	oblačno	slab	DA
11	niska	sunčano	slab	DA
12	ugodna	oblačno	umjeren	DA
13	niska	oblačno	umjeren	NE
14	visoka	sunčano	umjeren	NE
15	visoka	oblačno	slab	DA
16	niska	kišno	jak	NE

Kako sada imamo skup podataka u tablici 3.2 možemo klasificirati novi podatak, recimo da imamo dan koji nisko temperaturu, sunčano vrijeme i umjeren vjetar, računamo prema prethodnim jednadžbama:

$$\begin{aligned}
P(\text{biciklirati} = da) &= \frac{7}{16} \\
P(\text{biciklirati} = ne) &= \frac{9}{16} \\
P(\text{temperatura} = niska | \text{biciklirati} = da) &= \frac{1}{7} \\
P(\text{temperatura} = niska | \text{biciklirati} = ne) &= \frac{4}{9} \\
P(\text{vrijeme} = sunčano | \text{biciklirati} = da) &= \frac{3}{7} \\
P(\text{vrijeme} = sunčano | \text{biciklirati} = ne) &= \frac{4}{9} \\
P(\text{vjetar} = umjeren | \text{biciklirati} = da) &= \frac{3}{7} \\
P(\text{vjetar} = umjeren | \text{biciklirati} = ne) &= \frac{3}{9} \\
P(da)P(niska|da)P(sunčano|da)P(umjeren|da) &= \frac{9}{784} = \mathbf{0.011} \\
P(ne)P(niska|ne)P(sunčano|ne)P(umjeren|ne) &= \frac{1}{27} = \mathbf{0.037}
\end{aligned}$$

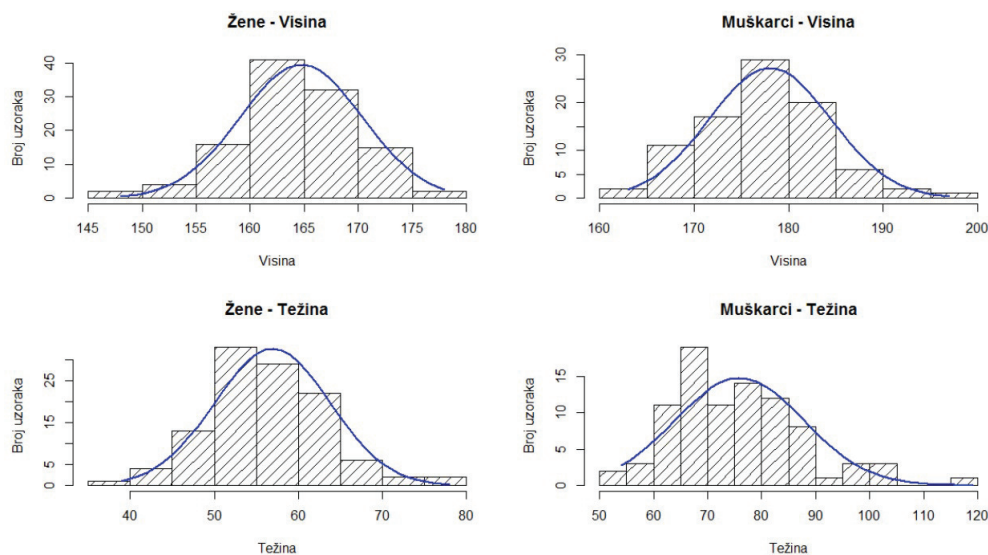
Na temelju izračunatih vrijednosti zaključujemo da za dani primjer će prijedlog

biti da ne idemo na bicikliranje. Takav rezultat svakako ima smisla jer iako je vrijeme sunčano, niska temperatura i umjeren vjetar će učiniti da nam bude hladno.

3.3. Primjena na kontinuiranom slučaju

Primjenu Bayesova klasifikatora objasniti ću na primjeru raspodjele visine, težine i veličine stopala kod muškaraca i žena. Raspoložemo s podacima o visini i težini muškaraca i žena te ću na temelju danih podataka prikazati postupak izračunavanja klasifikatore pomoću kojih će biti moguće odrediti najvjerojatniji spol prema naivnom Bayesovom klasifikatoru temeljem samo navedenih podataka.

Raspodjela visine i težine muškaraca i žena unutar populacije od 200 ljudi prikupljene tijekom ankete o samoprocjeni visine i težine (engl. *Self-Reports of Height and Weight*) je prikazana slikom 3.1



Slika 3.1: Raspodjele visine i težine u žena i muškaraca

Počnimo s pretpostavkom kako je vjerojatnost da je osoba muškog i ženskog spola jednaka, točnije: $P(M) = P(Z) = 0.5$

Imamo sljedeću tablicu već utvrđenih vrijednosti visine, težine i duljine stopala za pet osoba ženskog spola:

Analognu tablicu za pet osoba muškog spola s istim podacima:

Idući korak je računanje medijana i varijance za svaki stupac (tip podataka) koje imamo u tablicu za svaki spol odvojeno. Iznosi izračunatih vrijednosti su navedeni u tablici:

Tablica 3.3: Tablica pet uzoraka za žene

Spol	Visina (cm)	Težina (kg)	Stopalo (cm)
Ž	153	45	22
Ž	156	48	23
Ž	162	60	24
Ž	167	58	25
Ž	172	68	26

Tablica 3.4: Tablica pet uzoraka za muškarce

Spol	Visina (cm)	Težina (kg)	Stopalo (cm)
M	170	70	26
M	176	72	27
M	178	75	27.5
M	182	83	28.5
M	186	89	29

Uzimamo prvi probni uzorak, gdje je osoba visine 180 cm, težine 75 kg i veličine stopala 28 cm i za koju ćemo odrediti spol. Primjenom jednadžbe 3.6 ćemo izračunati posteriori vrijednost za spol Ž ili M koristeći podatke u probnom uzorku:

$$P(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_{y'=1}^C p(x|y')p(y')} \quad (3.6)$$

U jednadžbu ubacujemo podatke potrebne u našem primjeru te ona za računanje posteriori vrijednosti spola žena i muškarac glasi:

$$\begin{aligned} skup = & P(M)p(visina|M)p(težina|M)p(stopalo|M)+ \\ & P(Z)p(visina|Z)p(težina|Z)p(stopalo|Z) \end{aligned}$$

Tablica 3.5: Tablica izračunatih vrijednosti

Spol	$\mu(visina)$	$\sigma^2(visina)$	$\mu(težina)$	$\sigma^2(težina)$	$\mu(stopalo)$	$\sigma^2(stopalo)$
Ž	162	6.957	55.8	8.352	24	1.414
M	178.4	5.426	77.8	7.139	27.5	1.068

$$P(Z|uzorak) = \frac{P(Z)p(visina|Z)p(tezina|Z)p(stopalo|Z)}{skup} \quad (3.7)$$

$$P(M|uzorak) = \frac{P(M)p(visina|M)p(tezina|M)p(stopalo|M)}{skup} \quad (3.8)$$

U navedenim jednadžbama, nazivnik se može zanemariti jer je vrijednost jednadžbe 3.3 konstanta.

Nakon što izračunamo tražene podatke, dobijemo sljedeće podatke, prikazane u tablici 3.6:

Tablica 3.6: Tablica izračunatih posteriori vrijednosti za prvi primjer

X	$p(visina X)$	$p(tezina X)$	$p(stopalo X)$
Ž	1.166×10^{-11}	3.594×10^{-11}	1.171×10^{-3}
M	1.353×10^{-1}	8.622×10^{-2}	3.434×10^{-1}

Analizom podataka izračunatih u tablici 3.6 vidimo da su vrijednosti vezani uz muški spol značajno veće, kada bi uvrstili navedene vrijednosti u jednadžbe za računanje posteriori vrijednosti vidjeli bi da je vjerojatnost da su mjere u uzorku mjere muške osobe značajno veća i mogli bi s velikom dozom sigurnosti zaključiti da je osoba u uzorku muškog spola.

Ako uzmemo drugi primjer, koji nije toliko očit već se nalazi na granici između dva moguća spola, rezultat će zahtijevati proračun. Uzmimo da imamo uzorak visine 171 cm, težine 65 kg, veličine stopala 26 cm. Njegove posteriori vrijednosti dane tablicom 3.7

Tablica 3.7: Tablica izračunatih posteriori vrijednosti za drugi primjer

X	$p(visina X)$	$p(tezina X)$	$p(stopalo X)$
Ž	4.482×10^{-4}	8.698×10^{-4}	8.155×10^{-2}
M	1.102×10^{-3}	1.551×10^{-6}	1.346×10^{-1}

Kao što vidimo u drugom primjeru, ovaj put posteriori vrijednosti su mnogo bliže pa zaključak neće biti moguć bez konkretnih rezultata. I ovaj put možemo zanemariti nazivnik u jednadžbama jer je evidence konstanta na obje strane.

$$P(M) = 1.150 \times 10^{-10} \quad P(Z) = 1.590 \times 10^{-8}$$

Vidimo da je posteriori vrijednost veća za žene pa zaključujemo da je uzorak osoba ženskog spola.

4. Program za analizu sportskih rezultata Bayesovim klasifikatorom

U idućem poglavlju govorim o aplikaciji koja je predmet ovog rada. Na početku opisujem korištene tehnologije, a kasnije ću opisati implementirane funkcionalnosti, način za korištenje i na koji način se primjenjuje Bayesov klasifikator u postupku učenja.

4.1. Korištene tehnologije i arhitektura aplikacije

Završni rad je napravljen kao desktop aplikacija koja tijekom svog rada zahtjeva stalnu vezu s Internetom jer su svi podaci potrebni za rad aplikacije pohranjeni na oblaku (engl. *cloud*) i to na Microsoftovoj platformi Azure. Koristim server u sklopu Azure platforme i na tom serveru nalazi se potrebna baza podataka. Relevantan sadržaj baze podataka naveden je u tablici 4.1

Tablica 4.1: Sadržaj tablica baze podataka

Sadržaj	Naziv tablice	Broj zapisa	Podataka
Vremenski uvjeti	Conditions	23	8 KB
Utakmice	Games	150	24 KB

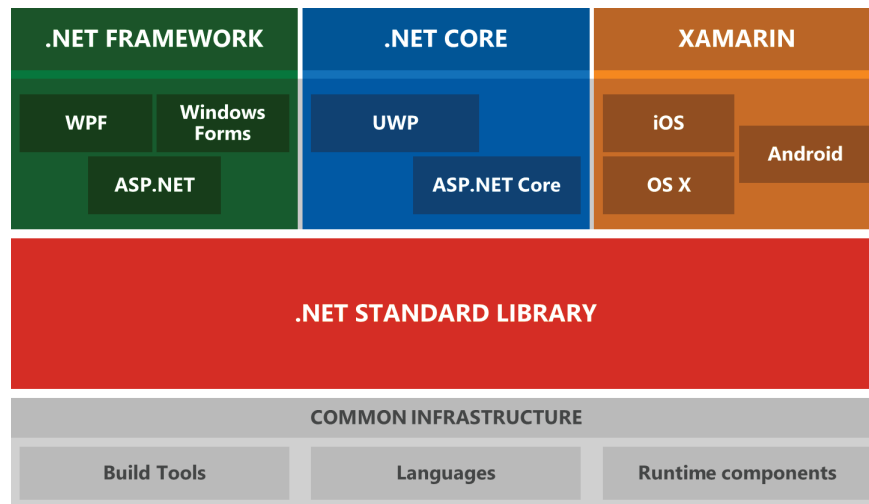
Tijekom razvoja korišteno je razvojno okruženje Visual Studio koje pruža napredne funkcionalnosti potrebne za jednostavan razvoj aplikacija. Korištena je konkretno verzija Visual Studio 2015 Community.

Korišten je .NET Entity Framework, koji je dio .NET platforme. Prema Community (b), .NET se sastoji od velikog broja ključnih komponenti. Ima standardnu biblioteku koja se zove *.NET Standard Library* i koja je veliki skup API-ja (engl. *Application programming interface*) koje se mogu pokrenuti svugdje. Ta standardna biblioteka

je implementirana od tri .NET okoline tijekom rada programa. (engl. *runtime*),

- .NET Framework,
- .NET Core,
- Mono (za Xamarin)

Svaki od .NET *runtime-a* radi s .NET jezicima. Dodatno, svaka platforma ima zasebne alate za *build* projekata. Ti alati su isti neovisno o odabranom *runtime-u*. Grafički prikaz je dan u 4.1.



Slika 4.1: .NET komponente arhitekture, prema Community (b)

.NET standardna biblioteka je skup API-ja koja su implementirana u .NET okruženju u trenutku izvođenja. Formalno rečeno to je skup API-ja koji čine uniformni set ugovora koji po kojima se kompajlira kod. Ti ugovori su građeni prema implementaciji za svaki od .NET okruženja. On omogućava prenosivost preko različitih okruženja i čini da kod može biti uspješno pokretan svugdje.

Navedeni .NET Framework je primjer .NET *runtimea*. Implementira spomenutu standardnu biblioteku i sadrži API-je koji su specifični za operacijski sustav Windows kao što su forme (engl. *Windows Forms*) i WPF. Radi se o okruženju koja je optimizirana za Windows desktop aplikacije.

.NET alati i zajednička infrastruktura:

- .NET jezici i pripadni kompajleri,
- Komponente korištene tijekom izvođenja kao što su JIT i Garbage Collector,
- .Net projektni sustav (poznat kao "csproj", "vsproj" i "fsproj"),
- MSBuild, (engl. *build engine*) za projekte,

- NuGet, Microsoftov menadžer paketa za .NET aplikacije,
- .NET CLI je sučelje komandne linije za stvaranje .NET projekata.

Postoji minimalno ograničenje na verziji Entity Frameworka, ono je postavljeno na verziji 4.5. Kao što piše u ent, Entity Framework je objektno odnosni mapper (ORM) (engl. *Object-relational mapper*) koji omogućuje ljudima koji razvijaju u .NET okolini da lakše rade s bazama podataka koristeći .NET objekte. Eliminira potrebu da se piše kod za pristup bazama, razne upite i slično što je obično potrebno pisati.

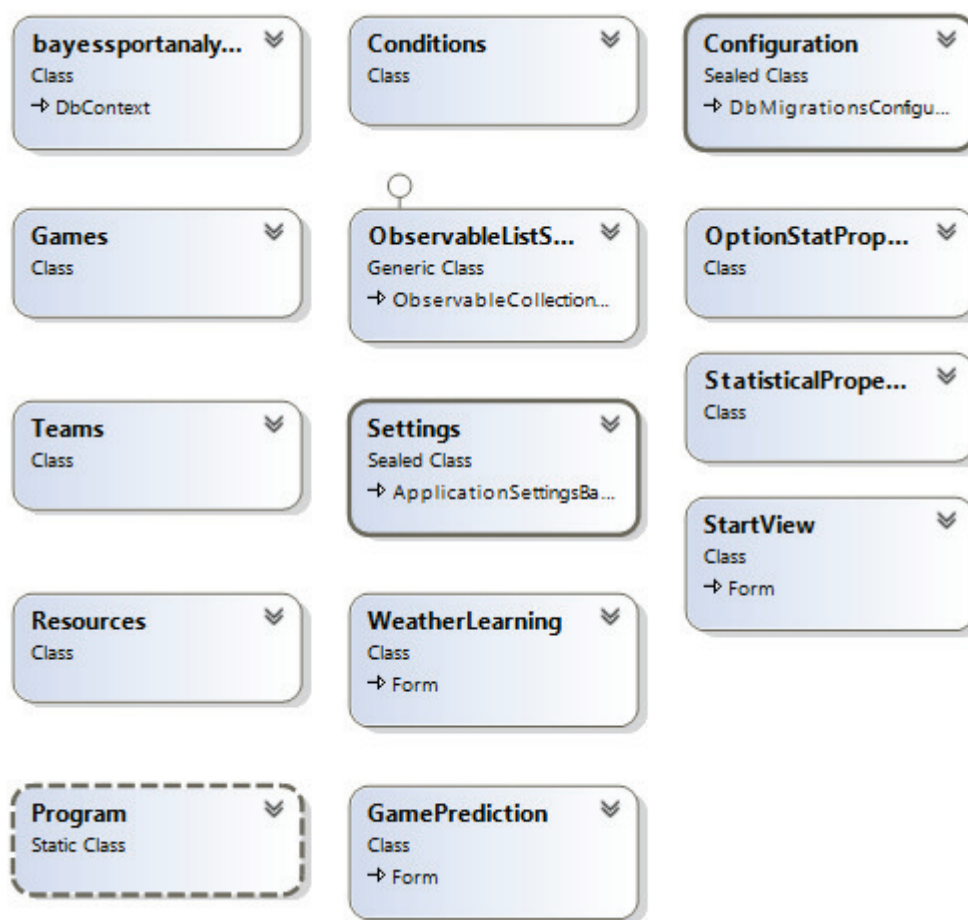
Kod je napisan u jeziku C#, Community (a) kaže da je to jezik koji je dizajniran za stvaranje različitih aplikacija na .NET Framework. Jezik je jednostavan, moćan, tip-ski siguran i objektno orijentiran. Tijekom godina razvoj C# je omogućio brzi razvoj aplikacija i zadržao obilježja C-ovskog tipa jezika. S obzirom na odabranu arhitekturu aplikacije i odabranu razvojnu okolinu, aplikacija je namijenjena za pokretanje na operacijskom sustavu Windows. Prvenstvena namjera kod izrade aplikacije je bila prikazati algoritam koji je tema rada i njegovu primjenu na skupu stvarnih podataka.

Prilikom pokretanja aplikacije prvo se učitava kontekst, kontekst dohvaća zapise i čini ih dostupnima aplikaciji koja se vrti na računalu. U prvom prozoru korisnik bira na koji će način koristiti aplikaciju. Prvi način je analiza diskretnih vrijednosti opisana u poglavlju 4.2 i taj način biramo pritiskom na lijevi gumb u početnom sučelju. Drugi način korištenja je analiza kontinuiranih vrijednosti, detaljno obrađena u poglavlju 4.3. Nakon što korisnik aplikacije odabere željeni proračun može se nakon zatvaranja prozora prebacivati na druge načine računa bez da mora ponovno pokretati aplikaciju.

Zapisi u bazi se pohranjuju na jednostavan način, kako se radi o programu koji samo treba popis podataka dovoljan je postojeći model prikazan na slici 4.3. U eventualnom kompleksnijem modelu u kojem uzimamo u obzir i druge faktore kao što je mjesto odigravanja, položaj na ljestvici i slično, možda bi bilo potrebno imati dodatnu tablicu sa potrebnim parametrima.

4.2. Analiza diskretnih vrijednosti

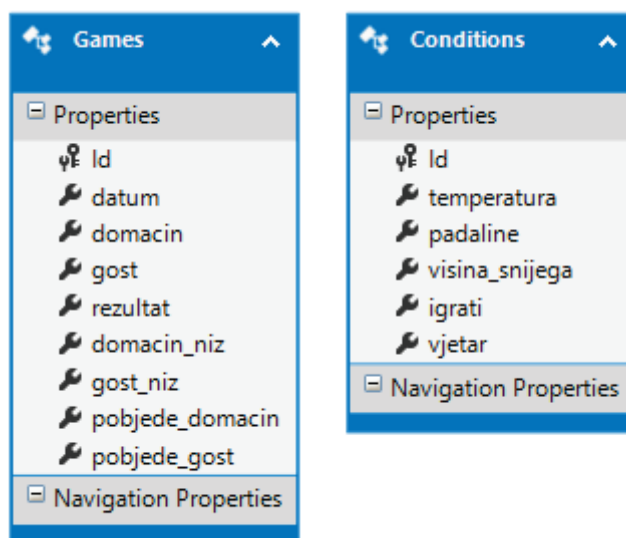
Prva od dvije funkcionalnosti programskog rješenja je analiza vrijednosti koje imaju diskretnu razdiobu, u ovom slučaju, vremenske uvjete koji uvjetuju bavljenje sportom na prirodnom ledu. Kod pokretanja forme učitavaju se podaci iz baze kako bi se u sučelju korisniku prikazao sve vrijednosti koje može iskoristiti. Korisnik mora unijeti vremenske prilike koje odabire od mogućih vrijednosti koje su ujedno navedene u tablici 1.3 na stranici 3. Na temelju danih vrijednosti računalo koristi izvedbu algoritma



Slika 4.2: Dijagram klasa

napisanog u poglavlju 3.2. Kao izvor podataka na temelju kojih donosi zaključak, program koristi zapise u bazi podataka koji se počinju dohvaćati kad korisnik pritisne gumb Izračunaj. Iz baze se zapisi prvo dohvate, izračunaju se kardinaliteti svih potrebnih skupova (npr. pretraže se svi zapisi gdje je igranje DA i vrijeme kišno). Na kraju proračuna, program vraća izlaz koji kaže DA ili NE ako su dobri uvjeti za biti na ledu. Metoda koja računa vrijednost ne vraća nikakav podatak već samo postavlja drugačiji tekst.

U priloženom primjeru vidimo primjer korištenja programa za primjer u kojem je temperatura oko nule, pada blagi snijeg, još uvijek se nije stvorio sloj snijega, a puše slab vjetar. Ovaj slučaj je rubni slučaj jer iako su svi parametri zadovoljavajući, zbog temperature oko nule, ponašanje ledene podloge nije sigurno pa zato i boravak na ledenoj površini nije preporučljiv. Sličan ishod bi bio da je jak vjetar, iako tad sam led ne bi bio opasan, boravak vani ne bi bio ugodan.



Slika 4.3: Model podataka u bazi

Koristeći ovo sučelje saznajte jesu li vremenske prilike pogodne za sport na ledu na otvorenom

Temperatura

Padaline

Visina snijega

Vjetar

NE

Slika 4.4: Primjer unosa diskretne raspodjele

4.3. Analiza kontinuiranih podataka

Drugo svojstvo ovog programa je predviđanje rezultata sportskih utakmica, u trenutnoj verziji prilagođena je samo za predviđanje ishoda utakmica pojedinih ekipa u hokeju na ledu, ali to je dovoljno da se vidi sposobnost programa. U bazi se nalaze zapisi međusobnih susreta nekih momčadi u zadnje dvije godine. Korisnik je dužan unijeti odgovarajuće podatke u sučelje na temelju kojih algoritam daje rezultat. Kao i kod diskretne analize, prvi korak je učitavanje početnih podataka koje će korisnik moći birati, prvo učitamo sve momčadi koje imamo pohranjene u bazi i čekamo daljnji korisnički unos.

Podržani podaci koji su potrebni za izračunavanje ishoda susreta su sljedeći:

- Odabrati momčad s popisa,
- Unijeti ukupan broj pobjeda momčadi u sezoni,
- Unijeti formu momčadi.

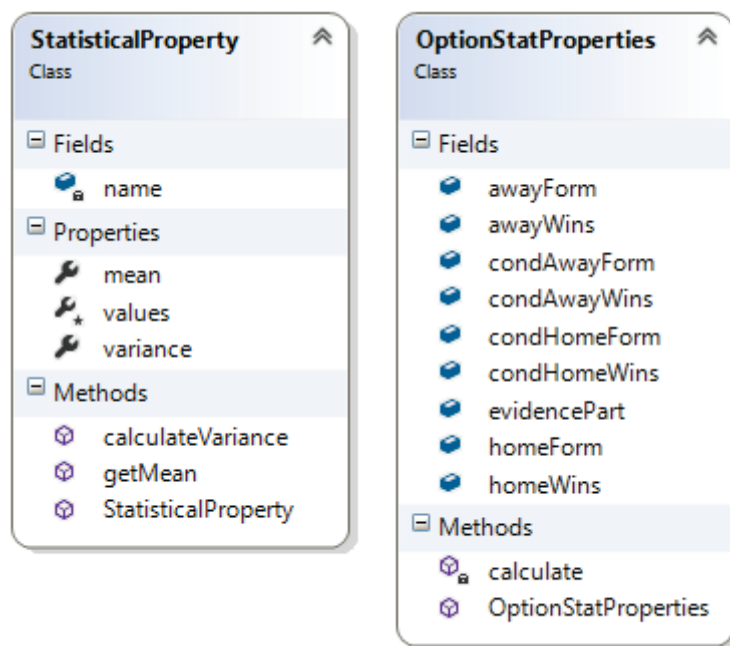
Pojam forme momčadi je jedini pojam koji nije jasan neupućenom čitatelju. Pojam forme označava niz zadnjih utakmica koje su završile istim ishodom (pobjeda ili poraz). Znači da je momčad koja je zabilježila niz od pet poraza u formi pet poraza, a momčad koja je nakon niza od deset poraza zabilježila pobjedu u formi od jedne pobjede.

Momčad s popisa se bira u listi padajućeg izbornika koji sadrži sve momčadi koje imaju zapise u bazi, ukupan broj pobjeda unosi se u okvir za tekst u obliku integera dok se forma unosi tako da se broj utakmica forme predznači s minus (-) ako se radi o nizu poraza, a bez predznaka ako je niz pobjeda. Detaljniju upotrebu objasniti ću na dva primjera.

Nakon što korisnik ispuni sve podatke koje želi i klikne gumb Izračunaj, prvo se provjerava unos kako ne bi odabrao istu momčad dvaput ili ako neki podatak nedostaje. Ako su svi uneseni podaci zadovoljavajući, algoritam započinje dohvaćanjem svih podataka iz baze koji su vezani uz unesene podatke. Dohvaćaju se svi susreti iz baze u kojima je odabrani domaćin bio domaćin i analogno za gosta. Pretpostavljamo da su vjerojatnosti pobjede obje momčadi iste, vjerojatno bi se mogao izraditi nekakav matematički model da oponaša razliku između boljih i lošijih, ali odabrane momčadi su podjednako jake pa možemo pretpostaviti da vjerojatnost pobjede obje momčadi iznosi isti.

$$P(domacin) = P(gost) = 0.5.$$

Slijedeći korak je parsiranje podataka iz baze, točnije, kako su u bazi pohranjeni rezultati susreta, a oni nama nisu bitni, trebamo odrediti pobjednika i to radimo tako da parsiramo zapis rezultata. Kako se radi o natjecanju u kojem nema izjednačenog rezultata, posao nam je olakšan. Nakon što smo sve podatke obradili možemo izvršavati potrebne matematičke operacije na skupu podataka. Za domaćina i za gosta računaju se medijani i varijance za svaki od parametara (forma i ukupni broj pobjeda) i pohranjuju se u varijable koje se kasnije koriste kod računanja posteriorne vjerojatnosti. Sav proračun obavlja se na analogan način onom u poglavlju 3.3. Dio dijagrama klasa 4.2 koji služi za matematičke operacije u dijelu analize kontinuiranih rezultata se nalazi u klasama `StatisticalProperty` i `OptionStatProperties`, klasa `StatisticalProperty` nam služi za računanje medijana i varijance specifičnog skupa podataka, s tim da sam za računanje varijance odabrao algoritam za varijancu uzorka. Jednadžba za varijancu uzorka



Slika 4.5: Klase za baratanje sa statističkim podacima

dana je izrazom 4.1

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m)^2 \quad (4.1)$$

gdje je m medijan skupa brojeva. Kod kojim računam varijancu je:

```
public double calculateVariance ()
{
    double ret = 0;
    if ( values.Count() > 0)
    {
        double avg = values.Average();
        double sum = values.Sum(d => Math.Pow(d - avg, 2));
        ret = ((sum) / ( values.Count() - 1));
    }
    if ( Double.IsNaN( ret ))
    {
        return 0.0;
    }
    return ret;
}
```

Klasa **OptionStatProperties** nam služi za kombiniranje i korištenje izračunatog u

StatisticalProperty da izračunamo posteriore svih vrijednosti kako bi dobili posterior numeratora tvrdnje za pobjedu domaćina odnosno pobjedu gosta. Izgled klasa prikazan je na slici 4.5. Navedene vrijednosti su izlaz metode koja reproducira bayes te ovisno o tome koja je veća vrijednost, tu momčad sustav vraća kao predviđenog pobjednika. Isto kao što je to bilo i u analizi diskretnog problema, korisnik može unutar jedne "sesije" neometano pokretati aplikaciju i na svaku naredbu za izračun će se jedina promjena događati u tekstu odgovora.

Primjer 1.: Promatramo susret dvije momčadi, gdje je prva navedena momčad domaćin te je u dosadašnjem toku sezone zabilježila 20 pobjeda, ali je zadnje dvije izgubila. Momčad navedena ispod je pobijedila 18 puta u tekućoj sezoni i pobijedila zadnje dvije, to zapisujemo na ovaj način:

Suparnici	Broj pobjeda	Forma
MIN	30	-2
TOR	30	2

Predviđam pobjedu: MIN

Slika 4.6: Primjer unosa kontinuirane raspodjele

Nakon postupka proračuna, sustav ispisuje predviđenog pobjednika koji je u ovom slučaju domaća momčad.

Primjer 2.: Zamijenimo suparnike tako da prijašnji domaćin postane gost i obrnuto. Novi domaćin je prvu utakmice sezone izgubio, a gost je u dosadašnjem toku sezone ubilježio jednu pobjedu i ima niz od jednog poraza što znači da je izgubio prethodnu utakmicu, to zapisujemo na ovaj način:

Suparnici	Broj pobjeda	Forma
TOR	0	-1
MIN	1	-1

Predviđam pobjedu: TOR

Slika 4.7: Drugi primjer unosa kontinuirane raspodjele

Za sve varijable koje su uneseno, program pretpostavlja Gaussovu razdiobu, računa na način koji je opisan u poglavlju 3.3 i na kraju izvršenog proračuna vjerojatnosti, nakon klika na „Izračunaj“, sustav daje izlaz predviđenog pobjednika.

4.4. Testiranje točnosti predviđanja

U ovom poglavlju prikazat ćemo popis testova koji su zapravo stvarne utakmice prilagođene za unos u aplikaciju. Utakmice koje su primjeri dolaze iz svih dijelova sezone. Izlazi aplikacije koji su točno predvidjeli ishod označeni su podebljanim tekstom, a pogrešni su označeni kurzivom.

Tablica 4.2: Tablica testnih primjera

Utakmica	Pobjede 1	Forma 1	Pobjede 2	Forma 2	Pobjednik	Bayes
STL-TOR	27	2	23	-2	STL	<i>TOR</i>
BOS-TOR	26	-1	23	-3	TOR	TOR
MTL-EDM	30	-2	28	-4	EDM	EDM
TOR-STL	25	1	27	2	STL	STL
MTL-STL	31	1	28	3	STL	STL
NSH-MTL	16	1	22	-1	MTL	<i>NSH</i>
BOS-MTL	28	2	31	-1	BOS	<i>MTL</i>
STL-EDM	31	-3	33	-2	EDM	EDM
MIN-STL	42	1	32	1	STL	STL
MIN-NSH	30	2	22	2	NSH	<i>MIN</i>
EDM-NSH	25	4	21	1	NSH	<i>EDM</i>
NSH-BOS	18	1	22	1	NSH	NSH

Analizom podataka u tablici 4.2 vidimo da je od 12 susreta za 7 točno predvidio pobjednika. Prema rezultatima ovog testa možemo reći da je preciznost predviđanja naivnog Bayesova klasifikatora koji u obzir uzima formu momčadi i broj pobjeda u sezoni 58%.

5. Zaključak

U ovom završnom radu testirao sam sposobnost naivnog Bayesova klasifikatora da točno predvidi rezultat utakmice analizom prethodnih utakmica. U prvim poglavljima završnog rada proučio sam teoriju u pozadini naivnog Bayesova klasifikatora, u drugom objasnio kako ću primijeniti naučeno na analizu diskretnih i kontinuiranih podataka i u trećem dijelu taj dio objasnio na primjeru aplikacije kojeg sam napravio u sklopu ovog završnog rada. Predviđanje ishoda utakmice je teško i za čovjeka pa je i teško naučiti računalo da točno predviđa rezultat kad se radi o aktivnosti koja utječe o mnogo faktora koje neke nije moguće kvantificirati.

Analiza sportskih rezultata metodom Bayesova klasifikatora dala je rezultate koji su slični kao i rezultati kod ispitivanja točnosti ljudskog predviđanja. Program je pokazao sposobnost izračunati i prepoznati uzorak za podatke koje ima u bazi, čim je rezultat iznenađujuć ili ne odgovara njegovim pretpostavkama, rezultat računa je pogrešan. Svi faktori uzeti u obzir bili su vezani uz ekipu kao cjelinu, individualni faktori nisu uzeti u obzir.

Očito je kako je još dosta faktora ostalo nepokriveno, iako treba biti svjestan da je sve gotovo nemoguće pokriti. Tijekom izrade ovog rada dodatno sam proučavao rezultate pa pretpostavljam da bi ubacivanje individualne komponente kroz račun s bazom statistika igrača u obzir dodatno povećao točnost.

Na kraju, zaključujem da je Bayesova metoda moćno svojstvo kada pokušamo predvidjeti nešto sa standardnim ponašanjem i kad se nalazi unutar okvira naših pretpostavki. Račun Bayesove metode je jak kao što je jaka njegova pretpostavka, a mnoge rezultate koje računalo ne može predvidjeti bi čovjek lako mogao i obrnuto. U ovom slučaju Bayesova metoda je primijenjena samo na broju pobjeda i formi ekipe, kada bi se uzeli i drugi podaci u obzir bi i model bio točniji, nažalost, neki podaci koji bi pomogli niti više nisu dostupni, ali čak i najbolji model ne bi bio otporan na iznenađenja zbog ljudske komponente u sportskim natjecanjima.

LITERATURA

Entity framework. URL <https://docs.microsoft.com/en-us/ef/>. Posljednji pristup: 30.5.2017.

Eli Ben-Naim, Federico Vazquez, Sidney Redner, et al. Parity and predictability of competitions. *Journal of Quantitative Analysis in Sports*, 2(4):1–12, 2006.

Microsoft Community. C#, a. URL <https://docs.microsoft.com/en-us/dotnet/csharp/csharp>. Posljednji pristup: 29.5.2017.

Microsoft Community. .net architectural components, b. URL <https://docs.microsoft.com/en-us/dotnet/standard/components>. Posljednji pristup: 29.5.2017.

Nenad Elezović. *Vjerojatnost i statistika: Diskretna vjerojatnost*. Element, Zagreb, 2016.a.

Nenad Elezović. *Vjerojatnost i statistika: Slučajne varijable*. Element, Zagreb, 2016.b.

K Ming Leung. Naive bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, 2007.

Kevin P Murphy. Naive bayes classifiers. *University of British Columbia*, 2006.

Željko Pauše. *Vjerojatnost : informacija, stohastički procesi : pojmovi - metode - primjene*. Školska knjiga, Zagreb, 1988.

Irina Rish. An empirical study of the naive bayes classifier. U *IJCAI 2001 workshop on empirical methods in artificial intelligence*, svezak 3, stranice 41–46. IBM New York, 2001.

Primjena Bayesovog klasifikatora u analizi sportskih rezultata

Sažetak

Tema rada je naivni Bayesov klasifikator. Potrebno je objasniti i implementirati primjenu Bayesovog klasifikatora na predviđanje ishoda utakmica na temelju zadane baze. Korisnik treba pristupati aplikaciji pomoću interaktivnog sučelja.

Ključne riječi: Bayes, klasifikator, rezultat, sport, utakmica.

Application of Bayes Classifier in Analysis of Sport Results

Abstract

Subject of this paper is naive Bayes classifier. It should explain and implement Bayes classifier method on predicting match outcome based on a given match database. There should be an interactive interface so that user can use this application.

Keywords: Bayes, classifier, match, result, sport.