



---

## Fazi klasifikacija

### *Računarska inteligencija*

---

Boris Cvitak

Jovan Stamenković

# Sadržaj

<b>1</b>	<b>Uvod</b>	<b>1</b>
<b>2</b>	<b>Fazi logika i rešenje problema</b>	<b>2</b>
2.1	Šta je fazi logika? . . . . .	2
2.2	Primer fazi logike . . . . .	2
2.3	Rešenje problema koristeći fazi logiku . . . . .	2
2.3.1	Unošenje ulaznih parametara i fazifikacija ulaznih podataka . . .	3
2.3.2	Fazi zaključivanje i fazi pravila . . . . .	4
2.3.3	Određivanje klase na osnovu funkcije pripadnosti . . . . .	4
<b>3</b>	<b>Različiti algoritmi klasifikacije</b>	<b>5</b>
3.1	Algoritmi klasifikacije . . . . .	5
3.2	Stablo izvođenja . . . . .	5
3.2.1	Rezultat klasifikacije Stabla izvođenja . . . . .	5
3.3	Stablo izvođenja sa unakrsnom validacijom . . . . .	6
3.3.1	Rezultat klasifikacije Stabla izvođenja sa unakrsnom validacijom	6
3.4	Naivni Bajes . . . . .	7
3.4.1	Rezultat klasifikacije Naivnog Bajesa . . . . .	7
3.5	K najbližih suseda . . . . .	8
3.5.1	Rezultat klasifikacije K najbližih suseda . . . . .	8
<b>4</b>	<b>Rezultat klasifikacije Fazi algoritma</b>	<b>9</b>
<b>5</b>	<b>Zaključak</b>	<b>10</b>
<b>6</b>	<b>Literatura</b>	<b>11</b>

# 1 Uvod

Klasifikacioni problem pripada grupi nadgledanih problema mašinskog učenja zato što je za svaku instancu u skupu podataka poznata njena ciljna grupa. U klasifikacionom problemu potrebno je pronaći vezu između vrednosti atributa instanci i ciljne promenljive tako da se na osnovu vrednosti atributa neke instance može prepoznati ciljna grupa te instance.

Fazi logika omogućava pisanje pravila na osnovu kojih se promenljivoj dodeljuje vrednost iz intervala  $[0, 1]$  koja govori o pripadnosti te promenljive određenom fazi skupu.

U našem radu, korišćen je skup podataka koji u sebi sadrži 692 fudbalera iz različitih delova sveta koji igraju na poziciji napadača. Za svakog fudbalera imamo informaciju o njegovom imenu, rejtingu, ocenu kontrole lopte, driblinga, brzine i završnice kao i klasu kojoj taj fudbaler pripada. Svakom fudbaleru je dodeljena klasa na osnovu njegovog rejtinga koji ima, a najbolji napadači pripadaju klasi 5. Naš zadatak je da na osnovu ocena za kontrolu lopte, driblinga, brzine i završnice predvidimo kojoj će klasi pripadati određeni fudbaler.

Ime	Rejting	Klasa	Kontrola lopte	Dribling	Brzina	Završnica
C.Ronaldo	94	5	93	92	92	93
L.Messi	93	5	95	97	87	95

Tabela 1: Skup podataka

Rejting	Klasa	Broj instanci
94-80	5-Svetksa klasa	70
79-75	4-Profesionalac	116
74-71	3-Poluprofeseionalac	184
70-66	2-Amater	199
66-55	1-Početnik	123

Tabela 2: Povezanost rejtinga i klase kao i broj instanci

## 2 Fazi logika i rešenje problema

### 2.1 Šta je fazi logika?

Lotfi Zadeh se smatra tvorcem fazi logike. On smatra da je fazi logika formalni sistem matematičke logike zasnovan na teoriji mogućnosti. Klasična logika se zasniva na jasnim i precizno utvđenim pravilima, a počiva na teoriji skupova. Neki element može da pripada nekom skupu ili da ne pripada. Skupovi imaju jasno određene granice. Za razliku od klasične logike, u fazi logici nije precizno definisana pripadnost jednog elementa određenom skupu, već se pripadnost meri u, recimo, procentima. Ove mere pripadnosti, skalirane, mogu da uzimaju vrednosti od 0 do 1. Ovakav pristup je prirodan, intuitivan jer nekad nismo u mogućnosti da precizno odredimo neko stanje.

### 2.2 Primer fazi logike

Kao primer uzećemo dane u nedelji i napraviti dva skupa. Skup radnih dana i skup vikenda. U klasičnoj logici bi se u skupu radnih dana našli: ponedeljak, utorak, sreda, četvrtak i petak, a u skupu vikend dana: subota i nedelja. Pripadnost elementa nekom skupu bi se izrazila brojem 1, a nepripadanje brojem 0. Međutim u fazi zaključivanju bi situacija bila nešto drugačija. Petak, kao dan koji je delom radni dan, a delom početak vikenda bio bi negde na granici ova dva skupa tj. njegova pripadnost prvom, skupu radnih dana bi se izražavala, recimo brojem 0,75 dok bi pripadnost drugom, skupu vikend dana bila cifra 0,25. Slično bi bilo i za nedelju kao dan koji jeste vikend, ali ne sasvim, celim svojim trajanjem, jer ipak se nedelja uveče doživljava kao priprema za novu radnu nedelju odnosno mnogi ljudi će ga okarakterisati kao ne sasvim vikend dan, jer posle njega dolazi ponedeljak.

### 2.3 Rešenje problema koristeći fazi logiku

Problem: Imamo skup fudbalera sa po četiri atributa za svakog od njih. Potrebno je, u odnosu na attribute, razvrstati sve fudbalere u pet klasa. Od najgore do najbolje klase redom idu Početnik, Amater, Poluprofesionalac, Profesionalac, Svetska klasa (eng. Beginner, Amater, Semi-professional, Professional, World class).

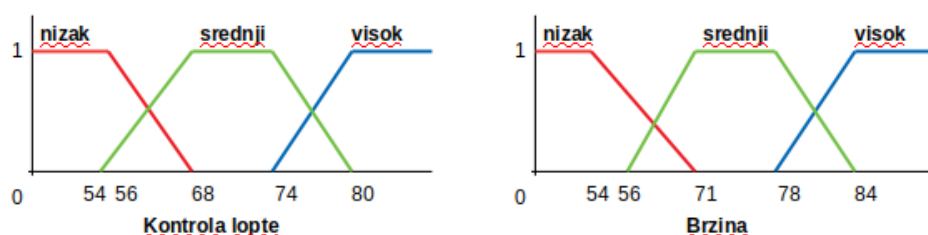
Rešenje: Kako bismo došli do rešenja moramo da prodemo kroz četiri koraka u fazi zaključivanju:

- Unošenje ulaznih podataka
- Fazifikacija ulaznih podataka
- Fazi zaključivanje
- Određivanje klase na osnovu funkcije pripadnosti

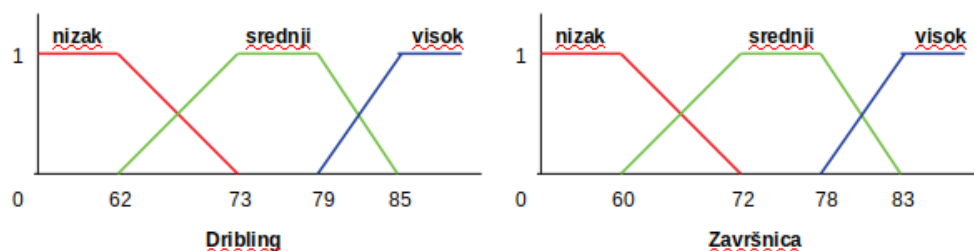
### 2.3.1 Unošenje ulaznih parametara i fazifikacija ulaznih podataka

Za svaki od četiri atributa imamo tri moguća stanja koja ih opisuju i to su nizak, srednji i visok (eng. low, medium i high). Za svaki od atributa, postupak fazifikacije prolazi iste korake. Unosimo vrednost atributa i zatim na osnovu te vrednosti određujemo pripadnost tog atributa stanju nizak, srednji ili visok.

Za stanje nizak koristimo takozvanu Z-funkciju. Imamo dve značajne granice. Za unetu vrednost  $x$ , koje se nalazi u intervalu do prve granice, vrednost funkcije pripadnosti je 1. Zatim, krećući se ka drugoj granici vrednost funkcije opada dok ne dođe do druge granice gde je vrednost 0. Od druge granice pa na dalje funkcija pripadnosti ima vrednost 0.



Za srednje stanje koristimo trapezoidnu funkciju pripadnosti. Ona ima četiri granice. Do prve granice vrednost je 0. Od prve do druge granice vrednost funkcije pripadnosti se kreće od 0 do 1. Između druge i treće granice ta vrednost je 1. Kada se krene od treće granice ka četvrtoj ta vrednost polako opada dok ne dođe do 0. Odatle pa na dalje funkcija pripadnosti ima vrednost 0.



Funkcija koju koristimo za stanje visok je S-funkcija. Ona, kao i Z-funkcija ima dve granice, ali se drugačije gleda (obrnute su vrednosti funkcije). Do prve granice vrednost je 0, zatim krećući se ka drugoj granici ta vrednost raste dok ne dostigne 1 i nakon toga, za svaku vrednost koja je veća od granice, funkcija pripadnosti ovom stanju ima vrednost 1.

### 2.3.2 Fazi zaključivanje i fazi pravila

Nakon što smo svakom atributu na osnovu ulaznih vrednosti odredili pripadnost stanju nizak, srednji i visok, sledeći zadatak je pisanje pravila koje će u stvari odrediti klasifikaciju.

Fazi pravila povezuju dva ili više iskaza. U našem slučaju imamo četiri atributa, tako da povežujemo četiri iskaza. U odnosu na vrednost iskaza konstruišemo pravila tako da za određene vrednosti dobijemo tačno određeno izlazno stanje. S obzirom da imamo pet mogućih izlaznih stanja postoji pet kategorija pravila:

- Početnik:
  - potrebno je da za tri atributa vrednost stanja bude nizak i jedan srednji
  - ili da vrednost za sva četiri atributa bude nizak
- Amater:
  - tri srednji i jedan nizak
  - dva srednji i dva nizak
  - tri nizak i jedan visok
  - dva nizak i jedan srednji i jedan visok
- Poluprofesionalac:
  - dva visok i dva nizak
  - jedan visok i jedan nizak i dva srednji
  - četiri srednji
- Profesionalac:
  - tri visok i jedan nizak
  - jedan visok i tri srednji
  - dva visok i jedan srednji i jedan nizak
- Svetska klasa:
  - četiri visok
  - tri visok i jedan srednji
  - dva visok i dva srednji

### 2.3.3 Određivanje klase na osnovu funkcije pripadnosti

Pošto su pravila zasnovana na logičkoj konjukciji, pripadnost klasi se određuje na osnovu minimalne vrednosti među vrednostima stanja atributa koja učestvuju u pravilu. Kada se prođu sva pravila, za svaku klasu se određuje pripadnost na osnovu maksimalne vrednosti među vrednostima koja su bila dodeljena toj klasi.

Fudbaler će na kraju pripadati onoj klasi čija je funkcija pripadnosti najveća.

## 3 Različiti algoritmi klasifikacije

### 3.1 Algoritmi klasifikacije

Kako bismo uporedili rešenja našeg algoritma fazi klasifikacije, potrebno je na istom skupu podataka primeniti i neke druge algoritme klasifikacije. Mi smo odabrali sledeće algoritme:

- Stablo izvođenja
- Stablo izvođenja sa unakrsnom validacijom
- Naivni Bajes
- K najbližih suseda

### 3.2 Stablo izvođenja

Stablo izvođenja (eng. Decision Tree) je jedan od najjednostavnijih algoritama za klasifikaciju. Svakom čvoru takvog stabla odgovara test nekog svojstva instance, a grane koje izlaze iz čvora različitim vrednostima tog svojstva. Listovima odgovaraju predviđene vrednosti ciljne promenljive. Instance su opisane vrednostima svojih svojstava. Predviđanje se vrši polazeći od korena, spuštajući se niz granu koja odgovara vrednosti testiranog svojstva instance za koju se vrši predviđanje. Predviđanje se dodeljuje instanci kad se dođe do lista. [1]

#### 3.2.1 Rezultat klasifikacije Stabla izvođenja

Prilikom primene algoritma Stabla izvođenja nad našim skupom fudbalera, mogli smo zaključiti da se veoma često na trening skupu algoritam prilagodi podacima, odnosno da kasnije ne može objektivno da rasazna klasu fudbalera (eng. overfitting).

	Početak	Amater	Poluprofesionalac	Profesionalac	Svetska klasa
Početak	29	7	0	1	0
Amater	12	37	10	1	0
Poluprofesionalac	2	9	34	10	0
Profesionalac	0	1	9	20	5
Svetska klasa	0	0	0	6	15

Tabela 3: Matrica konfuzije na test skupu

Uspešnost algoritma (eng. accuracy) je **0.64**.

### 3.3 Stablo izvođenja sa unakrsnom validacijom

Kako bi se unapredio osnovni algoritam stabla izvođenja, prilikom treniranja podataka u procesu evaluacije se koristi unakrsna validacija (eng. cross-validation). Ceo skup podataka kojim se raspolaže se deli na  $n$  približno jednakih podskupova. Jedan podskup se izdvaja i trening se vrši na ostalih  $n-1$  podskupova. Posle treninga, kvalitet naučenog znanja se ocenjuje na izdvojenom podskupu. Ovaj postupak se ponavlja za sve ostale izdvojene podskupove i kao finalna ocena kvaliteta se uzima prosek dobijenih ocena za svaki od podskupova. Za vrednost  $n$  se obično uzima broj 5 ili 10 i ne preporučuju se mnogo manje ili veće vrednosti. Ovakav postupak daje stabilniju ocenu kvaliteta. Pored toga, prednost ovog metoda je da se u svakom od  $n$  koraka unakrsne validacije koristi velika količina podataka pri treniranju, a da sve raspoložive instance u jednom trenutku budu iskorišćene za testiranje. [1]

#### 3.3.1 Rezultat klasifikacije Stabla izvođenja sa unakrsnom validacijom

Uz pomoć unakrsne validacije unapredili smo običan algoritam Stabla izvođenja, a prilikom same primene algoritma isprobali smo različite ulazne parametre kako bismo dobili što bolji rezultat. Za kriterijum kvaliteta podele po čvorovima upotrebljeni su gini koeficijent i entropija. Dubine drveća su se menjale po neparnim brojevima od tri do jedanaest. Kao mera preciznosti koja se gleda prilikom uspešnosti algoritma posmatrane su preciznost <sup>1</sup> i f1 mera. <sup>2</sup>

	Početak	Amater	Poluprofesionalac	Profesionalac	Svetska klasa
Početak	32	3	2	0	0
Amater	13	36	10	1	0
Poluprofesionalac	0	13	35	7	0
Profesionalac	1	0	9	21	4
Svetska klasa	0	0	0	4	17

Tabela 4: Matrica konfuzije na test skupu

Najbolji rezultati dobijeni su kada je dubina drveća pet, kriterijum podele gini, a mera preciznosti f1 mera. Uspešnost algoritma (eng. accuracy) je **0.67**.

<sup>1</sup>[https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html).

<sup>2</sup>[https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html)



### 3.4 Naivni Bajes

Algoritam Naivni Bajes (eng. Naive Bayes) je klasifikacioni algoritam koji je zasnovan na Bajesovoj teoremi iz verovatnoće.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \quad (1)$$

Vrlo je bitno da pretpostavimo da su atributi instanci međusobno nezavisni kako bismo mogli primeniti Bajesovu teoremu. Opšta formula Naivnog Bajesovog algoritma klasifikacije je:

$$P(B|A_1, A_2, \dots, A_n) = P(B) * \prod_{i=1}^n P(A_i|B) \quad (2)$$

Interpretacija algoritma nam pokazuje kolika je verovatnoća da će instanca pripasti klasi B ukoliko su vrednosti atributa te instance jednaki  $A_1, A_2, \dots, A_n$ .

$$B = \operatorname{argmax}_B P(B) \prod_{i=1}^n P(A_i|B) \quad (3)$$

#### 3.4.1 Rezultat klasifikacije Naivnog Bajesa

Prilikom primene algoritma Naivni Bajes, primetili smo najveća odstupanja među rezultatima. Uspešnost se prožimala između 0.65 i 0.75. Ako uzmemo u obzir da postoji mogućnost male zavisnosti između atributa instanci, onda ovaj rezultat moramo uzeti u razmatranje sa malom dozom rezervisanosti. Takođe, primetili smo da prilikom primene algoritma skoro uopšte ne dolazi do preprilagođavanja trening podacima.

	Početak	Amater	Poluprofesionalac	Profesionalac	Svetska klasa
Početak	25	11	1	0	0
Amater	10	40	10	0	0
Poluprofesionalac	0	6	44	5	0
Profesionalac	0	1	10	21	3
Svetska klasa	0	0	0	2	19

Tabela 5: Matrica konfuzije na test skupu

Uspešnost algoritma (eng. accuracy) je **0.71**.

## 3.5 K najbližih suseda

Metoda K najbližih suseda (eng. K nearest neighbours) zasniva se na vrlo jednostavnoj ideji — pronaći K instanci najslićnijih nepoznatoj instanci, takozvanih suseda i predvideti vrednost njene ciljne promenjive na osnovu vrednosti koje odgovaraju susedima. U slučaju klasifikacije, predviđanje odgovara najčešćoj klasi među klasama suseda. Pojam sličnosti instanci najjednostavnije se formalizuje preko funkcija rastojanja. Intuitivno, što je rastojanje između dva objekta veće, to je sličnost između njih manja i obratno. [1]

### 3.5.1 Rezultat klasifikacije K najbližih suseda

Algoritam K najbližih suseda pokazao se kao najbolji algoritam u našem slučaju. Intuitivno, to ima smisla. Ako znamo da su klase fudbalerima dodeljene veštački (odnosno da smo ih mi sami dodeli), onda znamo i da su fudbaleri koji se po osobinama nalaze na granici između dve klase veoma slični. Zato se ovim algoritmom pronalaze susedi jednog fudbalera i na osnovu klasa kojima pripadaju susedi nagađa klasa kojoj pripada fudbaler.

	Početaknik	Amater	Poluprofesionalac	Profesionalac	Svetska klasa
Početaknik	27	10	0	0	0
Amater	7	45	8	0	0
Poluprofesionalac	0	11	37	7	0
Profesionalac	0	0	7	27	1
Svetska klasa	0	0	0	3	18

Tabela 6: Matrica konfuzije na test skupu

Najbolji rezultati dobijeni su kada je broj suseda koji se posmatra 7, a distanca među susedima izračunava kao euklidsko rastojanje. Uspešnost algoritma (eng. accuracy) je **0.75**.

## 4 Rezultat klasifikacije Fazi algoritma

Na osnovu petnaest definisanih pravila fazi klasifikacije kao i na osnovu određenih granica za sve attribute datog skupa, fazi algoritam doneo je sledeće rezultate:

	Početnik	Amater	Poluprofesionalac	Profesionalac	Svetska klasa
Početnik	55	62	6	0	0
Amater	10	137	45	7	0
Poluprofesionalac	0	66	81	37	0
Profesionalac	1	11	38	49	17
Svetska klasa	0	0	0	6	64

Tabela 7: Matrica konfuzije fazi algoritma

Uspešnost algoritma(eng. accuracy) je **0.56**.

Možemo da zaključimo da se algoritam odlično ponaša prilikom klasifikacije fudbalera Svetske klase što ima smisla ako pogledamo da ovi fudbaleri imaju jako visoke vrednosti svojih atributa pa se relativno lako prepoznaju. Takođe, možemo primetiti da se fudbaleri klase Amater dosta dobro prepoznaju. Obzirom da je broj amatera najveći u datom skupu, možemo zaključiti da algoritam ima dovoljno podataka na osnovu kojih bi mogao da zaključi koje su vrednosti atributa potrebne da bi se fudbaler našao u toj klasi.

Prilikom klasifikacije fudbalera koji pripadaju klasama Poluprofesionalac i Profesionalac, vidimo da algoritam može tačno predvideti klasu tih fudbalera. Međutim, zbog vrlo uskog opsega vrednosti atributa koje uzimaju ovi fudbaleri, vrlo lako algoritam može da pogreši klasu tako što predvidi ili višu ili nižu klasu.

Prilikom predviđanja klase Početnik, algoritam sigurno određuje klasu onim fudbalerima čije su vrednosti attribute baš niske, ali čim se malo povećaju vrednosti, algoritam dolazi u zabunu da li je fudbaler početnik ili amater.

## 5 Zaključak

Početni skup fudbalera sadržao je pedeset i tri atributa, a od svih su zadržani samo: rejting, kontrola lopte, dribling, brzina i završnica. Na osnovu rejtinga, dodeljene su klase fudbalerima.

S obzirom na veliko sažimanje skupa podataka, na to da su klase dodeljene samo na osnovu jednog atributa i da je veza između rejtinga i klase dodeljena ručno bez analize podataka, možemo zaključiti da fazi algoritam uopšte nije loš, jer je na osnovu veoma malo informacija uspeo da pređe uspešnost od 50%.

U ovakvoj situaciji gde imamo samo četiri atributa, svaki atribut ima veliki uticaj na predviđanje klase. To možemo videti na primeru Lionela Messija. Naime, kada bismo pored ova četiri atributa imali i peti atribut "igra glavom", s obzirom da Messi ima veoma malu vrednost tog atributa, on se ne bi nalazio u Svetskoj klasi već bi bio raspoređen u Profesionalce ili Poluprofesionalce. Zbog toga su izabrani kontrola lopte, dribling, brzina i završnica kao atributi koji najbolje opisuju napadača.

Kako bismo unapredili preciznost algoritma, trebalo bi uzeti u obzir veći skup atributa, preciznije definisati početne klase fudbalera, još detaljnije ispitati granice atributa kao i dodeliti neko novo stanje atributu ukoliko je to potrebno.

## 6 Literatura

### Literatura

- [1] P. J. i Mladen Nikolić, *Veštačka inteligencija*.