Primjena strojnog učenja u sportskoj prediktivnoj analitici

Valentino, Poljak

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Valentino Poljak

Tehnike strojnog učenja u sportskoj prediktivnoj analitici

DIPLOMSKI RAD

Varaždin, 2021.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Valentino Poljak

Matični broj: 44165/15-R

Studij: Informacijsko i programsko inženjerstvo

Tehnike strojnog učenja u sportskoj prediktivnoj analitici

DIPLOMSKI RAD

Mentor/Mentorica:

Doc. dr. sc. Dijana Oreški

Varaždin, srpanj 2021.

Valentino Poljak

Izjava o izvornosti

Izjavljujem da je moj završni/diplomski rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor/Autorica potvrdio/potvrdila prihvaćanjem odredbi u sustavu FOI-radovi

Sažetak

U ovome radu provest će se prediktivna analiza rezultata španjolske nogometne lige najelitnijeg ranga, tzv. LaLige. Predikcija i praćenje rezultata je vrlo zanimljivo područje koje se već neko vrijeme istražuje. Da bi se provela ova analiza potreban je skup podataka koji je preuzet u obliku dataset-a s web stranice *Kaggle*. Uz tu web stranicu podaci su prikupljeni s web stranice *FlashScore*. U skupu podataka se nalaze najvažnije informacije o svakoj utakmici koja je odigrana u dvije sezone, 2014./2015. te 2018./2019. Te informacije sadrže ekipe koje su bile gostujuće, odnosno domaće, trenere jedne i druge ekipe, statističke podatke, kao što su broj udaraca, zaleđa, kornera i slično te konačan rezultat. Svaka sezona ima ukupno 380 utakmica. Pratit će se utjecaj pojedine informacije, tj. Atributa na konačan ishod svake utakmice. Provest će se nekoliko metoda za potonje, a one su: Stablo odlučivanja, kreiranje ekspertnog sustava, neuronske mreže te Bayesove mreže. Nakon izvršenja svih metoda, rezultati će se usporediti i pojasniti.

Ključne riječi: nogometne utakmice; LaLiga; Stablo odlučivanja; XRKB; Neuronske mreže; Analiza sportskih rezultata; BigML; Ekspertni sustavi; Bayesove mreže

Sadržaj

1. Uvod

U današnje vrijeme nogomet kao sport polako gubi smisao te počinje postajati sve veći biznis, bilo kroz razne sponzorske ugovore ili kao klađenje, sve više se ide u smjeru zarade u sportu umjesto o uživanja u njemu te samim time nogomet kao sport gubi smisao. Kako je nogomet vrlo popularan i zapravo kako mnogi smatraju „najvažnija sporedna stvar na svijetu“ predikcija rezultata je zanimljiva svima koji ga prate, posebno navijačima. Predikcija rezultata je vrlo zanimljiva za istraživanje, međutim ovisi o previše faktora na koje je nemoguće utjecati, kao što su: moral igrača, ozljede, postotak spremnosti za utakmicu, taktika i slično. Tako je moguće samo putem nekih općenitih podataka, tipa međusobni omjeri ekipa iz nekoliko prošlih sezona, pokušati predvidjeti kako će ti timovi odigrati sljedeću utakmicu.

U ovome će se radu najprije prikazati neka prethodna slična istraživanja, povezana sa sportskom prediktivnom analitikom. Nakon toga će se prikazati skup podataka iz španjolske LaLige koja je prvi razred nogometa u Španjolskoj te će se usporediti rezultati ekipa između dviju sezona, prve i zadnje iz skupa podataka, a to su sezone 2014./2015. te 2018./2019. Skup podataka je preuzet s internetske stranice te uređen za potrebe ovog rada. Koristiti će se nekoliko metoda za predviđanje, a to su: stablo odlučivanja, neuronske mreže te Bayesove mreže. Spomenute metode će se prvo teorijski objasniti kako bi bilo jasnije o čemu se točno radi te nakon toga pokazati primjeri svake metode. Nakon toga će se rezultati objasniti, usporediti te će se iznijeti zaključak.

2. Prethodna istraživanja

U ovome poglavlju obrađeno je nekoliko istraživanja koja imaju veze sa prediktivnom analitikom u sportu. Svako istraživanje se baziralo na nekom različitom skupu podataka. Pojedina istraživanja su pokušala odrediti kako će završiti rezultat između dvije ekipe, dok su neka istraživanja pokušala predvidjeti konačnu tablicu, točnije na kojem će mjestu ekipa na kraju završiti.

2.1. „Predikcija nogometnih rezultata Bayesovom mrežom“

Nogomet je jedan od najpopularnijih sportova u svijetu. Predviđanje rezultata nogometnih utakmica je zbog tog razloga zanimljivo mnogima, od navijača do običnih ljudi. Također, predviđanje rezultata je i vrlo zanimljiv istraživački problem, iako je taj problem vrlo težak za predviđanje jer ovisi o previše faktora na koje je nemoguće utjecati. Za provođenje istraživanja znanstvenici su koristili nekoliko prethodnih istraživanja od 2000. do 2010. godine koja su pokušavali otkriti različite stvari u nogometu, jedno istraživanje je pokušalo predvidjeti pobjednika na europskom prvenstvu 2000. godine, drugo je pokušalo odrediti koji će klub imati najviše pobjeda, neriješenih ishoda te poraza, itd (Owramipur et al., 2013).

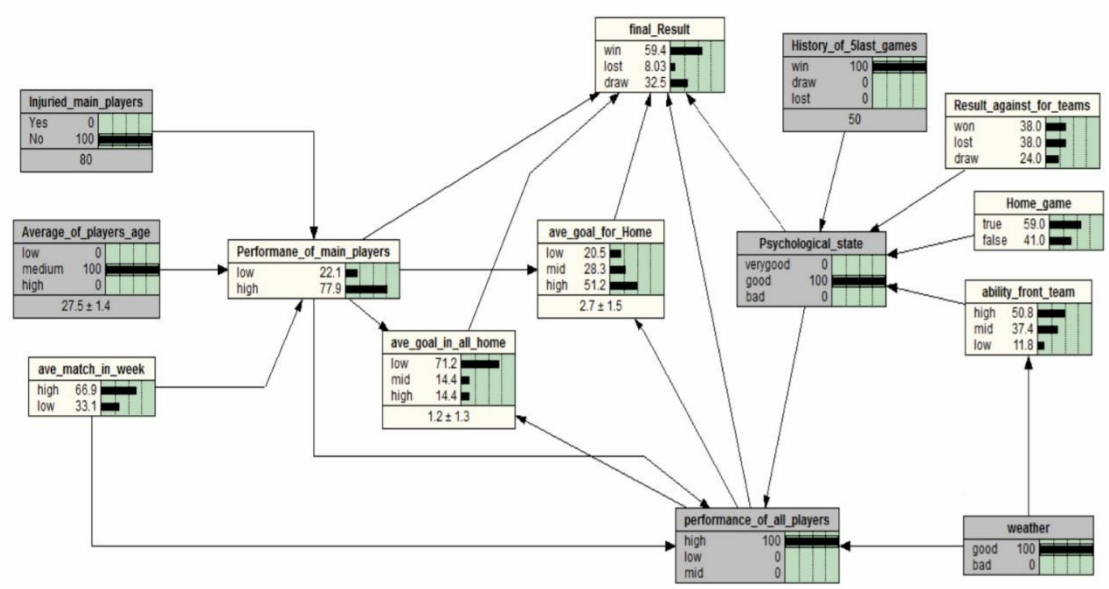
Ovo istraživanje pokušava odrediti kako će određeni tim, u ovom slučaju Barcelona, odigrati sve svoje utakmice u ligi kroz jednu sezonu, znači 38 utakmica. Za ovu predikciju korišteni su atributi koji su podijeljeni u 2 skupine: psihološka i ne-psihološka skupina. Ti atributi su korišteni zbog toga što ih u procjenama i kreiranju koeficijenata koriste eksperti u kladionicama (Owramipur et al., 2013).
Neki od glavnih psiholoških atributa su:

- *Weather*
- *History_of_5last_games*
- *Result_against_for_teams*
- *Home_game*
- *ability_front_team*
- *Psychological_state*

  Glavni ne-psihološki atributi su bili:
- *Average_of_players_age*
- *Injuried_main_players*
- *ave_match_in_week*
- *Performane_of_main_players*

- *performance_of_all_players*
- *ave_goal_in_all_home*
- *ave_goal_for_Home*

Podaci koje su prikupili za Barcelonine utakmice i sve statistike koje su koristili preuzete su s nekoliko relevantnih stranica koje se bave nogometnim statistikama.
Kako bi napravili predikciju i Bayesovu mrežu koristili su se alatom Netica. U Neticu su upisali podatke za pojedinu utakmicu te pokrenuli proces predviđanja. Primjer jedne utakmice u Netici je na sljedećoj slici (Owramipur et al., 2013).



Slika 1. Primjer Bayesove mreže – Netica prema (Owramipur et al., 2013)

Ovaj proces su radili i pokretali za svih 38 utakmica u sezoni. Rezultati svake pojedine utakmice su prikupljeni te su finalni rezultati i pozicija Barcelone odlučeni. Na kraju se ispostavilo da su ovim istraživanjem uspješno predviđeni rezultati u velikom broju utakmica, čak 92% njih, odnosno predvidjeli su hoće li Barcelona pobijediti, izgubiti ili odigrati neriješeno, a nisu predviđali točne rezultate (Owramipur et al., 2013).

2.2. „Otkrivanje glavnih uzroka nogometnih rezultata rudarenjem podataka“

Zadnjih nekoliko godina javlja se povećanje interesa za korištenje statističkih metoda s ciljem identificiranja faktora važnih za ishode sportskih natjecanja. Još 1970-ih godina se prvi puta spominje i dokumentira statističko razmišljanje u sportu. 2005. i 2008. godine dvije vrlo zapažene kolekcije statističkih analiza su primijenjene na širok raspon sportova, od američkog nogometa, bejzbola, košarke, nogometa, hokeja na ledu i slično. Kako bi se predviđali rezultati najčešće korištena metoda je bila složena poissonova distribucija (eng. double Poisson distribution). Može se kroz povijest vidjeti kako mnogi znanstvenici korištenjem upravo te metode pokušavaju napraviti predviđanje rezultata u nogometu. Najprije se koristilo u predviđanju engleske nogometne lige od strane engleskih znanstvenika, a istu su tehniku primijenili i njemački znanstvenici. Na kraju su isto pokušali i talijani.

U ovome istraživanju istraženo je u kakvoj su vezi krajnji rezultat utakmice (pobjeda, neriješeni ishod ili poraz) te određene varijable utakmice. Podaci za analizu su uzeti od četiri uzastopna natjecanja. Sami je cilj bio shvatiti koji su to faktori važni za pobijediti utakmicu. Točnije, cilj je bio saznati i izabrati, od mnogo različitih varijabli, varijable koje najviše utječu na vjerojatnost pobjede na utakmici. Za odabir najvažnijih varijabli korišten je algoritam slučajnih šuma (eng. Random Forest machine learning algorith), a kako bi se te varijable sažele u manje skupove kako bi se lakše mogle analizirati i vizualizirati korišten je algoritam glavnih komponenata (eng. Algorithm of Principal Component Analysis) ili skraćeno *PCA* (Carpita et al., 2016).

Ukupno su pronađene 482 različite varijable koje utječu na pobjedu, a kao podaci su uzete utakmice talijanskog najelitnijeg nogometnog razreda, „*Serie A“* od sezone 2008-09 do sezone 2011-12, znači ukupno 4 sezone. Od svih varijabli ukupno je izabrano 17 njih za koje se smatra da su najvažnije varijable koje utječu na krajnji rezultat utakmice. Tih 17 varijabli je sažeto u 6 manjih skupova, po tri za domaće utakmice i isto toliko za gostujuće utakmice. Tom analizom je omogućeno zaključiti određene akcije koje su najdjelotvornije protiv određenih klubova iz lige, odnosno koju akciju bi protivnička momčad trebala koristiti kako bi mogla zabiti više golova i samim time pobijediti na utakmici. Takva analiza se naravno može koristiti i u obrnutom slučaju, odnosno može se koristiti i kod tima nad kojim je analiza provedena kako bi saznao u kojem je polju nogometne igre trenutno najslabiji. Na kraju analize, zaključeno je kako je veoma stabilna i pouzdana kada je provedena na utakmicama te se može vidjeti da iako se svake godine timovi mijenjaju ispadanjem i ulaskom u elitni rang natjecanja, kao i igrački kadar koji se mijenja svake godine, neke stvari u taktici svakog tima ostaju iste, kao i njihove prednosti i mane (Carpita et al., 2016).

2.3. „Predikcija rezultata fakultetskog nogometa rudarenjem podataka“

Rudarenje podataka za cilj ima otkriti implicitne, otprije nepoznate i potencijalno korisne informacije ili znanja iz podataka. Kod mnogih istraživanja s rudarenjem podataka u sportu fokus je na raspored utakmica, vizualizaciju igre ili igrača, identifikaciju interesantnog momenta na utakmici ili igrača. Dok je sve navedeno ranije zanimljivo kao istraživanje, jednako tako je zanimljivo i predvidjeti konačne rezultate utakmica. Generalno, predikcija sama po sebi nije veliki problem. Stručni komentatori, treneri pa čak i igrači daju svoje predikcije gotovo svakodnevno ili barem prije određene utakmice. To rade na temelju svog iskustva, instinkta ili osjećaja u tom trenutku. Međutim, to sve su samo subjektivne procjene. Da bi se subjektivnost što više smanjila prilikom predviđanja, potrebno je koristiti neke dodatne podatke za samo predviđanje. Neki od tih podataka mogu biti prethodni rezultati između dvije momčadi, trenutna forma igrača i drugo. U današnje je vrijeme sama predikcija i statistika podataka u nogometu, a i sportu općenito u velikom porastu i vrlo je zanimljiva velikom broju znanstvenika, ali i zaljubljenika u sport (Leung & Joseph, 2014).

Za izradu ovog istraživanja, znanstvenici su koristili neke informacije o samom sportu za koji rade predikciju, a to je američki nogomet, točnije, srednjoškolski američki nogomet. Morali su naučiti sva pravila i neke određene, često korištene napade i obrane koje pojedini timovi koriste u svojim utakmicama. Uz to, pomagali su si i prijašnjim istraživanjima kojih je na tu temu bilo mnogo. Mnogo je problema kod američkog nogometa sa statistikom jer u slučaju da napadač ne uspije uloviti loptu koju mu je branič poslao, lošije bodove dobiva i branič čija je lopta možda bila upotrebljiva i dobra. Zbog tog i mnogo takvih sitnih statističkih problema je ovaj problem još i zanimljiviji za istraživanje (Leung & Joseph, 2014).

U mnogim današnjim sportovima gledatelji i navijači žele predvidjeti rezultat utakmice i gledati utakmicu kako bi uspjeli potvrditi svoje znanje tom sportu. Tradicionalni pristupi uključuju subjektivne predikcije, objektivne predikcije te jednostavne statističke metode. Ipak, ti pristupi možda nisu previše pouzdani u mnogim situacijama. Ovo istraživanje prikazuje pristup sportskog rudarenja podataka koji pomaže otkriti zanimljiva znanja te predvidjeti krajnje rezultate utakmica kao što je fakultetski nogomet (američki nogomet). Ovaj pristup daje predviđanje temeljeno na kombinaciji 4 različita mjerenja na povijesnim rezultatima. Sama evaluacija rezultata na primjeru iz stvarnog života, stvarnog natjecanja u fakultetskom američkom nogometu prikazuje kako ovaj pristup vodi do relativno visokih točnosti u predikciji rezultata (Leung & Joseph, 2014).

Podaci za istraživanje su prikupljeni sa stranice koja prati rezultate fakultetskog američkog nogometa kroz povijest te su uzeti podaci utakmica od sezone 2005-06 do sezone 2013-14. Također za svaku utakmicu je prikupljena krajnja statistika utakmice kao npr. Broj faulova, broj pokušaja dodavanja, broj uspješnih dodavanja, pokušaji pretrčavanja i slično. Kada su podaci za prve 4 sezone prikupljene napravio se model koji je pokušao predvidjeti sezonu 2011-12 te je imao vrlo visoku stopu uspješnosti, otprilike 91% točnosti. Isto je ponovljeno i za sljedeće sezone u kojima je algoritam pokazivao također vrlo visoke postotke točnosti. To je pokazalo kako je ovaj pristup vrlo pouzdan za predikciju krajnjeg rezultata sportskih utakmica (Leung & Joseph, 2014).

2.4. „Modeliranje podataka za predikciju nogometnih rezultata“

Otkrivanje umjetne inteligencije je dalo mogućnost izrade prediktivnih sustava s nevjerojatnom točnošću. Strojno učenje se koristi u gotovo svim područjima, zbog svoje visoke efektivnosti. Jedno od tih područja je predikcija sportskih rezultata. Ovo istraživanje demonstrira rad na stvaranju generaliziranog prediktivnog modela za predikciju rezultata engleske „Premier League“. Korištenjem raznih analiza podataka i inženjeringa kreiran je sustav za prepoznavanje najvažnijih faktora za predikciju rezultata nogometnih utakmica koji sukladno tome kreira vrlo precizne predikcije koristeći strojno učenje (Baboota & Kaur, 2019).

Vrlo zahtjevan problem tijekom modeliranja rezultata za ovo istraživanje javlja se u velikoj kompetitivnosti engleske nogometne lige, gdje zapravo svaka ekipa može pobijediti bilo koju drugu ekipu. Najbolji pokazatelj toga je i osvajanje engleske lige od strane Leicestera, protiv svih izgleda te na iznenađenje cijelog svijeta. Zato koristeći mnoga prijašnja istraživanja vezana uz analizu i kretanja rezultata u engleskoj ligi će se modelirati podaci za ovo istraživanje. Podaci su dakle preuzeti s nekoliko web stranica i uz pomoć prijašnjih istraživanja modelirani. Podaci se koriste za englesku ligu od 2005. do 2016. godine. Ili 11 sezona. Uz statističke podatke, model će imati i tzv. *Rating* statističke podatke koje će pomagati u predikciji na način da će svaki klub u svakom trenutku imati neki rating, što će i označavati zapravo trenutnu formu ekipe (Baboota & Kaur, 2019).

Table 2
Feature description.

Feature name	Feature abbreviation	Class category
Home form	*HForm*	*Class A*
Away form	*AForm*	*Class A*
Home streak	*HSt*	*Class A*
Away streak	*ASt*	*Class A*
Past *k* home shots on target	*HSTKPP*	*Class A*
Past *k* away shots on target	*ASTKPP*	*Class A*
Past *k* home goals	*HGKPP*	*Class A*
Past *k* away goals	*AGKPP*	*Class A*
Past *k* home corners	*HCKPP*	*Class A*
Past *k* away corners	*ACKPP*	*Class A*
Home attack rating	*HAttack*	*Class A*
Away attack rating	*AAttack*	*Class A*
Home midfield rating	*HMidField*	*Class A*
Away midfield rating	*AMidField*	*Class A*
Home defence rating	*HDefence*	*Class A*
Away defence rating	*ADefense*	*Class A*
Home overall rating	*HOverall*	*Class A*
Away overall rating	*AOverall*	*Class A*
Home goal difference	*HTGD*	*Class A*
Away goal difference	*ATGD*	*Class A*
Home weighted streak	*HStWeighted*	*Class A*
Away weighted streak	*AStWeighted*	*Class A*
Form differential	*FormDifferential*	*Class B*
Streak differential	*StDifferential*	*Class B*
Past *k* shots on target differential	*STKPP*	*Class B*
Past *k* goals differential	*GKPP*	*Class B*
Past *k* corners differential	*CKPP*	*Class B*
Attack rating differential	*RelAttack*	*Class B*
Midfield rating differential	*RelMidField*	*Class B*
Defence rating differential	*RelDefense*	*Class B*
Overall rating differential	*RelOverall*	*Class B*
Goal difference differential	*GDDifferential*	*Class B*
Weighted streak differential	*StWeightedDifferential*	*Class B*

Slika 2. Prikaz klasifikacije atributa za skup podataka (Baboota & Kaur, 2019)

Metode korištene za ovu analizu su sljedeće (Baboota & Kaur, 2019):

- Gaussov naivni Bayesov klasifikator (eng. Gaussian naive Bayes) – metoda koja je varijanta naivnog Bayesovog klasifikatora koja prati Gaussovu normalnu raspodjelu i podržava kontinuirane podatke. Korištena je za klasifikaciju podataka
- Stroj s potpornim vektorima (eng. Support vector machine (SVM)) – to je skup algoritama učenja koji se vrlo efektivni u rješavanju slučajeva s višedimenzionalnim prostorima.
- Slučajna šuma (eng. Random forest) – algoritam koji se sastoji od više tzv. Stabala odluke. Stablo odluke predstavlja način kojim računalo dolazi do zaključka. Sve kreće

od korijenskog čvora. Nakon što su algoritmu predani podatci s kojima će raditi, danji postupak se svodi na postavljanje pravih pitanja. Odgovori na pitanja moraju biti u obliku točno ili netočno to jest 0 ili 1. Tako da kod svakog pitanja svaki od čvorova stvara svoja 2 nova podčvora. Postavljanjem dovoljnog broja pravih pitanja dolazi se do konačnog čvora u stablu. Odabir pravih pitanja, to jest odabir atributa koji će se testirati u pojedinom čvoru vrši se izračunom informacijske dobiti.

- Pojačavanje gradijenta (eng. Gradient boosting) – Kao i slučajna šuma, pojačavanje gradijenta je tehnika temeljena na stablu odlučivanja. Obje tehnike koriste kombiniranje klasifikatora, međutim ovdje je pojačavanje klasifikatora trenirano sekvencijski na cijelom skupu podataka, s povećanjem točnosti modela nakon svake iteracije.

Nakon svih provedenih metoda, najtočnija i najbolja metoda je upravo ova potonja metoda, Pojačavanje gradijenta, nakon koje slijedi metoda Slučajnih šuma. Pojačavanje gradijenta je pokazalo rezultat predviđanja 0,60 na skali od 0 do 1 i to je najbolji rezultat ovog istraživanja. Kao i svako istraživanje vezano uz predikciju, tako je i ovo imalo problema sa skupljanjem podataka, jer uvijek određeni podaci fale ili nisu potpuni. Naravno, postoje i faktori koji se ne mogu mjeriti niti koristiti za istraživanje, a u ovom su sportu vrlo bitni, a to su: informacije o ozljedama, stanju igrača prije utakmice, razmišljanju igrača i ostali psihološki aspekti igrača i tima. Usprkos svemu tome, vrlo dobar rezultat istraživanja je postignut i sa podacima s kojima se na kraju raspolagalo (Baboota & Kaur, 2019).

3. Prikupljanje, priprema i obrada podataka

U ovom poglavlju opisati će se podaci korišteni za analizu te koraci koji su provedeni nad podacima kako bi sama analiza bila što točnija te dala željeni rezultat. Podaci su preuzeti za prvu nogometnu ligu Španjolske, pod nazivom *LaLiga Primera division.* Kako bi se dobili podaci bitni za ovaj projekt proveli su se određeni koraci, a oni su sljedeći:

- Prikupljanje podataka s raznih stranica
- Kombiniranje podataka i selekcija atributa u jedan skup podataka
- Obrada odabranog skupa podataka

Za prikupljanje podataka korišteni su već gotovi skupovi podataka (eng. Datasets) dostupni online na raznim stranicama koji prate povijest nogometa te rezultate i razne statistike nogometne igre.

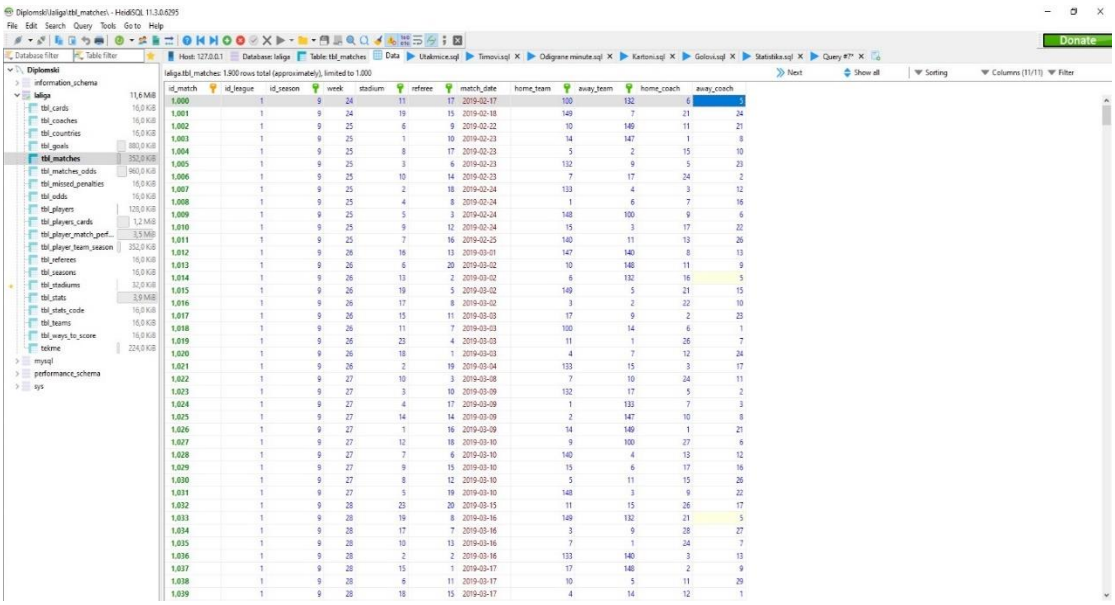
Bilo je potrebno pronaći pouzdane stranice koje imaju sve potrebne i točne podatke za provođenje ove analize. Tako su odabrane stranice kao što su *Kaggle* te *Football-Data* i *FlashScore*.

Glavni skup podataka je preuzet sa stranice *Kaggle* te sadrži sve informacije o utakmicama španjolskog prvenstva od sezone 2014./2015. do sezone 2018./2019. Sve informacije su poredane u zasebne datoteke tipa *csv* pa tako imamo posebne datoteke s utakmicama gdje su informacije o datumu utakmice, redni broj domaćina i gosta i slično, zatim postoje datoteke s podacima o igračima, sucima, statistici, načinima zabijanja pogodaka, pogocima, uglavnom vrlo opsežne informacije koje su već preuzete i spojene u cijeli taj skup podataka. Međutim, problem s podacima je taj što su ti podaci trebali biti u jednoj datoteci, u jednoj tablici kako bi se mogli analizirati za ovaj rad jer bi se inače mogle raditi samo analize po određenim kriterijima, dok ovako je moguće sve te kriterije spojiti i napraviti veliku analizu španjolske lige. Na kraju je odabrano 23 polja s informacijama koje su smatrane najbitnijima za analizu, od čega je velik broj iz već spomenutog skupa podataka, dok su ostala polja putem raznih upita stavljena u tablicu s drugih skupova podataka. Polja koja se koriste za konačnu tablicu su sljedeća:

1. ID utakmice
2. Datum
3. Sezona
4. Kolo
5. Sudac
6. Domaćin

7. Gost
8. Stadion
9. Trener domaćina
10. Trener gosta
11. Korneri domaćina
12. Faulovi domaćina
13. Zaleđa domaćina
14. Posjed lopte domaćina
15. Udarci domaćina u okvir gola
16. Udarci domaćina - ukupno
17. Korneri gosta
18. Faulovi gosta
19. Zaleđa gosta
20. Posjed lopte gosta
21. Udarci gosta u okvir gola
22. Udarci gosta - ukupno
23. Rezultat

U spomenutom konačnom skupu podataka upisano je ukupno 760 redaka s podacima što odgovara utakmicama dviju sezona koje su uspoređivane, obzirom da je po sezoni ukupno 380 utakmica. Svaka utakmica ima svoj jedinstveni ID po kojem se ona vodi i prema tom ID-u su vezani ostali podaci u tablicu. Kako je već gore navedeno, svaka utakmica ima razne statističke i opisne podatke, zna se tko je vodio koji klubu toj utakmici, tko je sudio utakmicu, na kojem se stadionu utakmica igrala, koliko je koji klub napravio faulova, zaleđa udaraca i slično te na kraju najvažniji dio svake utakmice, konačni rezultat. Kako su svi ti podaci bili u početnom skupu podataka razdijeljeni u različite manje tablice, bilo je potrebno putem upita doći do podataka koji su potrebni za ovu analizu. Da bi uopće bilo moguće pisati upite i izvoditi razne akcije nad tim datotekama, trebalo je koristiti neki sustav za upravljanje bazama podataka, a u ovom je slučaju korišten MySQL, dok je za administraciju koda korištena aplikacija HeidiSQL.



Slika 3. Primjer sučelja HeidiSQL-a

Na prethodnoj slici je moguće vidjeti sučelje HeidiSQL-a te jednu od mnogih tablica koje su nastale učitavanje „.csv“ datoteka u aplikaciju. Kako bi se mogli obrađivati svi podaci odjednom, bilo ih je potrebno spojiti u jednu veliku tablicu te odlučiti koji će podaci uopće ići u tu tablicu. U nastavku je napisani SQL upit za kreiranje nove tablice iz postojećih tablica s velikim brojem podataka koji su bitni za konačni skup.

„CREATE table tekme as SELECT DISTINCT m.id_match AS "ID utakmice", m.match_date AS "Datum", s.season AS "Sezona", m.week AS "Kolo", r.referee_name AS "Sudac", t1.team_name AS "Domaćin", t2.team_name AS "Gost", st.stadium_name AS "Stadion", c1.coach_name AS "Domaći trener", c2.coach_name AS "Gosti trener" FROM
tbl_matches m JOIN tbl_seasons s ON m.id_season = s.id_season
JOIN tbl_referees r ON m.referee = r.id_referee
JOIN tbl_stadiums st ON m.stadium = st.id_stadium
JOIN tbl_teams t1 ON m.home_team = t1.id_team
JOIN tbl_teams t2 ON m.away_team = t2.id_team
JOIN tbl_coaches c1 ON m.home_coach = c1.id_coach
JOIN tbl_coaches c2 ON m.away_coach = c2.id_coach
WHERE s.id_season IN (10,9)
ORDER BY m.match_date, m.id_match“

Tim je upitom dakle kreirana tablica s većinom podataka. Za dobivanje ostalih podataka napisano je više manjih upita da bi se došlo do traženih podataka te su onda ti manji upiti spojeni u jedan veći upit kojim su upisani u tablicu. Jedini problem je zapravo bila zadnja kolona

u konačnoj tablici a to je kolona s konačnim rezultatima. Iako je sami skup vrlo velik i opsežan te sadrži stvarno vrlo mnogo podataka, nigdje nije upisan konačni rezultat pojedine utakmice. Za tu priliku su sa stranice *FlashScore* prepisani rezultati svih utakmica u polje u koje pripadaju. Nakon toga je konačan skup napokon bio gotov te je spreman za daljnje korištenje.

4. Metodologija istraživanja

U ovom odjeljku opisati će se svaka od korištenih metoda te će se prikazati konkretan primjer za svaku odabranu metodu.

4.1. Stablo odlučivanja

Stablo odlučivanja je jedna od najčešće korištenih metoda prilikom klasifikacije, predviđanja, za procjenu vrijednosti, grupiranje, opisivanje podataka i vizualizaciju (Begičević, 2012). Elementi stabla odlučivanja su (Oreški, 2021c):

- Korijenski čvor
- Unutarnji čvorovi
- Čvorovi listova (završni čvor)

Svaki čvor stabla, osim završnog čvora, specificira test koji treba provesti na jednoj od vrijednosti deskriptivnog atributa, dok završni čvor specificira predviđenu klasifikaciju upita (Oreški, 2021c).

Stablo odlučivanja se u odlučivanju primjenjuje kao grafički model za vizualizaciju procesa odlučivanja kad se rješavanje problema odlučivanja svodi na donošenje više sukcesivnih odluka, a uz takav prikaz problema odlučivanja veže se i postupak računanja očekivanih vrijednosti inačica odluke u uvjetima rizika. Pogodno je one situacije kada se donošenje određene odluke sastoji od niza manjih odluka, a koje su vremenski slijedno povezane. Stablo odlučivanja za glavnu primjenu ima prikaz svih mogućnosti te definira koji je problem odlučivanja. Uglavnom se primjenjuje kada se treba donijeti neka odluka u rizičnim situacijama, koje se uglavnom odnose na poslovni svijet. Npr. Stablom odlučivanja može se donijeti i odluku o tome koji posao želimo raditi ili koju srednju školu upisati. Korištenjem metode stabla odlučivanja uvijek postoji nekoliko dijelova koji su poznati, a to su mogućnosti između kojih se potrebno odlučiti, poznate su posljedice koje bi se mogle dogoditi odabirom neke mogućnosti te je poznata i vjerojatnost da će se neka od posljedica dogoditi. Stablo odlučivanja se u pravilu sastoji od niza povezanih slijednih odluka i svaka od tih odluka je ovisna o svojoj prethodnoj. Stablo odlučivanja zapravo prikazuje kompletnu strukturu odlučivanja, jer je lakše donositi odluku kada je nešto slikovito prikazano nego u obliku tablice brojeva (Ekonomski fakultet Zagreb, 2011).

Postoje neke pretpostavke uporabe stabla odlučivanja. One su sljedeće (Saralegui, 2017):

- Donositelj odluke na raspolaganju ima većinu relevantnih inačica odluke

- Moguće posljedice inačica odluke se na neki način mogu kvantificirati
- Pri izboru se razmatraju samo ona obilježja inačica odluka koja se mogu kvantificirati
- Stablo odlučivanja se može analizirati ako postoje subjektivne vjerojatnosti nastupanja nesigurnih događaja

Stablo odlučivanja se stvara pomoću formule entropije. Formula entropije predstavlja mjerenje ukupnog nereda ili nehomogenosti iz baze podataka te ona glasi (Skladistenje.com, 2002):

$$E = \sum_{b} \left[\left(\frac{n_b}{n_t} \right) \cdot \left(\sum_{c} \left(-\frac{n_{bc}}{n_b} \right) \cdot \log_2 \left(\frac{n_{bc}}{n_b} \right) \right) \right]$$

Uobičajena metoda izrade stabla odlučivanja je rekurzivno particioniranje. To je metoda kod koje izrada modela kreće od korijena stabla. Proces učenja započinje usporedbom mogućih podjela na temelju jedne značajke. Podjela u korijenu predstavlja podjelu ulaznog prostora na dva podprostora s granicom paralelnom jednoj ulaznoj dimenziji. Ta podjela se ponavlja u svakoj sljedećoj grani dok sve grane ne postanu potpuno čiste ili dok se ne zadovolji neki od kriterija zaustavljanja (Skladistenje.com, 2002).

4.2. Izrada stabla odlučivanja u alatu BigML

Za izradu stabla odlučivanja koristio se alat BigML. BigML je zapravo programibilna, potrošna te skalabilna platforma za strojno učenje koja pomaže u rješavanju i automatizaciji klasifikacije, predviđanja vremenske serije, regresije, analize klastera, otkrivanja udruženja, detekcije anomalije i zadataka modeliranja uz određene teme. On pomaže velikom broju analitičara, znanstvenicima te developerima širom svijeta da riješe zadatke strojnog učenja od početka do kraja, pretvarajući podatke u djelotvorne modele koji se koriste kao udaljene usluge ili, koji se lokalno, ugrađuju u aplikacije kako bi se napravila predikcija. Napravljen je 2011. godine s ciljem da strojno učenje bude lako i lijepo za svakoga, a već nakon nekoliko godina izvrsnog i teškog rada stručnjaci iz BigML-a su uspjeli napraviti rješenje koje danas koriste različite organizacije svijeta svih veličina. Uz to, uspjeli su izgraditi mnoga sofisticirana rješenja temeljena na strojnom učenju koja izdaju prediktivne obrasce iz svojih podataka.

Dakle, kao što je već navedeno, za ovu analizu stablo odlučivanja je izrađeno u alatu BigML. Da bi se u njemu izradilo stablo, najprije je potrebno u web sučelje alata učitati .csv datoteku (može biti i .txt, .json, itd.) s podacima koje želimo imati na stablu odlučivanja. U ovom slučaju je to bila .csv datoteka koja je dobivena izvozom podataka iz tablica HeidiSQL-a. Za učitavanje samih podataka u alat potrebno je otići na „Sources“ te kliknuti gumb za dodavanje (učitavanje) nove datoteke u alat.

Kada se datoteka učita, potrebno je ući u datoteku te od nje napraviti „dataset“ kako bi se moglo upravljati tim podacima i koristiti ih za predviđanje. Navedeno je moguće učiniti pritiskom na tipku „Configure dataset“ i nakon toga „Create dataset“ što je vidljivo na sljedećoj slici.

Nakon što je „dataset“ kreiran iz njega je potrebno izraditi model, a da bi se model izradio potrebno je ući u nedavno izrađeni „dataset“ te kliknuti na opciju „Model“ kojom će se kreirati model. Naravno prilikom kreiranja modela, potrebno je izabrati atribut za koji će se stablo odlučivanja izraditi.

Nakon ovog zadnjeg koraka model je izrađen te je moguće vidjeti što stablo odlučivanja pokazuje.

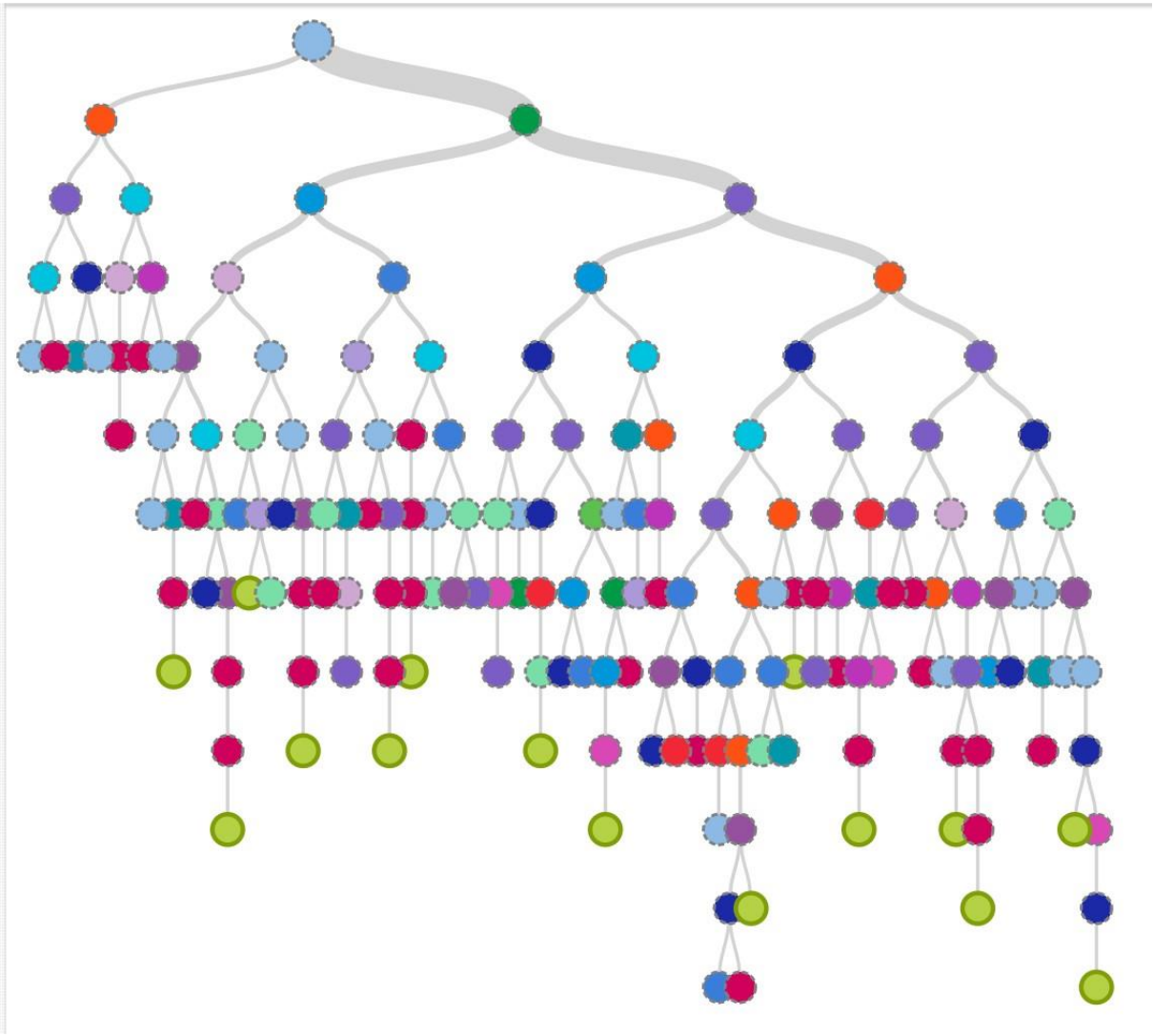
4.2.1. Prikaz stabla odlučivanja za LaLigu u sezonama 2014./2015. i 2018./2019.

Sada slijede prikazi i rezultati kreiranih stabla odlučivanja za španjolsku LaLigu u prvoj i zadnjoj godini iz skupa podataka koji je korišten za ovaj rad. Prikazati će se slikama te objasniti ti rezultati i donijeti zaključak između njih.

Za kreiranje stabla odlučivanja korišten je atribut „Rezultat“ koji je zapravo konačni rezultat svake utakmice u LaLigi. Time će se dobiti procjena koji je točno rezultat pod određenim uvjetima najčešći i usporediti će se vrijednosti između dviju sezona.

4.2.1.1. Stablo odlučivanja – LaLiga sezona 2014./2015.

Izgled stabla odlučivanja za sezonu 2014./2015. je vidljivo na sljedećoj slici.



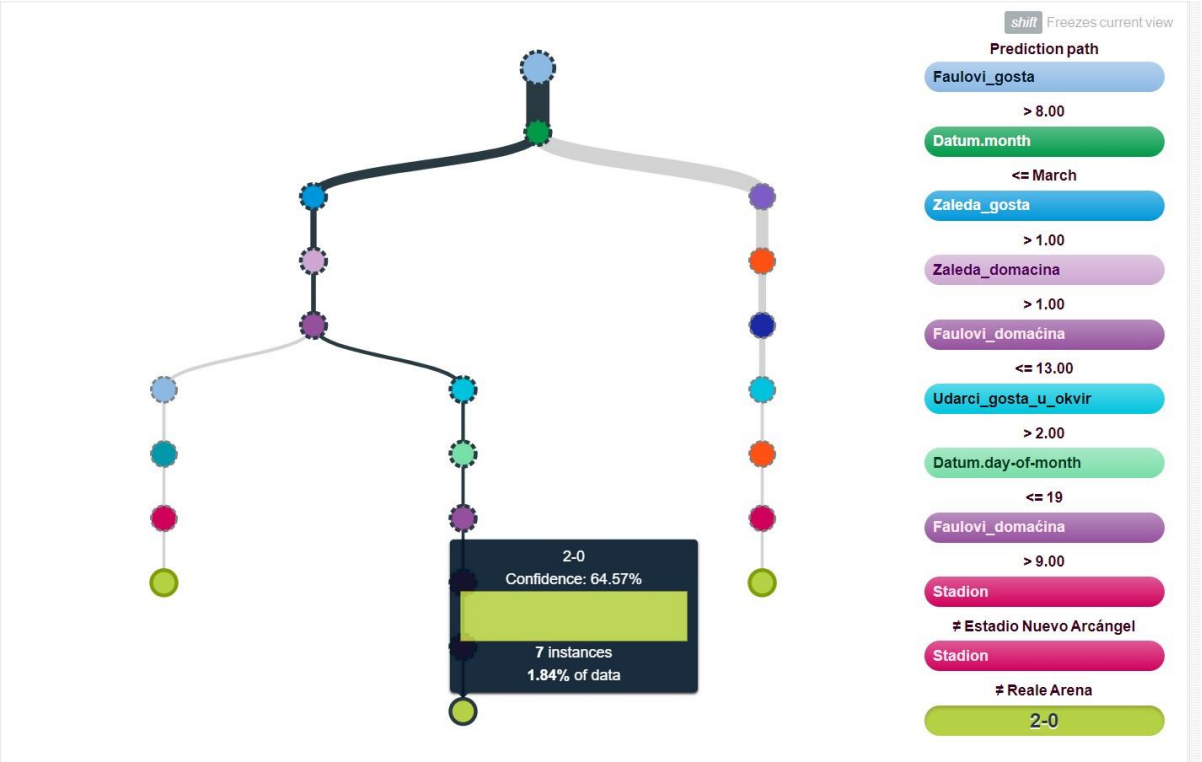
Slika 4. Prikaz klasifikacije atributa za skup podataka

Sada kada je stablo izrađeno, može se dobiti izvještaj koji govori o tome o kojim sve atributima je najviše ovisio atribut za koji je stablo izrađeno. U nastavku slijedi popis svih atributa i postotak njihovih ovisnosti.

- Faulovi_gosta: 18.80%
- Posjed_lopte_domacina: 18.17%
- Datum.month: 13.95%
- Kolo: 9.15%
- Udarci_domacina: 7.45%
- Zaleda_gosta: 7.31%
- Udarci_gosta_u_okvir: 4.77%
- Korneri_gosta: 4.32%
- Stadion: 3.46%
- Zaleda_domacina: 3.00%
- Faulovi_domaćina: 2.85%
- Datum.day-of-month: 2.26%
- Udarci_domacina_u_okvir: 1.12%
- Udarci_gosta: 1.10%
- Korneri_domacina: 0.76%
- Datum.day-of-week: 0.63%
- Sudac: 0.59%
- Gosti trener: 0.31%

Ovi podaci govore kako rezultat najviše ovisi o prekršajima koje je počinila ekipa koja igra u gostima te o posjedu lopte ekipe koja je domaćin. Ti podaci bi se mogli protumačiti tako da ukoliko domaća ekipa pokušava držati loptu u svom posjedu dulje vrijeme, gosti bi većim brojem faulova mogli doći do povoljnijeg rezultata za sebe jer ukoliko se prepusti lopta i kontrola domaćinima, doći će do obrnutog efekta.

Sljedeće je potrebno pronaći najpouzdanije pravilo unutar stabla odlučivanja, a to je ono pravilo koje ima najmanje odstupanje. U ovom slučaju postoji tri pravila koja su najpouzdanija, a ona su da će rezultat biti: 2-0, 1-1 ili 1-0. Pouzdanost ovog pravila je 64,57%. Na sljedećoj slici je prikazano jedno od triju najpouzdanijih pravila.



Slika 5. Najpouzdanije pravilo stabla odlučivanja - sezona 2014./2015.

Još jedna stvar koja je vrlo važna za usporedbu dviju sezona, a nalazi se unutar izvještaja stabla odlučivanja je broj utakmica koje su završile određenim rezultatom, tj. Odgovor na pitanje kojim je rezultatom završio najveći broj utakmica. Navest će se 5 najčešćih rezultata za sezonu 2014./2015., a oni su:

- 1-1, ukupno 43 utakmica završile tim rezultatom
- 1-0, ukupno 35 utakmica završilo tim rezultatom
- 0-1, ukupno 35 utakmica završilo tim rezultatom
- 2-0, ukupno 33 utakmice završile tim rezultatom
- 0-0, ukupno 30 utakmica završilo tim rezultatom

Vidljivo je kako u pet najčešćih rezultata postoje dva rezultata koja daju domaćeg pobjednika, dva koja daju neriješen ishod te jedan rezultat kojim pobjeđuje gostujuća ekipa. S nekim od ovih pet rezultata, završilo je ukupno 176 utakmica, što bi značilo da je gotovo 50% (točnije 46%) utakmica završavalo nekim od ovih rezultata.

4.2.1.2. Stablo odlučivanja – LaLiga sezona 2018./2019.

Za sezonu 18./19. postupak je isti, prvo će se prikazati izgled cjelokupnog stabla odlučivanja.

Slika 6. Stablo odlučivanja - Sezona 2018./2019.

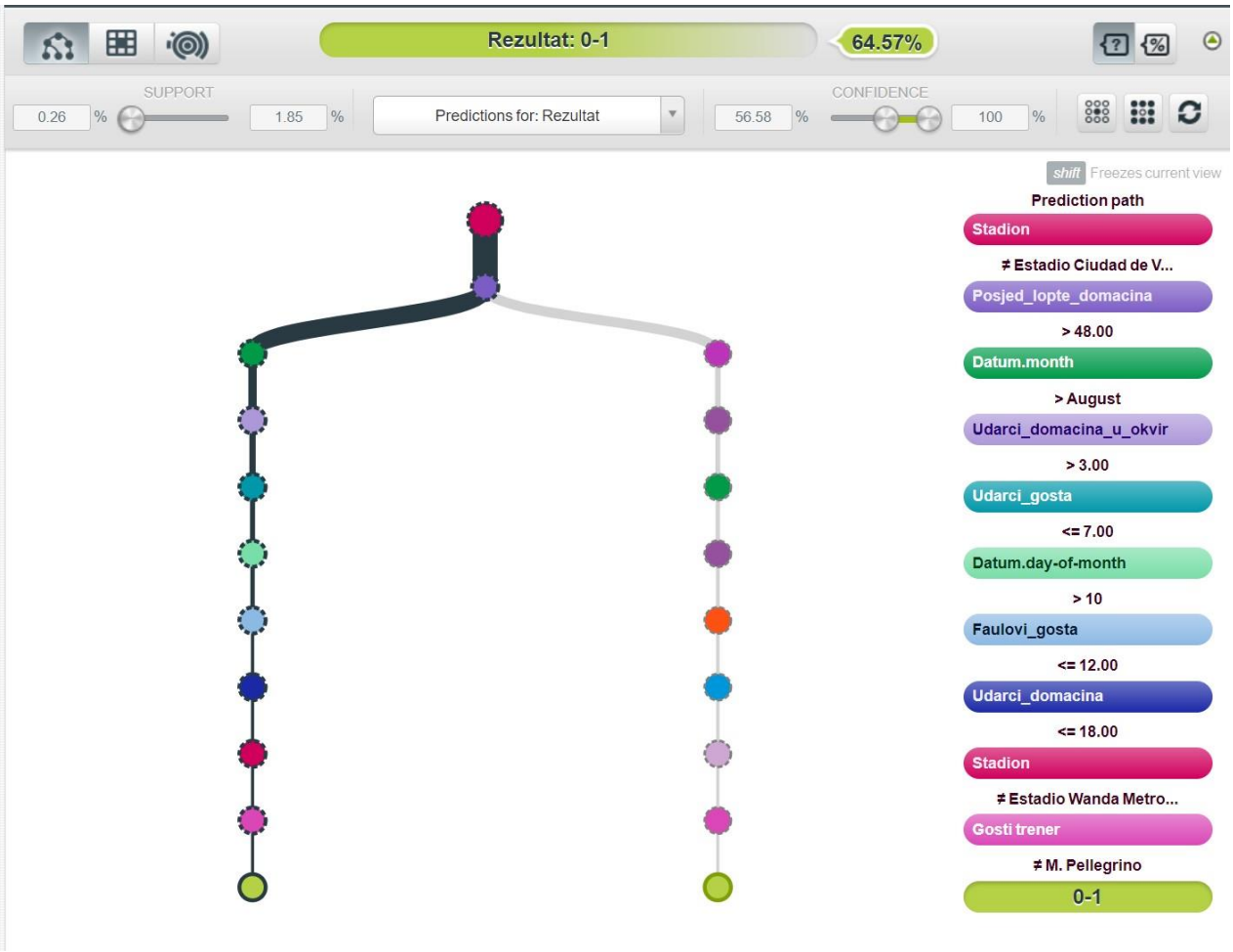
Kao i za sezonu 14./15., tako je i za ovu sezonu moguće vidjeti koji atributi najviše utječu na promatrani atribut za koji je stablo izrađeno. Ti su atributi sljedeći.

- Stadion: 18.01%
- Posjed_lopte_domacina: 15.03%
- Datum.month: 10.96%
- Faulovi_domaćina: 7.02%
- Datum.day-of-month: 6.60%
- Korneri_domacina: 6.33%
- Zaleda_gosta: 6.29%
- Udarci_domacina: 5.75%
- Korneri_gosta: 4.92%
- Kolo: 4.14%

- Udarci_domacina_u_okvir: 4.01%
- Faulovi_gosta: 3.72%
- Zaleda_domacina: 2.03%
- Udarci_gosta: 1.96%
- Udarci_gosta_u_okvir: 1.54%
- Gosti trener: 1.06%
- Datum.day-of-week: 0.57%
- Sudac: 0.08%

Ovdje je vidljiva razlika u odnosu na sezonu 2014./2015. gdje je najveća ovisnost o rezultatu bila u prekršajima gostujuće ekipe, dok je ovdje riječ o stadionu na kojem se igra utakmica. Uz stadion atribut o kojem najviše ovisi rezultat je posjed lopte domaćina. To bi se moglo protumačiti na način da kada ekipa igra na domaćem stadionu, traži veći posjed lopte kako bi stvarala veći broj šansi, upućivala veći broj udaraca te tako automatski sebi povećala šanse za povoljniji rezultat.

Sada dolazi red na pronalazak najpouzdanijeg pravila ovog stabla odlučivanja. Ovdje je riječ o dva pravila koja su najpouzdanija, a ona su za rezultate: 1-1 te 0-1. Zanimljivo je kako je pouzdanost tih pravila jednaka pouzdanosti pravila iz sezone 14./15., što znači da ona iznosi 64,57%.



Slika 7. Najpouzdanije pravilo stabla odlučivanja - sezona 2018./2019.

Također, za ovu sezonu postoji u izvještaju broj utakmica koje su završile određenim rezultatom. U nastavku slijedi pet najčešćih rezultata.

- 1-1, ukupno 50 utakmica završilo ovim rezultatom
- 1-0, ukupno 41 utakmica završila ovim rezultatom
- 2-1, ukupno 38 utakmica završilo ovim rezultatom
- 2-0, ukupno 35 utakmica završilo ovim rezultatom
- 0-1, ukupno 30 utakmica završilo ovim rezultatom

Dakle, u sezoni 18./19. u najčešćih pet rezultata ne postoji rezultat koji je završio bez pogodaka, što znači da je ta sezona bila atraktivnija i zanimljivija. Najčešći rezultat kao i u prethodno promatranom razdoblju je rezultat 1-1. Sljedeća tri rezultata su pobjede domaćina, dok jedan rezultat daje pobjedu gostujuće ekipe. Ovdje je s nekim od pet najčešćih rezultata ukupno završilo 194 utakmica, što daje 51% svih utakmica odigranih u sezoni.

4.3. Ekspertni sustav

Nastankom i razvojem umjetne inteligencije počeli su se stvarati ekspertni sustavi. Oni su zapravo veoma pametni (inteligentni) računalni programi koji koriste znanje i ponašanje čovjeka ili organizacije koji imaju ekspertno znanje i iskustvo u određenom području za rješavanje vrlo složenih problema iz tih istih područja. Prvi koncept ekspertnog sustava javlja se i razvijen je od strane Edwarda Feigenbauma, profesora i osnivača Laboratorija sustava znanja na Sveučilištu Stanford (Petersen, 2016).

Postoji nekoliko karakteristika ekspertnih sustava, a najvažnije su brzina rješavanja zadataka sa stopostotnom točnošću, objašnjavanje te davanje odgovora na temelju teorije ili heurističkih pravila. Također, oni mogu i informirati korisnika koji postavlja pitanja. Danas ti sustavi mogu koristiti i „machine learning“ koji im omogućuje da poboljšaju svoje performanse temeljeno na iskustvu, jednako kao i ljudi. Najviše se koriste u industrijama koje uključuju financijske usluge, telekomunikacije, zdravstveno osiguranje, korisničku podršku, video igrice, javnom prijevozu (Jurić et al., 2005).

U ovom će se radu za kreiranje ekspertnog sustava koristiti alat pod imenom XpertRule Knowledge Builder (XRKB). Na temelju stabla odlučivanja i najpouzdanijeg pravila koje je dobiveno tim stablom, izraditi će se ekspertni sustav koji će napraviti predikciju za to pravilo.

4.3.1. Izrada ekspertnog sustava SDT metodologijom u alatu XRKB

Kao što je već navedeno, u alatu XRKB će se izraditi ekspertni sustav. Međutim, valja napomenuti da će se zbog ograničenja u korištenju demo verzije alata moći koristiti samo 15 elemenata, odnosno prikazati samo najpouzdanije pravilo za svaku sezonu.

U alatu XRKB koristi se tzv. Metodologija inženjerstva znanja, pod nazivom *SDT Methodology*, punim nazivom Metodologija strukturiranih zadaća odlučivanja (eng. Structured Decision Tasks). SDT metodologija se sastoji od 4 koraka, a oni su (Oreški, 2021b):

- Modeliranje problema
- Strukturiranje znanja
- Akvizicija znanja
- Testiranje znanja

4.3.1.1. Modeliranje problema

Kod 1. koraka riječ je o određivanju prirode, opsega te ciljeva aplikacije temeljene na znanju u obliku modela rješavanja problema neovisnog o shemi prezentacije. Ti se problemi mogu prema ljudskom načinu donošenja odluka svrstati u šest kategorija, a one su (Oreški, 2021b):

- Dijagnoza
- Subjektivna procjena
- Kategorička procjena
- Preporuka/Odabir/Savjet
- Klasifikacija
- Interpretacija pravila i propisa

U XRKB alatu se za predstavljanje strukturiranog načina donošenja odluka može koristiti *Map View.* Metode strukturiranja zadaća mogu biti sljedeće (Oreški, 2021b):

- Slijedno ulančavanje unaprijed
- (uvjetno) ulančavanje unaprijed
- (uvjetno) ulančavanje unatrag
- Ponavljanje (petlje)
- Hibridni model

Prikaz ekspertnog sustava za najpouzdanije pravilo stabla odlučivanja u ove dvije sezone spada u kategoriju *Klasifikacija* zbog toga što se odabralo samo jedno pravilo stabla odlučivanja (jedna situacija) i to najpouzdanije pravilo, tj. Jedno od tri najpouzdanija pravila u prvom slučaju te jedno od dva najpouzdanija u drugom slučaju. Uz to, tijekom izvođenja se upisuju točni podaci kako bi se na kraju dobila najbolja opcija, odnosno krajnji rezultat koji je najpouzdanije pravilo. Što se tiče metoda strukturiranja zadaća za ovaj rad se koristila metoda (uvjetno) ulančavanje unaprijed, odnosno svaka sljedeća zadaća koja se izvršavala je ovisila o rezultatu prethodne. Npr. Ukoliko je broj zaleđa gosta bio veći od nekog raspona, nije se izvršavala akcija koja slijedi iza, a to je broj zaleđa domaćina, nego se izvršila druga akcija.

4.3.1.2. Strukturiranje znanja

Drugi korak je strukturiranje znanja i u njemu se radi pretvaranje problema u strukturiranu hijerarhiju zadaća odlučivanja. Strukturiranje zadaća uključuje sljedeće tri stvari (Oreški, 2021b):

- Odvajanje kontrolnih zadaća
- Određivanje atributa
- Grupiranje atributa i izlaza u podzadaće, pa zadaće

U radu je korišten skup atributa za prikaz najpouzdanijeg pravila, u jednom slučaju je to bilo devet atributa, u drugom deset. U oba slučaja većina je numeričkih atributa, kao broj kornera, broj zaleđa i slično, a ostali manji broj atributa su liste. Rezultat, odnosno izlaz koji se pokušava dobiti je rezultat najpouzdanijeg pravila stabla odlučivanja te svaki atribut je ulančan unaprijed, odnosno svaka zadaća koja slijedi, ovisi o tome kako će završiti prethodna te hoće li zadovoljiti određeni uvjet da se nastavi prema dobivanju najpouzdanijeg pravila ili će se dobiti neko manje pouzdano pravilo.

4.3.1.3. Akvizicija znanja

Treći korak SDT metodologije predstavlja proces prikupljanja pojedinačnih zadaća od eksperata, podataka ili dokumentacije. Ciljevi akvizicije su sljedeći (Oreški, 2021b):

- Razvoj pojedinačnih stabala odlučivanja
- Razvoj uzoraka pravila
- Dodavanje procedura logici zadaća odlučivanja

Za tu se svrhu koristi nekoliko metodologija, a one su (Oreški, 2021b):

- Metodologija tablice istine
- Metodologija stabla izuzetaka
- Metodologija vrednovane indukcije
- Automatsko generiranje stabla odlučivanja indukcijom
- Direktno unošenje stabla odlučivanja
- Odabir odgovarajuće metodologije akvizicije znanja

Metodologija korištena za akviziciju znanja u ovom radu je metodologija direktnog unošenja stabla odlučivanja zato što kako bi se dobio poželjan izlaz (najčešći rezultat prema najpouzdanijem pravilu) moraju se poklopiti svi izlazi zadaća. Ta je metodologija korištena prema prikupljenom skupu podatak koji se moraju poklopiti da se dobije zadovoljavajući izlaz. Ukoliko se samo jedna zadaća ne poklopi, tada će se dobiti drugačiji izlaz, što znači da su izlazi isključivi i samim time je ova metoda najprikladnija za korištenje u ovom radu.

4.3.1.4. Testiranje znanja

Zadnji korak je testiranje znanja. U njemu se događa provjera ispravnosti i potpunosti svog znanja, a to osigurava povratnu vezu koja služi za usavršavanje strukturiranja znanja i faza akvizicije do onog trenutka kada je traženi nivo ispravnosti i potpunosti znanja u sustavu postignut. Testiranje sustava također uključuje i provjeru sveukupnih performansi i

operabilnosti sustava, uključujući bazu znanja, strategiju zaključivanja te sučelja za podatke (Oreški, 2021b).

Testiranje ekspertnog sustava je provedeno na način da je sustav pokrenut nekoliko puta te su se koristili različiti testni podaci. Korištenjem tih podataka rezultat koji je dobiven je zadovoljavajuć te odgovara podacima prema kojima je sustav napravljen. Za obje promatrane sezone će biti prikazan jedan slučaj (primjer) testiranja u nastavku rada.

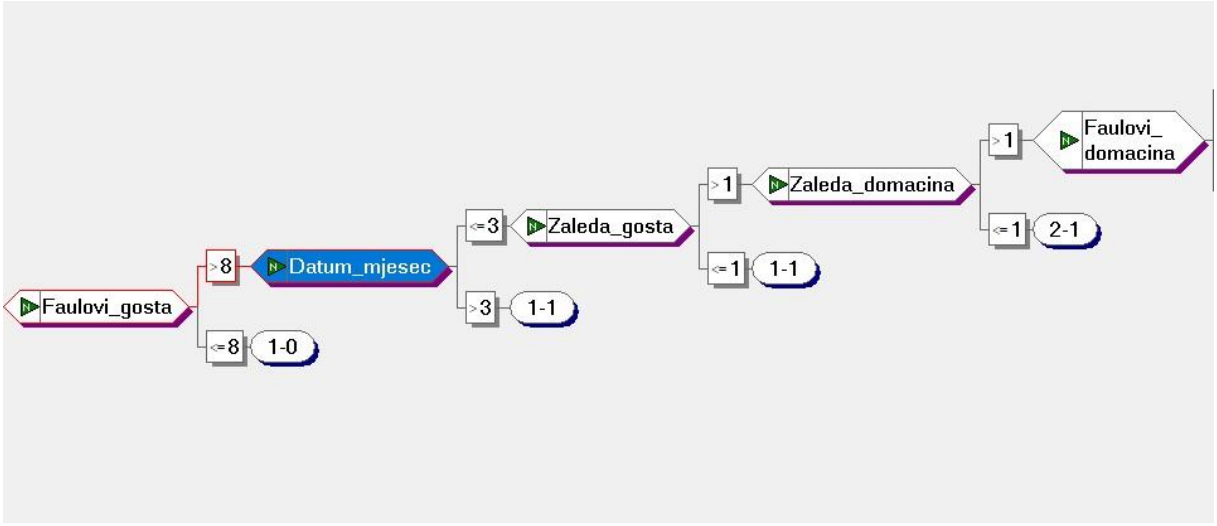
4.3.2. Prikaz ekspertnog sustava za sezonu 2014./2015.

Kako je stablo odlučivanja prikazalo 3 najpouzdanija pravila, ovdje će se prikazati samo jedno od njih, za rezultat 2-0.

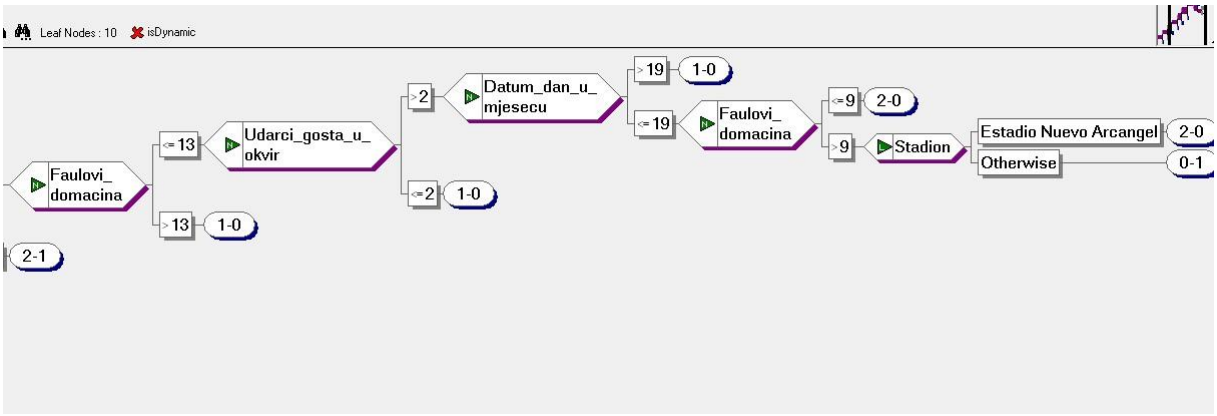
Ekspertni sustava se sastoji od ukupno 9 atributa, a oni su sljedeći:

- Faulovi_gosta
- Datum_mjesec
- Zaleda_gosta
- Zaleda_domacina
- Faulovi_domacina
- Udarci_gosta_u_okvir
- Datum_dan_u_mjesecu
- Stadion

Od tih devet polja, sedam ih je numeričkih s vrijednostima više ili manje od zadanog broja, a ostala dva polja su liste. Dakle, liste su polja „Rezultat“ i „Stadion“ te je zbog toga njima bilo potrebno upisati vrijednosti koje mogu poprimati. Na sljedećoj slici se nalazi izgled ekspertnog sustava za sezonu 2014./2015.

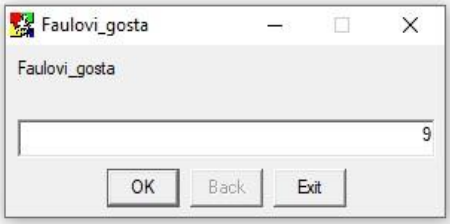


Slika 8. Ekspertni sustav - sezona 2014./2015. - dio 1

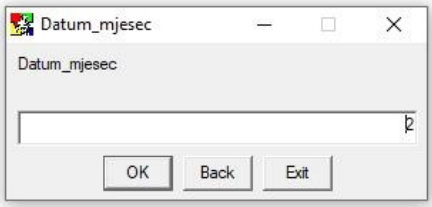


Slika 9. Ekspertni sustav - sezona 2014./2015. - dio 2

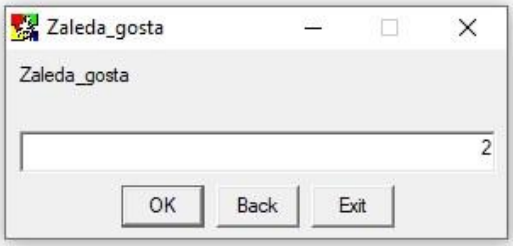
Sada kada je izrađeno najpouzdanije pravilo, potrebno je pokrenuti testiranje sustava, odnosno predviđanje na temelju pravila. U nastavku slijede slike koje prikazuju predviđanje.



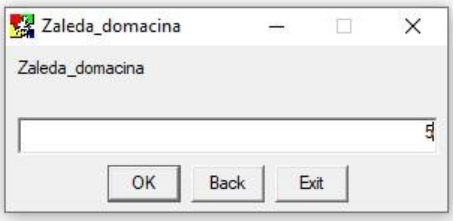
Slika 10. Predviđanje sustava - atribut 1



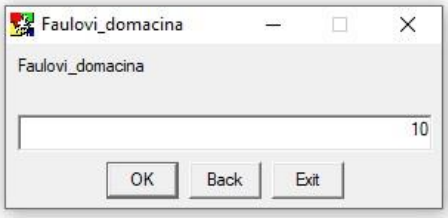
Slika 11. Predviđanje sustava - atribut 2



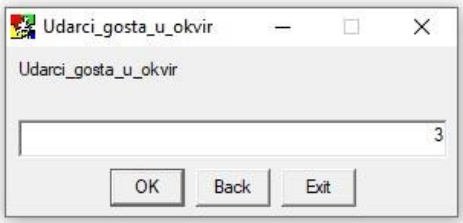
Slika 12. Predviđanje sustava - atribut 3



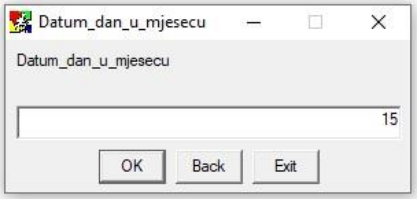
Slika 13. Predviđanje sustava - atribut 4



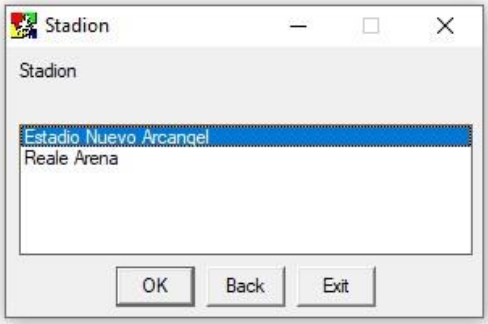
Slika 14. Predviđanje sustava - atribut 5



Slika 15. Predviđanje sustava - atribut 6



Slika 16. Predviđanje sustava - atribut 7



Slika 17.Predviđanje sustava - atribut 8

Slika 18. Predviđanje sustava - krajnji rezultat

Dakle, ovo je bio rezultat predviđanja za sezonu 2014./2015. putem XRKB-a na temelju podataka koji su uneseni u sustav.

4.3.2.1. Prikaz ekspertnog sustava za sezonu 2018./2019.

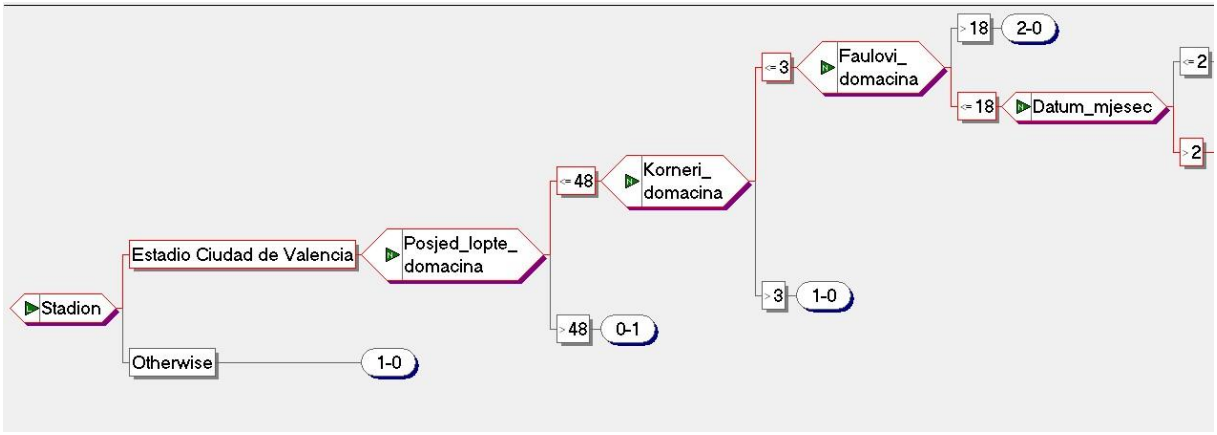
U ovom poglavlju će se ponoviti isti postupak iz prethodnog poglavlja, samo za sezonu i stablo odlučivanja iz sezone 2018./2019. Najpouzdanija pravila su dva pa će tako se opet uzeti samo jedno pravilo i to za rezultat 1-1.

Ekspertni sustav sastoji se od ukupno 10 atributa, a oni su sljedeći:

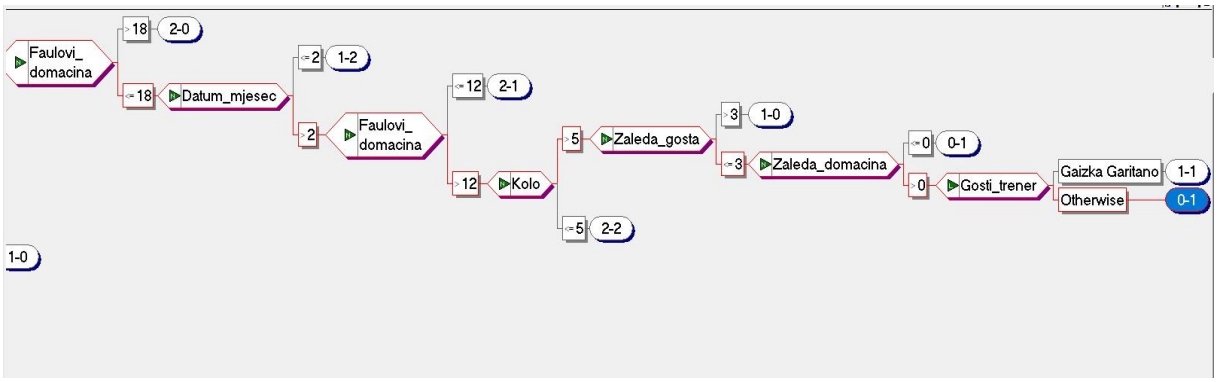
- Stadion
- Posjed_lopte_domaćina
- Korneri_domaćina
- Faulovi_domaćina
- Datum_mjesec
- Kolo
- Zaleđa_gosta
- Zaleđa_domaćina
- Gosti_trener
- Rezultat

Od tih deset polja, sedam je numeričkih te su samo tri polja liste, „Stadion“ i „Gosti_trener“ te „Rezultat“.

Na sljedećim slikama je prikaz ekspertnog sustava za predviđanje rezultata za sezonu 2018./2019.

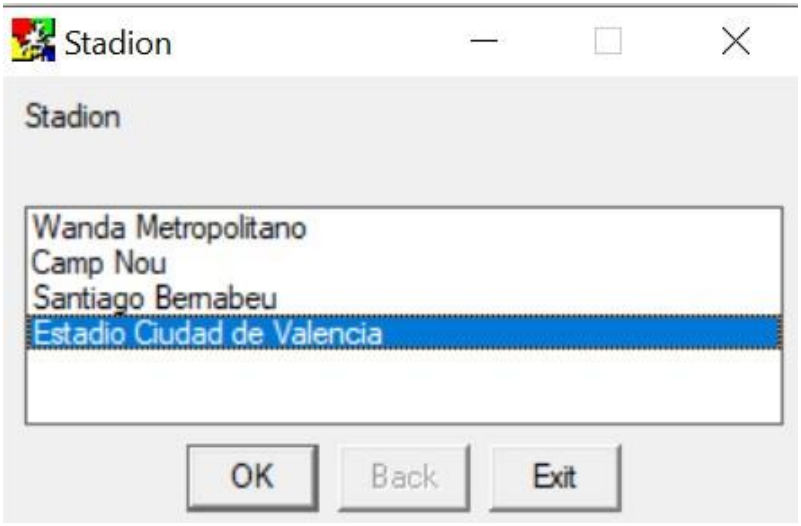


Slika 19. Ekspertni sustava - sezona 2018./2019. - dio 1

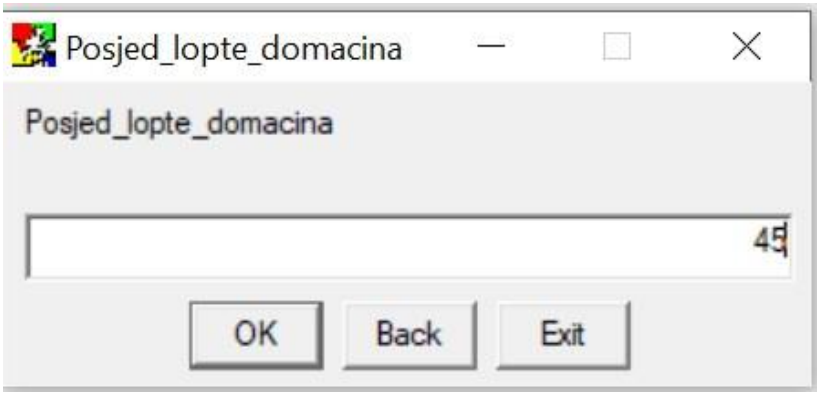


Slika 20. Ekspertni sustava - sezona 2018./2019. - dio 2

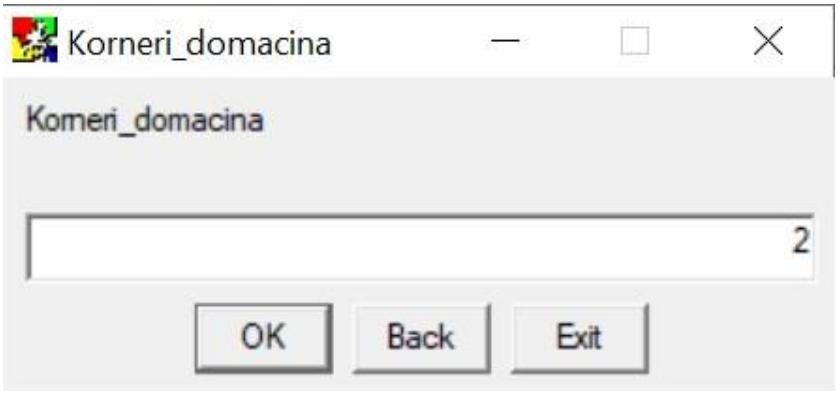
Kao i u prethodnom poglavlju, kada je izrađeno najpouzdanije pravilo, potrebno je pokrenuti testiranje sustava, odnosno predviđanje na temelju pravila. U nastavku slijede slike koje prikazuju predviđanje.



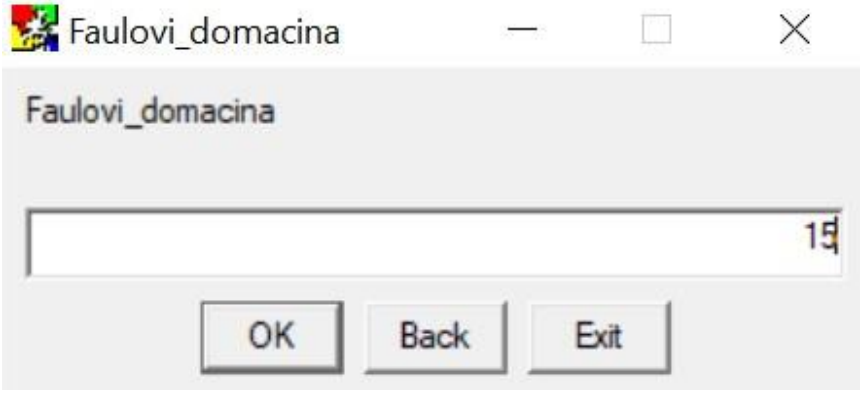
Slika 21. Predviđanje sustava sezona 2018./2019. - atribut 1



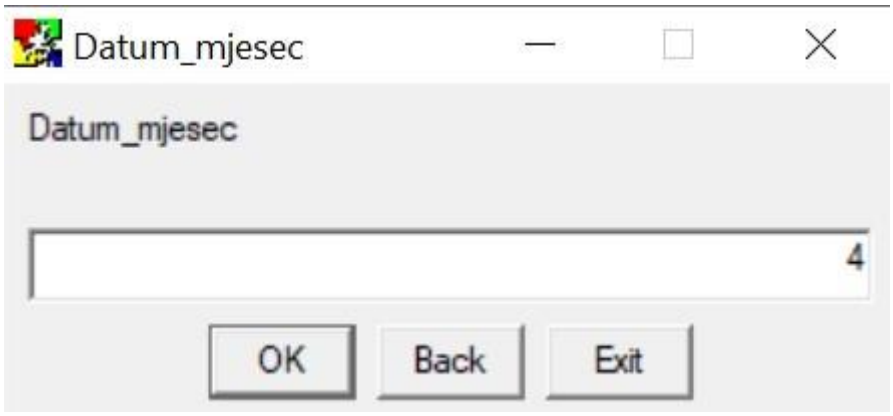
Slika 22. Predviđanje sustava sezona 2018./2019. - atribut 2



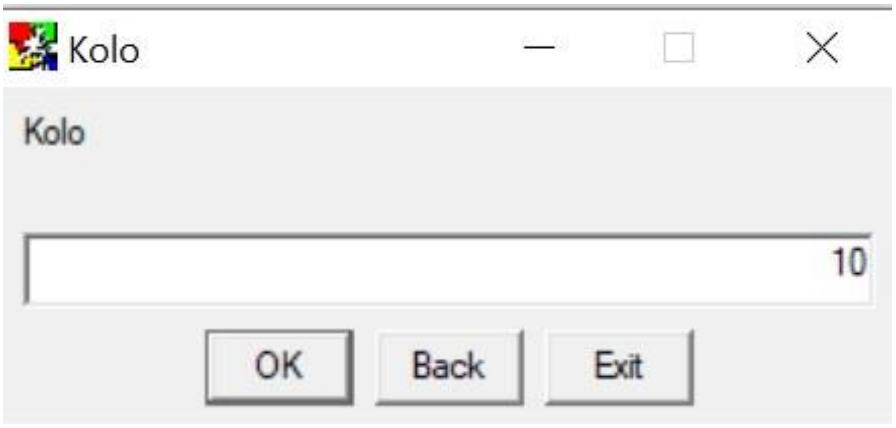
Slika 23. Predviđanje sustava sezona 2018./2019. - atribut 3



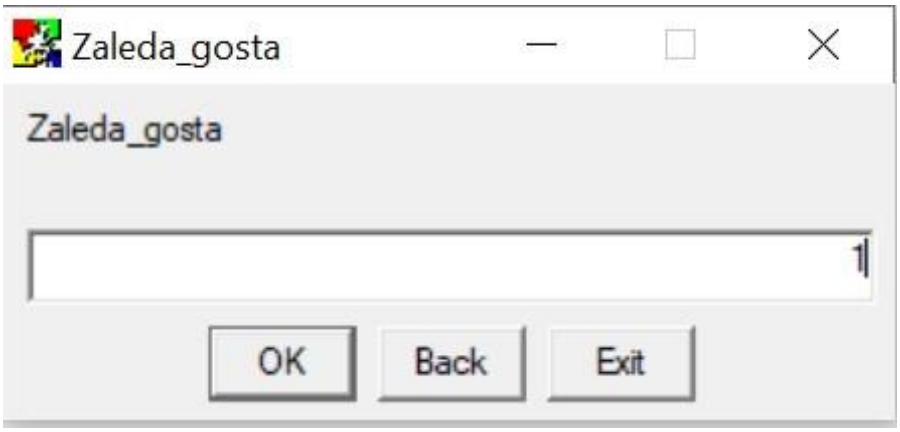
Slika 24. Predviđanje sustava sezona 2018./2019. - atribut 4



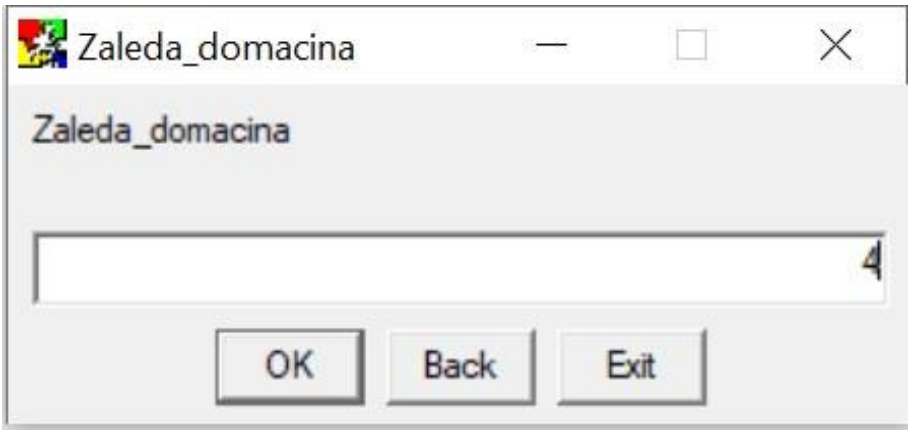
Slika 25. Predviđanje sustava sezona 2018./2019. - atribut 5



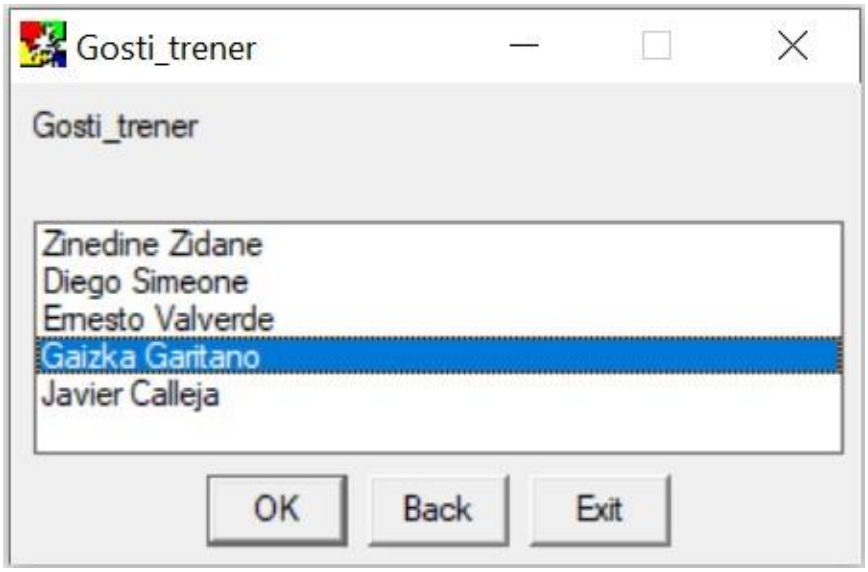
Slika 26. Predviđanje sustava sezona 2018./2019. - atribut 6



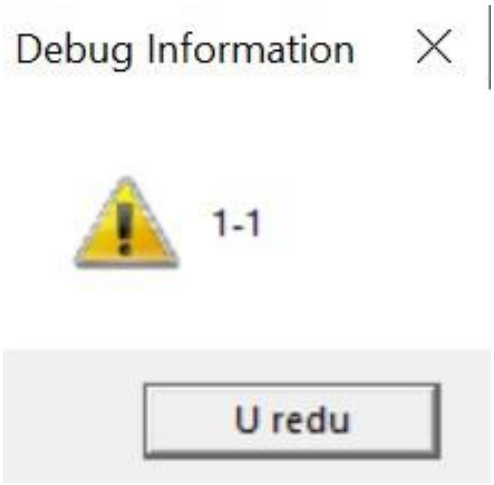
Slika 27. Predviđanje sustava sezona 2018./2019. - atribut 7



Slika 28. Predviđanje sustava sezona 2018./2019. - atribut 8



Slika 29. Predviđanje sustava sezona 2018./2019. - atribut 9



Slika 30. Predviđanje sustava sezona 2018./2019. - rezultat

Ovime je završeno predviđanje rezultata za sezonu 2018./2019. korištenjem ekspertnog sustava i alata XRKB.

4.4. Neuronske mreže

Od samog početka razvoja računala postavlja se pitanje hoće li ikada postojati računalo koje će imati mogućnost samostalnog razmišljanja, hoće li umjetna inteligencija uspjeti doseći tu razinu. Danas se pojam neuronskih mreža jako usko veže uz pojam umjetne inteligencije. Razlog tome je što su neuronske mreže već značajno promijenile internetske poslove kao sto su web pretraživanja, online reklame te kreiranje preporuka. neuronske mreže počinju primjenjivati i u medicini, prepoznavanju govora, prevođenju tekstova, prepoznavanju objekata, vožnji autonomnih vozila, procjeni rizika u bankarstvu i druge. Činjenica da je algoritam uspio nadmašiti radiologe s dugogodišnjim iskustvom u prepoznavanju upale pluća na temelju rendgenskog snimka, jedan je od pokazatelja u kakvom smjeru kreće napredak i da je to stvarno budućnost (AI, 2021).

Neuronska mreža je zbir umjetnih neurona (najčešće kao apstraktnih pojmova) koji su međusobno povezani i interaktivni kroz operacije obrade signala. Uređena je po uzoru na ljudski mozak. Mreža može imati niz ili jedan ulaz ulaza i uvijek jedan izlaz, između kojih se nalazi jedan ili više tzv. skrivenih slojeva (tzv. višeslojne mreže). Pojedinačni neuroni su, kao i slojevi, međusobno spojeni vezama kroz koje idu signali. Veze među njima se aktiviraju ako je zadovoljen uvjet postavljen tzv. aktivacijskom funkcijom (Židov, 2018).

Faze rada umjetnih neuronskih mreža moguće je podijeliti na (Židov, 2018):

- Fazu učenja
- Fazu selekcije
- Fazu testiranja
- Operativnu fazu

Dvije najvažnije faze su faza učenja i faza testiranja, u fazi selekcije nastoji se optimizirati duljina treniranja te broj neurona u skrivenom sloju. Operativna faza se odnosi na korištenje neuronske mreže.

Neuronske mreže primjenjuju se u raznim područjima, neka od njih su zdravstvo, obrazovanje, elektrotehnika, poslovanje, vojska, financije i brojne druge gospodarske grane. Kako s vremenom raste uporaba neuronskih mreža tako se i povećava broj algoritama za učenje neuronskih mreža. Neki od tih algoritama su prilagođeni za rješavanje samo jedne vrste problema, dok su neki univerzalni i moguće ih je koristi za sve tipove problema. U nastavku slijedi podjela algoritama (Židov, 2018).

Algoritmi se dijele na (Židov, 2018):

- Nadgledane – algoritmi za koje su potrebne poznate ulazne i izlazne vrijednosti
- Nenadgledane – algoritmi za koje su potrebne samo ulazne vrijednosti

U nadgledane algoritme ubrajamo (Židov, 2018):

1. Za probleme predviđanja
    - Mreža širenja unatrag
    - Modularna mreža
    - Mreža s radijalno zasnovanom funkcijom
    - Mreža opće regresije

2. Za probleme klasifikacije
    - Vjerojatnostna mreža
    - Mreža učeće vektorske kvantizacije

U nenadgledane algoritme ubrajaju se (Židov, 2018):

- Kohonenova mreža
- Mreža adaptivne rezonantne teorije

Algoritam širenja natrag je najpopularnije pravilo učenja. Način na koji algoritam radi je opisan u nastavku kroz pet koraka (Kliček, 2021a):

1. Prvo se postavljaju ulazne vrijednosti koje se preko skrivenog sloja prenose do izlaznog sloja i ostvaruju izlani vektor.
2. Za vrijeme prenošenja ulazno vektoraračunaju se ulazne i izlazne vrijednosti za svaki neuron u skrivenom sloju.
3. Za svaki neuron u izlaznom sloju računa se lokalna greška.
4. Za svaki sloj, počevši od predzadnjeg i sloja neposredno nakon ulaznog potrebno je izračunati skaliranu lokalnu grešku idelta težinu.
5. Obnoviti sve težine veza u mreži dodavanjem delta težina prijašnjim vrijednostima.

Prednost mreže širenja unatrag je što ima dodatne slojeve koji dopuštaju da se rezultat jednog sloja dodatno obrađuje, uređuje i stvara kompleksni sustav, kao nedostatak navodi se dugotrajno treniranje te osjetljivost na početne vrijednosti težina. Postoje ključni pojmovi koje je potrebno znati da bi se neuronske mreže razumjele, a oni su sljedeći: „Duboko učenje“ (eng. Deep learning), neuroni i tijela stanica te dendriti.

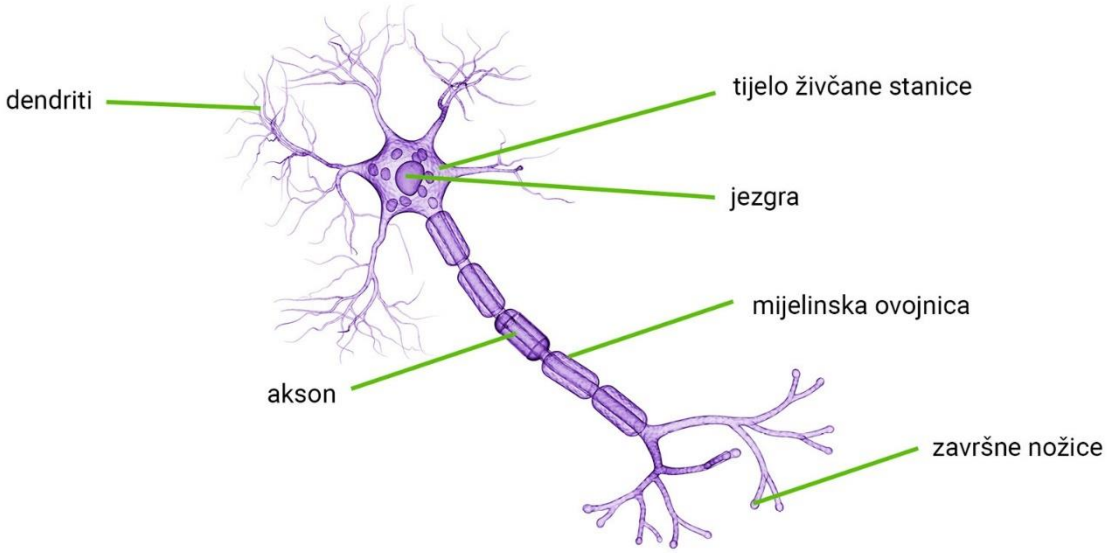
4.4.1. Neuroni

Neuronska mreža, bez obzira na to je li biološka ili umjetna, sastoji se od velikog broja jednostavnih jedinica, neurona, koje primaju signale jedna od druge i prenose ih jedna do druge. Neuroni su jednostavni procesori informacija koji se sastoje od tijela stanice i produžetaka koji neurone međusobno povezuju. Većinu vremena ne rade ništa, odnosno miruju i čekaju da kroz produžetke stignu signali (AI, 2021).

Neuron ili živčana stanica je osnovni element živčanog sustava. Sastoji se od 3 glavna elementa (Židov, 2018):

- tijelo s jezgrom,
- kratki ogranci (dendriti),
- dugi ogranak (akson).

Najlakše objašnjeno, neuron prima signal kroz dendrite, obraduje informacije u jezgri te ih šalje u sljedeći neuron pomoću aksona (Židov, 2018).
Na sljedećoj slici prikazuje se građa neurona.



Slika 31. Građa neurona (e-skole.hr, 2021)

4.4.2. Izrada neuronske mreže

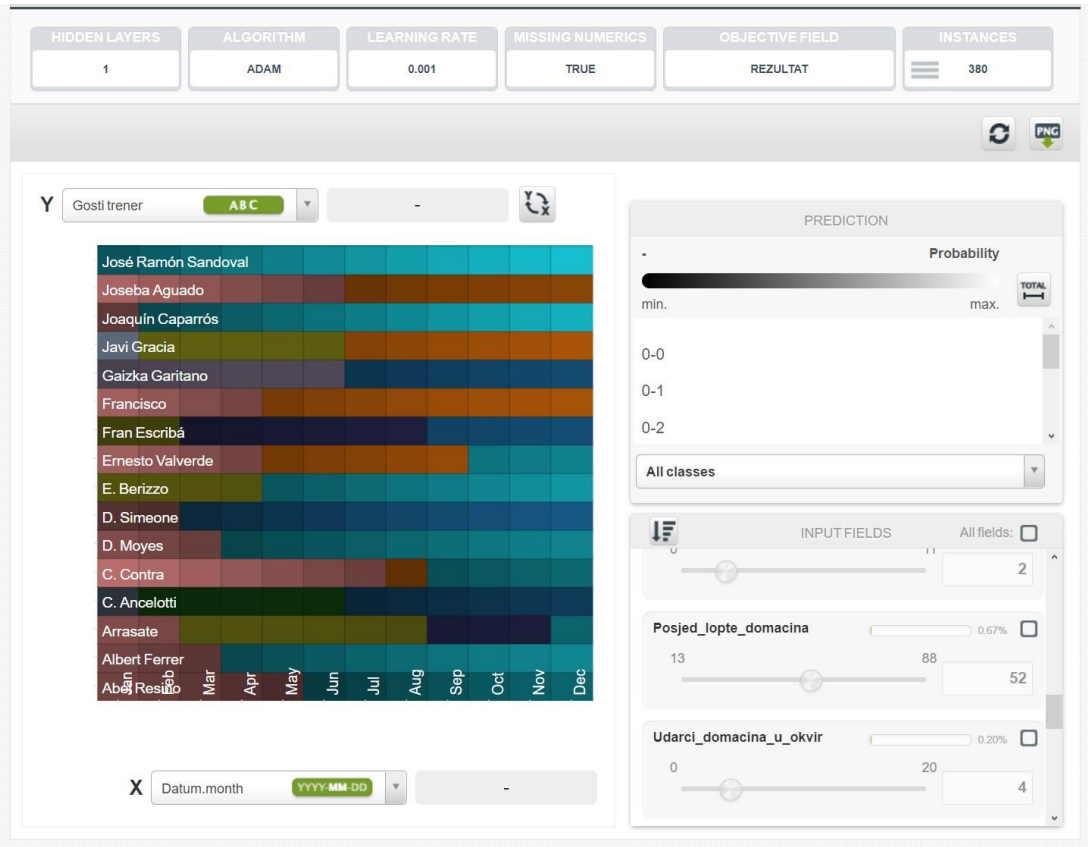
Neuronske mreže se izrađuju u alatu BigML, baš kao i stablo odlučivanja. Kako bi se izradila neuronska mreža u BigML-u, potrebno je otići na dataset te na gumb „Configure“ i u padajućem izborniku izabrati opciju „Deepnet“.

Nakon odabira opcije za kreiranje neuronske mreže, otvara se izbornik s konfiguracijom za neuronsku mrežu. Tamo se odabire koji atribut se želi promatrati kroz neuronsku mrežu te također postoje i ostale opcije vezane uz kreiranje neuronske mreže.

Kada se odabere atribut koji se želi promatrati i opcije koje odgovaraju za kreiranje neuronske mreže, ona je kreirana.

4.4.2.1. Prikaz neuronske mreže za sezonu 2014./2015.

Na sljedećoj slici se nalazi prikaz neuronske mreže za sezonu 2014./2015.



Slika 32. Neuronska mreža - sezona 2014./2015.

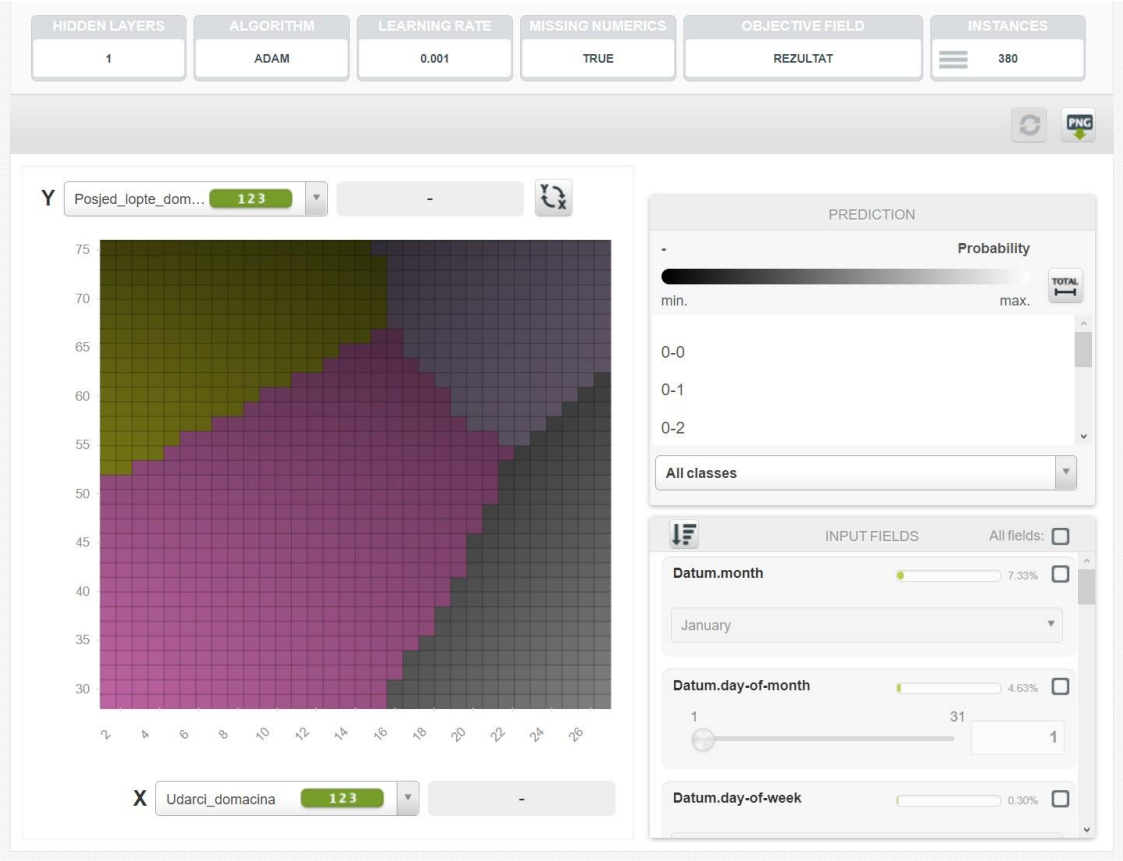
Na X i Y osi se odabiru atributi koji najviše utječu na promatrani atribut, a u ovom slučaju su to atributi koji se odnose na gostujućeg trenera te na mjesec u godini. Kao najčešći rezultat, obzirom na dva atributa koja najviše utječu na promatrani, pokazuje se rezultat 1-1 dok je

sljedeći najčešći rezultat 0-0. Treći rezultat koji se dobije raznim kombinacijama vrijednosti je rezultat 2-0. Rezultat 1-1 javlja se u slučaju kada je trener gostiju Joaquin Caparros, trener Granade, a mjesec je prosinac. Kada se uzmu u obzir rezultati koje je Granada kao gost odigrala u prosincu, dvije od 3 su završile 1-1, a u zadnjoj su izgubili 2-0. Ova je statistika zanimljiva jer pokazuje kako treneri kada gostuju kod neke ekipe igraju. Obzirom da su najčešći rezultati neriješeni ishodi ili porazi gosta, vidljivo je kako gosti očito kada nemaju veliku podršku publike igraju slabije, te su u podređenom statusu u odnosu na domaćina. Potpuna lista atributa o kojima najviše ovisi rezultat slijedi u nastavku:

- Gosti trener, 22,19%
- Datum.month, 17,03%
- Gost, 11,34%
- Domaći trener, 11,20%
- Sudac, 6,48%
- Stadion, 5,37%
- Zaleda_domacina, 4,76%
- Domaćin, 4,51%
- Posjed_lopte_domacina, 3,92%
- Udarci_gosta_u_okvir, 3,59%
- Posjed_lopte_gosta, 2,07%
- Faulovi_gosta, 1,74%
- Udarci_domacina, 1,27%
- Udarci_gosta, 1,11%
- Korneri_gosta, 1,08%
- Kolo, 0,51%
- Datum.day-of-week, 0,51%
- Faulovi_domaćina, 0,34%
- Korneri_domacina, 0,30%
- Zaleda_gosta, 0,28%
- Udarci_domacina_u_okvir, 0,19%
- Datum.year, 0,10%
- Datum.day-of-month, 0,10%

4.4.2.2. Prikaz neuronske mreže za sezonu 2018./2019.

Kao i za prethodnu sezonu, postupak će biti isti, prvo će se prikazati neuronska mreža za sezonu 2018./2019. te zatim prikazati i atributi o kojima najviše ovisi rezultat.



Slika 33. Neuronska mreža - sezona 2018./2019.

Kako je moguće vidjeti sa prethodne slike, na rezultat najviše utječu atributi „Posjed lopte domaćina“ te „Udarci domaćina“. Kombinacijom vrijednosti tih dvaju atributa dobiva se da je najčešći rezultat 1-1, a sljedeća dva su 1-0 te 1-2. Dva od tri rezultata su jednaka kao i za sezonu 2014./2015. iako su atributi o kojima to ovisi drugačiji. Zanimljivo je kako dok domaćini imaju manji posjed lopte, a veći broj udaraca, dobiva se najveći postotak točnosti za određeni za najčešći rezultat 1-1. Kod rezultata 1-0, posjed manji posjed lopte daje veću točnost, ali broj udaraca mora također biti manji. Što znači da kada se domaća ekipa brani, odnosno igra defanzivno, ali pokušava više pucati prema golu gostujuće ekipe, rezultat često završava neriješeno 1-1, a kada se domaćin brani, ali ne puca puno, odnosno očito traži prilike samo iz kontra napada, tada se očekuje pobjeda domaćina. To je vrlo zanimljivo jer bi bilo za očekivati da kada ekipa više šutira prema golu, zabije i više golova, odnosno pobjedi utakmicu, a kada ima mali broj udaraca te se brani odigra neriješeno ili izgubi. Međutim, ovdje je vidljivo kako situacija u stvarnosti nije takva, nije potpuno logična. Lista ovisnosti atributa o rezultatu je sljedeća:

- Udarci_domacina, 22,65%
- Posjed_lopte_domacina, 15,54%
- Domaći trener, 8,41%
- Datum.month, 7,33%
- Udarci_domacina_u_okvir, 7,02%
- Gosti trener, 5,81%
- Posjed_lopte_gosta, 5,23%
- Datum.day-of-month, 4,63%
- Domaćin, 3,34%
- Korneri_domacina, 3,03%
- Faulovi_domaćina, 2,51%
- Kolo, 2,25%
- Sudac, 2,21%
- Udarci_gosta, 1,91%
- Udarci_gosta_u_okvir, 1,54%
- Faulovi_gosta, 1,41%
- Stadion, 1,33%
- Korneri_gosta, 1,02%
- Zaleda_domacina, 1,02%
- Gost, 0,86%
- Zaleda_gosta, 0,67%
- Datum.day-of-week, 0,30%

4.5. Bayesove mreže

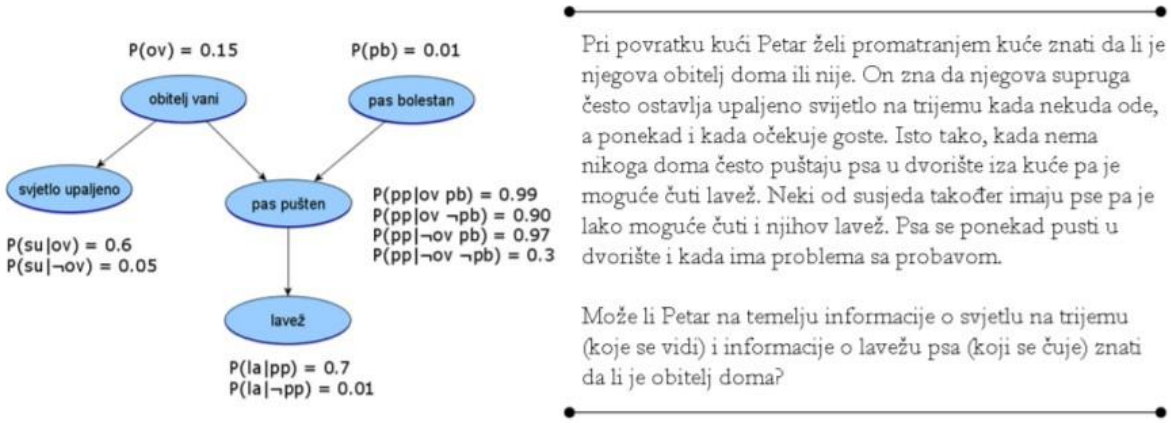
Treća metoda koja će se koristiti i prikazati je Bayesova mreža. Bayesove mreže su prediktivni grafički model kojim se prikazuju zavisnosti između varijabli. Ta mreža je usmjereni aciklički graf u kojem čvorovi predstavljaju varijable, a bridovi njihove međuzavisnosti. Graf je acikličan, što znači da ne smiju postojati ciklusi u grafovima, odnosno, varijable moraju biti nezavisne. Pomoću Bayesove mreže može se izračunati vjerojatnost bilo koje nepoznate varijable u sustavu temeljem drugih poznatih varijabli (Prcela, 2010).

Bayesove mreže imaju dugu povijest, a najvažniji događaji vezani uz njih su se dogodili osamdesetih godina prošlog stoljeća. Tada se formirala opće prihvaćena struktura mreže, mnogi alati za rasuđivanje su razvijeni te su se tada Bayesove mreže počele koristiti za razne primjene. Nešto kasnije su se počeli pojavljivati razni algoritmi za učenje Bayesovih mreža iz skupova podataka. Danas Bayesove mreže predstavljaju model primjenjiv u računalnim sustavima u raznim područjima ljudskog djelovanja (Soni, 2018).

Za definiranje Bayesove mreže, potrebno je definirati sljedeće atribute (Prcela, 2010):

- Čvorovi mreže – varijable u problemu
- Mogući ishodi svih čvorova – vrijednosti koje varijable mogu poprimiti
- Bridovi mreže – povezanost varijabli
- Združene distribucije vjerojatnosti ishoda u svakom pojedinom čvoru ovisno o njegovim roditeljima u mreži

Za one čvorove koji nemaju roditelje, potrebno je definirati samo „a priori“ očekivanja njihovih ishoda. „A priori“ očekivanja ishoda onih čvorova koji imaju roditelje definirana su posredno preko pripadnih tablica združene distribucije vjerojatnosti i preko „a priori“ očekivanja njihovih roditelja (Peraić, 2012).



Slika 34. Primjer Bayesove mreže (Peraić, 2012)

Također, Bayesova mreža ima nekoliko osnovnih svojstava, a ona su sljedeća (Kliček, 2021b):

- Kada model kodira ovisnosti između svih varijabli, on može rješavati situacije kada nedostaju neki unosi podataka.
- Bayesove mreže se mogu koristiti za učenje uzročnih ovisnosti, te se stoga mogu koristiti da poboljšaju razumijevanje o problemskoj domeni i predviđaju posljdice nekog djelovanja (npr. u poslovnim, društvenim, ekološkim i sličnim sustavima).
- Zbog toga što model ima i uzročnu i vjerojatnosnu semantiku, model je idealan prikaz kombinacijom prethodnog znanja (koji često dolazi u uzročnom obliku) i podataka.
- Bayesove statističke metode povezane s Bayesovim mrežama nude učinkovit i principjelan pristup za izbjegavanje pretreniranosti podataka.

4.5.1. Bayesova formula

Bayesova mreža je naziv dobila po britanskom matematičaru Thomasu Bayesu (1702.-1761.) koji je u svom radu opisao matematičku formulu koja danas ima veliku važnost kod teorije vjerojatnost. Formula je opisana u 18. stoljeću, međutim pravu primjenu je dobila tek u 20. stoljeću kada se krenulo razvijati područje umjetne inteligencije i to u području strojnog učenja. Formula je sljedeća (Perić, 2012):

$$P(H_i|A) = \frac{P(H_i)P(A|H_i)}{P(A)}$$

gdje je $\{H_1, H_2, \dots, H_n\}$ potpun skup događaja na vjerojatnosnom prostoru, a *A* je događaj za koji vrijedi da je $P(A) > 0$ u vjerojatnosnom prostoru. Bayesova formula računa vjerojatnost da se potvrdio skup početnih hipoteza, ali uz uvjet da je ostvaren događaj A. Za primjenu formule je potrebno poznavati vjerojatnosti $P(A)$ i $P(H_i)$ te je potrebna i statistika kojom se određuje vjerojatnost $P(A|H_i)$. Formula vrijedi u slučaju da nema međusobne ovisnosti između događaja A i niza hipoteza H (Perić, 2012).

4.5.2. Zaključivanje u Bayesovoj mreži

Bayesova vjerojatnosna mreža može se koristiti za probabilističko zaključivanje o vjerojatnostima bilo kojeg čvora u mreži ako su poznate tablice uvjetnih vjerojatnosti. U većini slučajeva čvor kojemu tražimo razdiobu vjerojatnosti nema poznate vjerojatnosti za neposredne čvorove pretke, ali je ipak moguće izračunati vjerojatnosti. Problematični su izračuni sa velikim brojem varijabli jer je točan proračun vjerojatnosti za proizvoljnu Bayesovu vjerojatnosnu mrežu težak kombinatorički problem. Predložene su mnogobrojne metode za probabilističko zaključivanje u Bayesovim mrežama, pri čemu mnoge preferiraju učinkovitost naspram preciznosti. Preciznost implicira točna rješenja (eng. exact solutions), no u tom slučaju čak i mreže od 10-ak čvorova zahtijevaju puno vremena. Alternativa su približna

rješenja (eng. approximate solutions) sa kojima će se dobiti rješenja sa vrlo malim odstupanjima od točnih rješenja. U nastavku će se pokazati način zaključivanja koji vrijedi u svakoj Bayesovoj mreži. Naime, računanje „a priori“ vjerojatnosti nekog čvora podrazumijeva računanje ishoda kad u mreži ne postoji nijedan dokaz. Suprotno tome, propagacija unaprijed/unatrag služi za računanje ishoda nekog proizvoljnog čvora kad su u mreži postavljeni dokazi (Soni, 2018).

4.5.3. Izrada Bayesove mreže u alatu Netica

Kako bi se izradila Bayesova mreža koristio se alat Netica. Netica je vrlo jednostavan alata namijenjen upravo za izradu Bayesove mreže. Samo da se napomene, kao i kod XRKB alata, u Netici je također korištena limitirano izdanje programa u kojem je moguće koristiti samo 15 atributa.

Za izradu Bayesove mreže potrebno je u Netici izabrati *File –> New Network* da bi se kreirala nova mreža.

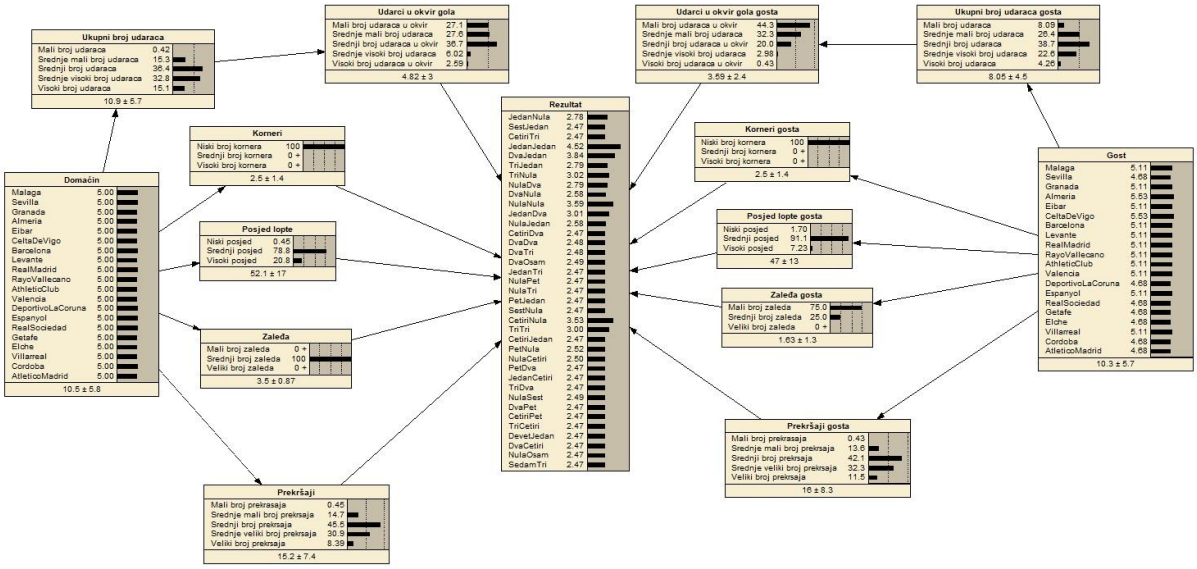
Kad je izrađena Bayesova mreža, potrebno je dodati čvorove koji će se koristiti u Bayesovoj mreži. To se radi na način da se iz taba *Modify* izabere *Add -> Nature Node.*

Kada su dodani svi čvorovi u mrežu, potrebno je tim čvorovima dodati vrijednosti koje mogu imati te imena, koja moraju biti jednaka imenima polja iz skupa podataka koji će se kasnije učitati u Bayesovu mrežu. Ime te vrijednosti se dodaju dvoklikom na čvor u polja *Name*, odnosno *State.*

Kada su sve vrijednosti dodane, strelicom se spoje čvorovi u mrežu. Kada je mreža napokon kreirana, može se učitati skup podataka u Neticu te se može pokrenuti kompilacija, odnosno predviđanje na temelju skupa podataka. Podaci se učitavaju na sljedeći način: *Cases -> Learn -> Incorp Case File*.

4.5.4. Prikaz Bayesove mreže za sezonu 2014./2015.

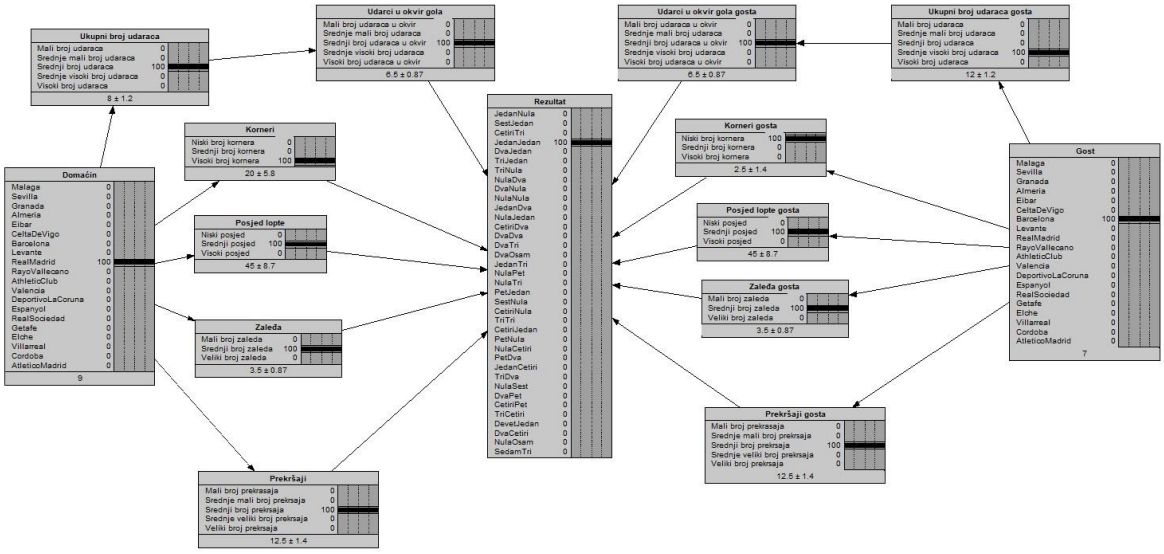
Bayesova mreža za sezonu 2014./2015. je kreirana od 15 čvorova koliko i program u limitiranom izdanju dopušta. Slijedi prikaz Bayesove mreže.



Slika 35. Bayesova mreža za sezonu 2014./2015.

Dakle, u ovoj mreži postoje čvorovi *Domaćin, Gost, Rezultat* te statistika domaćina i gosta, odnosno njihovi korneri, broj udaraca, broj zaleđa i drugo. Svi ti statički čvorovi ovise o čvoru *Domaćin,* odnosno o čvoru *Gost,* ovisno za kojeg se od njih dvoje ti podaci odnose. A o svim tim statističkim podacima ovisi zapravo i čvor *Rezultat*, koji sadrži sve rezultate koji su se dogodili u sezoni 2014./2015.

Sada slijedi primjer predviđanja zavisne varijable, odnosno jedne hipotetske situacije.

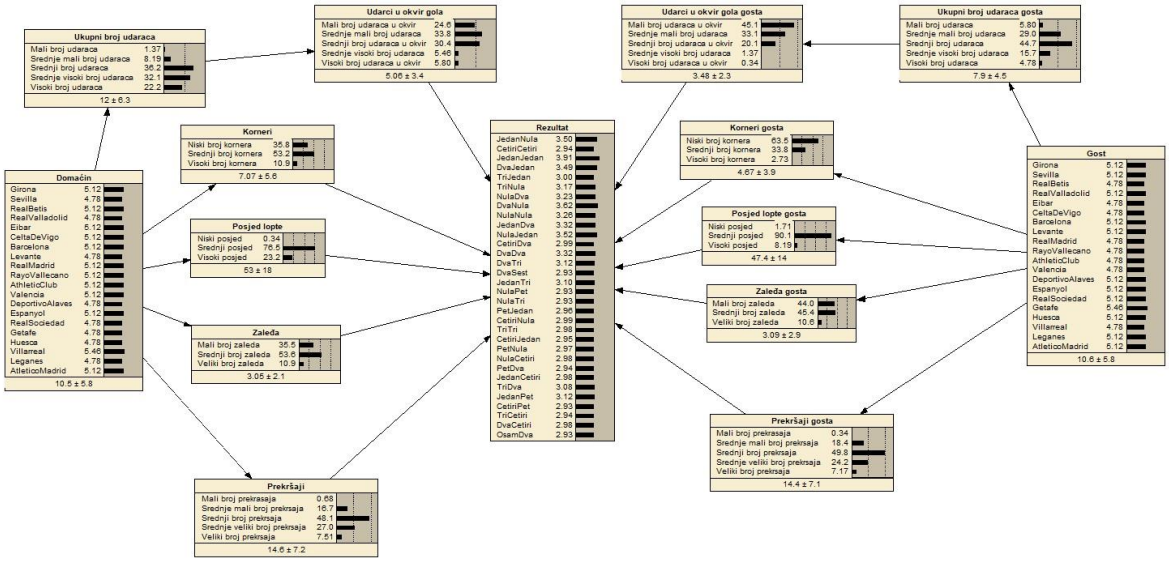


Slika 36. Primjer predviđanja zavisne varijable za sezonu 2014./2015.

Dakle, u ovom slučaju kao domaćin je uzet klub *Real Madrid,* koji na svom stadionu često vrši pritisak, radi dobar broj prekršaja te ne daje protivniku previše prostora za igru. Međutim, s druge strane kao gost je *Barcelona,* ekipa koja voli imati te ima jako veliki posjed lopte, bez obzira protiv koga igrali, sa slabijim brojem kornera, ali više kreiranih akcija i udaraca. Tako u ovom slučaju domaćin *Real Madrid* ima srednji broj udaraca prema golu *Barcelone* te isto tako srednji broj udaraca u okvir gola. *Barcelona* pak ima veći broj udaraca što odgovara njihovom stilu, ali isto kao i domaćin, srednji broj udaraca u okvir gola. Što se tiče kornera, tu je u prednosti *Real Madrid*, koji ima visoki broj kornera, u odnosu na *Barcelonu*, koja ima niski broj kornera. *Posjed lopte* je kod obje ekipe srednji, iako je kod *Barcelone* posjed zasigurno nešto veći od *Real Madrida.* U broj ***zaleđa,*** obje ekipe su jednake i imaju srednji broj zaleđa što znači oko 3 zaleđa svaki. Ista je situacija i sa brojem prekršaja, obje ekipe rade dobar broj prekršaja te on iznosi otprilike 13 prekršaja svaki. Na kraju dolazimo do *Rezultata,* koji u ovom hipotetskom slučaju za ovo predviđanje daje rezultata 1-1, što je i čest rezultat, a ujedno je često i najrealniji, obzirom na ono što se vidi na terenu između ove dvije ekipe.

4.5.5. Prikaz Bayesove mreže za sezonu 2018./2019.

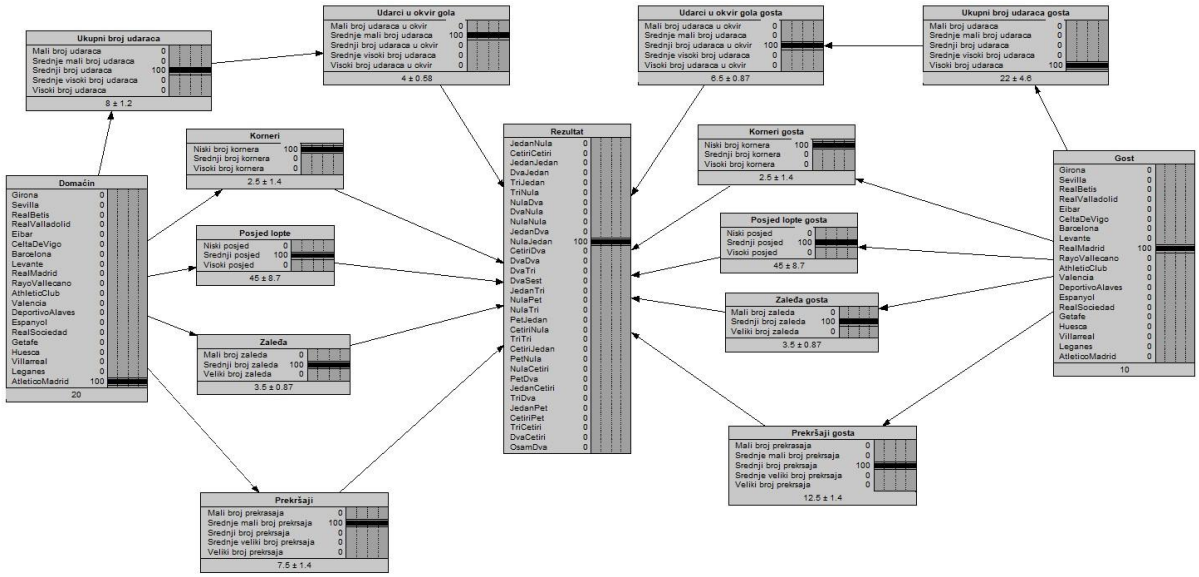
Za sezonu 2018./2019. se ponavlja isti postupak kao i ranije, što znači da prvo u nastavku slijedi prikaz Bayesove mreže za sezonu 2018./2019.



Slika 37. Prikaz Bayesove mreže za sezonu 2018./2019.

Kao i u prethodnoj mreži, u ovoj mreži postoje čvorovi *Domaćin, Gost, Rezultat* te statistika domaćina i gosta, odnosno njihovi korneri, broj udaraca, broj zaleđa i drugo. Svi ti statički čvorovi ovise o čvoru *Domaćin,* odnosno o čvoru *Gost,* ovisno za kojeg se od njih dvoje ti podaci odnose. A o svim tim statističkim podacima ovisi zapravo i čvor *Rezultat*, koji sadrži sve rezultate koji su se dogodili u sezoni 2018./2019.

Nakon prikaza Bayesove mreže slijedi primjer predviđanja zavisne varijable za sezonu 2018./2019.



Slika 38. Primjer predviđanja zavisne varijable za sezonu 2018./2019.

Za sezonu 2018./2019. kao primjer predviđanja je uzet gradski derbi iz Madrida, Atletico Madrid – Real Madrid. *AtleticoMadrid* je domaćin te mu je statistika kao u nastavku. *Ukupan_broj_udaraca* je srednji, oko 8 udaraca na utakmici, dok su *Udarci_u_okvir_gola* srednje mali, svega 4 udarca u okvir gola protivnika, što je zapravo i česta stvar kod Atletico Madrida jer uglavnom igraju obrambeno i napadaju samo preko kontra napada. *Korneri* su niski, tek 2 kornera na utakmici, dok su *Posjed_lopte* i *Zaleda* srednji s 45% postotka lopte u nogama i 3 zaleđa na utakmici. *Prekrsaji* su srednje mali, 7 prekršaja na utakmici što je zapravo neočekivano jer je poznato da Atletico igra vrlo grub nogomet s puno prekršaja i kartona, pogotovo u ovakvom derbiju.

S druge strane je Real Madrid. *RealMadrid* je na ovoj predikciji izabran kao gostujuća ekipa, sa sljedećom statistikom. *Ukupan_broj_udaraca* je vrlo visok, čak 22 udarca na gol protivnika, od čega *Udarci_u_okvir_gola* su srednji i iznose tek 6. To je s jedne strane vrlo dobro jer se vidi da se šanse stvaraju i pokušava se postići pogodak, međutim u isto vrijeme i zabrinjava činjenica da od 22 udarca samo njih 6 ide u okvir gola. *Korneri*, kao i kod domaćina niski, 2 kornera na utakmici. *Posjed_lopte*, *Zaleda_gosta* te *Prekrsaji_gosta* su srednji i to redom 45, 3, te 12. Ti podaci su za Real Madrid normalni, obzirom da posjedom lopte nikada ne grade svoju igru te su uvijek tu oko 50%. Prekršajima su isto uvijek u sredini i rade onoliko koliko je potrebno da spriječe protivničke napade.

Krajnji rezultat ovog predviđanja je 0-1, što znači pobjedu gosta, odnosno Real Madrida, koji je stvorio čak trostruko više šansi i udaraca, dok su u ostalim segmentima igre obje momčadi podjednake.

4.6. Sumiranje rezultata dobivenih kroz metode

Kroz ovo istraživanje provedene su sljedeće metode: Stablo odlučivanja, ekspertni sustavi te neuronske mreže. Svaka od njih je dala svoje rezultate, a oni će se sada sumirati. Što se tiče rezultata koji je najčešći kroz sezonu 2014./2015., postoje 3 najpouzdanija pravila, s točnošću od 64,57%, a ta točnost se odnosi na rezultate: 2-0, 1-1 te 1-0. Kod sezone 2018./2019. javlja se nešto drugačiji ishod, postoje 2 najpouzdanija pravila, a ona su za rezultate: 0-1 te 1-1. Vidljivo je da se rezultat 1-1 ponavlja kod obje sezone. Osim najpouzdanijih pravila, može se vidjeti i koji su rezultati najčešći bez obzira na pravila. Najčešći rezultati u sezoni 2014./2015. su bili sljedeći:

- 1-1, ukupno 43 utakmica završile tim rezultatom
- 1-0, ukupno 35 utakmica završilo tim rezultatom
- 0-1, ukupno 35 utakmica završilo tim rezultatom

- 2-0, ukupno 33 utakmice završile tim rezultatom
- 0-0, ukupno 30 utakmica završilo tim rezultatom

Za sezonu 2018./2019. ti rezultati su izgledali ovako:

- 1-1, ukupno 50 utakmica završilo ovim rezultatom
- 1-0, ukupno 41 utakmica završila ovim rezultatom
- 2-1, ukupno 38 utakmica završilo ovim rezultatom
- 2-0, ukupno 35 utakmica završilo ovim rezultatom
- 0-1, ukupno 30 utakmica završilo ovim rezultatom

Dakle, kod obje sezone je najčešći rezultat bio 1-1 ukoliko se gleda samo najčešći rezultat bez određenog pravila. Zanimljiva je činjenica da je 2018./2019. palo više golova te su utakmice bile atraktivnije, iako je to prva godina kada je jedna od najvećih zvijezda, Cristiano Ronaldo, koji je godišnje zabijao u prosjeku po 40-tak golova u sezoni, otišao iz LaLige.

Uz stablo odlučivanja, napravljeni su i ekspertni sustavi na temelju najpouzdanijih pravila iz stabla odlučivanja za obje sezone koje su bile uspoređivane. U tim sustavima je moguće vidjeti kako ima dosta sličnosti u najpouzdanijem pravilu između dviju sezona, iako se u nekim atributima sustavi razlikuju.

Metoda koja je korištena uz dvije već navedene su neuronske mreže. Njima se pokazalo koji atributi najviše utječu na promatrani atribut te kako se pomicanjem različitih vrijednosti dobivaju drugi rezultati kao predikcija. Tako je za sezonu 2014./2015. najčešći rezultat 0-1, odnosno pobjeda gostiju. To je veoma zanimljivo jer je španjolska liga poznata kao „domaćinska liga“ odnosno da u njoj uvijek veliku prednost imaju ekipe koje su domaćini. U sezoni 2018./2019. najčešći rezultat je bio, korištenjem ove metode, rezultat 1-1 koji se slaže i sa predikcijom provedenom stablom odlučivanja. Taj je rezultat nekako i logičniji od onoga kojeg je metoda pokazala za sezonu 2014./2015. jer su utakmice u španjolskoj ligi vrlo „tvrde“, domaćini igraju grubo te čuvaju rezultat, a većinom gosti imaju problema s probijanjem domaće obrane.

Na kraju je napravljeno predviđanje metodom Bayesove mreže gdje se može vidjeti predviđati krajnji rezultat pomicanjem raznih statističkih podataka. Također, napravljeni su i primjeri predviđanja za zavisnu varijablu za obje sezone za koje se uspoređuju podaci.

5. Zaključak

U radu je provedeno istraživanje na temu sportske prediktivne analitike. Prikazano je nekoliko prethodnih sličnih istraživanja koja su služila kao primjer za izradu ovog istraživanja. Podaci koji su korišteni su preuzeti s raznih stranica koje spremaju nogometne statistike i rezultate kroz povijest, no zapravo ne postoji neki gotovi „dataset“ sa svim potrebnim podacima pa je bilo potrebno podatke prvo urediti i prilagoditi.

Istraživanje je provedeno korištenjem stabla odlučivanja te neuronskih mreža u alatu BigML te je također načinjen i ekspertni sustav u alatu XpertRule Knowledge Builder. Uz njih je napravljeno i predviđanje Bayesovim mrežama u alatu Netica. Svaka metoda je drugačija, koristi drugačije načine predviđanja, međutim najčešći rezultati su sa svakom od njih bili približni. Kombinacijom ovih metoda može se doći do predviđanja rezultata u pojedinoj nogometnoj ligi, s vrlo visokim postotkom točnosti. Istina, stablo odlučivanja je dalo rezultat pouzdanosti od samo 64,57% u obje sezone za najpouzdanija pravila, međutim u analizi sportskih rezultata je vrlo teško dobiti bolje rezultate, ti skupovi podataka bi trebali imati jako puno podataka, profile igrača s njihovim trenutnim razmišljanjima, formom, s uvjetima na terenu i slično. A baš zbog toga što je to trenutno još uvijek nemoguće, ni ne postoji takav sustav koji bi mogao savršeno predvidjeti sportske rezultate. No, u tome i je ljepota sporta, ali i samog pokušaja istraživanja i predviđanja rezultata.

Naravno, moguće je ove podatke dopuniti sa još više statističkih podataka s pojedinih utakmica, a to bi bili podaci o kartonima, rezultatima po poluvremenima, broju kartona, zamjenama, prvim postavama i slično. Tada bi ova predikcija imala još više smisla te bi zasigurno dala još točnije i kvalitetnije rezultate.

Popis literature

AI, E. of. (2021). *Osnove neuronskih mreža*. https://course.elementsofai.com/hr/5/1

Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, *35*(2), 741–755. https://doi.org/10.1016/j.ijforecast.2018.01.003

Begičević, N. (2012). *Uvod u stabla odlučivanja. Nastavni materijal za kolegij "Poslovno odlučivanje."*

Carpita, M., Sandri, M., Simonetto, A., & Zuccolotto, P. (2016). Discovering the drivers of football match outcomes with data mining. *Quality Technology and Quantitative Management*, *12*(4), 561–577. https://doi.org/10.1080/16843703.2015.11673436

e-skole.hr. (2021). *Od živčane stanice do živčanoga sustava*. https://edutorij.e-skole.hr/share/proxy/alfresco-noauth/edutorij/api/proxy-guest/3b8a4b4e-84b0-4580-aa6f-e38efe028ed9/biologija-8/m03/j01/index.html

Ekonomski fakultet Zagreb. (2011). *Stablo odlučivanja Stablo odlučivanja Primjer :* 1–11.

Jurić, Z., Račić, N., & Radica, G. (2005). EKSPERTNI SUSTAV INTELIGENTNOGA DIZELSKOG MOTORA. *NAŠE MORE*, *52*, 81–87. https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=12790

Kliček, B. (2021a). *Inteligentni sustavi, nastavni materijali na predmetu Inteligentni sustavi [moodle]*.

Kliček, B. (2021b). *Inteligentni sustavi - nastavni materijali na kolegiju Inteligentni sustavi [Moodle]*. Sveučilište u Zagrebu, Fakultet organizacije i informatike, Varaždin.

Leung, C. K., & Joseph, K. W. (2014). Sports data mining: Predicting results for the college football games. *Procedia Computer Science*, *35*(C), 710–719. https://doi.org/10.1016/j.procs.2014.08.153

Oreški, D. (2021a). *Inteligentni sustavi*. Sveučilište u Zagrebu, Fakultet organizacije i informatike, Varaždin.

Oreški, D. (2021b). *RAZVOJ SUSTAVA TEMELJENIH NA ZNANJU SDT METODOLOGIJOM 1. Uvod*.

Oreški, D. (2021c). *Stablo odlučivanja*.

Owramipur, F., Eskandarian, P., & Mozneb, F. S. (2013). Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team. *International Journal of*

Computer Theory and Engineering, *5*(5), 812–815. https://doi.org/10.7763/ijcte.2013.v5.802

Peraić, I. (2012). *BAYESOVE MREŽE U MODELIRANJU UČENIKA*. https://mapmf.pmfst.unist.hr/~ani/radovi/diplomski/Peraic_Ivan_2012.pdf

Petersen, S. (2016). Definition - expert system. *Expert System*. https://searchenterpriseai.techtarget.com/definition/expert-system

Prcela, M. (2010). *PREDSTAVLJANJE ZNANJA ZASNOVANO NA INTEGRACIJI ONTOLOGIJA I BAYESOVIH MREŽA*. http://lis.irb.hr/MLAA/prcela-doktorska-disertacija.pdf

Saralegui, U. (2017). *Occupancy Estimation and People Flow Prediction in Smart Environments*.

Skladistenje.com. (2002). *Stabla odlučivanja*. http://www.skladistenje.com/stabla-odlucivanja/

Soni, D. (2018). Introduction to Bayesian Networks. *Towards Data Science*. https://towardsdatascience.com/introduction-to-bayesian-networks-81031eeed94e

Židov, I. (2018). *Uvod u neuronske mreže*. https://repozitorij.mathos.hr/islandora/object/mathos:256

Popis slika