Razvoj i usporedba prediktivnih modela nogometnih rezultata primjenom strojnog učenja

Živičnjak, Petar

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Petar Živičnjak

Razvoj i usporedba prediktivnih modela nogometnih rezultata primjenom strojnog učenja

DIPLOMSKI RAD

Varaždin, 2022.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Petar Živičnjak

Matični broj: 0016133762

Studij: Organizacija poslovnih sustava

Razvoj i usporedba prediktivnih modela nogometnih rezultata primjenom strojnog učenja

DIPLOMSKI RAD

Mentorica:

Izv.prof.dr.sc. Oreški Dijana

Varaždin, rujan 2022.

Petar Živičnjak

Izjava o izvornosti

Izjavljujem da je moj 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 potvrdio prihvaćanjem odredbi u sustavu FOI-radovi

Sažetak

U ovome radu biti će provedena prediktivna analiza rezultata engleske Premier lige. Da bi se takva analiza mogla provesti potreban je određeni skup podataka koji je za potrebe ovog rada preuzet s web stranice *Kaggle*. Taj skup podataka sadrži sve najvažnije informacije za odigrane utakmice. Podaci ukupno obuhvaćaju 29 sezona odnosno sezone od 1993/1994 pa do 2021/2022. Ti podaci sadrže informacije poput koja je ekipa bila domaća, a koja gostujuća, ime glavnog suca te statističke podatke kao što su broj kornera, slobodnih udaraca, broj postignutih golova kao i krajnji rezultat. U svakoj sezoni odigrano je ukupno 380 utakmica. U teorijskom djelu rada biti će napravljen sažetak prethodnih istraživanja i rezultata. Isto tako biti će detaljnije objašnjene metode koje će biti korištene u radu kao što su stablo odlučivanja, neuronske mreže i Bayesove mreže. Nadalje u radu biti će izrađeni navedeni modeli nad skupom podataka te će biti uspoređeni dobiveni rezultati provedenih metoda.

Ključne riječi: Stablo odlučivanja; Neuronske mreže; Bayesove mreže; Strojno učenje; Analiza nogometnih rezultata; Premier liga;

Sadržaj

1. Uvod

U današnje vrijeme precizno predviđanje rezultata nogometnih utakmica je u velikom predmetu interesa među ljudima, a pogotovo među mladima. Razlog tome je što nogomet kao sport gubi na svojoj važnosti odnosno na nogomet se sve više gleda kao na posao. Iz toga razloga predviđanje nogometnih rezultata postaje sve važnija stvar u svijetu biznisa odnosno kao izvor zarade. Predviđanje rezultata bilo kojeg sporta je kompleksan postupak jer ovisi o mnogo faktora na koje se teško može utjecati kao što su taktika, moral igrača, ozljede pojedinih igrača i sl. Zbog toga se modeli predviđanja rezultata izrađuju na temelju statističkih podataka iz prethodnih utakmica. Međutim takvi modeli mogu pružiti ograničena predviđanja jer se ne uzimaju u obzir svi faktori odnosno prethodno navedeni faktori (moral,taktika i sl.).

U ovom radu biti će proučena prethodna istraživanja koja se bave predviđanjem sportskih rezultata te će na temelju istih biti zabilježene korištene metode, podaci, izvori i tipovi podataka te zaključci i rezultati do kojih su pojedini autori došli. U nastavku rada biti će prikazan konkretan skup podataka koji obuhvaća podatke o odigranim utakmicama engleske Premier lige. Taj skup podatak preuzet je sa internetske stranice *Kaggle* i biti će dodatno uređen u svrhu izrade ovog rada. U ovome radu koristiti će se ukupno tri metode za izradu prediktivnih modela a to su: stablo odlučivanja, neuronske mreže i Bayesove mreže. Kako bi se bolje razumjele navedene metode biti će obrađena teorijska podloga za svaku pojedinu metodu. Na kraju će se dobiveni rezultati modela usporediti i objasniti te će se iznijeti zaključak.

2. Prethodna istraživanja

U ovom poglavlju biti će obrađeno pet sličnih i relevantnih istraživanja različitih autora vezanih uz ovu temu. Svako istraživanje se temeljilo na različitom skupu podataka, a sva istraživanja koriste slične metode za predviđanje rezultata. Sva ta istraživanja fokusirala su se na predviđanje rezultata između nekih dviju momčadi.

2.1. Predviđanje nogometnih rezultata korištenjem Bayesovih mreža

Nogomet je jedan od najpoznatijih sportova na svijetu. Predviđanje rezultata nogometnih utakmica je privuklo jako puno ljudi koji imaju puno strasti prema nogometu od menadžera nogometnih klubova pa do navijača. Predviđanje nogometnih rezultata postao je intrigirajući istraživački problem zbog njegove težine odnosno na ishod utakmice može utjecati puno čimbenika kao što su timski rad, vještine vrijeme i sl. Vrlo je teško predvidjeti konačne rezultate jer se sve može dogoditi unutar 90 minuta. Jedan od potencijalnih faktora koji može odlučiti utakmicu je i sreća. Zbog svih tih različitih čimbenika koji utječu na nogometnu utakmicu ovo istraživanje koristi Bayesove mreže jer je prethodno dokazano da su primjenjive u predviđanju vremena i sporta kao i mnogih drugih predikcija (Razali et al., 2017).

U ovom istraživanju predviđanja su se provodila pomoći softvera WEKA. To je softver otvorenog koda koji pruža najsuvremeniju implementaciju algoritma za rudarenje podataka i strojno učenje. Svaka utakmica je promatrana zasebno jer svaka utakmica ima različite vrijednosti faktora koji variraju ovisno o utakmici. Eksperimenti predviđanja se ponavljaju za svih 380 utakmica u svakoj od sezona te se onda rezultati uspoređuju između sezona (Razali et al., 2017).

Ovo istraživanje koristilo je podatke za sezone 2010-2011, 2011-2012 i 2012-2013. Natjecanje se ukupno sastoji od 20 momčadi, a svaka momčad igra međusobno dva puta u sezoni. Na slici 1 mogu se vidjeti atributi odnosno glavni čimbenici koji su se koristili u ovom istraživanju (Razali et al., 2017).

Attributes	Sample Values
Home Team	Manchester United
Away Team	Wigan
Home Team Shots	17
Away Team Shots	8
Home Team Shots on Target	10
Away Team Shots on Target	4
Home Team Corners	5
Away Team Corners	5
Home Team Fouls Committed	10
Away Team Fouls Committed	14
Home Team Yellow Cards	2
Away Team Yellow Cards	2
Home Team Red Cards	0
Away Team Red Cards	0
Half Time Home Team Goals	0
Half Time Away Team Goals	0
Full Time Home Team Goals	4
Full Time Away Team Goals	0

Slika 1: Popis korištenih atributa (Razali et al., 2017)

U ovom istraživanju podaci o 380 utakmica svake sezone podijeljeni su u 10 skupova jednake veličine. Nakon toga svaki skup podijeljen je na 90% podataka koji su se koristili za učenje modela dok se ostalih 10% koristilo za predviđanje rezultata. Na slici 2 se mogu vidjeti postoci točnosti predviđanja po sezonama (Razali et al., 2017).

Overall Prediction Results across Three Seasons in EPL

Season	Prediction Accuracy (%)
2010-2011	75.26
2011-2012	79.47
2012-2013	70.53

Slika 2: Točnost predviđanja modela po sezonama (Razali et al., 2017)

Prema točnostima koje se mogu vidjeti na slici 2 autori su došli do modela koji u prosjeku predviđa sa 75.09% točnosti.

Kvaliteta ovog sustava je poprilično visoka iz razloga što su se vršile predikcije rezultata za tri uzastopne godine nogometnog natjecanja *Premier Lige* te na kraju sezone svaki od tih godina nije imao manji postotak od 70% pogođenih rezultata što je zapravo dosta impresivno jer autori navoda da su prethodna istraživanja imala točnost od 59.21%. Srednja vrijednost modela je 75.09%. Autori ovog istraživanja navode kako se uporabom *Bayesovih mreža* znatno povećao postotak predvidljivosti rezultata, točnije za 16% u odnosu na ranija istraživanja. U zaključku istraživanja navodi se kako bi se dobiveni rezultati mogli koristiti za buduća istraživanja kod predviđanja rezultata (Razali et al., 2017).

2.2. Predviđanje nogometnih rezultata u Engleskoj ligi

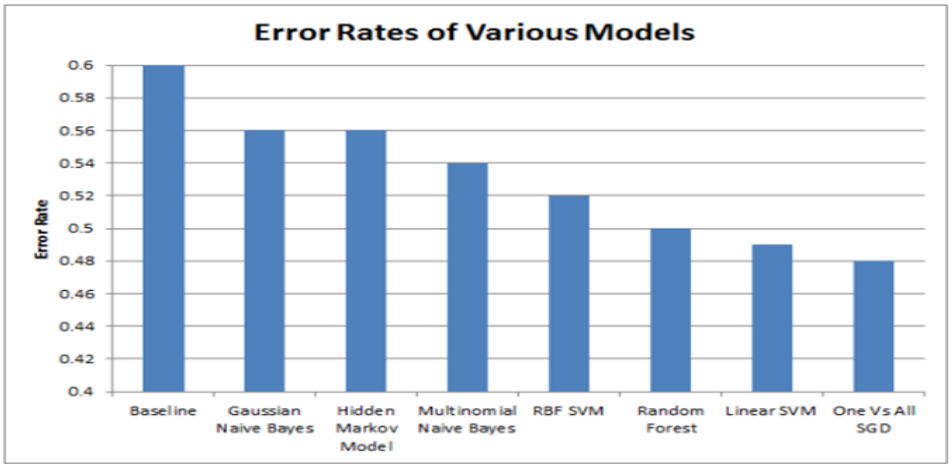
Tijekom svjetskog nogometnog prvenstva 2010 godine bilo je mnogo zanimljivih trenutaka. Autori rada motivaciju su pronašli u hobotnici Paul-u koja je tijekom trajanja prvenstva točno predvidjela svih osam utakmica. Ovakvi modeli imaju primjenu u stvarnom svijetu u obliku kockanja, poboljšanja treniranja nogometnih momčadi kao i za novinarstvo (Ulmer i Fernandez, 2014).

Ovdje su se koristila dva skupa podataka odnosno skup podataka za trening koji sadrži podatke od 10 sezona (od sezone 2002/2003 do sezone 2011/2012) i skupa podataka na kojem se vrši predviđanje i koji sadrži podatke za 2 sezone (sezone 2012/2013 i 2013/2014). Budući da svaka momčad igra sa svim ostalim momčadima po dva puta u sezoni ukupno imaju 3800 utakmica u skupu podataka za treniranje i 760 utakmica u skupu podataka za predviđanje (Ulmer i Fernandez, 2014).

Svaka utakmica sadrži sljedeće podatke odnosno atribute (Ulmer i Fernandez, 2014):

- Home team
- Away team
- Score
- Winner
- Goals of home team
- Goals of away team

U ovom istraživanju korišteni su razni modeli za predviđanje rezultata. Na slici 3 mogu se vidjeti ti modeli te stopa pogreške za te modele (Ulmer i Fernandez, 2014).



Slika 3: Modeli i stopa pogreške (Ulmer i Fernandez, 2014)

Autori su došli do zaključka da je najbolji model *One Vs All SGD* jer taj model ima najmanju stopu pogreške. Ali s druge strane isto tako su naveli da ovaj model ne predviđa dobro neriješene rezultate te da stopa pogreške dosta ovisi o veličini skupa za učenje odnosno kako se skup smanjuje tako se greška povećava (Ulmer i Fernandez, 2014).

Ovo istraživanje daje odgovor na pitanje koji od modela ima najveću uspješnost prilikom predviđanja. Provedeno je istraživanje na 8 modela od kojih *One Vs All SGD* daje najbolje rezultate. *One vs All* klasifikacija je metoda koja uključuje treniranje različitih binarnih klasifikatora od kojih je svaki dizajniran za prepoznavanje određene klase. Nakon toga se ti klasifikatori zajedno koriste za višeklasnu klasifikaciju. Navodi se da je najveći problem bio ogromna količina podataka na temelju kojih se došlo do rezultata. Ali ujedno i ta ogromna količina podataka dala je preciznije rezultate. Kako bi sustav dao još preciznije rezultate savjetuje se korištenje još većeg broja ulaznih podataka (Ulmer i Fernandez, 2014).

2.3. Predviđanje sportskih rezultata korištenjem strojnog učenja

Ovaj rad se bavi predviđanjem rezultata kriketa odnosno IPL lige. U svakoj IPL sezoni postoji ukupno 8 momčadi koje igraju međusobno u prvoj fazi. Nakon prve faze, 4 pobjedničke momčadi napreduju u drugu fazu gdje opet igraju međusobno. Dvije najbolje momčadi napreduju u finale gdje se onda odlučuje o pobjedniku. Rezultat svake utakmice ovisi o različitim uvjetima kao što su mjesto, izvedba igrača, izbačaj i sl. Autori u ovom radu predviđaju rezultate koristeći tri algoritma strojnog učenja odnosno koriste *Support Vector Machine (SVM), NaiveBayes i Ctree* (Agrawal et al., 2018).

Skup podataka koji je korišten u ovom istraživanju prikupljen je sa internetskih izvora i sadrži podatke o ukupno 500 odigranih utakmica. Svaki red u skupu podataka sadrži 21 atribut koji se nazivaju podaci o isporuci. Ti atributi se mogu vidjeti na slici 4. Isto tako skup podataka sadrži i dodatnih 14 atributa koji govore o statistici utakmice i nazivaju se rezultati utakmice. Ovi atributi se mogu vidjeti na slici 5 (Agrawal et al., 2018).

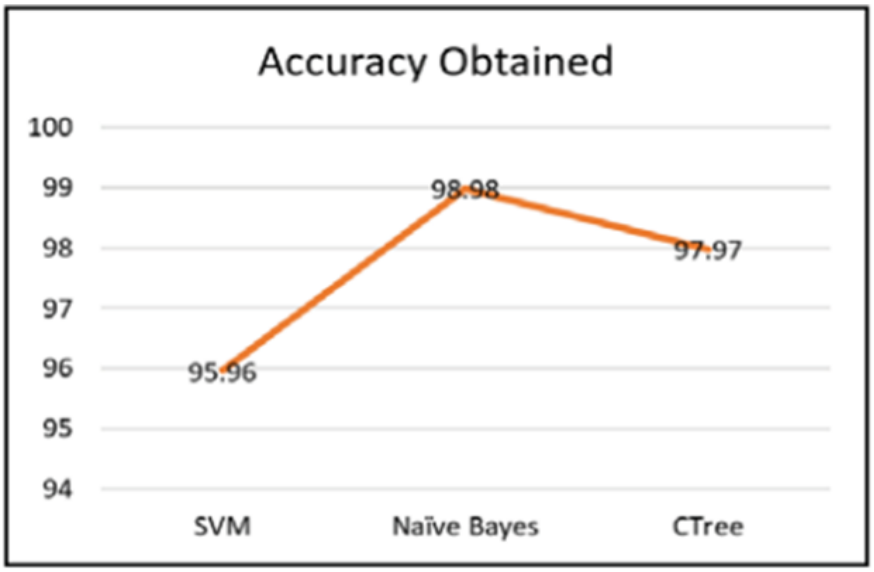
Match Feature		
Match ID	Inning	Batting Team
Bowling Team	Over	Ball
Batsman	Non Striker	Bowler
Is Super Over	Wide Runs	Bye Runs
Legbye Runs	Noball Runs	Penalty Runs
Batsman Runs	Extra Runs	Total Runs
Player Dismissed	Dismissal Kind	Fielder

Slika 4: Atributi naziva Podaci o isporuci (Agrawal et al., 2018)

Match Feature		
Match ID	Season	City
Team1	Team2	Toss Winner
Toss Decision	Result	Winner
Win By Runs	Win By Wickets	Player of the Match
Venue		

Slika 5: Podaci naziva Rezultati utakmice (Agrawal et al., 2018)

Točnost prethodno navedenih metoda koje su korištene u ovom istraživanju mogu se vidjeti na slici 4. Iz slike se vidi da metoda SVM daje točnost od 95.96%, metoda naivni Bayes točnost od 98.98% a metoda Ctree točnost od 97.97% (Agrawal et al., 2018).



Slika 6: Usporedba točnosti metoda (Agrawal et al., 2018)

Autori su došli do zaključka da Naivni Bayes daje najbolje rezultate predviđanja odnosno ima najveću točnost (98.98%) između testiranih metoda (Agrawal et al., 2018).

U ovom istraživanju svaki od tri korištena modela daje postotak predviđanja veći od 95%. Taj rezultat je jako impresivan. Model je implementiran na način tako da se može proširiti sa različitim okruženjima što bi trebalo davati skoro pa isti postotak rezultata ako se redovito unesu svi traženi parametri. Na taj način se mogu identificirati novi vektori parametra te napraviti preciznije predviđanje (Agrawal et al., 2018).

2.4. Predviđanje nogometnih rezultata korištenjem metoda strojnog učenja

Sportski događaji oduvijek su bili vrlo zanimljivi širokom krugu ljudi. Jedan od najpopularnijih sportova je nogomet, a Liga prvaka je najteže i najprestižnije klupsko nogometno natjecanje na svijetu. Zbog svoje popularnosti i malog broja mogućih ishoda utakmica, predviđanje rezultata je vrlo zanimljiv i naizgled jednostavan izazov. Ali to ipak nije tako odnosno vrlo je teško predvidjeti konačan ishod utakmice jer on ovisi o jako puno čimbenika. Liga prvaka se sastoji od dvije faze. U prvoj fazi natjecanja ekipe su podijeljene u skupine, a svaka skupina ima 4 momčadi od kojih svaka igra po dvije utakmice sa ostalim ekipama u grupi. Dvije najbolje ekipe iz grupa prelaze u drugu fazu gdje se natjecanje nastavlja po nokaut sustavu turnira (Hucaljuk i Rakipović, 2011).

Postoji puno faktora koji utječu na konačan rezultat utakmice, a postoje dva načina kako odabrati potrebna obilježja odnosno atribute. Prvi način implicira da osoba posjeduje određeno znanje o problemu i da se na temelju toga znanja odaberu oni atributi koji će imati najviše utjecaja na finalan rezultat. Drugi način implicira da osoba ne posjeduje znanje o problemu pa se iz tog razloga odabiru svi atributi i onda se na temelju testiranja pokušavaju izdvojiti atributi koji imaju najveći utjecaj. Autori ovog istraživanja odabrali su prvi način. Ovdje su odabrane sljedeće skupine atributa (Hucaljuk i Rakipović, 2011):

- Trenutna forma momčadi na temelju rezultata postignutih u posljednjih šest utakmica
- Ishod prethodnog susreta ekipa koje igraju utakmicu
- Trenutna pozicija na ljestvici
- Broj ozlijeđenih igrača iz prve postave
- Prosječan broj postignutih i primljenih golova po utakmici

Nadalje autori navode da kako bi se dobio optimalan spoj prethodno navedenih skupina atributa da je potrebno provesti testiranja. Došli su do zaključka da se optimalan rezultat dobije ako se prethodne skupine atributa dodatno rastave. Kao primjer možemo uzeti prvu navedenu kategoriju odnosno trenutnu formu momčadi koja se dalje rastavlja na broj pobjeda, poraza i neriješenih utakmica. Finalan set podataka sastoji se od ukupno 20 atributa te su podaci koje sadrži prikupljeni sa različitih internetskih izvora. Ukupan broj zapisa u setu podataka iznosi 96 iz tog razloga jer se ukupno odigra 96 utakmica u grupnoj fazi lige prvaka. Isto tako korišteni su dva seta podataka: basic (bez subjektivne procjena kvalitete tima odnosno sastavljen iz različitih internetskih izvora) i expert (skup podataka sastavljen od strane stručnjaka) (Hucaljuk i Rakipović, 2011).

U ovom istraživanju autori su podijelili skup podataka na sljedeća 3 načina (Hucaljuk i Rakipović, 2011):

- Skup za treniranje modela sadrži utakmice od prve 3 runde, skup za testiranje sadrži preostale 3 runde
- Skup za treniranje modela sadrži utakmice od prve 4 runde, skup za testiranje sadrži preostale 2 runde
- Skup za treniranje modela sadrži utakmice od prvih 5 runde, skup za testiranje sadrži preostalu rundu

U istraživanju testirano je ukupno čak 6 metoda za predviđanje rezultata. Metode te preciznost metoda grupirane po skupu podataka i podjeli podataka mogu se vidjeti na slici 7 (Hucaljuk i Rakipović, 2011).

Feature set	Training/test set size		Naive Bayes		Bayessian net		LogitBoost		k-NN		Random forest		ANN		Most common	
	Trainng	Test	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
Expert	3	3	0.5	47.9%	0.546	56.3%	0.5	50.0%	0.535	54.2%	0.511	52.1%	0.541	54.2%	0.22	50.0%
	4	2	0.542	53.1%	0.546	53.1%	0.594	59.4%	0.526	53.1%	0.551	53.1%	0.692	68.8%	0.24	56.3%
	5	1	0.554	56.3%	0.552	56.3%	0.675	68.8%	0.548	56.3%	0.468	50.0%	0.578	56.3%	0.22	50.0%
Basic	3	3	0.543	52.1%	0.567	56.3%	0.46	47.9%	0.569	56.3%	0.534	56.3%	0.587	58.3%	0.22	50.0%
	4	2	0.542	53.1%	0.534	56.3%	0.641	62.5%	0.638	62.5%	0.644	65.6%	0.692	68.8%	0.24	56.3%
	5	1	0.554	56.3%	0.517	56.3%	0.547	56.3%	0.548	56.3%	0.468	50.0%	0.578	56.3%	0.22	50.0%

Slika 7: Usporedba točnosti metoda (Hucaljuk i Rakipović, 2011)

Autori su došli do zaključka da naivni Bayes-ovi klasifikatori imaju najlošiju preciznost predviđanja dok umjetna neuronska mreža ima najbolje rezultate predviđanja. Neuronska mreža korištena u ovom istraživanju sastoji se od 5 skrivenih neurona. Ovakav rezultat je iznenađujući jer se skup podataka sastoji od jako malog broja zavisnih atributa pa bi takav skup trebao biti idealan za metodu naivnog Bayesa (Hucaljuk i Rakipović, 2011).

U zaključku se navodi da su rezultati ovog istraživanja slični ili čak malo bolji nego u prethodnim istraživanjima. Od korištenih metoda najbolja se pokazala umjetna neuronska mreža koja ima postotak uspješnosti od 68%. Svoj postotak, istraživanje opravdava na temelju velikog niza faktora koji se moraju uzeti u obzir i modelirati sukladno. Jako je teško napraviti model koji će pokriti sve slučajeve. Uspješnost bi se mogla poboljšati tako da se napravi model forme za svakog igrača (Hucaljuk i Rakipović, 2011).

2.5. Predviđanje sportskih rezultata korištenjem Bayesovih mreža i sličnih metoda

Bayesove mreže zapravo pružaju način za prikaz i korištenje znanja koje je često dobiveno od stručnjaka u pojedinom području. To znanje se može temeljiti na subjektivnim prosudbama kao i na podacima. Predviđanje nogometnih ishoda idealno je za primjenu Bayesove mreže jer ima mnogo složenih čimbenika. Ovdje se izvedba Bayesovih mreža promatra na utakmicama kluba Tottenham Hotspur i to za dvije sezone (1995/1996 i 1996/1997) (Joseph et al., 2006).

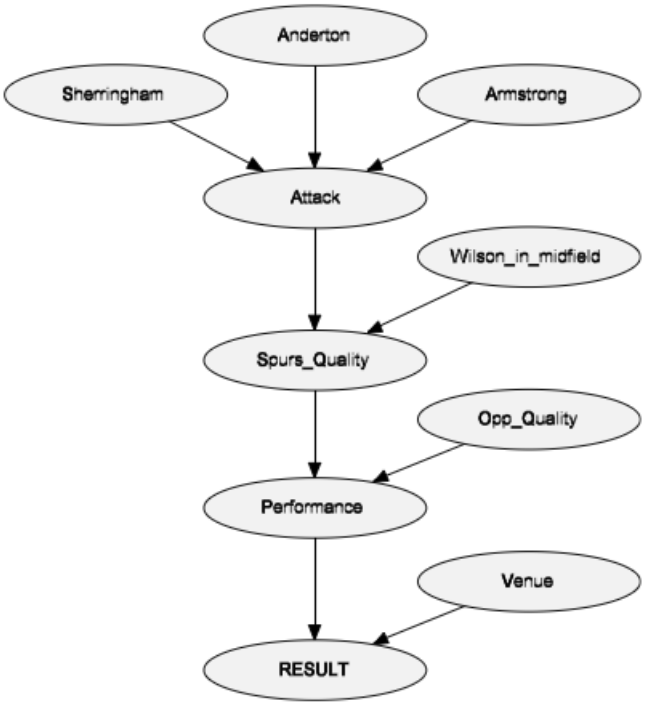
Prethodno su već spomenute metode kod prikupljanja podataka odnosno odabira obilježja. U ovom istraživanju koristi se drugi pristup odnosno pristup koji pretpostavlja da osoba nema znanja o problemu. Za konstruiranje ekspertne Bayesove mreže koristilo se nekoliko sljedećih značajki (Joseph et al., 2006):

- Prisutnost ili odsutnost tri igrača: Sherringhama, Andertona i Armstroga. Ova vrijednost je true ili false
- Wilsonova igračka pozicija
- Kvaliteta protivničke momčadi (visoka, srednja ili niska)
- Mjesto odigravanja (domaćin ili gost)

Kako bi se pojednostavila struktura Byesove mreže koristila su se tri dodatna čvora (Joseph et al., 2006):

- Napad - predstavlja kvalitetu napada ekipe Tottenham (niska, srednja ili visoka)
- Kvaliteta_Spursa - sveukupna kvaliteta momčadi Tottenhem (niska, srednja ili visoka)
- Performance - kakve će performanse imati momčad Tottenham uzimajući u obzir vlastitu kvalitetu i kvalitetu protivnika

Ukupan skup podataka koji se koristi ima samo 76 zapisa jer se gledaju samo utakmice momčadi Tottenhem za dvije sezone. Na slici 8 može se vidjeti izgled konstruiranog ekspertnog stabla koji prikazuje sve prethodno navedene kategorije i čvorove (Joseph et al., 2006).



Slika 8: Dijagram ekspertne Bayesove mreže (Joseph et al., 2006)

U ovom istraživanju se još i koristi model k-najbližih susjeda što je jedan od najosnovnijih, ali najvažnijih klasifikacijskih algoritama u strojnom učenju. Ovaj model pripada domeni nadziranog učenja i nalazi intenzivnu primjenu u prepoznavanju uzoraka, rudarenju podataka i otkrivanju upada. Široko je dostupan i primijenjen u stvarnim scenarijima jer je neparametarski model, što znači da ne daje nikakve temeljne pretpostavke o distribuciji podataka (Joseph et al., 2006).

Ovdje se zapravo uspoređuju ekspertne Bayesove mreže (konstruirane pomoću stručnjaka) sa još četiri modela a to su (Joseph et al., 2006):

- Naivni Bayes
- Bayesove mreže učene na temelju podataka
- K-najbližih susjeda
- Stablo odlučivanja

Rezultati istraživanja odnosno točnost pojedinih modela se mogu vidjeti na slici 8.

Train period–Test period	Number of correct predictions by learner					
	Most common	MC4	Naive BN	Hugin BN	Expert BN	KNN
95/96–95/96 season	16 (42.11%)	28 (73.68%)	26 (68.42%)	21 (55.26%)	20 (52.63%)	37 (97.37%)
96/97–96/97 season	18 (47.37%)	30 (78.95%)	31 (81.58%)	26 (68.42%)	25 (65.79%)	37 (97.37%)
Average for full seasons	17 (44.74%)	29 (76.32%)	28.5 (75.00%)	23.5 (61.84%)	22.5 (59.21%)	37 (97.37%)
Period 1–period 234 95/96	12 (42.86%)	8 (28.57%)	9 (32.14%)	8 (28.57%)	14 (50.00%)	12 (42.86%)
Period 12–period 34 95/96	7 (38.89%)	6 (33.33%)	6 (33.33%)	3 (16.67%)	10 (55.56%)	7 (38.89%)
Period 123–period 4 95/96	2 (25.00%)	2 (25.00%)	2 (25.00%)	2 (25.00%)	3 (37.50%)	2 (25.00%)
Sum for 1995/1996 periods	21 (38.89%)	16 (29.63%)	17 (31.48%)	13 (24.07%)	27 (50.00%)	21 (38.89%)
Period 1–period 234 96/97	11.5 (41.07%)	10 (35.71%)	13 (46.43%)	11 (39.29%)	19 (67.86%)	11 (39.29%)
Period 12–period 34 96/97	7.5 (41.67%)	7 (38.89%)	10 (55.56%)	3 (16.67%)	10 (55.56%)	5 (27.78%)
Period 123–period 4 96/97	5 (62.50%)	2 (25.00%)	5 (62.50%)	2 (25.00%)	3 (37.50%)	1 (12.50%)
Sum for 96/97 periods	24 (44.44%)	19 (35.19%)	28 (51.85%)	16 (29.63%)	32 (59.26%)	17 (31.48%)
Period 23 95/96–period 4/1 95/97	6 (33.33%)	4 (22.22%)	6 (33.33%)	Unavailable	9 (50.00%)	7 (38.89%)
Period 234 95/96–period 1 96/97	4 (40.00%)	2 (20.00%)	4 (40.00%)	3 (30.00%)	6 (60.00%)	3 (30.00%)
Period 34 95/96–period 12 96/97	8 (40.00%)	6 (30.00%)	8 (40.00%)	11 (55.00%)	15 (75.00%)	7 (35.00%)
Period 4 95/96–period 123 96/97	6 (20.00%)	8 (26.67%)	6 (20.00%)	10 (33.33%)	22 (73.33%)	8 (26.67%)
Period 4/1 95/97–period 23 96/7	6.67 (33.33%)	7 (35.00%)	8 (40.00%)	7 (35.00%)	16 (80.00%)	7 (35.00%)
Season 95/96–season 96/97	13 (34.21%)	8 (21.05%)	13 (34.21%)	20 (52.63%)	25 (65.79%)	15 (39.47%)
Sum for cross season periods	43.67 (32.11%)	35 (25.74%)	45 (33.09%)	51 (43.22%)	93 (68.38%)	47 (34.56%)
Overall average percentage	40.05%	41.72%	47.86%	39.69%	59.21%	50.58%
Overall disjoint training/data	38.48%	30.19%	38.81%	32.31%	59.21%	34.98%

Slika 8: Usporedba točnosti predviđanja metoda (Joseph et al., 2006)

Autori su došli do zaključka da kada se koristi isti skup podataka za učenje modela i za testiranje modela da je onda model KNN odnosno k-najbližih susjeda daje daleko najveću točnost predviđanja koja iznosi čak 97.3%. Međutim relevantniji je drugi slučaj odnosno slučaj u kojem se originalni skup podataka dijeli na skup podataka za učenje i skup podataka za predviđanje. Kada se koristite dva skupa podataka tada su autori zaključili da je najbolji model ekspertna Bayes-ova mreža koja daje točnost od 59.21%. Za takav pristup možemo reći da je odličan pristup ako želimo poboljšati svoje rezultate pomoću strojnog učenja. Pomoću takvog modela možemo vidjeti koji točno atributi najviše utječu na predviđanje te ih na temelju toga varirati kako bismo dobili što bolje rezultate (Joseph et al., 2006).

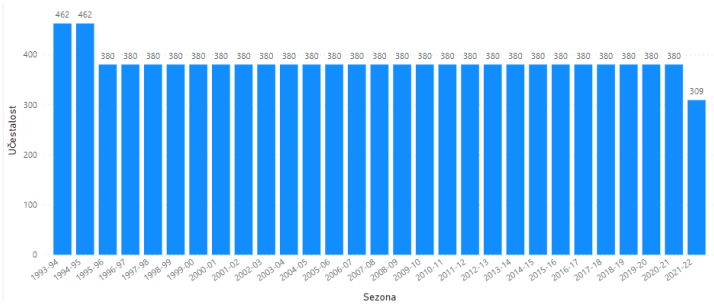
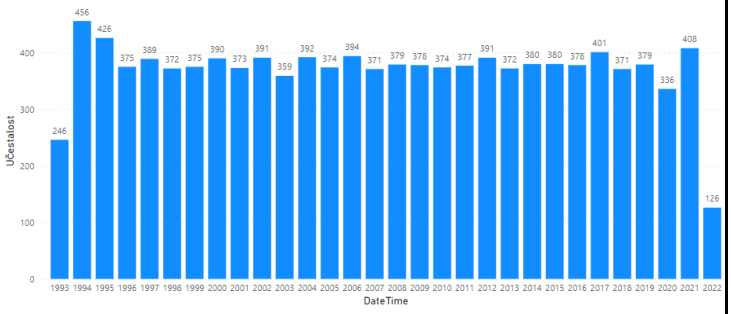
3. Opis i priprema podataka

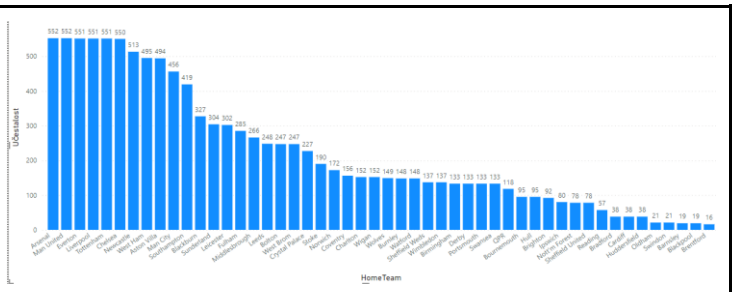
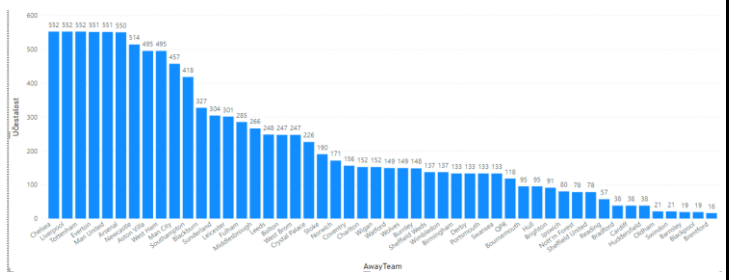
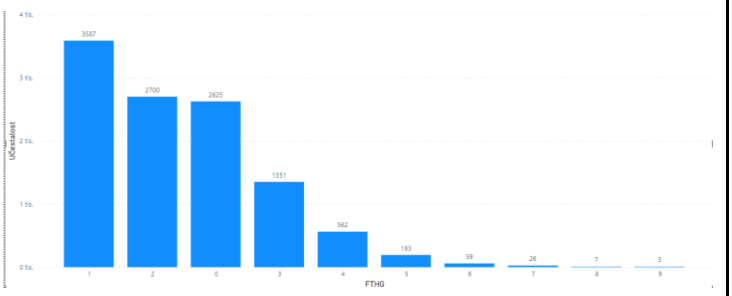
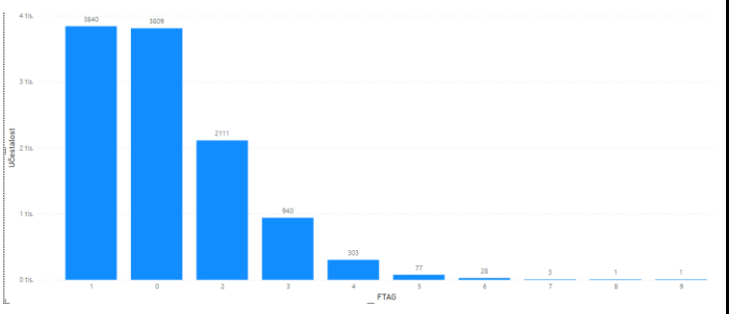
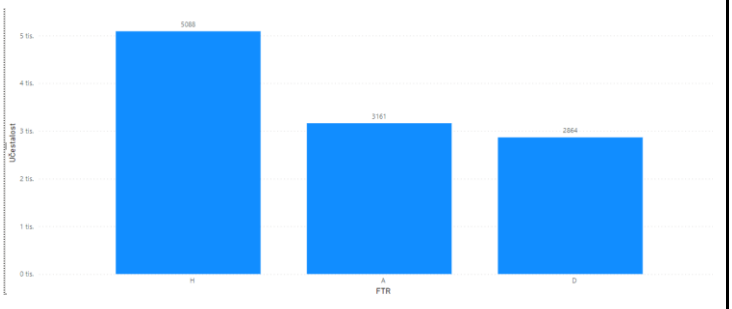
Kako bi se mogli izraditi bilo kakvi modeli potreban je određeni skup podataka. U ovom poglavlju biti će prikazan originalni skup podataka koji je korišten, njegovo uređivanje te finalno uređeni skup podataka.

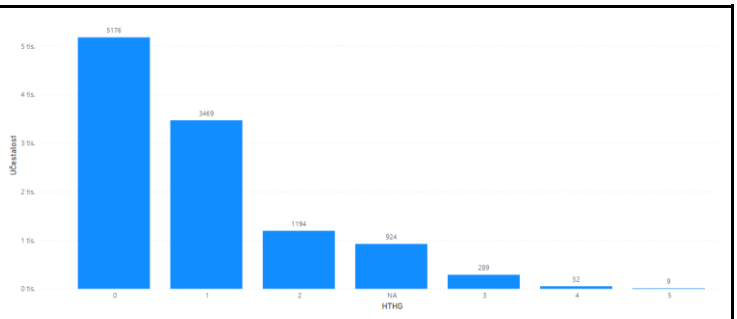
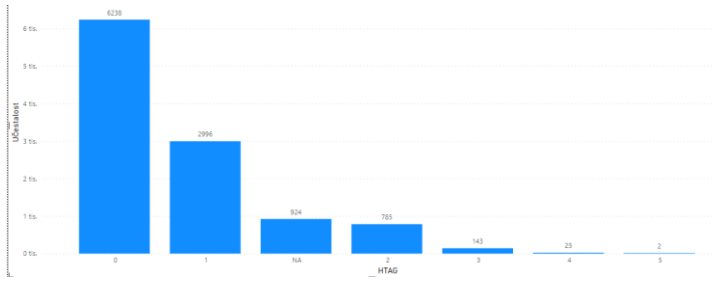
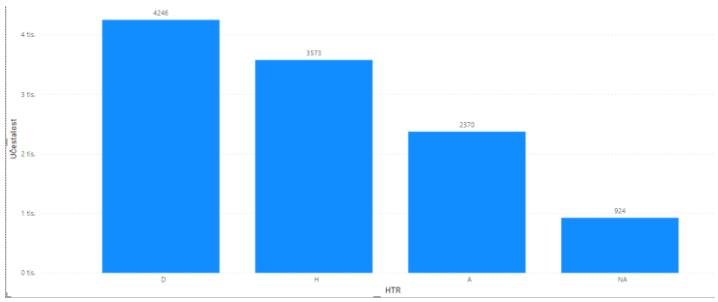
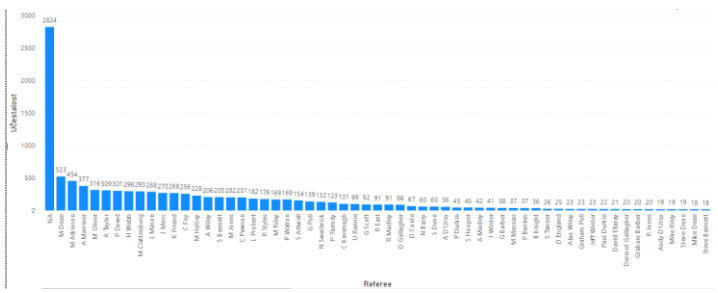
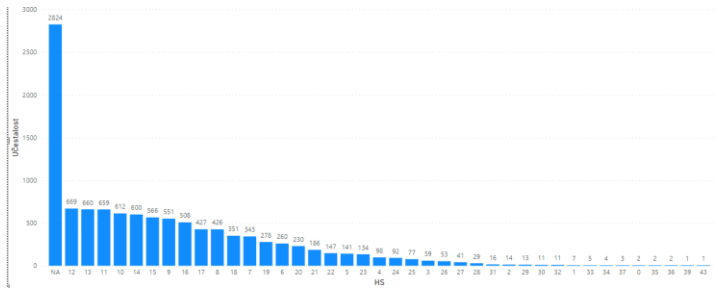
3.1. Opis podataka

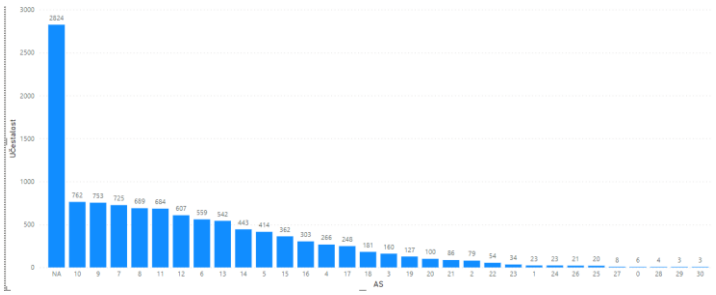
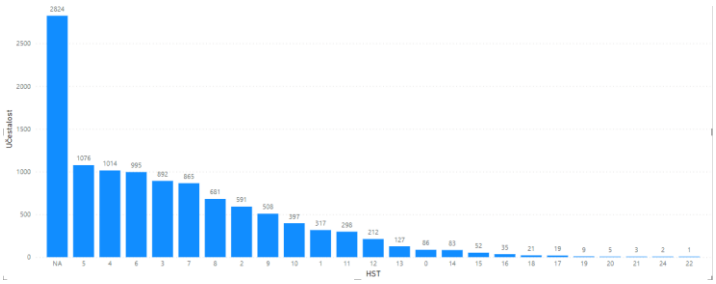
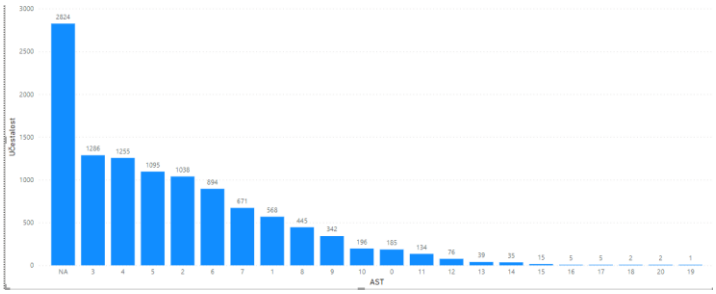
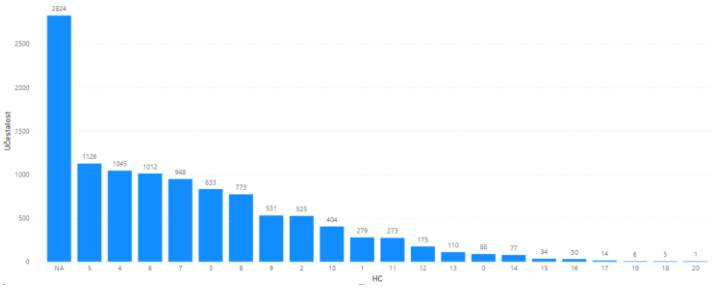
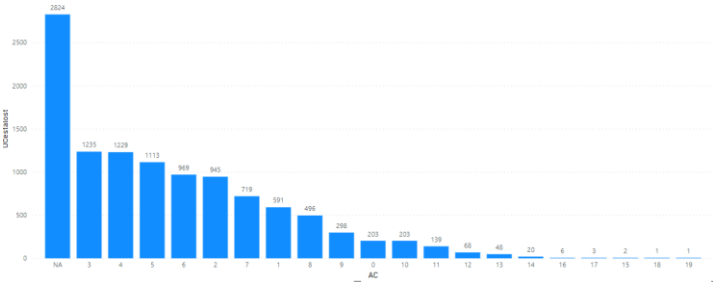
U ovom projektu koristi se gotov skup podataka koji je preuzet s internetske stranice *Kaggle*. Ti podaci opisuju odigrane utakmice engleske Premier lige te sam skup podataka ima naziv *„English Premier League (EPL) Results“*. Uz same podatke, autor podataka dao je i opis pojedinih atributa. U tablici 1 mogu se vidjeti svi atributi koji su sadržani u skupu podataka kao i njihovo pojašnjenje te distribucija vrijednosti.

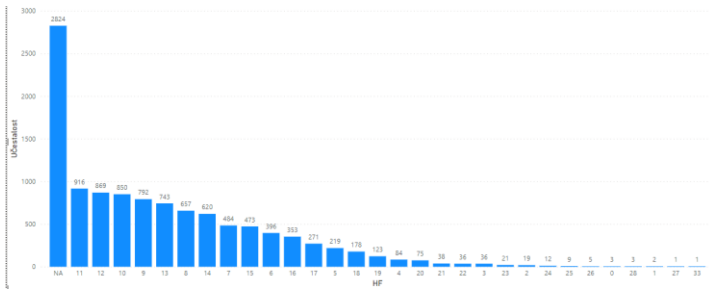
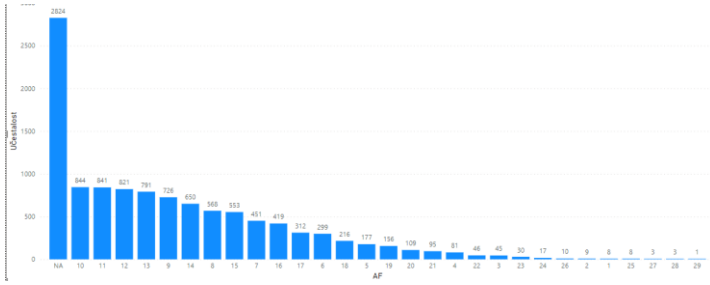
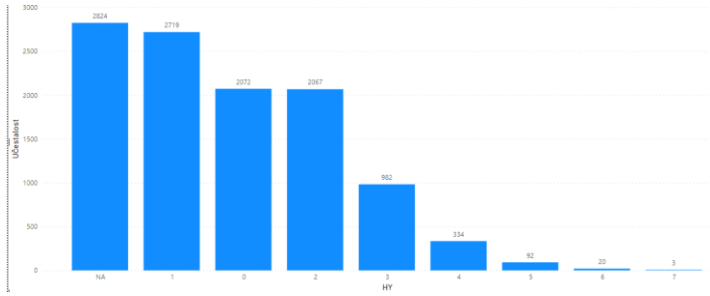
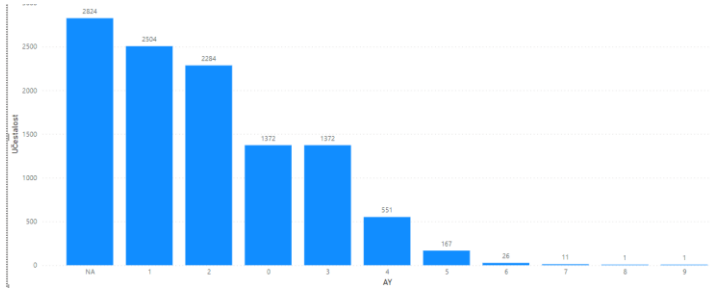
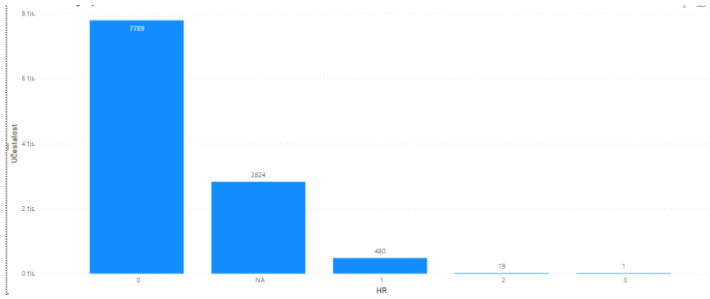
Tablica 1: Prikaz varijabli

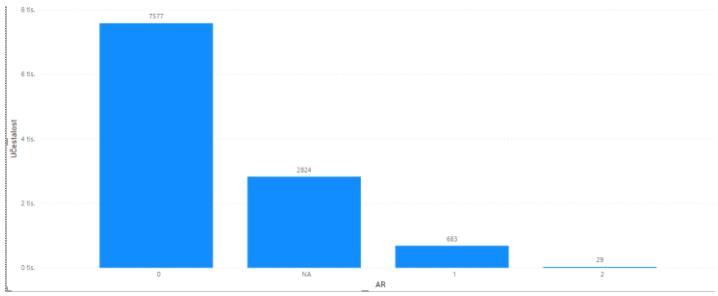
Atribut	Opis atributa	Distribucija vrijednosti
Season	Sezona u kojoj je odigrana utakmica	
DateTime	Datum i vrijeme odigravanja utakmice (yyyy-mm-dd hh:mm:ss)	

HomeTeam	Domaći tim	
AwayTeam	Gostujući tim	
FTHG	Ukupan postignuti broj golova na utakmici domaćeg tima	
FTAG	Ukupan postignuti broj golova na utakmici gostujućeg tima	
FTR	Rezultat utakmice (H=pobjeda domaćina, D=neriješeno, A=pobjeda gostiju)	

HTHG	Broj pogodaka u prvom poluvremenu domaćeg tima	
HTAG	Broj pogodaka u prvom poluvremenu gostujućeg tima	
HTR	Rezultat prvog poluvremena (H=pobjeda domaćina, D=neriješeno, A=pobjeda gostiju)	
Referee	Sudac utakmice	
HS	Broj udaraca domaćeg tima	

AS	Broj udaraca gostujućeg tima	
HST	Broj udaraca u okvir domaćeg tima	
AST	Broj udaraca u okvir gostujućeg tima	
HC	Broj kornera domaćeg tima	
AC	Broj kornera gostujućeg tima	

HF	Broj prekršaja domaćeg tima	
AF	Broj prekršaja gostujućeg tima	
HY	Broj žutih kartona domaćeg tima	
AY	Broj žutih kartona gostujućeg tima	
HR	Broj crvenih kartona domaćeg tima	

AR	Broj crvenih kartona gostujućeg tima	

3.2. Priprema podataka

Ova faza uključuje procese čišćenja podataka te njihovu obradu za daljnje korištenje. Ova faza se zapravo provodi kako bi se dobili što precizniji rezultati. Za pripremu podataka odabrani su svi atributi u skupu podataka. Istraživanje podataka se temelji na podacima koji su dostupni na stranici izvora podataka. Sama kvaliteta podataka se čini u redu.

Na slici 9 može se vidjeti originalni skup podataka u .csv formatu otvoren pomoću programa Microsoft excel.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Season,DateTime,HomeTeam,AwayTeam,FTHG,FTAG,FTR,HTHG,HTAG,HTR,Referee,HS,AS,HST,AST,HC,AC,HF,AF,HY,AY,HR,AR											
2	1993-94,1993-08-14T00:00:00Z,Arsenal,Coventry,0,3,A,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
3	1993-94,1993-08-14T00:00:00Z,Aston Villa,QPR,4,1,H,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
4	1993-94,1993-08-14T00:00:00Z,Chelsea,Blackburn,1,2,A,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
5	1993-94,1993-08-14T00:00:00Z,Liverpool,Sheffield Weds,2,0,H,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
6	1993-94,1993-08-14T00:00:00Z,Man City,Leeds,1,1,D,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
7	1993-94,1993-08-14T00:00:00Z,Newcastle,Tottenham,0,1,A,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
8	1993-94,1993-08-14T00:00:00Z,Oldham,Ipswich,0,3,A,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
9	1993-94,1993-08-14T00:00:00Z,Sheffield United,Swindon,3,1,H,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
10	1993-94,1993-08-14T00:00:00Z,Southampton,Everton,0,2,A,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
11	1993-94,1993-08-14T00:00:00Z,West Ham,Wimbledon,0,2,A,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
12	1993-94,1993-08-15T00:00:00Z,Norwich,Man United,0,2,A,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
13	1993-94,1993-08-16T00:00:00Z,Tottenham,Arsenal,0,1,A,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
14	1993-94,1993-08-17T00:00:00Z,Everton,Man City,1,0,H,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
15	1993-94,1993-08-17T00:00:00Z,Ipswich,Southampton,1,0,H,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
16	1993-94,1993-08-17T00:00:00Z,Leeds,West Ham,1,0,H,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
17	1993-94,1993-08-17T00:00:00Z,Wimbledon,Chelsea,1,1,D,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
18	1993-94,1993-08-18T00:00:00Z,Blackburn,Norwich,2,3,A,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
19	1993-94,1993-08-18T00:00:00Z,Coventry,Newcastle,2,1,H,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											
20	1993-94,1993-08-18T00:00:00Z,Man United,Sheffield United,3,0,H,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA											

Slika 9: Originalni skup podataka

Skup podataka koji se koristi u ovom projektu se sastoji od ukupno od 23 atributa od kojih su konkretno 4 atributa kategorička, 16 numeričkih te jedan tekstualni i jedan dateTime. Ukupno ovakav neobrađeni skup podatak ima 11 113 zapisa.

Za čišćenje podataka koristili su se softveri SQL server te SQL Server Management Studio(SSMS). Čišćenje podataka započelo je preuzimanjem izvornog skupa podataka te je taj skup podataka uvezen pomoću SQL Server Management Studia(SSMS) u bazu podataka kao tablica. Nakon učitavanja podataka u tablicu primijećeno je da za određene sezone nedostaje velika količina podataka. To je provjereno pomoću SQL upita koji se može vidjeti na slici 10.

```
Select DISTINCT (Season)
from dbo.results_premier_league
where HR ='NA' and AR='NA'
```

Slika 10: SQL upit za pretraživanje praznih vrijednosti

Rezultat toga upita je bio da sezone od 1993/1994 pa do 1999/2000 nemaju zapravo nikakve statističke podatke osim krajnjeg rezultata. Razlog tome je što su to stariji podaci i prije se nisu bilježili svi statistički podaci kao danas. Iz razloga jer nedostaje jako puno podataka te sezone su izbrisane. Brisanje se izvršilo pomoću SQL upita koji se nalazi na slici 11.

```
Delete
from dbo.results_premier_league
where HR ='NA' and AR='NA'
```

Slika 11: SQL upit za brisanje određenih sezona

Pomoću funkcije filtriranja provjerena je ispravnost podataka odnosno da li neki zapis ima praznu vrijednost. Nisu pronađeni zapisi koji imaju praznu vrijednost u bilo kojem atributu. Isto tako ovaj skup podataka nije sadržavao duplikate. Primijećeno je da kolona „DateTime“ zapravo ne posjeduje informaciju kada je utakmica odigrana već samo datum odigravanja utakmice odnosno samo sezone od 2019/2020 sadrže podatke o vremenu. Iz tog razloga odlučeno je da se izbriše vremenska komponenta i ostane samo datum. To je odrađeno pomoću SQL upita koji se nalazi na slici 12.

```
Update results
Set DateTime = REVERSE (PARSENAME(REPLACE (Reverse (DateTime), 'T', '.'), 1))
```

Slika 12: Ažuriranje kolone DateTime

Primijećeno je i da sezona 2021/2022 nema sve utakmice odnosno ima samo 309 zapisa. Razlog tome je taj da je to trenutno aktualna sezona te još uvijek nisu odigrane sve utakmice. Ovi podaci će ostati u finalnom skupu podataka jer su podaci za svaki zapis kompletni, a što je više zapisa to će modeli biti precizniji.

Nakon provedenog prethodnog uređivanja podaci su očišćeni i spremni za korištenje te se mogu vidjeti na slici 13. Finalan skup podataka ima ukupno 8289 zapisa, 380 utakmica po sezoni osim aktualne sezone koja ima 309 odigranih utakmica u vrijeme pisanja ovog rada. Ukupno su tu podaci o 22 odigrane sezone. Može se zaključiti da su svi podaci ispravni te da je skup podataka spreman za daljnje korištenje u sljedećoj fazi.

Season	DateTime	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee
2000-01	2000-08-19	Charlton	Man City	4	0	H	2	0	H	Rob Harris
2000-01	2000-08-19	Chelsea	West Ham	4	2	H	1	0	H	Graham Bar...
2000-01	2000-08-19	Coventry	Middlesbro...	1	3	A	1	1	D	Barry Knight
2000-01	2000-08-19	Derby	Southampton	2	2	D	1	2	A	Andy D'Urso
2000-01	2000-08-19	Leeds	Everton	2	0	H	2	0	H	Dermot Gall...
2000-01	2000-08-19	Leicester	Aston Villa	0	0	D	0	0	D	Mike Riley
2000-01	2000-08-19	Liverpool	Bradford	1	0	H	0	0	D	Paul Durkin
2000-01	2000-08-19	Sunderland	Arsenal	1	0	H	0	0	D	Steve Dunn
2000-01	2000-08-19	Tottenham	Ipswich	3	1	H	2	1	H	Alan Wiley
2000-01	2000-08-20	Man United	Newcastle	2	0	H	1	0	H	Steve Lodge
2000-01	2000-08-21	Arsenal	Liverpool	2	0	H	1	0	H	Graham Poll
2000-01	2000-08-22	Bradford	Chelsea	2	0	H	1	0	H	Mark Halsey
2000-01	2000-08-22	Ipswich	Man United	1	1	D	1	1	D	Jeff Winter
2000-01	2000-08-22	Middlesbro...	Tottenham	1	1	D	0	1	A	Peter Jones
2000-01	2000-08-23	Everton	Charlton	3	0	H	0	0	D	Andy Hall
2000-01	2000-08-23	Man City	Sunderland	4	2	H	2	0	H	David Ellaray
2000-01	2000-08-23	Newcastle	Derby	3	2	H	1	1	D	Dermot Gall...
2000-01	2000-08-23	Southampton	Coventry	1	2	A	0	1	A	F Taylor
2000-01	2000-08-23	West Ham	Leicester	0	1	A	0	0	D	Rob Styles

Slika 13: Izgled očišćenih podataka

4. Razvoj prediktivnih modela

U nastavku će biti opisani procesi modeliranja problema kao i izrade pojedinih modela. Tehnike modeliranja koje će se koristiti su stablo odlučivanja, neuronske mreže te Bayesove mreže. Za izradu stabla odlučivanja i neuronske mreže koristit će se alat BigML, dok će se za izradu Bayesove mreže koristiti alat Netica. Svaki pojedini model imati će svoje poglavlje odnosno postupak za izradu svakog model biti će prikazan zasebno.

4.1. Stablo odlučivanja

Stablo odlučivanja je u suštini alat za podršku odlučivanju koji koristi model odluka vizualno sličan stablu te njihove moguće posljedice. Ovaj model uključuje ishode slučajnih događaja, troškove resursa te korisnost. Kod korištenja stabla odlučivanja moramo imati na umu da je to jedan od načina prikazivanja algoritma koji sadrži samo uvjetne naredbe kontrole (Agrawal et al., 2018).

Stablo odlučivanja je zapravo uobičajena metoda rudarenja podataka koja se koristi za uspostavljanje sustava klasifikacije na temelju više kovarijanta ili za razvoj algoritma predviđanja za neku ciljanu varijablu. Ova metoda klasificira podatke u dijelove nalik granama koje konstruiraju obrnuto stablo s korijenskim čvorom, unutarnjim čvorovima i čvorovima lista. Algoritam koji se koristi je neparametraski što zapravo znači da se može učinkovito nositi s velikim i kompliciranim skupovima podataka. Kada je taj skup podataka dovoljno velik tada se on može podijeliti na skupove podataka za obuku odnosno treniranje i skup podataka za validaciju odnosno testiranje. Stablo odlučivanja je jedna od najučinkovitijih metoda rudarenja podataka te je prvi put predstavljena 1960-ih godina. Ova metoda ima široku primjenu jer je laka za korištenje, nema dvosmislenosti i robusna je čak i kada nedostaju određene vrijednosti. Kod stabla odlučivanja diskretne i kontinuirane varijable mogu se koristiti kao ciljane ili nezavisne varijable (Song i Lu, 2015).

Stablo odlučivanja kao metoda pojavljuje se u dva područja znanosti. Prvo područje je analiza odlučivanja i tu se koriste kako bi vizualno predstavio način donošenja odluka od strane eksperta odnosno koriste se kod ekspertnih sustava. Takva stabla odlučivanja opisuju način na koji ljudski eksperti dolaze do odluke kod nekog problema odlučivanja kao što je npr. izbor lokacije tvrtke, dijagnoza u medicini i sl. Drugo područje je strojno učenje gdje su stabla odlučivanja zapravo prediktivni modeli koji na temelju podataka izvode njihove veze s ciljem dobivanja izlaznih vrijednosti. Takva stabla koriste se u rudarenju podataka za otkrivanje skrivenih veza između podataka.Takva stabla temelje se na podacima, a ne na odluci eksperta

i dijele se na klasifikacijska stabla i regresijska stabla. Klasifikacijsko stablo je rezultat događaja koji se trebaju dogoditi, a rezultati ovise o rezultatu prethodnog događaja. Ovo stablo se najčešće koristi kad se za svaki podatak zna kojoj klasi pripada. Kod regresijskog stabla koristi se zbroj kvadrata i analiza regresije za predviđanje vrijednosti ciljane varijable (Zekić-sušac, bez dat.).

Stablo odlučivanja određuje tijek razvoja ekspertnog sustava. Isto tako radi predikciju na temelju slijeda testova na deskriptivnim atributima upita. Stablo odlučivanja se sastoji od sljedećih elementa (Oreški, 2021):

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

Korijenski čvor koji se isto tako zove i čvor odluke ili početni čvor, predstavlja izbor koji će rezultirati podjelom svih podataka na dva ili više međusobno isključivih podskupova. Unutarnji čvorovi predstavljaju jedan od mogućih izbora dostupnih u toj točki strukture stabla. Taj čvor je povezan s roditeljskim čvorom (čvor iznad njega) i sa čvorom dijete (čvor ispod njega). Čvorovi lista ili završni čvorovi predstavljaju konačan rezultat kombinacije odluka ili događaja (Song i Lu, 2015).

Stablo odlučivanja je zapravo alternativni način prikazivanja i analize situacije odlučivanja. To je zapravo slikoviti model koji predstavlja čitavu strukturu odlučivanja. Definirane su neke pretpostavke kod korištenja stabla a to su (Efzg, 2011):

- Donositelj odluke ima na raspolaganju većinu relevantnih inačica odluke
- Moguće posljedice (ishodi) inačica odluke mogu se na neki način kvantificirati
- Pri izboru se razmatraju samo ona obilježja inačica odluka koja se mogu kvantificirati
- Stablo odlučivanja može se analizirati ako postoje subjektivne vjerojatnosti nastupanja nesigurnih događaja

Stablo odlučivanja kao i svaka druga metoda ima određene prednosti i nedostatke. Prednosti stabla odlučivanja su sljedeće (Oreški, 2021):

- Interpretabilnost
- Rad s kategorijskim i kontinuiranim atributima
- Mogu se modelirati interakcija između deskriptivnih atributa
- Robusno je na kurs dimenzionalnosti
- Robusno je na šum u skupu, ako se radi obrezivanje

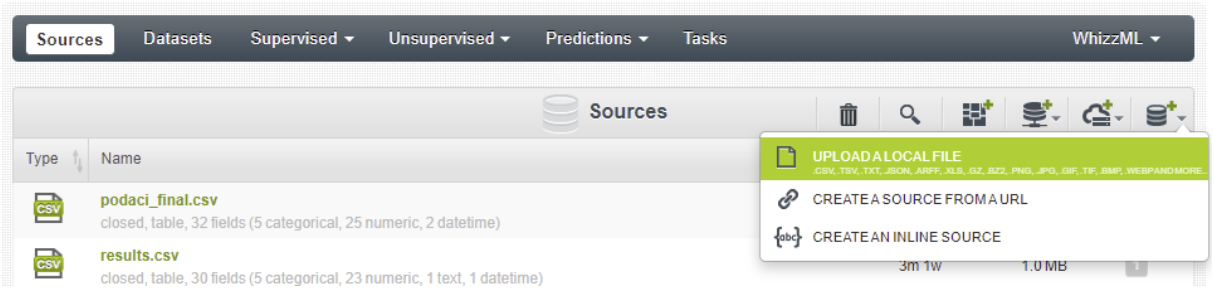
Nedostaci stabla odlučivanja su (Oreški, 2021):

- Stalo postaje veliko kada se radi s kontinuiranim atributima
- Jako je osjetljivo na skup podataka s kojim se radi
- Male promjene u skupu podataka mogu uzrokovati temeljne promjene u stablu

4.1.1. Izrada stabla odlučivanja

Za izradu modela stabla odlučivanja koristio se alat BigML. BigML je ustvari programabilna i skalabilna platforma za stojno učenje koja olakšava rješavanje i automatizaciju zadataka klasifikacije, regresije, analize klastera, otkrivanje anomalija i modeliranje tema. BigML koristi tisuće analitičara, programera i znanstvenika diljem svijeta za rješavanje zadatka strojnog učenja pretvarajući podatke u djelotvorne modele koji se koriste kao udaljene usluge ili lokalno, ugrađene u aplikacije radi predviđanja. Ovaj alat nudi veliki izbor algoritma za strojno učenje koji dokazano rješavaju probleme u stvarnom svijetu primjenom jedinstvenog, standardiziranog okvira. BigML također olakšava neograničene prediktivne aplikacije u svim industrijama kao što su zrakoplovstvo, automobilska industrija, energija, IoT i sl. Alat nudi jednostavno sučelje i interaktivnu vizualizaciju što prediktivne model čini razumljivijim (BigML, bez dat.).

Kako bi mogli izraditi stablo odlučivanja u alata BigML najprije je potrebno učitati datoteku s podacima u alat. Alat podržava različite formate datoteka, a skup podataka koji je ovdje korišten je u .csv formatu te je dobiven izvozom podataka iz SSMS-a. Za učitavanje podataka potrebno je otići na karticu „Sources“, odabrati gumb za učitavanje nove datoteke u alat te odabrati opciju „Upload local file“. Taj postupak se može vidjeti na slici 14.

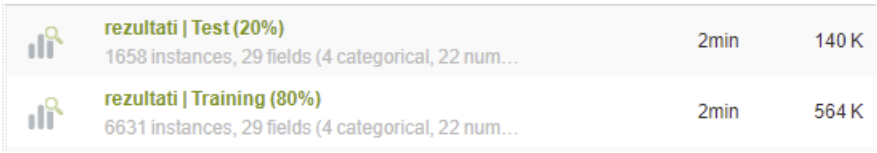


Slika 14: Učitavanje skupa podataka u alat BigML

Sljedeći korak je izrada skupa podataka odnosno „dataset-a“ iz učitane datoteke. To se može učinit tako da se odabere učitana datoteka i nakon toga se odabere opcija „Configure dataset“. Nakon toga se može promijeniti naziv skupa podataka te je potrebno odabrati gumb „Create dataset“.

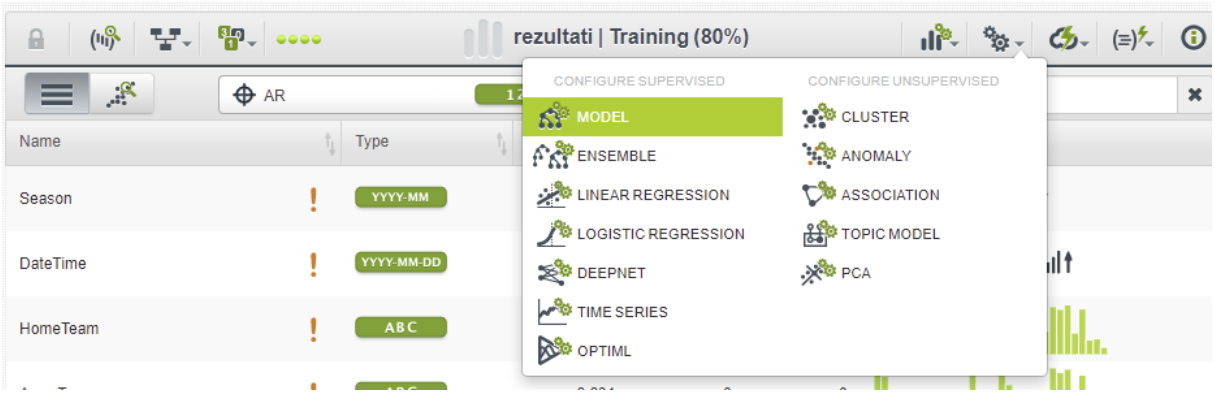
Alat je nakon kreiranja „dataset-a“ dodao još nekoliko atributa odnosno rastavio je datumsku komponentu u atribute naziva DateTime.Year, DateTime.Month i DateTime.DayOfMonth. Zbog toga razloga ovako kreiran „dataset“ ima nešto više atributa nego originalni skup podataka. Nakon što je kreiran „dataset“ potrebno ga je podijeliti na skup za treniranje i skup za testiranje. U alatu BigML postoji implementirana funkcionalnost koja omogućava da se skup podataka slučajni odabirom podijeli na 20% za testiranje i 80% za treniranje odnosno stvaraju se dva skupa podataka. Postupak je prikazan na slici 15, a na slici 16 se mogu vidjeti dva nova „dataset-a“ koji su stvoreni nakon primijenjene funkcionalnosti.

Slika 15: Random split funkcionalnost



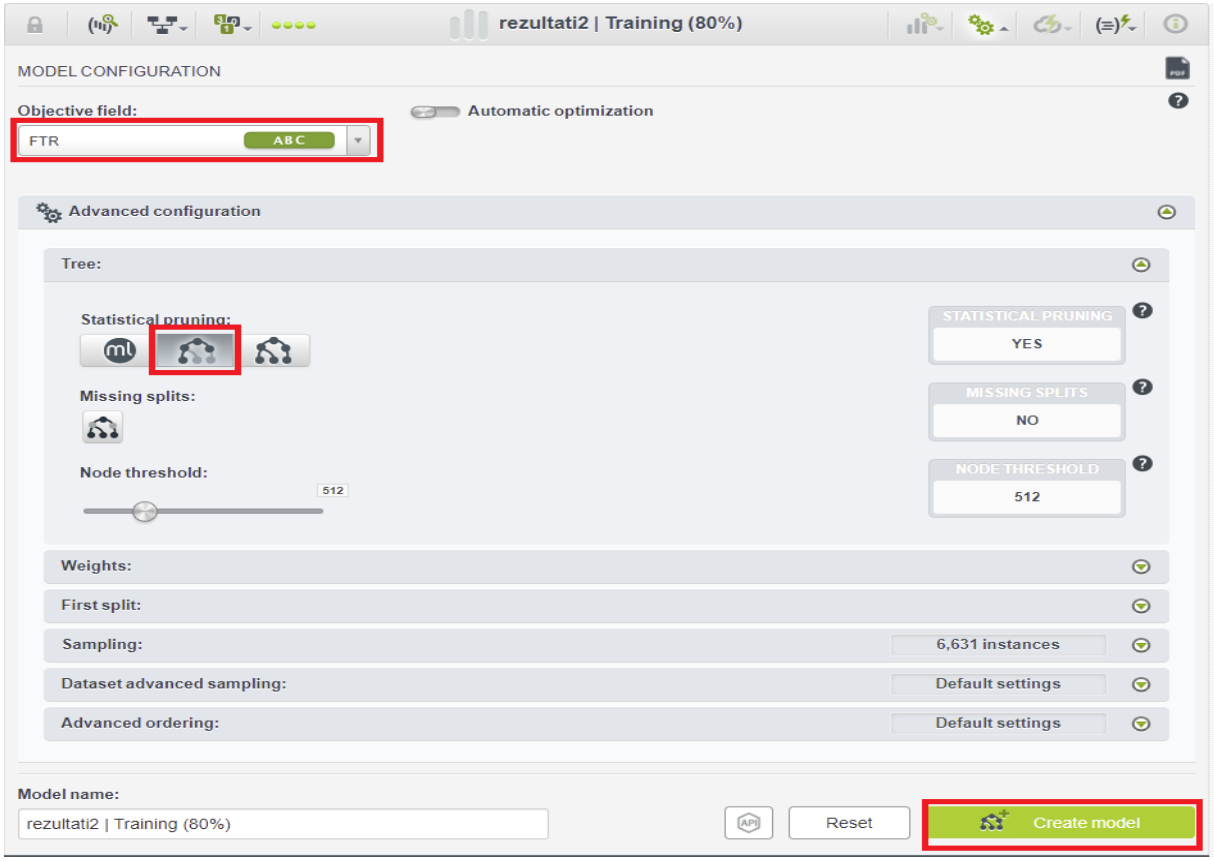
Slika 16: Novokreirani skupovi podataka

Alat je u skup za treniranje smjestio ukupno 6631 instanca dok je u skup za testiranje smjestio 1658 instanca. Nakon uspješnog učitavanja podataka te dijeljenja podataka na dva skupa sljedeći korak je izrada stabla odlučivanja nad skupom podataka koji je namijenjen za treniranje. Kako bi se izradilo stablo odlučivanja potrebno se u alatu pozicionirati na „dataset“ unutar „dataset-a“ odabrati opciju „Configure“ pa „Model“. To se može vidjeti na slici 17.



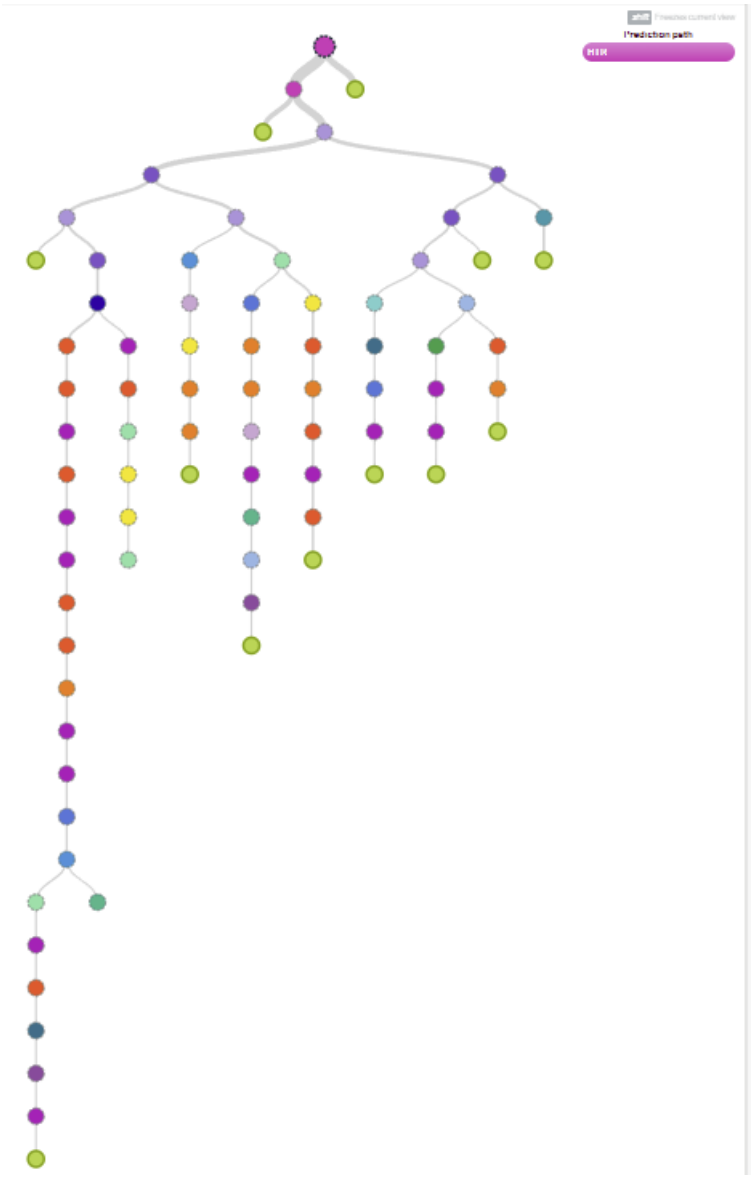
Slika 17: Postupak izrade stabla odlučivanja

Isto tako potrebno je i podesiti opcije prije stvaranja samog modela. To podešavanje opcija se može vidjeti na slici 18. Najprije je potrebno odabrati atribut koji će se predviđati, a u ovom slučaju to je FTR odnosno rezultat utakmice. Ishod atributa FTR ima tri moguća stanja, a to su A (pobjeda gostujućih), H (pobjeda domaćih) i D (neriješeno) koji nam govore o ishodu utakmice odnosno da li je pobijedio domaćin, gost ili je utakmica završila neriješeno. Zatim je potrebno odabrati koju vrstu obrezivanje će program koristiti. Za potrebe ovog rada isprobane su različite vrste obrezivanja te je u nastavku prikazano stablo koje je imalo najveću pouzdanost. Ovdje je odabrana opcija „Smart prunning“. Na kraju je potrebno odabrati opciju „Create model“.



Slika 18: Konfiguriranje stabla odlučivanja

Stablo odlučivanja koje je dobiveno iz navedenih parametara se može vidjeti na slici 19.



Slika 19: Stablo odlučivanja

Nakon što je stablo odlučivanja izrađeno može se prikazati koji atributi imaju najveću važnost kod finalnog ishoda pomoću opcije „Model Summary Report“. U nastavku slijedi popis atributa te njihovih važnosti:

- HTR: 34.3%
- AST: 15.17%
- HST: 11.94%
- AwayTeam: 9.98%
- Referee: 8.09%

- HomeTeam:4.1%
- DateTime.DayOfMonth: 3.11%
- Season: 2.29%
- AC: 1.76%
- HC: 1.53%
- AF: 1.42%
- HR: 1.39%
- HF: 1.09%
- HY: 0.91%
- AS: 0.9%
- DateTime.Year: 0.6%
- HS: 0.54%
- AR: 0.52%
- DateTime.Month: 0.37%

Iz navedenih podataka može se vidjeti da je najvažniji atribut HTR. To zapravo govori da finalan rezultat utakmice najviše ovisi o rezultatu na poluvremenu koji može imati tri stanja: A(pobjeda gostujućih), H(pobjeda domaćih) i D (neriješeno). Isto tako finalan rezultat dosta ovisi i o AST i HTS što su zapravo udarci u okvir gola domaćeg odnosno gostujućeg tima. To se može protumačiti kao da što više pojedina ekipa puca u okvir gola to su joj veće šanse za pobjedu.

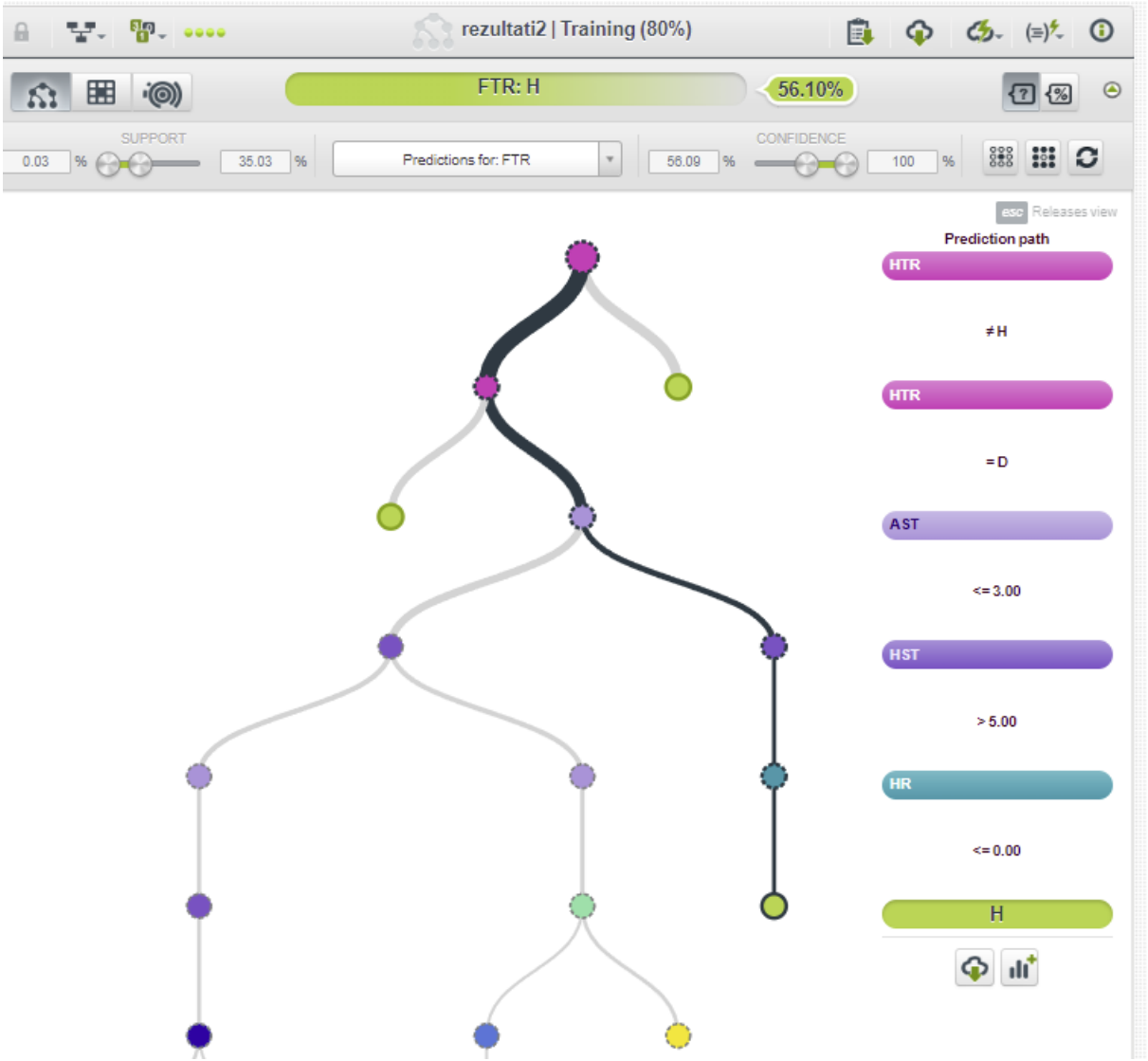
Pomoću opcije „Model Summary report“ može se vidjeti distribucija podataka kao i distribucija predviđanja stabla odlučivanja. Te distribucije nalaze se na slici 20. Može se vidjeti da stablo odlučivanja predviđa više pobjeda domaćina kao i pobjeda gostiju nego to ustvari je. S druge strane stablo predviđa daleko manje neriješenih rezultata od stvarnih.

```
Data distribution:
    A: 29.45% (1953 instances)
    D: 24.75% (1641 instances)
    H: 45.80% (3037 instances)


Predicted distribution:
    A: 34.17% (2266 instances)
    D: 12.67% (840 instances)
    H: 53.16% (3525 instances)
```

Slika 20: Distribucija podataka

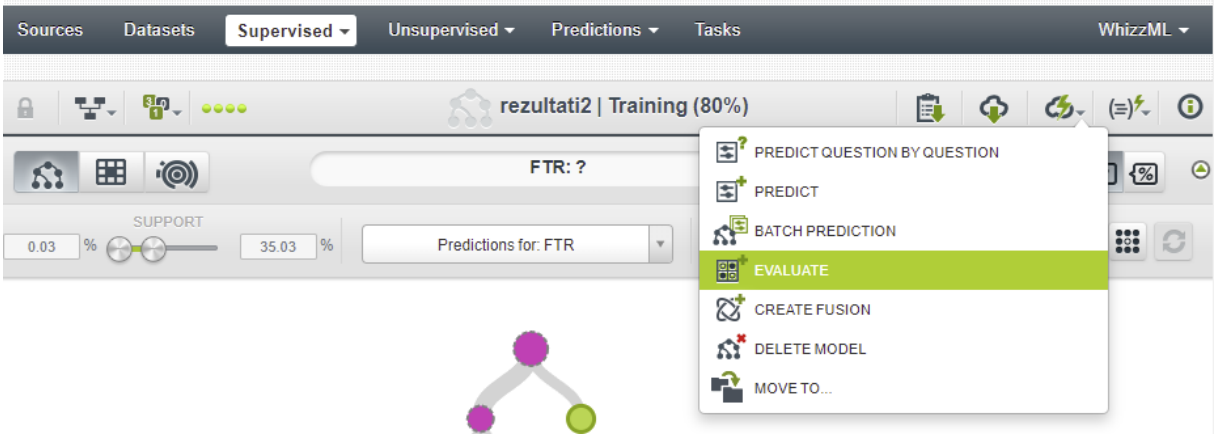
Kod stabla odlučivanja potrebno je još i prikazati pravilo odnosno granu koja ima najvišu pouzdanost. To pravilo se može vidjeti na na slici 21. Pouzdanost ovog pravila iznosi 56.10%.



Slika 21: Najpouzdanije pravilo u stablu odlučivanja

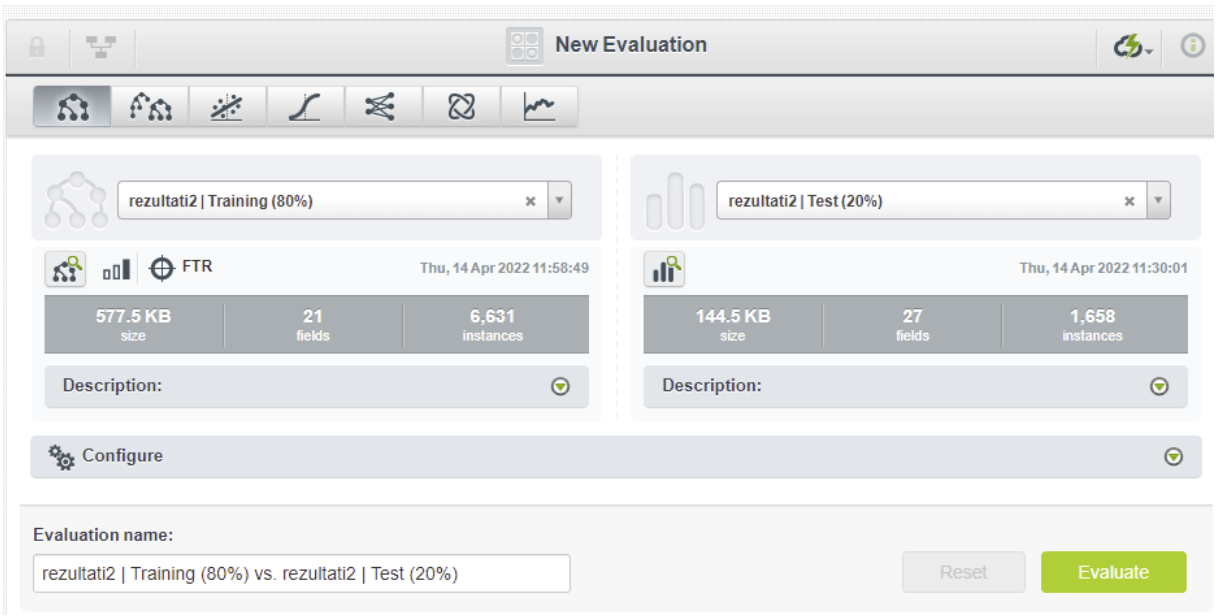
Pravilo prikazano na slici 21 zapravo govori ako je rezultat na poluvremenu neriješeno te ako gosti nemaju više od tri udarca u okvir a domaći imaju manje od pet udarca u okvir te ako domaći tim ne dobije crveni karton tada će pobijediti domaći tim.

Sljedeći korak je evaluacija stabla. Ovdje zapravo evaluiramo koliko je stablo točno odnosno koliko dobro predviđa rezultate. Uzima se prethodno kreirani skup za treniranje te na tom skupu primjenjuje stablo odlučivanja. Rezultat će biti postotak točnog predviđanja. Postupak izrade evaluacije se može vidjeti na slikama 22 i 23.



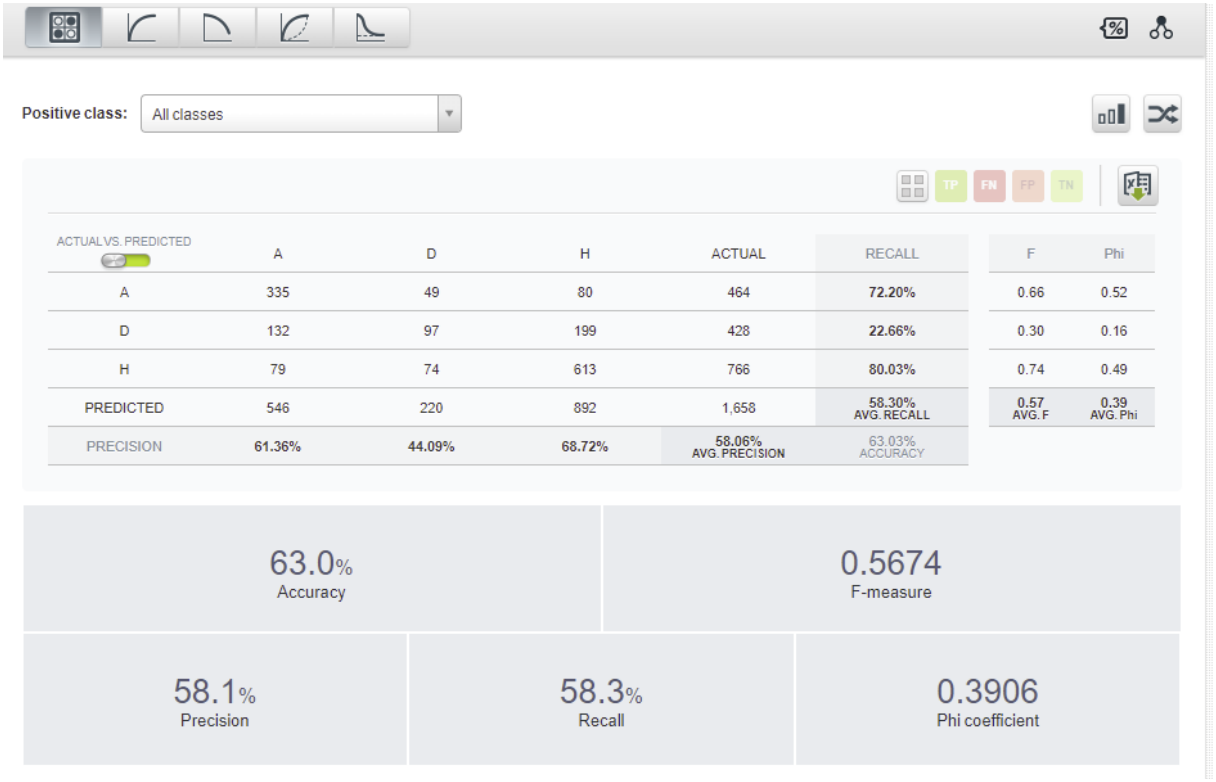
Slika 22: Evaluacija stabla

Nakon što se odabere opcija „Evaluate“ , otvara se novi prozor kao na slici 23. U tom prozoru potrebno je odabrati odgovarajuće stablo odlučivanja kao i skup podataka za testiranje. Nakon toga potrebno je odabrati gumb „Evaluate“.



Slika 23: Konfiguriranje evaluacije stabla

Nakon izvršene evaluacije dobije se matrica konfuzije. Matrica konfuzije je tablica koja se koristi kako bi se opisale performanse modela klasifikacije na određenom skupu podataka koji ima poznate vrijednosti. Matrica konfuzije za ovo stablo odlučivanja može se vidjeti na slici 24 (Data School, 2014).



Slika 24: Matrica konfuzije stabla odlučivanja

Iz matrice konfuzije može se vidjeti da je točnost stabla odlučivanja 63% odnosno to znači da je stablo dobro klasificiralo 63% instanci. Isto tako može se vidjeti da stablo odlučivanja najlošije predviđa neriješene rezultate. Stablo je predvidjelo samo 220 neriješenih rezultata dok je u stvarnosti ukupno 428 neriješenih rezultata. S druge strane stablo klasificira više instanci u razrede A i H od stvarnog broja instanci. Prosječna preciznost stabla je 58.01%. Preciznost zapravo govori o tome koliko su mjerenja iste stavke blizu jedno drugog. Može se reći da je u usporedbi s prethodnim istraživanjima ukupno gledajući stablo imalo otprilike identičnu ukupnu točnost, ali relativno lošu točnost u predviđanju neriješenih rezultata.

4.2. Neuronske mreže

Umjetna neuronska mreži služi za rješavanje problema umjetne inteligencije. Model neuronskih mreža paradigma je za obradu informacija inspirirana načinom na koji biološki neuronski sustavi obrađuju podatke (Bunker i Thabtah, 2019).

Ljudski mozak je zapravo sustav za prepoznavanje uzoraka. On radi tako da procesuira neki ulaz iz svijeta te na temelju toga ulaza stvara neki izlaz. Taj proces radi automatski i gotovo bez ikakvog napora. Neuronske mreže pokušavaju oponašati način na koji ljudski mozak radi odnosno rješava probleme. Neuronska mreža u širem smislu je zapravo umjetna

replika ljudskog mozga kojom se nastoji simulirati postupak učenja i obrade podataka. Konkretnije i preciznija definicija je da je neuronska mreža skup međusobno povezanih jednostavnih elementa čija funkcionalnost se temelji na biološkom neuronu i koje služe paralelnoj obradi podataka. Neuronske mreže jako dobro rješavaju probleme klasifikacije i predviđanja odnosno sve probleme kod kojih postoji složena veza ulaza i izlaza. Najčešće se koriste za prepoznavanje uzoraka, obrade slike i govora, nelinearno upravljanje, simulacije i sl. (Dalbelo Bašić et al., 2012).

Umjetne neuronske mreže pripadaju u inteligentne metode rudarenja podataka čiji je cilj pronaći skrivene veze između podataka. Kao što je već navedeno, neuronska mreža je međusobno povezana nakupinama jednostavnih elemenata koji se nazivaju čvorovima. Rad neuronskih mreža se temelji na načinu djelovanja neurona kod živih bića. Sposobnost obrade podataka koju mreža posjeduje je posljedica jačine veze između neurona, a ta jačina veze se postiže kroz proces učenja iz skupa podataka. Drugim riječima neuronske mreže su programi koji najčešće iterativnim putem pokušavaju pronaći vezu između ulaznih i izlaznih varijabli u prošlim podacima kako bi mogli za nove vrijednosti ulaznih varijabli predvidjeti vrijednost izlaznih varijabli. Neuron od kojih se neuronska mreža sastoji je zapravo jedinica za obradu podataka koja prima ponderirane ulazne vrijednosti od drugih varijabli te prema nekoj formuli transformira tu ulaznu vrijednost te šalje izlaz drugim varijablama. Razlikuju se mnogobrojni algoritmi neuronskih mreža (Zekić-Sušac i Frajman-Jakšič, 2008).

Faze rada neuronskih mreža može se podijeliti na (Dumančić, 2014):

- Fazu učenja (treniranja)
- Fazu selekcije (unakrsne validacije)
- Fazu testiranja
- Operativnu fazu (fazu opoziva)

Od prethodno navedenih faza, faza učenja i faza testiranja se navode kao dvije najvažnije faze. Te dvije faze su zapravo temeljne faze rada neuronskih mreža. U fazi selekcije se optimizira duljina treniranja te broj skrivenih neurona. Operativna faza se odnosi na primjenu neuronske mreže na nove skupove podataka s nepoznatim rezultatima i fiksiranim težinama (Dumančić, 2014).

Kod učenja neuronske mreže, težine veza između neurona se modificiraju najčešće iterativno, da se postigne traženi izlaz. Algoritam koji se koristi za podešavanje težina se naziva pravilo učenja. Jedno od najpoznatijih pravila učenja je širenje prema natrag koji se ujedno i najčešće koristi. Širenje prema natrag se odnosi na algoritam za podešavanje težina veza u

višestrukim slojevima mreže. Rad algoritma širenje prem natrag opisan je kroz sljedećih pet koraka (Kliček, 2021):

1. Postaviti ulaznu vrijednost i prenosti tu vrijednost do izlaznog sloja te ostvariti izlazni vektor
2. Za vrijeme prenošenja ulaznog vektora, određuju se ulazne i izlazne vrijednosti svakog neurona u mreži
3. Za svaki element obrade u izlaznom sloju računa se skalirana lokalna greška i delta težina.
4. Za svaki sloj počevši od predzadnjeg sloja prije izlaznog i sloja neposredno nakon ulaznog računa se skalirana lokalna greška i delta težina
5. Obnoviti sve težine veza u mreži dodavanjem delta težina prijašnjim vrijednostima

Prednosti algoritma širenja prema natrag su da ima dodatne slojeve koji dopuštaju da se rezultat jednog sloja dodatno obrađuje te da uređuje i stvara kompleksni sustav. Nedostaci ovog algoritma su dugotrajno treniranje, osjetljivost na početne vrijednosti težina, algoritmi treniranja mreže su dugotrajni i ne osiguravaju konvergenciju (Kliček, 2021).

Algoritmi neuronske mreže se mogu podijeliti po nekoliko kriterija (Zekić-Sušac, bez dat.) :

- Algoritmi prema broju slojeva
- Algoritmi prema tipu učenja

Algoritmi prema broju slojeva se dalje dijele na (Zekić-Sušac, bez dat.):

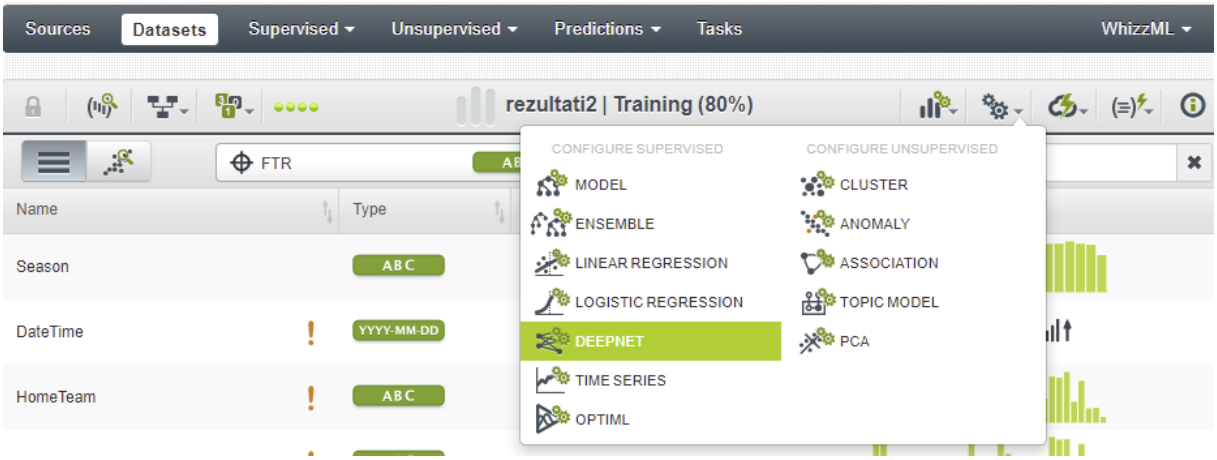
- Dvoslojne- samo ulazni i izlazni sloj (Perceptron, ADALINE)
- Višeslojne- ulazni, izlazi i jedna ili više skrivenih slojeva (višeslojni perceptron)
- Deep learning- mreže s nekoliko stotina slojeva, izvode se na BigData platformama

Algoritmi prema tipu učenja se dijele na (Zekić-Sušac, bez dat.):

- Nadgledani- poznate su vrijednosti ulaznih i izlaznih varijabli na skupu podataka za učenje mreže (Višeslojni perceptron, General Regression, LVQ)
- Nenadgledani- poznate su ulazne vrijednosti, ali nisu poznate vrijednosti izlaznih varijabli na skupu podataka za učenje mreže (ART mreža, Kohonenova samoorganizirajuća mreža)

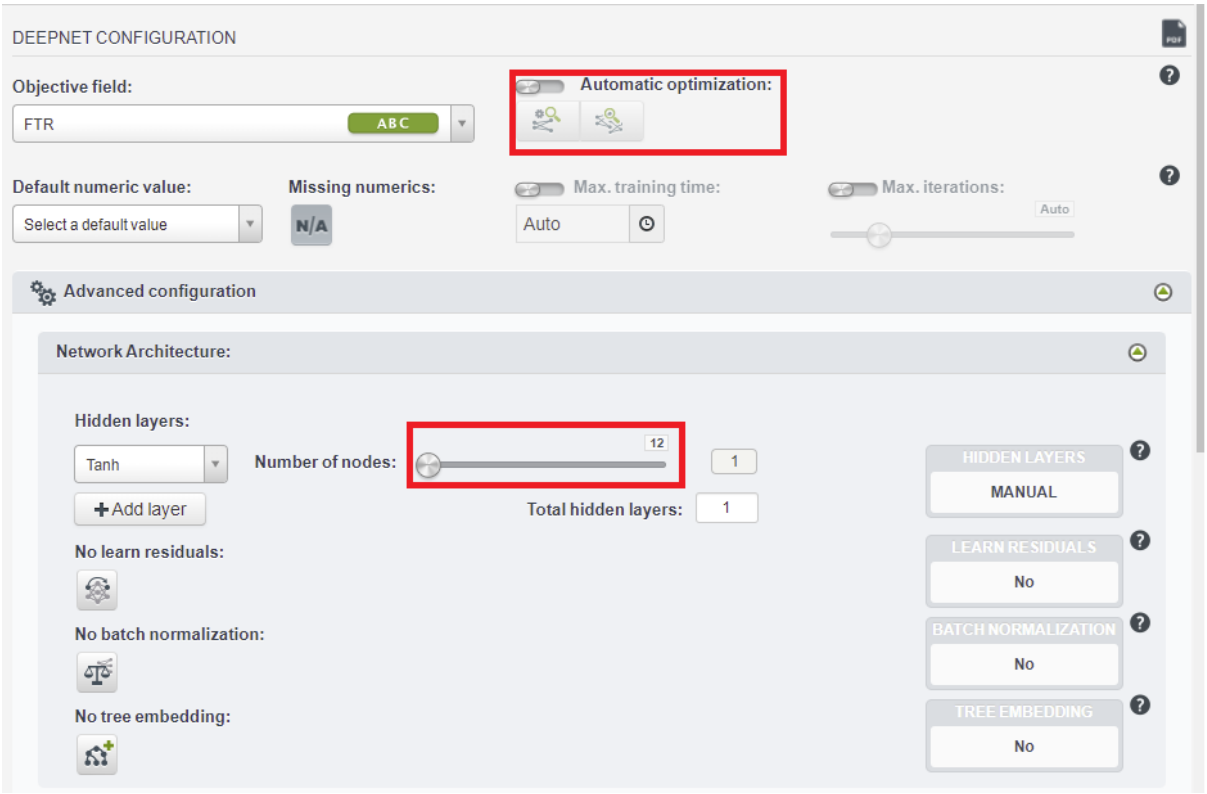
4.2.1. Izrada neuronske mreže

Nakon izrade stabla odlučivanja potrebno je izraditi neuronsku mrežu za isti skup podataka. Neuronska mreža će se kao i stablo odlučivanja izraditi u alatu BigML. Kako bi se izradila neuronska mreža potrebno je u izborniku „*Configure*“ odabrati opciju neuronskih mreža ili „*Deepnet*“. To se može vidjeti na slici 25. Nakon toga otvara se novi izbornik gdje je potrebno konfigurirati postavke za izradu neuronske mreže.



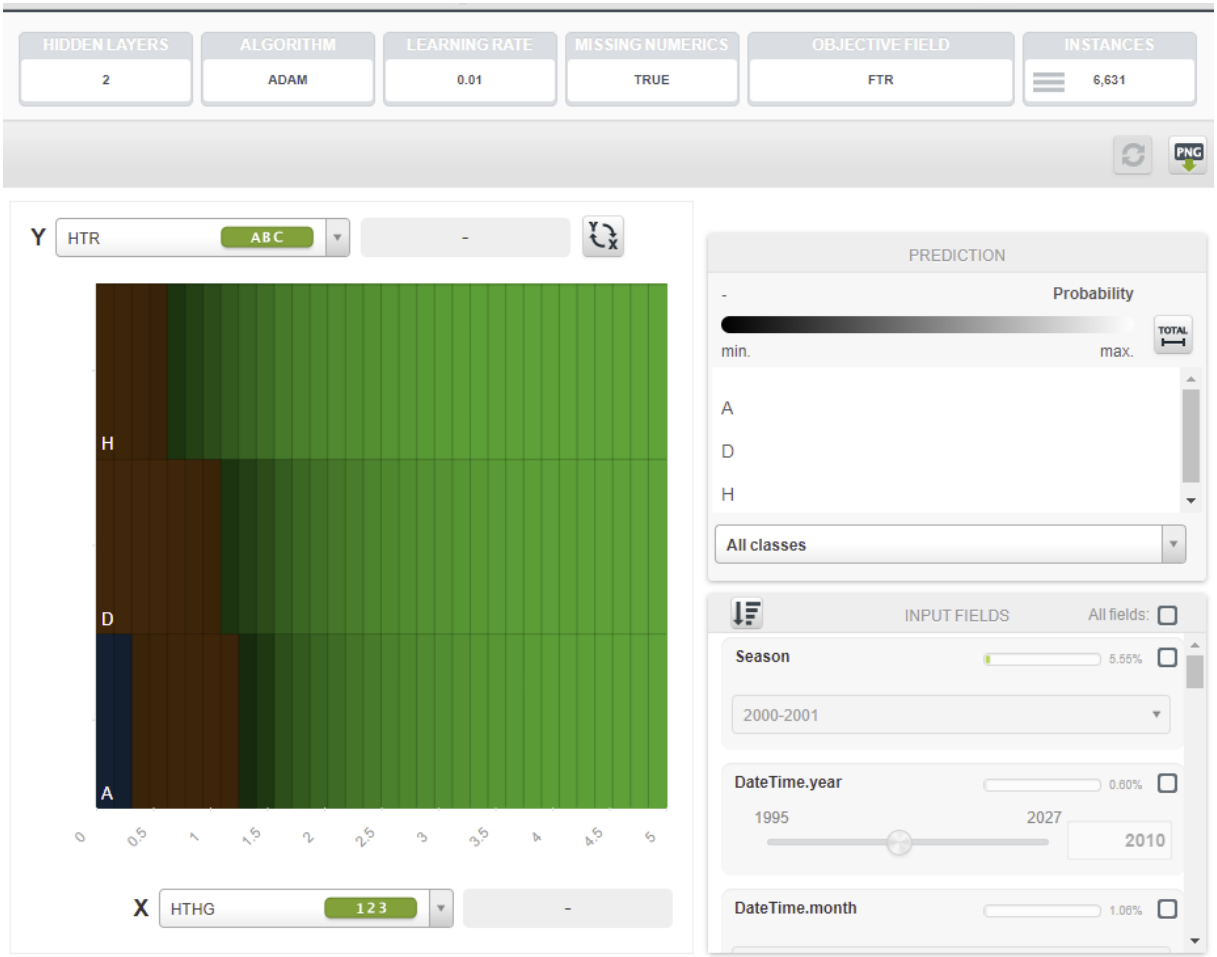
Slika 25: Stvaranje neuronske mreže

Kod konfiguriranje neuronske mreže potrebno je odabrati broj neurona unutar mreže. Općenito pravilo glasi da kod uzimanja skrivenih neurona uzima se aritmetička sredina ukupnih atributa te od aritmetičke sredine smije biti najviše odstupanje od 3 do 5 neurona. To konkretno znači da bi neuronska mreža trebala imati 12 skrivenih neurona jer se za kreiranje neuronske mreže koristi ukupno 24 atributa. U ovom slučaju izraditi će se više neuronskih mreža te će se prikazati samo ona koja daje najbolje performanse. Izraditi će se mreža s 11, 12 i 13 skrivenih neurona kao i mreža sa automatskom optimizacijom broja neurona. Na slici 26 se može vidjeti konfiguriranje neuronske mreže s 12 skrivenih neurona.



Slika 26: Konfiguriranje neuronske mreže

Nakon izrađenih više neuronskih mreža ispostavilo se da je neuronska mreža koja koristi auto optimizaciju najviše precizna. Ovakva kreirana mreža ima dva skrivena sloja. Na slici 27 može se vidjeti neuronska mreža s dva skrivena sloja.



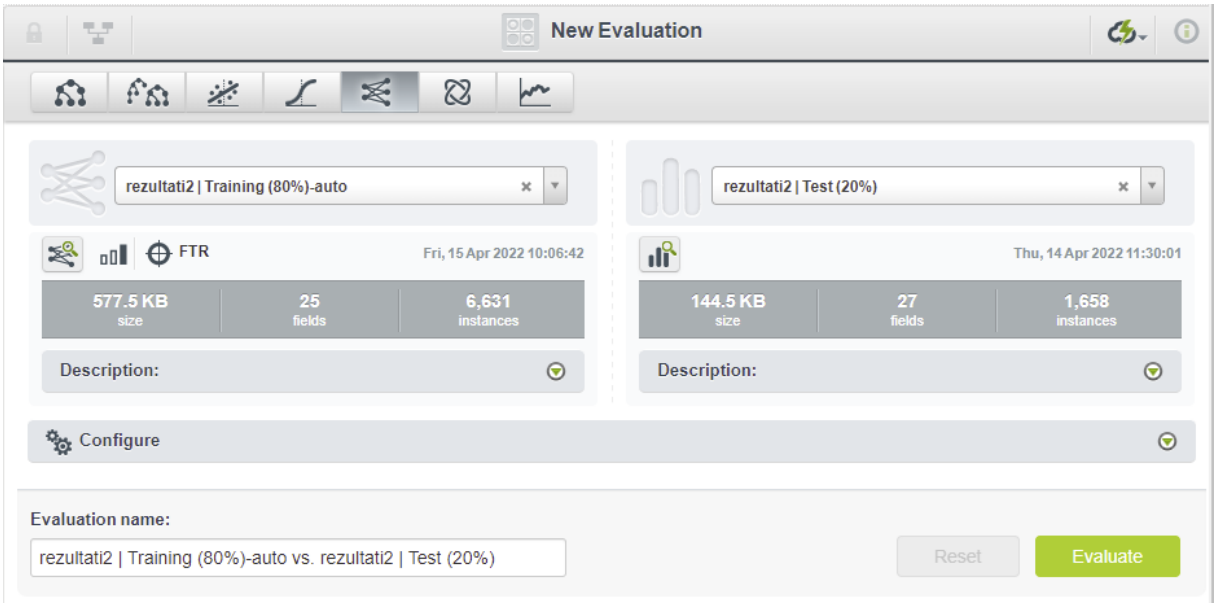
Slika 27: Neuronska mreža

Na osi X i Y se odabiru najvažniji atributi odnosno oni atributi koji najviše utječu na promatrani atribut što je u ovom slučaju FTR. U ovom slučaju najviše utječe atribut Referee ali zbog tipa atributa „text“ on se ne može prikazati na osi. Sljedeća dva najvažnija atributa su HTR (rezultat utakmice na poluvremenu) i HTHG (broj udaraca u okvir domaće momčadi). Prema gore navedenoj neuronskoj mreži najčešći ishod utakmice je H odnosno pobjeda domaćina, nakon toga slijedi D odnosno neriješeni rezultat dok je najmanje slučajeva A odnosno pobjeda gostujuće ekipe. U nastavku se nalazi popis atributa i njihovih postotaka važnosti na ciljani atribut:

- Referee: 11.21%
- HTHG: 11.07%
- HTR: 8.65%
- HTAG: 8.18%
- AwayTeam: 7.92%
- HomeTeam: 7.66%
- HST: 7.44%

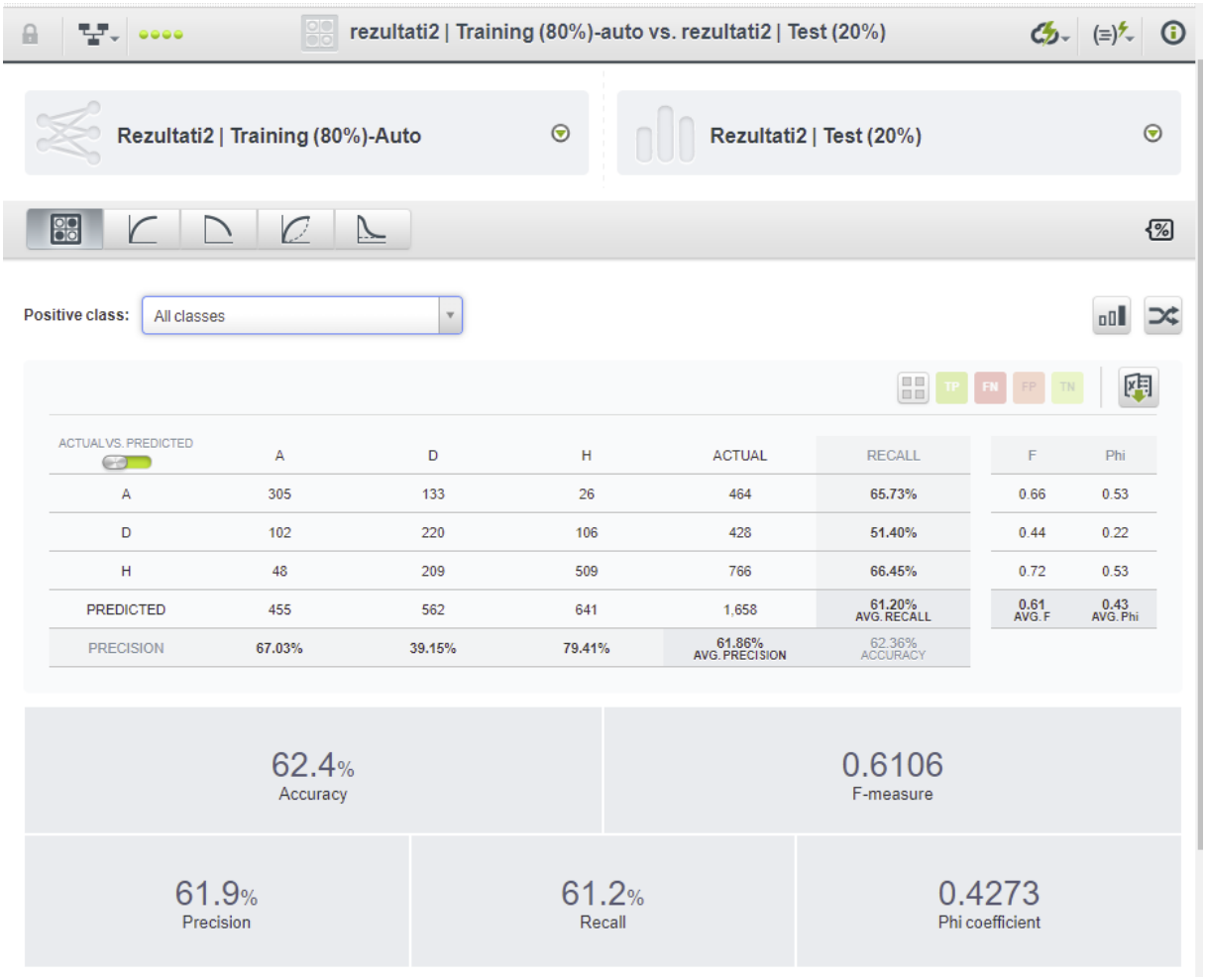
- AST: 7.2%
- Season: 5.55%
- HY: 3.75%
- HR: 3.27%
- AC: 2.36%
- HF: 2.31%
- HS: 2.24%
- HC: 2.01%
- AS: 1.74%
- AR: 1.65%
- AF: 1.33%
- AY: 1.15%
- DateTime.Month: 1.05%
- DateTime.DayOfMonth: 1.05%
- DateTime.Year: 0.6%

Slično kao i stablo, neuronsku mrežu potrebno je evaluirati. Postupak je isti kao i kod stabla odlučivanja. Evaluirati će se koliko dobro neuronska mreža predviđa rezultate odnosno koliko dobro svrstava instance u određeni skup. Konfiguracija evaluacije neuronske mreže može se vidjeti na slici 28.



Slika 28: Konfiguracija evaluacije neuronske mreže

Na slici 29 može se vidjeti matrica konfuzije za neuronske mreže. Kao što je već navedeno matica konfuzije govori o performansama modela odnosno neuronske mreže.



Slika 29: Matrica konfuzije za neuronske mreže

Iz matrice konfuzije može se vidjeti da je točnost neuronske mreže 62.4% što je neznatno lošije nego kod stabla odlučivanja. S druge strane preciznost iznosi 61.9% što je bolje nego kod stabla odlučivanja. Isto tako može se vidjeti da se najtočnije predviđaju pobjede gostujuće ekipe odnosno krivo su svrstane samo 9 instanci. Najlošije predviđanje je za neriješene rezultate međutim taj postotak je puno bolji nego kod stabala odlučivanja. Predviđanje za pobjedu domaćih je nešto lošije od predviđanja gostujuće pobjede ali opet bolje nego kod stabla odlučivanja. U usporedbi s prethodnim istraživanjima ova neuronska mreža ima nešto lošije performanse.

4.3. Predikcija korištenjem Bayesovih mreža

Zadnja metoda koja će se koristiti u ovome radu su Bayesove mreže. Bayesova mreža je zapravo vjerojatnosti grafički model koji predstavlja grupu slučajnih varijabli i njihovu ovisnost uz pomoć neperiodičnog grafa odnosno uz pomoć usmjerenog acikličkog grafa. Usmjereni aciklički graf se sastoji od grupe čvorova koje predstavljaju varijable dok bridovi

predstavljaju ovisnost između tih varijabli. Ta ovisnost između varijabli prikazana je pomoću strukture čvorova, te ta struktura osigurava kvalitativni dio slučajnog zaključivanja u Bayesovim mrežama (Lale et al., 2014).

Bayesove mreže se sastoje od tri vrste čvorova (Lale et al., 2014):

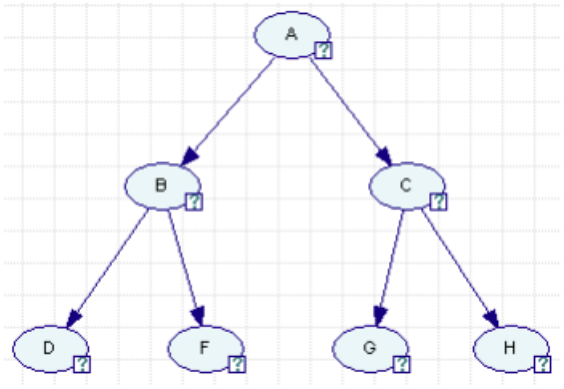
- Čvor slučajnosti
- Čvor odluke/rješenja
- Čvor korisnosti/sredstva

Bayesove mreže doživjele su najveći razvoj tijekom osamdesetih godina prošlog stoljeća. Tijekom tih godina formirala se opće prihvaćena struktura mreže, razvili su se mnogi alati za rasuđivanje te su se Bayesove mreže počele koristiti u raznim područjima. Tijekom devedesetih godina razvili su se mnogi algoritmi za učenje Bayesovih mreža iz raznih skupova podataka. Danas se one koriste kao model u računalnim sustavima kao i u raznim područjima ljudskog djelovanja. Ovaj model je idealan za uzimanje događaja koji se dogodio i predviđanje vjerojatnosti da je bilo koji od nekoliko mogućih poznatih uzroka bio čimbenik (Lale et al., 2014).

Kako bi se Bayesova mreža mogla definirati potrebno je definirati sljedeće parametre (Prcela, 2010):

- Čvorove u mreži (varijable u problemu)
- Moguće ishode svih čvorova (vrijednosti koje varijable mogu poprimiti)
- Bridove u mreži (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 čvorova koji imaju roditelje definirana su preko tablica združene distribucije vjerojatnosti i preko a priori očekivanja njihovi roditelja. Iz tog razloga nepotrebno je u definiciji mreže eksplicitno navoditi a priori vjerojatnosti ishoda za čvorove koji imaju roditelje (Peraić, 2012).



Slika 30: Primjer jednostavne Bayesove mreže (Peraić, 2012)

Kada se Bayesove mreže koriste zajedno sa statističkim varijablama tada one imaju nekoliko prednosti (Kliček, 2021):

- Kada model kodira ovisnost između svih varijabli, on može rješavati situacije kada nedostaju neki unosi podataka
- Bayesove mreže se mogu koristiti za učenje uzročnih ovisnosti odnosno mogu se koristiti za poboljšanje razumijevanja u problemskoj domeni kao i za predviđanje posljedica djelovanja
- Model je idealan za prikaz kombinacija prethodnog znanja i podataka jer ima uzročnu i vjerojatnosnu semantiku
- Bayesove statističke metode povezane s Bayesovim mrežama nude učinkovit pristup za izbjegavanje pretreniranosti podataka.

Bayesove mreže mogu se izrađivati pomoću tehnika strojnog učenja, ali i preko znanja ljudi eksperta pa ovakav dvostrani pristup daje velike prednosti u odnosu na ostale tehnike, kako strojnog učenja, tako i prikaza nesigurnih informacija (Kliček, 2021).

4.3.1. Bayesova formula

Britanski matematičar Thomas Bayes po kojemu su i Bayesove mreže dobile naziv je u svojem radu opisao matematičku formulu koja danas ima veliku važnost kod teorije vjerojatnosti. Ta formula glasi:

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

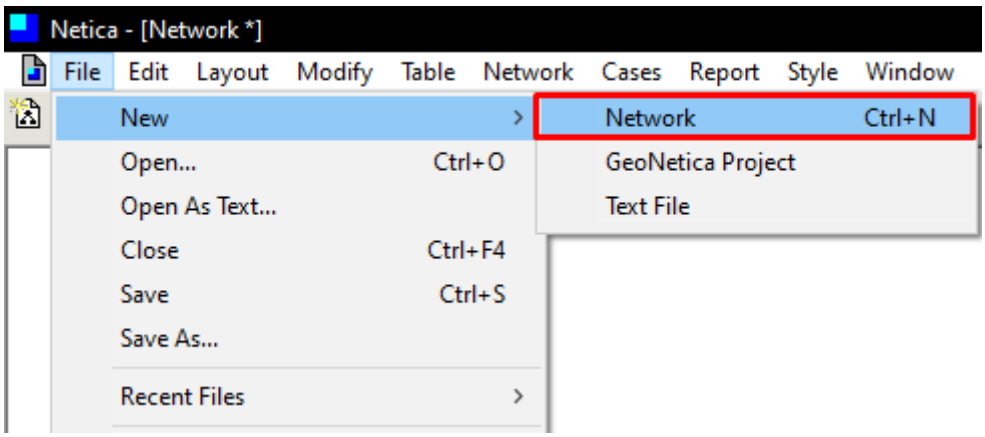
Gdje je $\{H_1, H_2, \dots, H_n\}$ potpun skup događaja na vjerojatnosnom skupu, a *A* je događaj za koji vrijedi da je *P(A) > 0* u vjerojatnosnom prostoru. Bayesova formula računa vjerojatnost da se ukoliko je ostvaren događaj *A,* potvrdio skup početnih hipoteza. Za primjenu formule potrebno je poznavati 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 hipoteza *H*. Iako je formula opisana još u 18. stoljeću pravu primjenu doživjela je tek u 20. stoljeću razvojem područja umjetne inteligencije (Perić, 2012).

4.3.2. Izrada Bayesove mreže

Kako bi se izradila Bayesova mreža korišten je alat Netica. Netica je moćan i cjelovit program za rad s mrežama uvjerenja i dijagramima utjecaja. Jednostavan je za korištenje te ima intuitivno korisničko sučelje za crtanje mreža, a odnosi između varijabli mogu se unijeti kao pojedinačne vjerojatnosti, u obliku jednadžbi ili naučiti iz skupova podataka. Netica može koristiti mreže za izvođenje različitih vrsta zaključivanja koristeći najbrže i najmodernije algoritme. Ako se učita mreža sa ograničenim znanjem odnosno podacima , Netica može pronaći odgovarajuće vrijednosti ili vjerojatnosti za sve nepoznate varijable. Bitno je napomenuti da će se u ovom radu limitirano izdanje koje dopušta najviše 15 atributa (Norsys, bez dat).

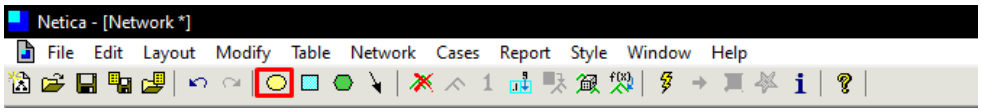
Nakon što se pokrene aplikacija potrebno je izraditi novu mrežu. Kako bi se izradila nova mreža potrebno je odabrati *File-> New->Network* kao što se može vidjeti na slici 31.



Slika 31: Izrada nove mreže u alatu Netica

Kako bi se dodali novi čvorovi unutar Bayesove mreže potrebno je odabrati *Modify->Add->Nature Node* ili jednostavno odabrati ikonu žutog kružića na alatnoj traci. To se može vidjeti na slici 32 i 33.

Slika 32: Dodavanje čvorova

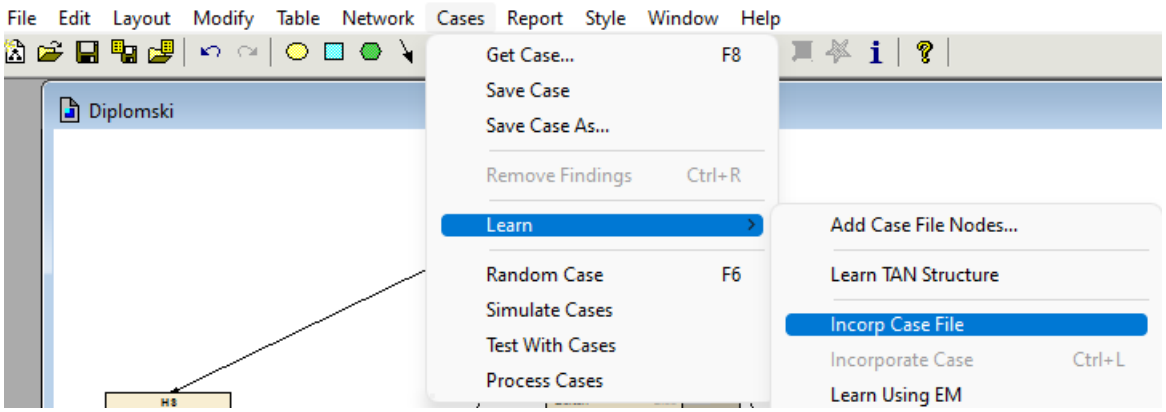


Slika 33: Dodavanje čvorova (drugi način)

Sada je potrebno dodati odgovarajuće čvorove te njihove pripadajuće vrijednosti. Imena čvorova kao i pripadajuće vrijednosti moraju biti identične kao i u skupu podataka. Svaki čvor u mreži odgovara nekom podatku za pojedinu utakmicu. Tako npr. čvor HTHG predstavlja broj golova domaćina na kraju prvog poluvremena. Zbog ograničenja alata koristit će se 15 atributa iz originalnog skupa. U nastavku se može vidjeti popis korištenih atributa kod kreiranja Bayesove mreže:

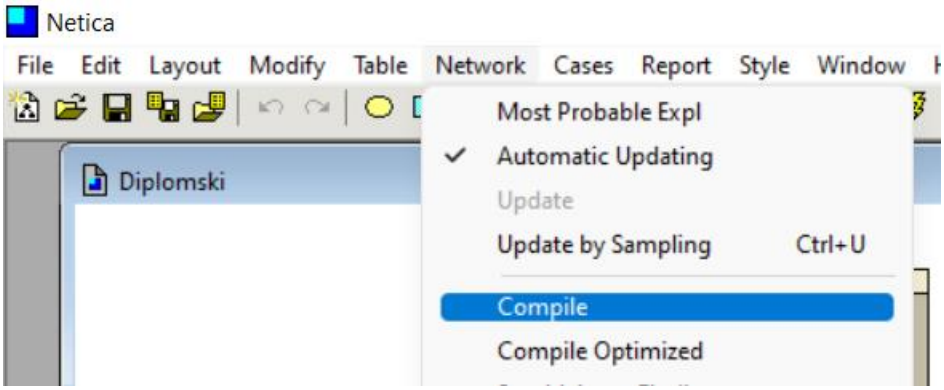
- HomeTeam – Domaći tim
- AwayTeam – Gostujući tim
- Referee – Sudac utakmice
- HTHG - Broj pogodaka u prvom poluvremenu domaćeg tima
- HTAG - Broj pogodaka u prvom poluvremenu gostujućeg tima
- HS – Broj udarca domaćeg tima
- HST – Broj udarca u okvir domaćeg tima
- AS - Broj udarca gostujućeg tima
- AST - Broj udarca u okvir gostujućeg tima
- HF – Broj prekršaja domaćeg tima
- AF - Broj prekršaja gostujućeg tima
- AY - Broj žutih kartona gostujućeg tima
- HY - Broj žutih kartona domaćeg tima
- HC – Broj kornera domaćeg tima
- AC – Broj kornera gostujućeg tima
- FTR - Rezultat utakmice

Nakon što su unesene sve potrebne vrijednosti, čvorove je potrebno međusobno povezati opcijom strelice unutar programa. Sada kada je mreža napokon kreirana potrebno je učitati skup podataka u Neticu kako bi se mogla pokrenuti kompilacija odnosno učenje mreže. Kako bi se podaci uspješno učitali potrebno je iz originalnog skupa podataka kreirati *Case* datoteku. *Case* datoteka se kreira tako da se u prvu ćeliju u originalnom skupu podataka unese sljedeća naredba: *// ~->[CASE-1]->~.* Ova naredba je zapravo Netica *Case file* zaglavlje na temelju kojeg Netica zaključuje da je te datoteka *Case file*. Kako bi se pokrenula kompilacija potrebno je odabrati *Cases->Learn->Incorp Case File.* Taj postupak se može vidjeti na slici 34.



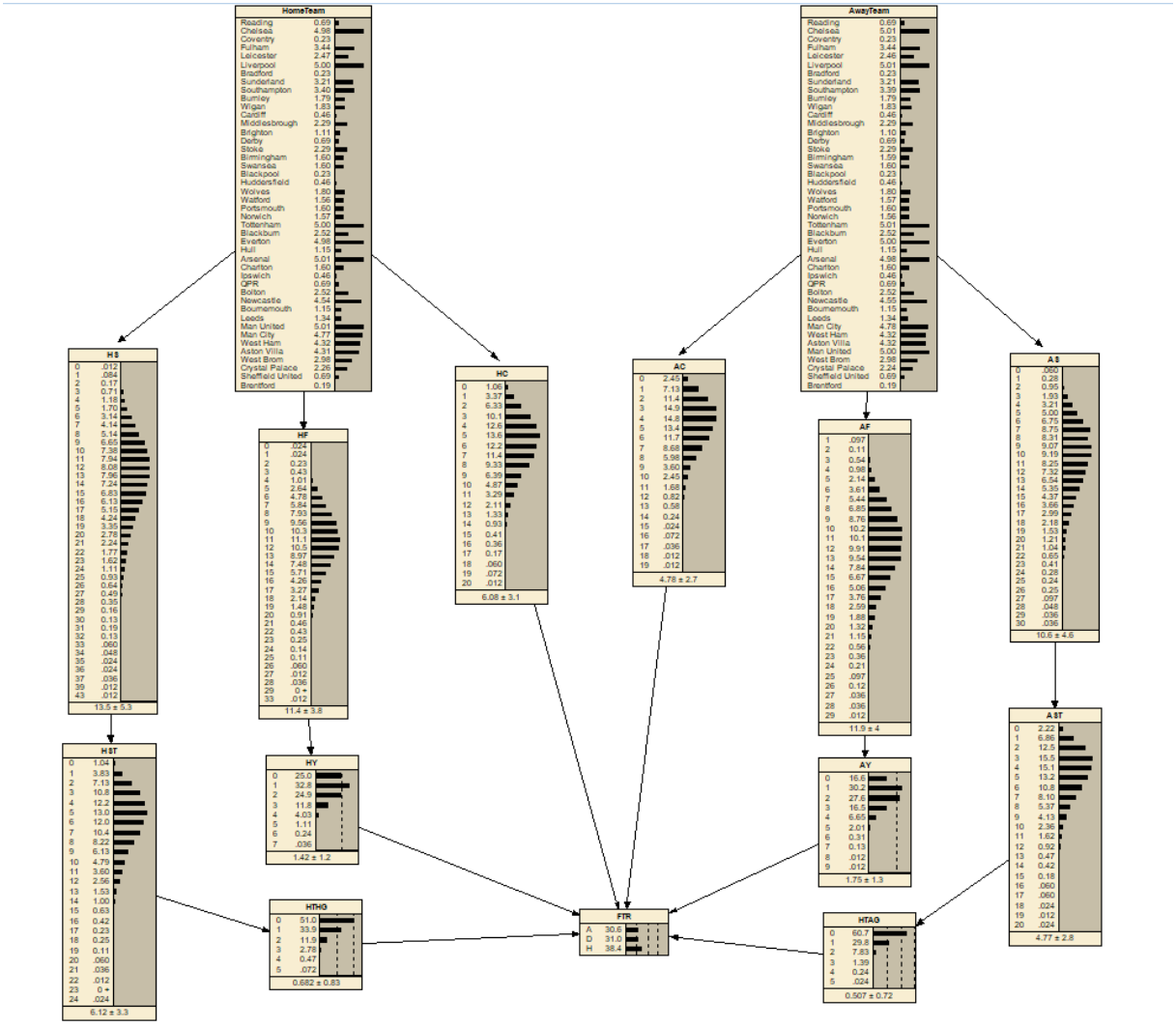
Slika 34: Učitavanje skupa podataka

Kada se uspješno učita skup podataka u izrađeni model potrebno je još mrežu naučiti predviđati iz podataka. To se može napraviti tako da se odabere *Network ->Compile.* Taj postupak se može vidjeti na slici 35.



Slika 35: Učenje mreže

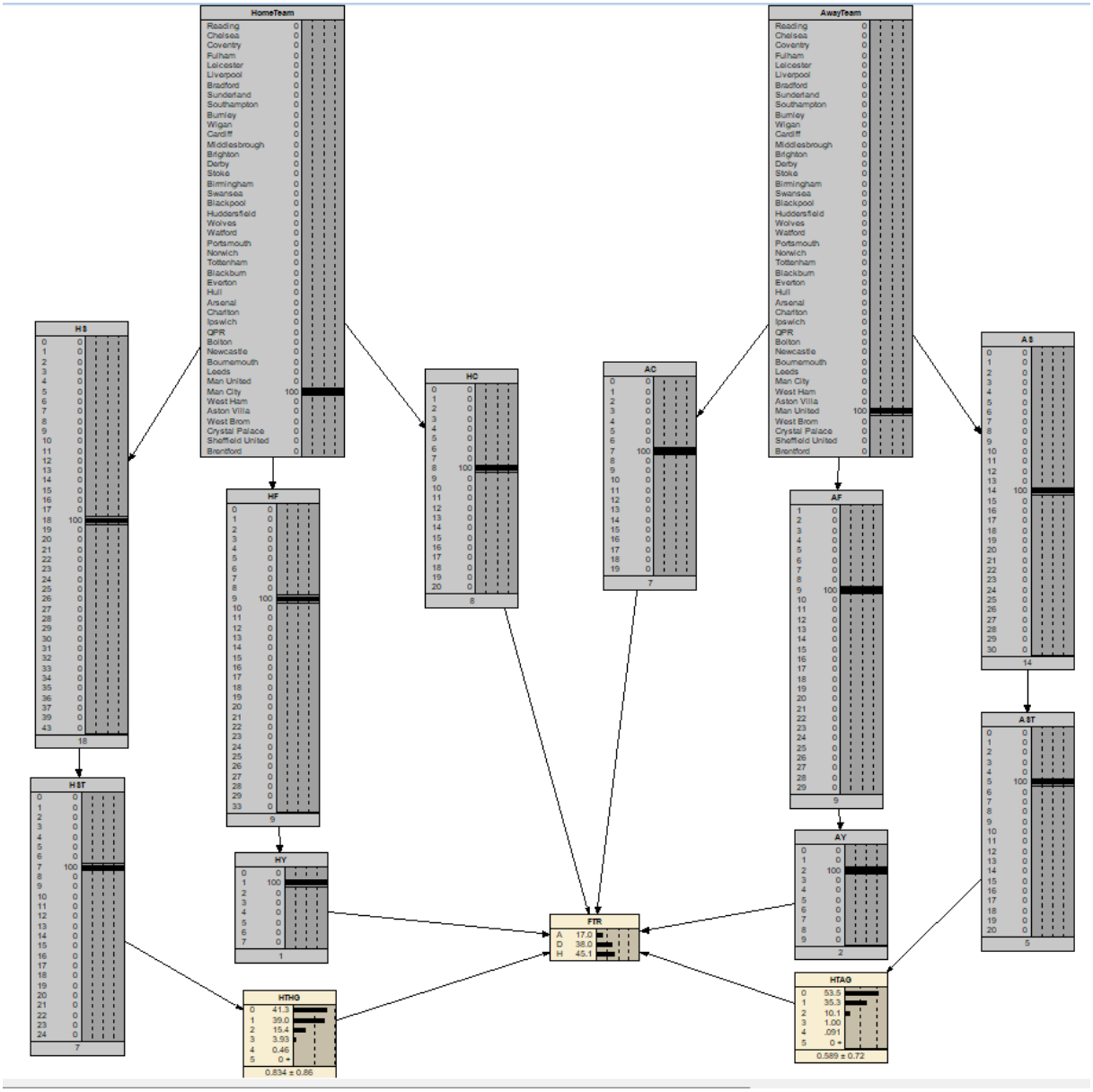
Nakon što su odrađeni svi prijašni koraci dobiva se Bayesova mreža kao na slici 36.



Slika 36: Izgled Bayesove mreže

Ovaj model Bayesovih mreža nam govori kako čvorovi HS, HF i HC ovise o čvoru HomeTeam. Isto tako čvorovi AS, AF i AS ovise o čvoru AwayTeam. Čvor HS utječe na čvor HST koji dalje utječe na čvor HTHG. Ista situacija je i na drugoj strani Bayesove mreže samo za statistike gostujućeg tima. Čvor HF još utječe na čvor HY. Svi ti čvorovi direktno ili indirektno utječu na finalan čvor odnosno na čvor FTR. Na temelju učitanih stvarnih podataka dobiveno je da je najvjerojatniji finalni ishod utakmice pobjeda domaćina (38.4%) zatim neriješeno (31%) te pobjeda gostujućih (30.6%). S druge strane pretpostavka za rezultate na poluvremenu je drugačija gdje se pretpostavlja da je najveća šansa za izjednačen rezultat odnosno vidljivo je da najveći postotak za postignute golove na poluvremenu za obje momčadi nula (51% kod domaćeg tima i 60.7% kod gostujučeg tima).

U nastavku slijedi predikcija rezultata odnosno predviđanje zavisne varijable. Tu će se mijenjati vrijednosti čvorova kako bi se moglo vidjeti kako oni utječu na finalan rezultat odnosno na čvor FTG. Ta predikcija se može vidjeti na slici 37.



Slika 37: Primjer predviđanja pomoću Bayesove mreže

U ovom konkretnom slučaju predviđanja pretpostavilo se da je domaći tim Manchester City dok je gostujući tim Manchester United. Svi statistički podaci u ovom slučaju predviđanja uzeli su se kao prosjek statistika u odigranim utakmicama aktualne sezone. Iz slike 37 se može vidjeti da Bayesove mreže predviđaju da je najvjerojatniji ishod utakmice pobjeda domaćina sa čak 45.1%. Zatim slijedi da će utakmica završiti neriješenim rezultatom u 38% slučajeva. Najmanje šanse se predviđaju za pobjedu gostujućeg tima odnosno imaju samo 17% šansa za pobjedu. Isto tako vidljivo je da je najvjerojatniji rezultat na poluvremenu 0:0 odnosno vjerojatnost da domaćin postigne 0 golova je 41.3% dok je vjerojatnost da gost postigne 0 golova je 53.5%.

Potrebno je još odraditi analizu osjetljivosti za jedan čvor koji nije korijenski. Analiza osjetljivosti pomaže utvrditi kako promjene određenih varijabli utječu na vrijednosti rezultata određenog događaja. U ovom konkretnom slučaju odradila se analiza osjetljivosti za čvor AF odnosno koliko čvor AF utječe na ostale čvorove što se može vidjeti na slici 38. Alat Netica ima ugrađenu opciju za izradu analize osjetljivosti. Kako bi se ona odradila potrebno je odabrati jedan čvor te zatim odabrati *Network -> Sensitivity to Findings.*

Sensitivity of 'AF' to a finding at another node:

Node	Variance Reduction	Percent	Mutual Info	Percent	Variance of Beliefs
AF	15.66	100	4.00914	100	0.8606219
AY	1.942	12.4	0.11040	2.75	0.0005666
AwayTeam	0.7587	4.85	0.12411	3.1	0.0005360
AS	0.01884	0.12	0.00146	0.0363	0.0000054
AST	0.008202	0.0524	0.00051	0.0127	0.0000020
AC	0.004152	0.0265	0.00044	0.011	0.0000015
HTAG	0.0006406	0.00409	0.00004	0.000991	0.0000002
FTR	0.0005684	0.00363	0.00004	0.00102	0.0000001
HTHG	0	0	0.00000	0	0.0000000
HomeTeam	0	0	0.00000	0	0.0000000
HST	0	0	0.00000	0	0.0000000
HY	0	0	0.00000	0	0.0000000
HC	0	0	0.00000	0	0.0000000
HS	0	0	0.00000	0	0.0000000
HF	0	0	0.00000	0	0.0000000

Slika 38: Analiza osjetljivosti čvora AF

Iz analize osjetljivosti može se zaključiti da čvor AF djeluje na sljedeće čvorove:

- AY - 12.4%
- AwayTeam - 4.85%
- AS – 0.12%
- AST – 0.052%
- AC – 0.027%
- HTAG – 0.004%
- FTR – 0.0036%

Vidljivo je da čvor AF najviše utječe na čvor AY i to u postotku od 12.4%. To je i logično jer je čvor AY čvor dijete od čvora AF. Isto tako ako neki tim napravi više prekršaja veća je vjerojatnost da će dobiti i više žutih kartona. S druge strane čvor AF jako malo utječe na ishod konačne utakmice odnosno na čvor FTR.

Zadnji dio koji će biti odrađen je prikaz točnosti predviđanja za Bayesove mreže. Skup podataka sadrži podatke o utakmicama sve do 10.4.2022. pa kako bi se izradila tablica predviđanja odabrati će se 20 utakmica koje su odigrane nakon toga datuma te će se prikazati

točnost predviđanja na tom uzorku. Potrebne informacije biti će prikupljene sa različitih sportskih internetskih izvora. Točnost se računa tako da se broj točnih predviđanja podijeli sa ukupnim brojem predviđanja te pomnoži sa 100 kako bi se dobio postotak. Prikaz točnosti predviđanja može se vidjeti u tablici 2.

Tablica 2: Točnost predviđanja Bayesove mreže

Utakmica	Zapravo	Predviđeno
Totenham - Brighton	A	D
Watford - Brentford	A	A
Southampton - Arsenal	H	H
Manchester Utd - Norwich	H	A
West Ham - Burnley	D	H
Newcastle - Leicester	H	A
Liverpool – Manchester Utd	H	D
Newcastle – Crystal Palace	H	H
Everton - Leicester	D	D
Chelsea - Arsenal	A	A
Manchester City - Brighton	H	H
Burnley - Southampton	H	H
Arsenal – Manchester Utd	H	H
Leicester – Aston Villa	D	D
Manchester City - Watford	H	H
Norwich - Newcastle	A	D
Brentford - Tottenham	D	H
Brighton - Southampton	D	D
Burnley - Wolves	H	H
Chelsea – West Ham	H	D
Prosječna točnost predviđanja	60%	

5. Usporedba dobivenih rezultata

U ovom radu primijenjene su metode stabla odlučivanja, neuronske mreže te Bayesove mreže. Svaka od tih metoda dala je svoje rezultate koji se u nekoj mjeri razlikuju od ostalih rezultata. U nastavku ovog poglavlja dati će se pregled dobivenih rezultata te će se oni međusobno usporediti. Dobiveni rezultati provedenih metoda mogu se vidjeti u tablici 3.

Tablica 3: Prikaz točnosti različitih metoda

	Rezultati ovog istraživanja	Istraživanje autora Hucljak i Rakipović	Rezultati istraživanja autora Razali et al	Istraživanje autora Ulmer i Fernandez	Istraživanje autora Joseph et al
Stablo odlučivanja	63%	53.1%	-	50%	76.3%
Neuronske mreže	62.4%	68.8%	-	-	-
Bayesove mreže	60%	53.1%	75.09%	-	61.8%

Iz tablice 3 je vidljivo da u ovom provedenom istraživanju najbolje rezultate predviđanja daje stablo odlučivanja sa ukupnom točnosti od 63%. Neznatno lošije rezultate daje neuronska mreža koja ima točnost od 62.4% što je lošije od stabla odlučivanja za samo 0.6%. Najlošije rezultate daje Bayesova mreža koja ima točnost od 60%. U usporedbi sa prethodnim istraživanjima može se vidjeti da stablo odlučivanja u prosjeku daje bolje rezultate odnosno dobiveni su bolji rezultati nego u dva od tri prethodna istraživanja. Isto tako vidljivo je da neuronske mreže imaju nešto lošiju točnost od prethodnog istraživanja i to za 6.4%. Bayesove mreže su u prosjeku imale nešto lošije performanse odnosno točnost. U dva od tri prethodna odlučivanja Bayesove mreže su dale bolje rezultate. Ako se gleda prosjek svih istraživanja navedenih u ovom radu tada najbolje performanse imaju neuronske mreže sa 65.6%. Nakon njih slijede Bayesove mreže sa 62.5% te stablo odlučivanja sa 60.6%. Može se zaključiti da svaka od metoda daje drugačije rezultate na različitim skupovima podataka odnosno neće uvijek ista metoda dati najbolje rezultate na različitim skupovima podataka.

6. Zaključak

Glavna tema ovog rada bio je razvoj prediktivnih modela primjenom strojnog učenja na temelju nogometnih podataka te njihova usporedba. Na samom početku rada prikazano je nekoliko prethodnih sličnih istraživanja koja su koristile različite metode te dobile različite rezultate. Skup podataka koji je korišten u radu preuzet je sa internetske stranice *Kaggle* koja sadrži veliki broj skupova podataka na različite teme. Prije same izrade prediktivnih modela bilo je potrebno urediti skup podataka kako bi bio što relevantniji i kako bi izrađeni modeli pružili što je moguće bolje rezultate. To uređivanje odrađeno je pomoću alata SSMS.

Kroz rad korištene su tri metode strojnog učenja: stablo odlučivanja, neuronske mreže i Bayesove mreže. Svaka od navedenih metoda se razlikuje od ostalih, ali su krajnji rezultati svih metoda bili slični odnosno dale su približno iste rezultate. Postotak točnog predviđanja i kod stabla odlučivanja i kod neuronskih mreža iznosi oko 63% dok je kod Bayesovih mreža on nešto niži i iznosi 60%. Uzimajući u obzir prethodna istraživanja, skup podataka i samu temu rada, rezultati su očekivani. Naravno rezultati bi se mogli i poboljšati kada bi se skup podataka nadopunio i proširio dodatnim informacijama kao što su posjed lopte, zamjene i sl. Isto tako mogli bi se i dodati podaci o svakom pojedinom igraču. Tada bi te metode zasigurno dale bolje i preciznije rezultate, ali bi sam skup podataka bio puno veći te bi i izrada istraživanja i modela bila daleko kompliciranija.

Kroz rad korištena su dva alata: BigML te Netica. BigML je alat koji se koristio za izradu stabla odlučivanja i neuronske mreže dok se alat Netica koristila za izradu Bayesove mreže. Oba alata su vrlo jednostavna za korištenje te su lagana za shvatiti, a pružaju jednostavan način za izradu prediktivnih modela odnosno za implementaciju strojnog učenja.

Popis literature

Agrawal S., Pal Singh S., Kumar Sharma J.,(2018). *Predicting Results of Indian Premier League T-20 Matches using Machine Learning.* Preuzeto 6.4.2022. s https://www.researchgate.net/publication/335572825_Predicting_Results_of_Indian_Premier_League_T-20_Matches_using_Machine_Learning

BigML (bez dat.). *About BigML*. Preuzeto 12.4.2022 s https://bigml.com/about/

Bunker R.P., Thabtah F., (2019). *A machine learning framework for sport result prediction*. Preuzeto 14.4.2022 s https://www.sciencedirect.com/science/article/pii/S2210832717301485

Dalbelo Bašić B., Čupić M., Šnajder J., (2012). *Umjetne neuronske mreže*. Preuzeto 14.4.2022 s http://degiorgi.math.hr/~singer/ui/ui_1415/UI_12_UmjetneNeuronskeMreze%5b1%5d.pdf

Data School (2014). *Simple guide to confusion matrix terminology*. Preuzeto 13.4.2022 s https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

Dumančić S., (2014). *Neuronske mreže*. Preuzeto 14.4.2022 s http://www.mathos.unios.hr/~mdjumic/uploads/diplomski/DUM05.pdf

Efzg (2011). *Stablo odlučivanja*. Preuzeto 12.4.2022 s https://www.efzg.unizg.hr/UserDocsImages/OIM/mdarabos/4-Stablo%20odlu%C4%8Divanja.pdf

Hucaljuk J., Rakipović A., (2011). *Predicting football scores using machine learning techniques.* Preuzeto 6.4.2022 s Predicting football scores using machine learning techniques | IEEE Conference Publication | IEEE Xplore

Joseph A., Fenton N.E., Neil M., (2006). *Predicting football results using Bayesian nets and other machine learning techniques.* Preuzeto 7.4.2022 s Predicting football results using Bayesian nets and other machine learning techniques - ScienceDirect

Kliček B. (2021). *Znanje i strojno učenje*. Inteligentni sustavi [Moodle]. Sveučilište u Zagrebu, Fakultet organizacije i informatike, Varaždin.

Lale S., Ćavar S., Lale D., Krile S., (2014). *Primjena Bayesove mreže u analizama rizika u*

pomorstvu. Preuzeto 20.4.2022. s https://hrcak.srce.hr/file/216417

Norsys (bez dat.). *Netica Application*. Preuizeto 21.4.2022 s https://www.norsys.com/netica.html

Oreški D. (2021). *Algoritmi strojnog učenja.* Inteligentni sustavi [Moodle]. Sveučilište u Zagrebu, Fakultet organizacije i informatike, Varaždin.

Prcela M., (2010). *Predstavljanje znanja zasnovano na integraciji ontologija i Bayesovih mreža*. Preuzeto 20.4.2022. s http://lis.irb.hr/MLAA/prcela-doktorska-disertacija.pdf

Preraić I., (2012). *Bayesove mreže u modeliranju učenika*. Preuzeto 20.4.2022. s https://mapmf.pmfst.unist.hr/~ani/radovi/diplomski/Peraic_Ivan_2012.pdf

Razali N., Mustapha A., Yatim F.A., Aziz R.A., (2017). *Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)*. Preuzeto 7.4.2022 s https://iopscience.iop.org/article/10.1088/1757-899X/226/1/012099

Song Y., Lu Y., (2015), *Decision tree methods:applications for classification and Prediction.* Preuzeto 12.4.2022 s https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/

Ulmer B., Fernandez M., (2014). *Predicting Soccer Match Results in the English Premier League*. Preuzeto 6.4.2022 s Ben Ulmer, Matt Fernandez, Predicting Soccer Results in the English Premier League.pdf (stanford.edu)

Zekić-Sušac M. (bez dat.). *Stabla odlučivanja*. Preuzeto 12.4.2022 s http://www.efos.unios.hr/poslovni-informacijski-sustavi/wp-content/uploads/sites/192/2017/10/P4_Stabla-odlucivanja-2017.pdf

Zekić-Sušac M., Frajman-Jakšić A., (2008). *Neuronske mreže i stabla odlučivanja za predviđanje uspješnosti studiranja*. Preuzeto 14.4.2022 s https://hrcak.srce.hr/file/73924

Popis slika

Popis tablica