Upotreba metoda nadziranog strojnog učenja u predviđanju sportskih ishoda te vizualizacija podataka

Vujčić, Mateo

Sveučilište
Sjever

Završni rad br. 2/RINF/2024

Upotreba metoda nadziranog strojnog učenja u predviđanju sportskih ishoda te vizualizacija podataka

Mateo Vujčić, 0314022129

Đurđevac, lipanj 2024. godine

Sveučilište
Sjever

Računarstvo i informatika

Završni rad br. 2/RINF/2024

Upotreba metoda nadziranog strojnog učenja u predviđanju sportskih ishoda te vizualizacija podataka

Student

Mateo Vujčić, 0314022129

Mentor

Doc.dr.sc. Tomislav Horvat

Đurđevac, lipanj 2024. godine

Sveučilište Sjever
Sveučilišni centar Varaždin
104. brigade 3, HR-42000 Varaždin

Prijava završnog rada

Definiranje teme završnog rada i povjerenstva

ODJEL: Odjel za računarstvo i informatiku

STUDIJ: Prijediplomski stručni studij Računarstvo i informatika

PRISTUPNIK: Mateo Vujčić | MATIČNI BROJ: 0314022129

DATUM: 4.4.2024. | KOLEGIJ: Strojno učenje i umjetna inteligencija

NASLOV RADA: Upotreba metoda nadziranog strojnog učenja u predviđanju sportskih ishoda te vizualizacija podataka

NASLOV RADA NA ENGL. JEZIKU: Using of supervised machine learning methods in sport outcome prediction and data visualization

MENTOR: Tomislav Horvat | ZVANJE: docent

ČLANOVI POVJERENSTVA:
1. mr.sc. Vladimir Stanisavljević, v. pred. - predsjednik
2. Dražen Crčić, predavač - član
3. doc.dr.sc. Tomislav Horvat - mentor
4. doc.dr.sc. Domagoj Frank - zamjenski član
5.

Zadatak završnog rada

BROJ: 2/RINF/2024

OPIS:

U današnjem digitalnom okruženju, sposobnost predviđanja ishoda postaje ključna u gotovo svim sferama života. Bilo da se radi o ekonomskim odlukama, zdravstvenim dijagnozama ili sportskim strategijama, sposobnost anticipacije budućnosti pruža neprocjenjivu prednost. Međutim, ostvarivanje preciznih predviđanja zahtijeva duboko razumijevanje i pravilnu interpretaciju relevantnih podataka koji mogu dolaziti s različitih izvora, ali i u različitim oblicima.

U radu je potrebno:
- istražiti područje primjene metoda nadziranog strojnog učenja u predviđanju ishoda u sportu s posebnim naglaskom na temu samog rada
- definirati i opisati pojmove znanosti o podacima, s posebnim naglaskom na strojno učenje
- proučiti različite tehnike i algoritme nadziranog strojnog učenja koji se koriste u predviđanju ishoda.
- razviti model strojnog učenja za predviđanje ishoda temeljen na prikupljenim podacima
- implementirati vizualizaciju podataka radi boljeg razumijevanja rezultata
- provjeriti pouzdanost i točnost predviđanja modela kroz evaluacijske mjere te usporediti rezultate s rezultatima ostalih istraživača
- razmotriti mogućnosti optimizacije i daljnjeg poboljšanja razvijenog modela

ZADATAK URUČEN: 22.5.2024 | POTPIS MENTORA:

SVEUČILIŠTE SJEVER

Predgovor

Ovim riječima želim izraziti duboku zahvalnost svome cijenjenom mentoru, doc. dr. sc. Tomislavu Horvatu, čija su me stručna uputstva, dragocjeni savjeti, inovativne ideje i neizmjerna podrška tijekom izrade završnog rada neprestano vodila i inspirirala. Njegovo mentorstvo omogućilo mi je stjecanje vrijednih znanja i motiviralo me za daljnje usavršavanje u području strojnog učenja. Zahvaljujući njemu, razvio sam strast prema istraživačkom radu i inženjerskom načinu razmišljanja, a ovaj rad postavlja čvrst temelj za moj budući profesionalni put. Također, iskrenu zahvalnost upućujem i svim profesorima Sveučilišta Sjever koji su svojim predanim radom, ekspertizom i entuzijazmom znatno obogatili moje obrazovanje. Njihova spremnost na dijeljenje znanja bila je neprocjenjiva.

Neizmjerno sam zahvalan svojoj obitelji i prijateljima za njihovu bezuvjetnu podršku, razumijevanje i vjeru u mene. Njihova podrška bila je ključna u trenucima kada je bilo najteže, pružajući mi snagu i ohrabrenje na svakom koraku.

Zahvaljujem se i svojim kolegama s kojima sam dijelio kako učionice, tako i ideje, tijekom naših studijskih godina. Vaša suradnja, prijateljstvo i kolegijalnost učinili su moje studentske dane izuzetno vrijednima i nezaboravnima.

Nadam se da će ovaj rad predstavljati vrijedan doprinos u području strojnog učenja i poslužiti kao izvor inspiracije budućim generacijama studenata koji teže izvrsnosti u ovom dinamičnom i uzbudljivom polju.

Sažetak

Završni rad istražuje primjenu nadziranog strojnog učenja u predviđanju ishoda nogometnih utakmica u engleskoj Premier ligi, a posebno poglavlje bavi se i vizualizacijom podataka. U uvodu se razmatraju značaj i primjenjivost strojnog učenja u sportskoj analizi. Detaljno se opisuju metode i algoritmi nadziranog učenja, kao i alati za obradu i vizualizaciju podataka, koji omogućavaju dublje razumijevanje dinamike utakmica.

Kroz različite faze analize predviđanja, od prikupljanja podataka do evaluacije modela, rad demonstrira kako tehnike strojnog učenja mogu identificirati uzorke i predviđati sportske ishode. Zaključno, rad naglašava potencijal strojnog učenja u sportu, uzimajući u obzir i izazove kao što su prilagodba modela i interpretacija podataka.

U kontekstu šire primjene, rad razmatra kako bi istraživanje moglo biti prošireno na druge sportske lige i disciplinu, naglašavajući značaj interdisciplinarnog pristupa i daljnjeg razvoja analitičkih metoda. Ovo istraživanje potvrđuje kako napredne analitičke metode mogu poboljšati predviđanja i donošenje odluka u sportu, te kako strojno učenje može služiti kao temelj za razvoj efikasnijih strategija u sportskom sektoru.

Ključne riječi: algoritmi strojnog učenja, klasifikacija podataka, nadzirano učenje, nenadzirano učenje, podaci o sportskim rezultatima, regresijska analiza, sportsko prognoziranje, učenje uz podršku, vizualizacija podataka

Summary

The thesis explores the application of supervised machine learning in predicting the outcomes of football matches in the English Premier League, with an emphasis on data visualization. The introduction discusses the significance and applicability of machine learning in sports analytics. It thoroughly describes the methods and algorithms of supervised learning, as well as tools for data processing and visualization, which enable a deeper understanding of match dynamics.

Through various phases of predictive analysis, from data collection to model evaluation, the work demonstrates how machine learning techniques can identify patterns and predict sports outcomes. In conclusion, the paper emphasizes the potential of machine learning in sports, considering challenges such as model adaptation and data interpretation.

In the context of broader application, the work considers how the research could be extended to other sports leagues and disciplines, highlighting the importance of an interdisciplinary approach and further development of analytical methods. This research confirms how advanced analytical methods can improve predictions and decision-making in sports, and how machine learning can serve as a foundation for developing more effective strategies in the sports sector.

Keywords: supervised learning, unsupervised learning, reinforcement learning, sports forecasting, data visualization, machine learning algorithms, regression analysis, data classification, sports result data.

Sadržaj

1. Uvod ... 1

1.1. Pregled istraživanja vezanih uz predviđanje ishoda u sportu ... 2
1.1.1. Odabir i ekstrakcija značajki ... 5
1.1.2. Evaluacija rezultata ostalih istraživača u nogometu ... 6

2. Znanost o podacima i strojno učenje ... 9

2.1. Definicija znanosti o podacima ... 9
2.1.1. Uloga znanosti o podacima ... 10
2.1.2. Znanost o podacima u predviđanju ishoda ... 11
2.2. Strojno učenje u znanosti o podacima ... 11
2.2.1. Nadzirano učenje ... 12
2.2.2. Nenadzirano učenje ... 14
2.2.3. Učenje uz podršku ... 15
2.2.4. Algoritmi za klasifikaciju ... 16
2.2.5. Algoritmi za regresiju ... 17
2.3. Metode evaluacije u strojnom učenju ... 19
2.3.1. Podjela skupa podataka ... 20
2.3.2. Unakrsna validacija ... 20
2.3.3. Drugi relevantni parametri evaluacije ... 21

3. Prikupljanje podataka ... 22

3.1. Izvor podataka ... 22
3.2. Alati za prikupljanje podataka s weba ... 23
3.3. Etika i pravila u prikupljanju podataka ... 24
3.4. Preuzimanje podataka o utakmicama engleske Premier lige ... 24
3.5. Programski kod za preuzimanja i pohranu podataka s weba ... 25

4. Predviđanje ishoda ... 28

4.1. Alati za predviđanje ishoda ... 28
4.2. Koraci i programski kod ... 30

5. Vizualizacija podataka ... 34

5.1. Alati za vizualizaciju podataka ... 34
5.2. Izrada vizualizacija i programski kod ... 35
5.3. Interpretacija rezultata ... 38
5.4. Rasprava ... 40

6. Zaključak ... 44
7. Literatura ... 45

1. Uvod

U suvremenom digitalnom okruženju, sposobnost predviđanja budućih ishoda postaje sve važnija u gotovo svim aspektima života. Bilo da je riječ o donošenju ekonomskih odluka, postavljanju dijagnoza u zdravstvu ili osmišljavanju sportskih strategija, prednost koju pruža anticipacija budućnosti je neprocjenjiva. Međutim, postizanje preciznih predviđanja iziskuje temeljito razumijevanje i adekvatnu interpretaciju relevantnih podataka. Ovaj uvod posvetit će se razmatranju značaja predviđanja ishoda, uz naglasak na neophodnost pristupa relevantnim podacima te na različite vrste podataka, strukturirane, polustrukturirane i nestrukturirane.

Sposobnost predviđanja ishoda nalazi primjenu u raznolikim segmentima života. U poslovnom sektoru predviđanja omogućavaju kompanijama da donose strateške odluke, prilagođavaju se tržišnim uvjetima i minimiziraju rizike. U medicinskom kontekstu točna predviđanja doprinose pravodobnim dijagnozama i personaliziranom pristupu liječenju, znatno poboljšavajući kvalitetu zdravstvene skrbi. U području sporta analitička obrada podataka ključna je za taktičko planiranje i optimizaciju učinkovitosti. Ističe se općeniti značaj predviđanja ishoda te ključna uloga relevantnih podataka u ostvarivanju preciznih predviđanja.

Neovisno o području primjene, pristup relevantnim podacima presudan je za postizanje optimalnih rezultata. Navedeni podaci omogućuju dublje razumijevanje trenutnih stanja i temelj su za informirano donošenje odluka. U poslovnom svijetu relevantni podaci uključuju informacije o potrošačkim navikama, u medicini podatke o biološkim i genetskim markerima, dok u sportu analiza performansi momčadi ima ključnu ulogu. Stoga, uspjeh u predviđanju ishoda ovisi o učinkovitom prikupljanju, analizi i interpretaciji relevantnih podataka.

Raznolikost podataka otvara izazove i pruža prilike za njihovo pravilno prikupljanje, analizu i tumačenje. Ključno je razumijevanje različitih tipova podataka kako bi se razvile učinkovite metode za predviđanje ishoda. U daljnjem tekstu detaljno se razmatraju karakteristike i primjene strukturiranih, polustrukturiranih i nestrukturiranih podataka. Strukturirani podaci su organizirane informacije pohranjene u definiranim strukturama poput tablica ili baza podataka, s jasno određenim atributima i odnosima među podacima. Oni omogućuju lako pretraživanje, filtriranje i analizu, čineći osnovu za kvantitativne i statističke analize u poslovanju i znanosti.

Polustrukturirani podaci, iako dijelom organizirani, ne posjeduju strogo definirane atribute ili odnose. Formati poput XML-a, JSON-a i YAML-a služe za predstavljanje ovih podataka, pružajući veću fleksibilnost u odnosu na strogo strukturirane podatke, ali zahtijevajući složenije metode analize. Nestrukturirani podaci pak obuhvaćaju informacije bez

jasne organizacije, poput tekstualnih dokumenata, audiozapisa, slika i videozapisa. Navedeni podaci ne slijede unaprijed definirane obrasce, što zahtijeva primjenu naprednih tehnika obrade za izvlačenje relevantnih informacija.

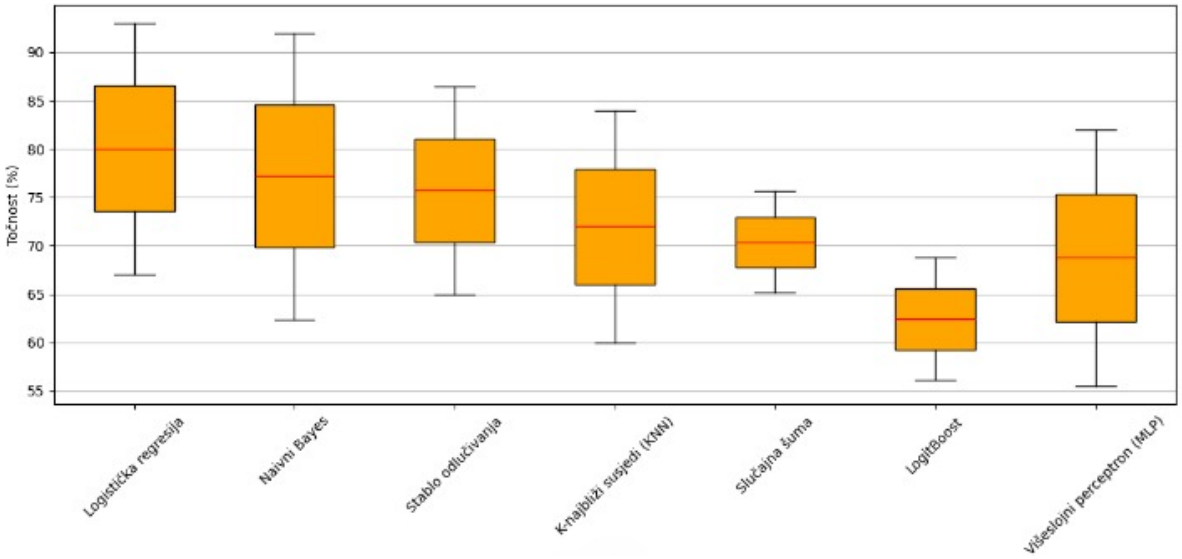
Raznolikost tipova podataka pruža obilje informacija, ali isto tako predstavlja izazove u njihovoj obradi. U kontekstu predviđanja ishoda, kombinacija različitih vrsta podataka često se koristi za temeljitu analizu i dublje uvide. U nastavku rada istražit će se kako navedeni tipovi podataka doprinose procesima predviđanja ishoda, posebice kroz prizmu znanosti o podacima i strojnog učenja.

1.1. Pregled istraživanja vezanih uz predviđanje ishoda u sportu

Tablica 1 u sklopu istraživanja prikazuje analizirane radove drugih istraživača koji se bave predviđanjem ishoda sportskih događaja iz kojih su izvučeni određeni zaključci. Autori su otkrili da istraživači najčešće koriste klasifikacijske metode nadziranog strojnog učenja. Međutim, glavni problem je što različiti radovi koriste različite skupove podataka, čime je usporedba rezultata otežana. Najčešće korištena metoda validacije je podjela skupa podataka, dok manji broj istraživača koristi unakrsnu provjeru. No, zbog povezanosti sportskih događaja, unakrsna provjera nije uvijek pogodna za sportska predviđanja. Najbolji rezultati se obično postižu korištenjem manjih skupova podataka. Pregledom radova drugih istraživača uočeno je kako se za predviđanje ishoda nogometnih utakmica koriste različiti algoritmi strojnog učenja. Analizirano je ukupno 16 radova iz perioda od 2000. do 2024. godine koji se bave nogometom, a među najčešće proučavanim ligama dominira Liga prvaka s udjelom od 55,56% u istraživanjima. Algoritmi uključeni u istraživanje su logistička regresija (engl. *logistic regression*), Naivni Bayes (engl. *Naive Bayes*), stablo odluke (engl. *decision trees*), k-najbližih susjeda (engl. *k-nearest neighbours*), slučajna šuma (engl. *random forest*), LogitBoost i višeslojni perceptron (engl. *multilayer perceptron*, kraće MLP). Svaki algoritam je testiran na skupu podataka koji obuhvaća povijesne rezultate utakmica, statistike momčadi te faktore domaćih i gostujućih utakmica. Većina ovih radova koristi više algoritama strojnog učenja, što znači da analizirani uzorak obuhvaća više od 50 različitih rezultata predviđanja sportskih ishoda. Međutim, usporedba tih rezultata je izazovna ili gotovo nemoguća zbog korištenja različitih skupova podataka i liga različitih konkurentnosti. Kada autori koriste različite skupove značajki ili podatkovne skupove, razmatraju se samo najbolji rezultati. Slika 1 prikazuje kutijasti dijagram (engl. *box plot*) raspona dobivenih rezultata analiziranih radova u predviđanju ishoda nogometnih utakmica.

Tablica 1. Analizirani radovi poredani po godini objave.

#	Godina objave	Skup podataka	Korišteni algoritmi	Najbolji rezultat	Korištene značajke
[1]	2008	English Premier League EPL, sezone 2002 – 2007	neuronska mreža	58,40%	Povijesni rezultati, statistike igrača, vremenski uvjeti, domaći teren
[2]	2011	Zadnjih 15 sezona nizozemskih nogometnih liga	linearna regresija	57,00%	Povijesni rezultati, trenutna forma momčadi, statistike igrača
[3]	2011	Liga prvaka	neuronska mreža, LogitBoost	68,80%	Statistike igrača, podaci o momčadi, povijesni rezultati
[4]	2013	Španjolska nogometna liga (Primera), sezona 2008	Bayesova mreža	92,00%	Statistike igrača, podaci o momčadi, povijesni rezultati
[5]	2014	EPL, sezona 2014	logistička regresija	93,00%	Povijesni rezultati, trenutna forma momčadi, statistike igrača
[6]	2015	Dutch Eredivisie, sezone 2000 – 2012	LogitBoost	56,10%	Javne baze podataka, povijesni rezultati, statistike igrača
[7]	2016	EPL, sezone 2010 – 2015	logistička regresija	69,50%	Statistike igrača, podaci o momčadi, povijesni rezultati
[8]	2018	Španjolska nogometna liga, sezone 2012 – 2016	logistička regresija	71,63%	Povijesni rezultati, trenutna forma momčadi, statistike igrača
[9]	2020	EPL, Ligue 1, Bundesliga, Seria A, Primera Division, sezone 2013 – 2017	slučajna šuma	75,62%	Statistike igrača, podaci o momčadi, povijesni rezultati
[10]	2021	5 europskih nogometnih liga i pripadajuće druge lige, sezone 2006 – 2017	integrirani model	81,77%	Karakteristike igrača, podaci o momčadi, povijesni rezultati
[11]	2021	UEFA liga prvaka, sezona 2016 - 2017	logistička regresija	81,00%	Povijesni rezultati, trenutna forma momčadi, statistike igrača
[12]	2021	Engleska Premier liga, sezona 2005 - 2006	šuma odluke, neuronska mreža	88,00%	Povijesni rezultati, trenutna forma momčadi, statistike igrača
[13]	2022	Lige petice sezone 2014-2017	Bayesova mreža	92,01%	Povijesni rezultati, trenutna forma momčadi, statistike igrača
[14]	2022	Engleska Premier Liga, sezone 2018-2019	SVM, Bayesova mreža	61,32%	Povijesni rezultati, trenutna forma momčadi, statistike igrača
[15]	2022	Engleska Premier Liga, sezone 2019-2021	logistička regresija, KNN, slučajna šuma	70,00%	Povijesni rezultati, trenutna forma momčadi, ELO rating, napadačke i obrambene sposobnosti, koeficijenti kladionica, domaći teren
[16]	2024	Svijetsko prvenstvo 2022	slučajna šuma	69,20%	Povijesni podaci o utakmicama, rangiranje momčadi, napadačke i obrambene sposobnosti momčadi



Slika 1. Kutijasti dijagram analize rezultata ostalih istraživača.

Kutijasti dijagram pruža detaljan uvid u točnost različitih algoritama strojnog učenja u predviđanju nogometnih rezultata. Analizirani algoritmi uključuju logističku regresiju (engl. *logistic regression*), Naivni Bayes(engl. *Naive Bayes*), stabla odluke (engl. *decision trees*), *k*-najbližih susjeda (engl. *k-nearest neighbours*), slučajnu šumu (engl. *random forest*), LogitBoost i višeslojni perceptron (engl. *multilayer perceptron*, kraće MLP). Rezultati pokazuju značajne razlike u točnosti među algoritmima. Logistička regresija ističe se širokim rasponom točnosti, što ukazuje na njezinu prilagodljivost različitim skupovima podataka i metodama obrade. Naivni Bayes pokazuje ujednačene i stabilne rezultate, dok stabla odluke i *k*-najbližih susjeda imaju širi raspon točnosti, što sugerira da njihova izvedba može ovisiti o specifičnim uvjetima i značajkama podataka. Slučajna šuma često postiže visoku točnost, čineći je pouzdanim izborom za predviđanje nogometnih ishoda. LogitBoost također pokazuje dobre rezultate, ali ne doseže najviše razine točnosti kao neki drugi algoritmi. Višeslojni perceptron ima konzistentnu točnost, što može ukazivati na potrebu za dodatnim finim podešavanjem modela ili prilagođavanjem značajki.

Dobiveni rezultati naglašavaju važnost pažljivog odabira algoritama i metoda obrade podataka kako bi se postigli optimalni rezultati u predviđanju sportskih ishoda. Korištenje odgovarajućih značajki i prilagodba modela ključni su za postizanje visoke točnosti i pouzdanosti predviđanja. Analiza rezultata također pokazuje da iako neki algoritmi mogu postići visoku točnost u specifičnim uvjetima, konzistentnost i prilagodljivost različitim skupovima podataka često određuju njihov ukupni uspjeh u praksi. Kao što je već istaknuto, svakodnevno se generira velika količina podataka, bilo strukturiranih ili nestrukturiranih, vezanih uz sportske događaje. Uz rast količine podataka, raste i broj relevantnih baza podataka

koje sadrže razne sportske statistike. Pregledavajući radove drugih istraživača, jasno je da su glavni izvori informacija uglavnom službene stranice sportskih organizacija. U radu je analizirano više studija koje se bave predviđanjem sportskih ishoda li izvlačenjem korisnih informacija i pravilnosti vezanih uz sport. Nadalje, kako se povećava dostupnost i volumen sportskih podataka, tako raste i interes za korištenje istih podataka za analitičke svrhe i predviđanje budućih ishoda. Sportske organizacije često pružaju detaljne statistike i povijesne podatke putem svojih službenih kanala, što istraživačima omogućuje pristup vrijednim informacijama za modeliranje i analizu. Također, mnogi radovi koriste te izvore podataka za razvijanje i testiranje različitih algoritama strojnog učenja, nastojeći unaprijediti točnost predviđanja u sportu. Postoji i velik broj drugih istraživanja, znanstvenih i preglednih radova, ne striktno povezanih s predviđanjem ishoda u nogometu, koja daju vrijedne rezultate i zaključke u predviđanju ishoda u sportu [17] [18] [19] [20] [21] [22] [23] [24] [25].

1.1.1. Odabir i ekstrakcija značajki

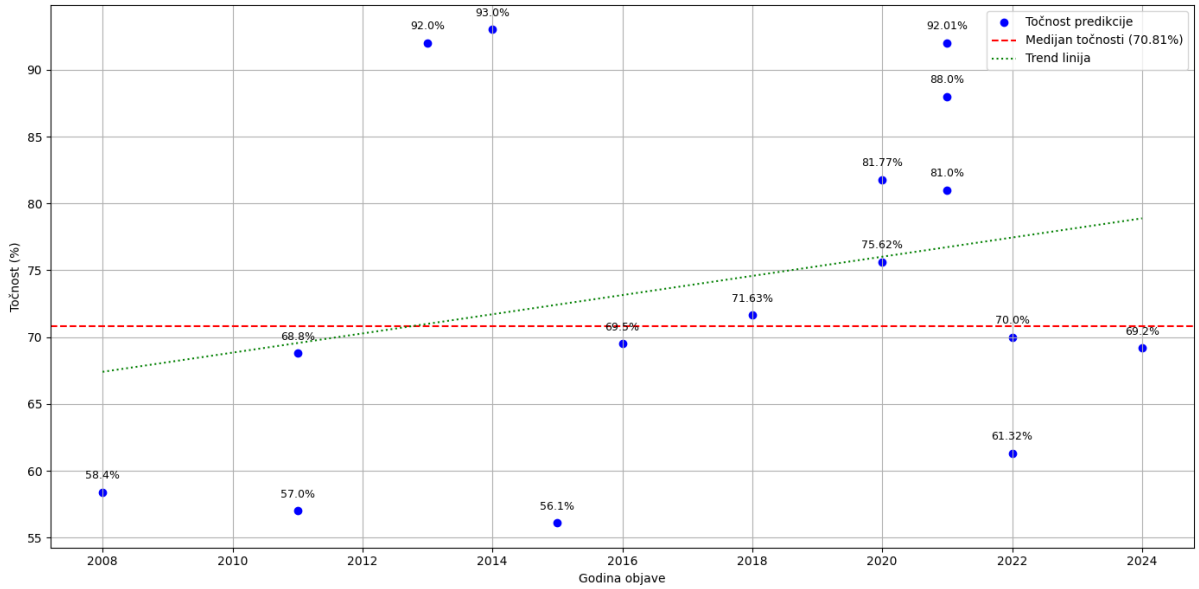
Odabir i ekstrakcija značajki ključni su koraci u analizi nogometnih podataka, jer značajno utječu na uspjeh algoritama za predviđanje. Proces uključuje prepoznavanje i uklanjanje nebitnih i suvišnih značajki kako bi se smanjila složenost problema i omogućilo brže i učinkovitije djelovanje algoritama strojnog učenja. U nogometu, odabir značajki obično započinje prikupljanjem raznih statističkih podataka poput broja golova, asistencija, posjeda lopte, broja udaraca, obrana vratara, prekršaja i kartona. Osim osnovnih statistika, mogu se koristiti i napredne metrike kao što su očekivani golovi, prosječna pretrčana udaljenost po igraču, i performanse tijekom domaćih i gostujućih utakmica. Jedna od čestih metoda za odabir značajki je korištenje filter metoda, koje karakteriziraju rangiraju značajke na temelju njihovog odnosa s ciljanom varijablom, bez obzira na modeliranje. Metode omotača koriste algoritam strojnog učenja kako bi ocijenile i odabrale značajke koje daju najbolje rezultate. Ugrađene metode, s druge strane, kombiniraju odabir značajki s procesom učenja modela, kao što je slučaj s nekim varijantama stabala odluke i linearnih modela. Nakon odabira značajki, ekstrakcija značajki uključuje transformaciju ili kombinaciju postojećih podataka kako bi se stvorile nove značajke koji bolje reprezentiraju informacije sadržane u podacima. Primjerice, kombiniranjem broja udaraca i posjeda lopte može se dobiti nova značajka koja opisuje učinkovitost napada momčadi. Analiza radova pokazuje da je uporaba metoda odabira i ekstrakcije značajki ključna za poboljšanje performansi modela predviđanja u nogometu.

Uvođenjem dodatnih značajki, kao što su psihološko stanje igrača ili sentiment analiza objava na društvenim mrežama, moguće je dodatno povećati točnost predviđanja. Optimalan skup značajki može značajno poboljšati rezultate predviđanja i omogućiti bolje razumijevanje dinamike nogometne igre.

1.1.2. Evaluacija rezultata ostalih istraživača u nogometu

Evaluacija rezultata istraživanja drugih autora u nogometu od ključne je važnosti za razumijevanje učinkovitosti različitih metoda predviđanja. Pregled radova pokazuje varijacije u rezultatima zbog različitih pristupa i metodologija. Rezultati pokazuju da su algoritmi poput logističke regresije, LogitBoost i MLP postigli najvišu prosječnu točnost, s vrijednostima oko 65%. S druge strane, drugo istraživanje pokazuje znatno više maksimalne točnosti za iste algoritme. Logistička regresija u drugom istraživanju postiže maksimalnu točnost od 93%, dok Bayes Network postiže 92%. Ove razlike u rezultatima mogu biti uzrokovane različitim skupovima podataka koji se koriste u istraživanjima. Na primjer, jedno istraživanje može biti fokusirano isključivo na englesku Premier ligu, dok drugo uključuje širi spektar sportskih događaja iz različitih liga i natjecanja. Također, metodologija obrade podataka može značajno utjecati na performanse modela. Različite metode odabira i obrade podataka, kao i različite značajke korištene u analizama, mogu dovesti do značajnih varijacija u točnosti predviđanja. Pored toga, korištenje različitih metoda validacije, kao što su unakrsna validacija (engl. *cross-validation*) ili jednostavna podjela podataka na trening i testne skupove, može utjecati na procjenu točnosti modela. Unakrsna validacija obično daje robusnije rezultate jer model testira na više različitih podskupova podataka. Međutim, kod predviđanja ishoda u sportu, unakrsna validacija nije uvijek prikladna jer su sportski događaji zavisni događaji. Rezultati budućih utakmica često ovise o rezultatima prethodnih utakmica, što znači da unakrsna validacija može testirati model na utakmicama koje još nisu odigrane, a to može dovesti do netočnih procjena. S druge strane, podjela podataka može dovesti do preoptimističnih procjena ako uzorci nisu reprezentativno odabrani. Autori su također dali nekoliko preporuka za poboljšanje rezultata predviđanja, kao što su unapređenje metoda učenja, korištenje većeg broja algoritama za strojnog učenja za pronalaženje optimalnih rješenja, poboljšanje metoda odabira značajki, optimizacija parametara strojnog učenja, upotreba relevantnih i optimalnih skupova značajki te pronalaženje pravilnosti među podacima. Istraživanje je pokazalo da uporaba novih, alternativnih algoritama strojnog učenja može donijeti dobre, a ponekad i bolje rezultate

predviđanja. Također je naglašeno da je glavni izazov razviti univerzalnu metodu koja će uspješno predviđati ishode u više sportova. Općenito, evaluacija rezultata istraživanja ukazuje na potrebu za standardizacijom pristupa kako bi se omogućila bolja usporedba i repliciranje studija. Kontinuirano poboljšanje metodologija i razmjena najboljih praksi može doprinijeti značajnom napretku u predviđanju nogometnih ishoda i boljem razumijevanju dinamičnosti igre.



Slika 2. Napredak algoritama strojnog učenja u nogometu po godinama.

Slika 2 prikazuje napredak algoritama strojnog učenja u predviđanju nogometnih ishoda kroz godine. Trend linija, izračunata metodom najmanjih kvadrata, sugerira stalni napredak u točnosti algoritama. Ovaj napredak je očekivan s obzirom na to da istraživači koriste prethodne radove kako bi unaprijedili postojeće metode, što dovodi do razvoja sve preciznijih algoritama. Trend linija pokazuje blagi porast točnosti kroz godine, što je pokazatelj kontinuiranog razvoja. Važno je napomenuti da, iako postoji stalan napredak, mogućnosti za daljnje značajno poboljšanje mogu biti ograničene. Daljnji napredak u algoritmima za predviđanje sportskih ishoda može varirati ovisno o specifičnostima pojedinih sportova. Osobito u nogometu, ljudski faktor igra ključnu ulogu, uključujući raspoloženje i motivaciju igrača te odluke trenera i uprave kluba, koje su često nepredvidljive za algoritme. Osim toga, vidljivo je da, unatoč napretku, postoji inherentna neizvjesnost u sportskim događajima koja se ne može u potpunosti eliminirati. Istraživanja pokazuju da, iako algoritmi mogu postići visoku točnost, uvijek postoji element nepredvidljivosti zbog dinamične prirode sporta. Na primjer, neočekivane ozljede, vremenski uvjeti, te individualne performanse igrača na dan utakmice mogu značajno utjecati

na ishode i predstavljaju izazov za čak i najsofisticiranije modele strojnog učenja. Analiza također pokazuje da su noviji algoritmi strojnog učenja sve više u stanju integrirati različite izvore podataka. Navedeni izvori uključuju statistike igrača, povijesne rezultate, podatke s društvenih mreža, te informacije o zdravstvenom stanju. Kombinacijom ovih različitih podataka, algoritmi mogu dodatno poboljšati točnost predviđanja. Međutim, krajnji cilj razvoja ovih algoritama ostaje ne samo povećanje točnosti već i razumijevanje kompleksnosti i nepredvidljivosti sporta kao dinamične i interaktivne aktivnosti.

2. Znanost o podacima i strojno učenje

S naglaskom na interdisciplinarnu prirodu znanosti o podacima, ovo poglavlje elaborira kako kombinacija statistike, matematike, informatike i specifičnog domenskog znanja omogućava ekstrakciju značajnih uvida iz širokog spektra podataka, bilo da su oni strukturirani, polustrukturirani ili nestrukturirani. Ističući važnost ove discipline u kontekstu neprestano rastućeg volumena podataka u digitalnom dobu, naglašava se njezina uloga u donošenju informiranih odluka kroz različite sektore – od poslovanja do zdravstva.

U fokusu su i temeljne metode strojnog učenja koje se koriste unutar znanosti o podacima, predstavljajući kako ove tehnike doprinose razvoju algoritama sposobnih za učenje iz podataka bez eksplicitnog programiranja. Razmatraju se različite vrste učenja – nadzirano, nenadzirano i učenje uz podršku – te njihova primjena u kreiranju modela koji mogu predviđati ishode, identificirati uzorke i donositi odluke na temelju analize podataka. Nadalje, detaljno se opisuju specifični algoritmi strojnog učenja, uključujući one za klasifikaciju i regresiju, pružajući uvid u njihove karakteristike, prednosti i potencijalne izazove. Posebna pažnja posvećena je metodama evaluacije učinkovitosti ovih modela, uključujući podjelu skupa podataka, unakrsnu validaciju i različite evaluacijske metrike, što je ključno za ocjenjivanje njihove sposobnosti generalizacije na novim, nepoznatim podacima. Ovo poglavlje, stoga, služi kao temelj za razumijevanje kako znanost o podacima i strojno učenje oblikuju procese donošenja odluka i razvoj strategija u različitim industrijskim sektorima, istovremeno naglašavajući složenost i interdisciplinarnost potrebnu za efikasno upravljanje i analizu podataka u svrhu ostvarivanja dubljih spoznaja i predviđanja budućih trendova.

2.1.1. Definicija znanosti o podacima

Znanost o podacima, poznata i kao podatkovna znanost, predstavlja interdisciplinarno područje koje se bavi prikupljanjem, analizom, interpretacijom i prezentacijom podataka kako bi se stekle spoznaje i informacije. Kombinacija statistike, matematike, računalne znanosti i domenskog znanja u ovoj disciplini omogućuje otkrivanje uzoraka, identifikaciju trendova te donošenje informiranih odluka. Ključna karakteristika znanosti o podacima je sposobnost ekstrakcije znanja iz različitih izvora podataka, bilo da su strukturirani, polustrukturirani ili nestrukturirani. U okviru šireg izlaganja o znanosti o podacima, ključno je razumjeti da ovo područje predstavlja temelj za analizu i interpretaciju velikih podataka (engl. *big data*), koristeći se načelima, definicijama problema, algoritmima i procesima za otkrivanje suptilnih,

no značajnih uzoraka ponašanja unutar obimnih podatkovnih skupova. Znanost o podacima ne samo da se oslanja na tehnike rudarenja podataka i strojnog učenja, već svojim djelokrugom nadmašuje ova područja, nudeći sveobuhvatan pristup u razumijevanju i korištenju podataka za informirano donošenje odluka [26]. Ovaj interdisciplinarni pristup omogućuje duboko prodiranje u složene podatkovne strukture, iz čega proizlaze uvidi koji mogu značajno utjecati na svakodnevni život pojedinaca i društva u cjelini. Primjene znanosti o podacima prožimaju različite aspekte suvremenog života, utječući na personalizaciju digitalnog oglašavanja, preporuke za multimedijski sadržaj, upravljanje elektroničkom poštom, prilagodbu ponuda u telekomunikacijama i zdravstvenom osiguranju, optimizaciju urbanih infrastruktura kao što su semafori, razvoj farmaceutskih proizvoda, pa čak i strategije policijskog nadzora [27].

Razumijevanje i primjena znanosti o podacima zahtijeva solidnu osnovu u statističkim metodama, algoritmima strojnog učenja, tehnologijama obrade i skladištenja podataka, kao i sposobnost analitičkog razmišljanja za interpretaciju rezultata. Uz to, etička pitanja imaju ključnu ulogu, posebno u kontekstu privatnosti, sigurnosti podataka i transparentnosti algoritama [28]. Stoga, u kontekstu prethodnog izlaganja, znanost o podacima predstavlja ne samo tehničku disciplinu već i katalizator za društvene i tehnološke promjene, čija primjena u svakodnevnom životu odražava široki spektar njenog utjecaja i potencijala za poboljšanje kvalitete života i donošenje temeljenih odluka [26].

2.1.1. Uloga znanosti o podacima

Znanost o podacima zauzima centralno mjesto u procesu donošenja odluka. Njena uloga je neizmjerno važna u brojnim sektorima, uključujući poslovnu analitiku, medicinska istraživanja, urbanističko planiranje i mnoge druge. Metodična analiza podataka koju provodi znanost o podacima omogućava otkrivanje skrivenih uzoraka i trendova, doprinoseći time dubljem razumijevanju kompleksnih fenomena i unapređenju operativnih performansi u različitim oblastima [27] [29].

U poslovnom sektoru, znanost o podacima omogućava tvrtkama da temeljem detaljne analize podataka o potrošačima, tržišnim trendovima i operativnoj efikasnosti donose strateške odluke usmjerene na optimizaciju resursa, povećanje prihoda i ostvarivanje konkurentske prednosti. Primjerice primjenom analize predviđanja, tvrtke mogu anticipirati buduće potražnje za svojim proizvodima i uslugama te time učinkovitije upravljati zalihama i logistikom. U medicinskom sektoru, znanost o podacima ima ključnu ulogu u napretku personalizirane

medicine i preciznih dijagnoza. Analizom kompleksnih setova podataka, uključujući genetske informacije i elektroničke zdravstvene zapise, znanstvenici mogu identificirati uzorke koji pridonose boljem razumijevanju bolesti, razvoju ciljanih terapija i predviđanju ishoda liječenja za pojedinačne pacijente. I u urbanom planiranju i upravljanju, znanost o podacima pomaže u optimizaciji infrastrukturnih projekata i gradskih usluga. Kroz analizu podataka o prometnim tokovima, potrošnji energije i demografskim promjenama, gradski planeri mogu dizajnirati učinkovitije i održivije gradove, poboljšavajući kvalitetu života građana [30]. Međutim, uspjeh u primjeni znanosti o podacima zahtijeva ne samo tehničku ekspertizu u obradi i analizi podataka, već i duboko razumijevanje specifičnih domena unutar kojih se podaci primjenjuju. To uključuje etičku odgovornost u zaštiti privatnosti i sigurnosti podataka, kao i sposobnost kritičkog promišljanja o implikacijama koje otkriveni uzorci mogu imati na društvo.

2.1.2. Znanost o podacima u predviđanju ishoda

U suvremenom poslovnom i društvenom okruženju, sposobnost predviđanja budućih ishoda postala je neophodna za informirano donošenje odluka. Znanost o podacima, u tom kontekstu, igra ključnu ulogu u razvoju strategija i modela koji mogu anticipirati buduće trendove i ishode. Analizom relevantnih podataka, znanstvenici o podacima su u mogućnosti identificirati ključne faktore i varijable koje utječu na različite ishode, uključujući ekonomske trendove, ponašanje potrošača, medicinske dijagnoze, i mnoge druge [31]. Primjenom sofisticiranih tehnika istraživanja podataka, poput klasifikacije, regresije, i grupiranja, znanstvenici o podacima mogu otkriti skrivene uzorke i veze između varijabli koje nisu očite na prvi pogled. Ove tehnike omogućavaju stvaranje preciznih modela za predviđanje, koji se oslanjaju na statističke metode i napredne tehnike strojnog učenja, uključujući neuronske mreže, slučajne šume i strojeve potpornih vektora [26]. Osim što pruža temelje za razvoj modela predviđanja, znanost o podacima također doprinosi optimizaciji postojećih procesa i strategija. Kroz kontinuiranu analizu i prilagodbu, organizacije mogu dinamički odgovarati na promjene u okruženju, poboljšavajući svoje poslovne operacije i povećavajući zadovoljstvo klijenata [30].

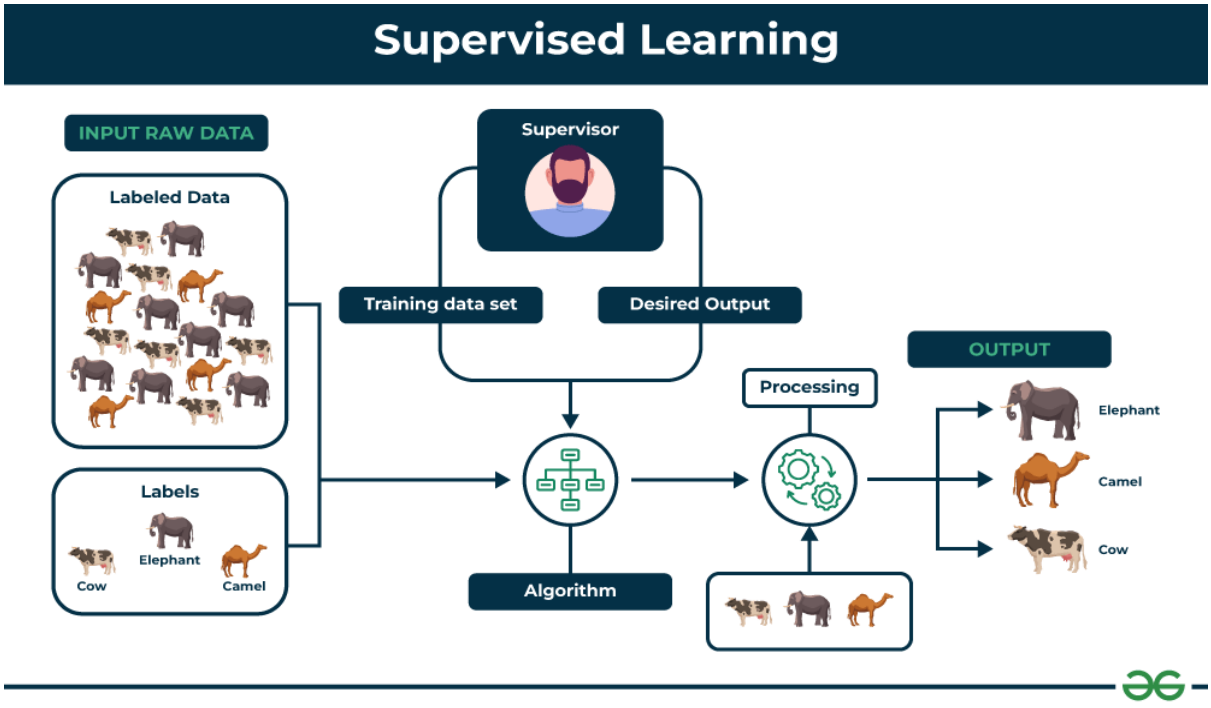
2.2. Strojno učenje u znanosti o podacima

Strojno učenje, kao integralni dio znanosti o podacima, odnosi se na razvoj algoritama i modela koji su sposobni učiti iz podataka, čime se omogućava donošenje predviđanja i odluka

bez potrebe za eksplicitnim programiranjem. Temelji se na ideji da sustavi mogu učiti iz podataka, identificirati uzorke i donositi odluke s minimalnom ljudskom intervencijom. S obzirom na njegovu svestranost, strojno učenje nalazi primjenu u širokom spektru područja, uključujući računalni vid, obradu prirodnog jezika, medicinsku dijagnostiku, financijsku analizu, robotiku, te mnoga druga, pružajući rješenja za raznovrsne probleme, od klasifikacije i regresije do detekcije obrazaca i grupiranja podataka [32] [33].

2.2.1. Nadzirano učenje

Nadzirano učenje (engl. *supervised learning*) je temeljna metoda strojnog učenja koja koristi označene podatke za treniranje modela. Proces uključuje obuku na podacima gdje su ulazi povezani s izlazima te omogućujući modelu da generalizira na nove podatke.



Slika 3. Koncept nadziranog učenja

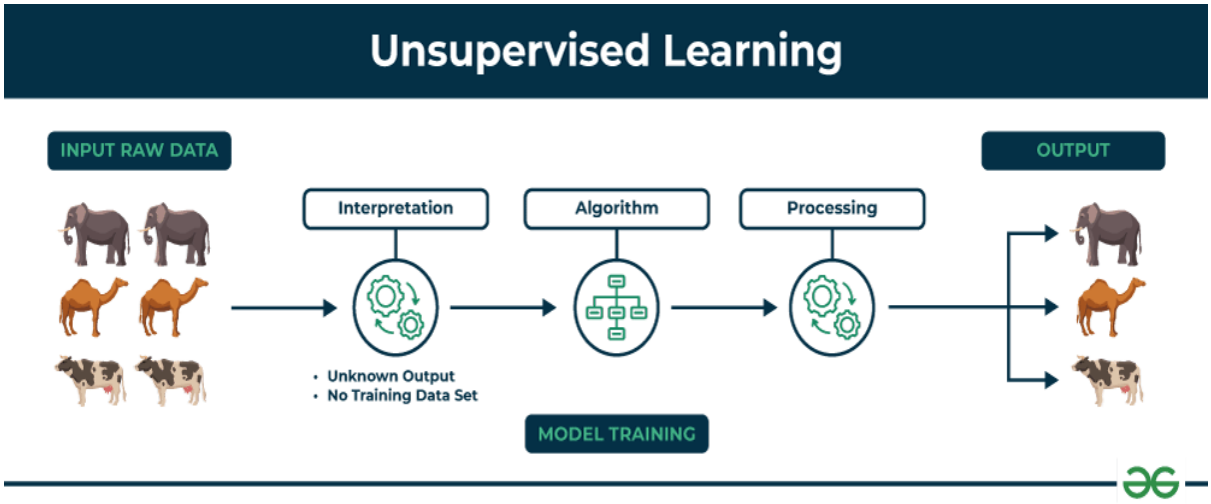
Izvor: preuzeto sa https://media.geeksforgeeks.org/wp-content/uploads/20231121154747/Supervised-learning.png

Riječ je o pristupu koji omogućava modelima da iz prethodno označenih podataka izvlače znanje, koristeći ga za donošenje odluka ili izvođenje predviđanja o novim, ranije nepoznatim podacima [27]. Ujedno, ovaj pristup učenja zahtijeva skup podataka s jasno definiranim ulazima (nezavisne varijable) i pripadajućim izlazima (ovisne varijable), omogućujući time modelu da uspostavi veze između ulaza i očekivanih izlaza. Slika 3 prikazuje

vizualni prikaz procesa nadziranog učenja u strojnom učenju, stiliziran u edukativne svrhe. U gornjem lijevom dijelu su dva slojevita spremišta podataka ili baze podataka, koje simboliziraju mjesto na kojem su pohranjeni podaci prije nego što ih model procesuira. Oni mogu predstavljati skupove podataka potrebne za treniranje i testiranje modela. Oznake na bazi podataka, poput „DATA“ i sličnih, mogu ukazivati na različite vrste podataka koji su pohranjeni unutar. Centralni mehanizam je srce strojnog učenja, mjesto gdje se obrađuju podaci i razvijaju modeli. S desne strane nalaze se tri oznake koje predstavljaju izlaz iz modela strojnog učenja: Deva, krava i slon. One su ciljani izlazni rezultati ili predviđanja koja model nadziranog učenja treba ostvariti. Ove ilustracije životinja ukazuju na to da model može klasificirati i prepoznati različite vrste životinja na temelju podataka s kojima je treniran. Na dnu su prikazani primjeri ulaznih podataka ili oznaka koje se koriste za treniranje modela. U donjem desnom kutu nalazi se dijagram s ciljem da prikaže ciklus testiranja modela, gdje model prolazi kroz iteracije učenja i poboljšavanja svojih predviđanja. Cjelokupna kompozicija pruža jasnu i intuitivnu vizualnu edukaciju o tome kako modeli uče iz označenih podataka i kako su sposobni primjenjivati to znanje za identifikaciju i klasifikaciju novih podataka. Nadalje, ključan aspekt nadziranog učenja je njegova sposobnost generalizacije, odnosno primjene naučenih veza na nove podatke koji nisu bili dio početnog trening skupa, čime se modelu omogućava da efikasno funkcionira u raznovrsnim i dinamičkim okruženjima. Prema tome, primjene nadziranog učenja su široke i raznovrsne, uključujući klasifikaciju, gdje modeli kategoriziraju ulazne podatke u predefinirane grupe ili klase, i regresiju, gdje se predviđaju kontinuirane vrijednosti na osnovu ulaznih podataka. Klasifikacija i regresija predstavljaju temeljne zadaće nadziranog učenja, omogućavajući modelima da rješavaju probleme kao što su dijagnostika bolesti, financijsko prognoziranje, prepoznavanje objekata na slikama i mnoge druge. Važnost odabira kvalitetnog i reprezentativnog trening skupa ne može se dovoljno naglasiti, budući da kvaliteta trening podataka direktno utječe na sposobnost modela da nauči relevantne veze i pravilno generalizira svoje znanje na nove primjere [27]. Stoga, proces pripreme podataka, uključujući čišćenje podataka, obradu nedostajućih vrijednosti i osiguravanje reprezentativnosti uzorka, ključan je korak u stvaranju uspješnih modela strojnog učenja.

2.2.2. Nenadzirano učenje

Nenadzirano učenje (engl. *unsupervised learning*), kao specifičan segment strojnog učenja, karakterizira istraživanje podataka bez unaprijed definiranih oznaka. Razlikujući se od nadziranog učenja, gdje je svaki ulazni podatak povezan s odgovarajućom izlaznom etiketom, nenadzirano učenje fokusira se na otkrivanje unutarnjih uzoraka i struktura unutar samih podataka, bez prethodnog znanja o ishodima [34].



Slika 4. Koncept nenadziranog učenja

Izvor: preuzeto sa https://media.geeksforgeeks.org/wp-content/uploads/20231124111325/Unsupervised-learning.png

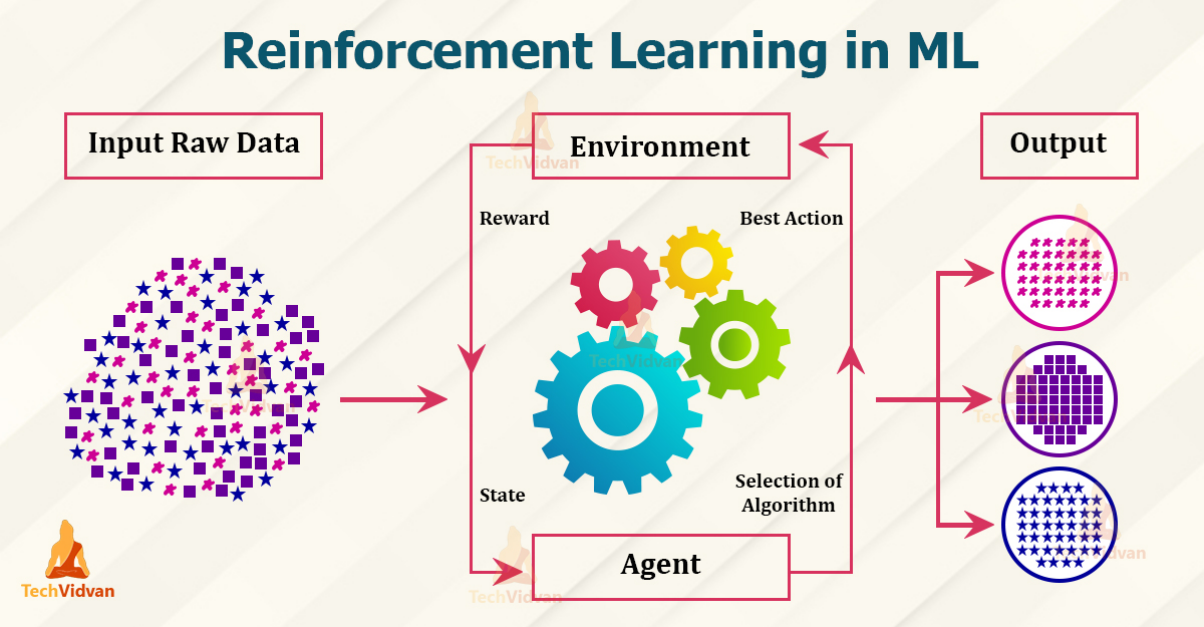
Osnovni cilj nenadziranog učenja jest detekcija skrivenih informacija, koje nisu eksplicitno očite. Ovaj metodološki pristup se široko koristi za analitičko istraživanje velikih skupova podataka, gdje algoritmi autonomno identificiraju strukture, često neophodne u kontekstima gdje specifični ciljevi istraživanja nisu jasno definirani. Grupiranje predstavlja jednu od ključnih tehnika u okviru nenadziranog učenja, gdje se podaci organiziraju u klastere na osnovu međusobne sličnosti. Ova tehnika nalazi primjenu u marketingu za segmentaciju tržišta, kao i u biomedicinskim istraživanjima za klasifikaciju bolesti na osnovu fenotipskih obilježja. Smanjenje dimenzionalnosti je još jedna bitna tehnika koja doprinosi pojednostavnjenju podataka, eliminacijom redundantnih ili manje važnih varijabli, što doprinosi efikasnijoj analizi i vizualizaciji podataka [27] [35]. Slika 4 predstavlja koncept nenadziranog učenja u strojnom učenju. Pri vrhu slike nalazi se niz objekata koji simboliziraju različite aspekte analize podataka kao što su kante za smeće (možda aluzija na čišćenje podataka), listovi s tablicama (mogući prikaz podataka ili algoritama), fascikli (koji mogu predstavljati organizaciju podataka) te nekoliko knjiga (što može ukazivati na teorijsku

podlogu ili skup znanja koji stoji iza algoritama nenadziranog učenja). Na lijevoj strani ilustracije nalazi se skup raznih životinja koji predstavljaju ulazne, neoznačene podatke koji se uvode u sustav. Ovo bi moglo predstavljati početni skup podataka koji se obrađuje. Proces se sastoji od tri glavne faze: interpretacija, algoritam i obrada. Tijekom faze interpretacije, podaci se analiziraju kako bi se razumjeli njihovi osnovni obrasci. Algoritmi, poput metoda grupiranja, primjenjuju se na ove podatke kako bi ih organizirali prema sličnostima. Konačno, kroz obradu, podaci se kategoriziraju, što je prikazano desno na slici gdje su životinje razvrstane u tri grupe: slonovi, deve i krave. Ovaj proces simbolizira nenadzirano učenje, gdje se bez prethodnih oznaka podaci analiziraju i grupiraju na temelju unutarnjih sličnosti i razlika.

Primjene nenadziranog učenja također obuhvaćaju prepoznavanje obrazaca i analizu teksta te generativno učenje gdje modeli kreiraju nove instance podataka, simulirajući distribuciju originalnih podataka. Usprkos potencijalu koji nenadzirano učenje pruža, izazovi ostaju u interpretaciji ishoda i validaciji struktura koje modeli otkrivaju. Zbog toga je razvoj algoritama koji mogu točno prepoznati relevantne podatke predmet neprestanih istraživanja [27] [36].

2.2.3. Učenje uz podršku

Učenje uz podršku (engl. *Reinforcement learning*) je metodologija unutar spektra strojnog učenja koja se usmjerava na razvoj algoritama sposobnih za samostalno učenje temeljem interakcije s okolinom. Ta se tehnika inspirira biološkim procesima učenja kod kojih se ponašanje usmjerava putem sistema nagrada i kazni. U srži ove paradigme leži koncept da algoritmi, odnosno agenti, stječu znanje kroz iskustveno učenje, reagirajući na povratne informacije iz okoline kako bi optimizirali svoje odluke i postupke u cilju maksimiziranja kumulativne nagrade [36]. Slika 5 prikazuje koncept strojnog učenja, posebno učenja uz podršku. Na lijevoj strani nalazi se skup neobrađenih podataka koji predstavljaju ulazne informacije. Ovi podaci ulaze u okruženje koje interagira s agentom. U okruženju, agent prima stanje i odabire najbolju akciju prema algoritmu. Ova akcija se zatim evaluira i daje povratnu informaciju u obliku nagrade, koja se vraća agentu kako bi poboljšao svoje buduće odluke. Na desnoj strani slike prikazan je izlazni rezultat koji pokazuje različite grupe podataka, što simbolizira organizirane podatke ili postignute ciljeve nakon procesa učenja. Zupčanici unutar okruženja simboliziraju različite faze obrade i donošenja odluka tijekom učenja. Također, prikazani su stupci koji bi mogli simbolizirati podatke ili znanje koje je prikupljeno i organizirano tijekom učenja.



Slika 5. Koncept učenja uz podršku

Izvor: preuzeto sa https://editor.analyticsvidhya.com/uploads/981063.jpg

Elementi učenja uz podršku obuhvaćaju agenta koji djeluje unutar specifičnog konteksta, okolinu koja pruža reakcije na akcije agenta, same akcije koje agent poduzima te nagrade koje agent prima kao odgovor na svoje postupke. Agent je programiran s ciljem maksimizacije akumulirane nagrade, a uspjeh je mjerljiv kroz sposobnost agenta da se prilagodi i efektivno djeluje unutar varijabilnih parametara okoline. Dakle, primjena ove vrste učenja je raznolika, prostirući se od robotike, gdje roboti uče manipulirati objektima, do optimizacije internet oglašavanja i razvoja strategija za igranje složenih igara. Jedan od ključnih izazova učenja uz podršku jest balansiranje između istraživanja novih strategija i korištenja već poznatih akcija za postizanje optimalnih rezultata [36]. Učenje uz podršku predstavlja napredak u umjetnoj inteligenciji, obećavajući značajna unapređenja u autonomiji i fleksibilnosti algoritama, što rezultira agentima koji su sposobni navigirati i prilagođavati se složenim i nepredvidivim okruženjima. Ovo područje je predmet intenzivnih istraživanja s ciljem unapređenja algoritama koji bi učinkovito usklađivali istraživanje i eksploataciju, omogućavajući kontinuiranu adaptaciju i učenje u dinamičnim kontekstima [34].

2.2.4. Algoritmi za klasifikaciju

Unutar domene strojnog učenja, algoritmi klasifikacije neophodni su za svrstavanje podataka u jasno određene kategorije ili klase. Njihova upotreba seže od dijagnostičkih

postupaka u medicini do analize osjećaja izraženih na društvenim mrežama. Različiti pristupi klasifikacijskim problemima omogućavaju modelima da se obuče na temelju etiketiranih podataka te da primijene stečeno znanje na nove, nepoznate instance [31].

Logistička regresija (engl. *logistic regression)* se, usprkos svom nazivu koji implicira veze s regresijskim modelima, često upotrebljava u klasifikacijske svrhe. Ovaj algoritam nudi procjenu vjerojatnosti pripadnosti određenih primjera specifičnoj klasi. Vrednuje se zbog jednostavne implementacije i intuitivnog načina tumačenja rezultata, no ograničenje predstavlja njezina linearna narav koja može biti nedostatna kod obrade nebalansiranih skupova podataka [27].

Stroj potpornih vektora (engl. *Support Vector Machine*) efikasno uspostavlja jasnu razdjelnu liniju među klasama, teži maksimiziranju margine između njih. Pokazuje dobre rezultate kod skupova podataka s visokim brojem dimenzija te kod slučajeva gdje granice između klasa nisu oštro definirane. Međutim, zahtijeva detaljno podešavanje parametara, posebno kod velikih skupova podataka, te može biti izazovan za interpretaciju [27].

Algoritam *k*-najbližih susjeda (engl. *k-nearest neighbours*) klasificira primjere na osnovi prevladavanja klase unutar *k*-najbližih susjeda u prostoru značajki. Zahvaljujući jednostavnosti implementacije, algoritam je popularan, iako može biti osjetljiv na šumove u podacima te zahtijeva čuvanje cijelog skupa podataka za svoje funkcioniranje [27].

Stabla odluke (engl. *decision trees*) klasificiraju podatke kroz hijerarhijski sustav pitanja i odgovora koji vode do krajnje klasifikacije. Njihove prednosti uključuju jednostavnu interpretaciju i sposobnost rada s kompleksnim, nelinearnim odnosima, dok su sklona prekomjernom prilagođavanju, posebno u skupovima podataka koji pokazuju visoku razinu varijabilnosti [27]. Ove metode predstavljaju tek dio bogatog spektra dostupnih algoritama za klasifikaciju. Izbor prikladnog algoritma zavisi od karakteristika skupa podataka, njegove veličine, te specifičnih zahtjeva analize [31]. Neprestano istraživanje i eksperimentiranje ključni su za postizanje optimalnih rezultata. Uz konstantan razvoj u polju strojnog učenja, novi algoritmi i rješenja neprestano proširuju mogućnosti preciznijih analiza podataka u različitim primjenjivim domenama.

2.2.5. Algoritmi za regresiju

Algoritmi regresije su ključni elementi u području strojnog učenja, usmjereni na predviđanje kontinuiranih vrijednosti na temelju raspoloživih ulaznih značajki. Oni

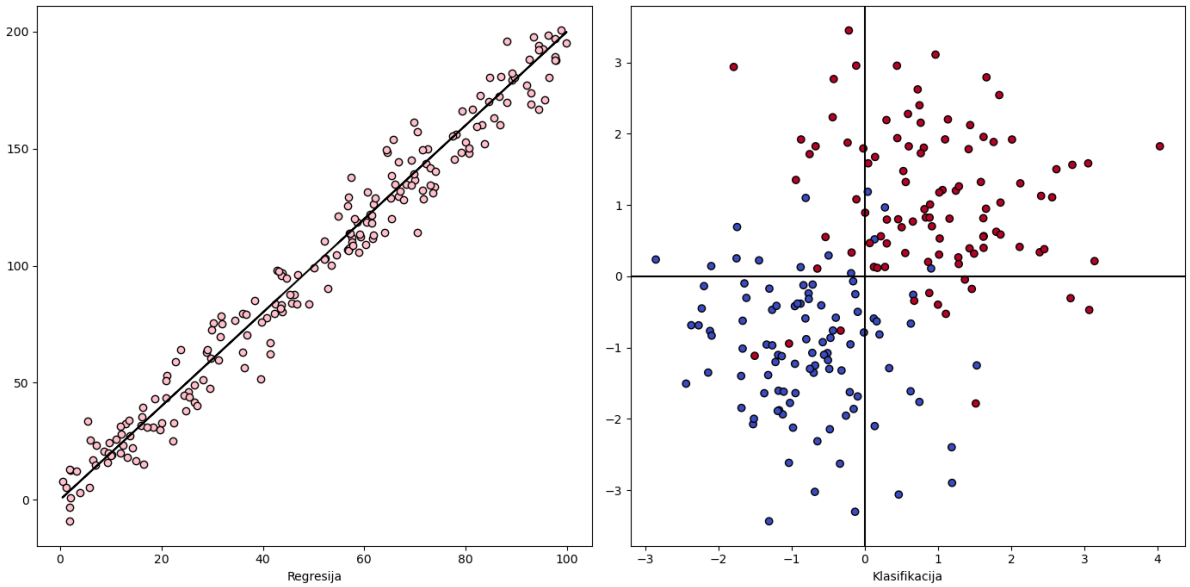
omogućavaju izgradnju modela koji kvantificiraju veze između nezavisnih i zavisnih varijabli te nude raznolike metodologije za predviđanje numeričkih ishoda [31].

Linearna regresija (engl. *linear regression*) služi kao temeljna metoda za modeliranje linearnih veza između prediktora i odgovora, s ciljem stvaranja optimalnog pravca koji najtočnije predstavlja distribuciju podataka. Njene prednosti uključuju laku interpretaciju i brzu primjenu. Međutim, njeno ograničenje je inherentna pretpostavka linearnosti odnosa, što može umanjiti njenu učinkovitost u obradi kompleksnijih podatkovnih struktura [35].

Regresija s potpornim vektorima (engl. *Support Vector Machine,* kraće SVR) prilagođava koncept stroja potpornih vektora iz klasifikacijskog konteksta za regresijske svrhe. Ona teži definiranju funkcije koja minimizira grešku između predviđenih i stvarnih vrijednosti, čime se izvlači optimalna predviđanja. SVR je posebno učinkovit u visokodimenzionalnim prostorima, no njegova učinkovitost ovisi o precizno odabranim hiperparametrima [35].

Regresijska stabla (engl. *Regression trees*) implementiraju model putem hijerarhijskih odluka koje vode do zaključivanja o vrijednostima izlazne varijable. To su modeli koji omogućuju modeliranje složenih nelinearnih odnosa i nude intuitivnu interpretaciju. Nedostatak im je tendencija ka prekomjernoj prilagodbi, osobito u situacijama s ograničenim skupom podataka. Nadalje, regresija slučajnih šuma (engl. *Random Forest Regression*) koristi ansambl više regresijskih stabala kako bi se smanjila varijanca i povećala točnost predviđanja. Ovaj pristup je robustan prema prekomjernoj prilagodbi i nudi skalabilnost, iako može otežati razumijevanje specifičnih veza unutar analiziranih podataka [35].

Slika 6 prikazuje ilustraciju koja predstavlja dva pojma iz strojnog učenja. Na lijevoj strani prikazan je grafikon s točkama i linijom trenda koja pokazuje regresijski model. Regresija se koristi za predviđanje kontinuiranih vrijednosti temeljem ulaznih podataka, a linija trenda predstavlja približan odnos između varijabli. Na desnoj strani prikazan je grafikon klasifikacije s točkama grupiranim u klase. Klasifikacija podataka uključuje razvrstavanje točaka u zasebne kategorije ili klase, što se obično prikazuje pomoću granice odlučivanja koja razdvaja točke u različite grupe.



Slika 6. Prikaz regresije i klasifikacije

Između dva grafikona nalazi se romb koji simbolizira usporedbu ili prijelaz između ova dva koncepta, sugerirajući da su iako su regresija i klasifikacija različiti zadaci, oni su oba sastavni dijelovi strojnog učenja. Oba grafikona imaju strelice koje ukazuju prema gore i udesno, što je standardna orijentacija za prikazivanje osi X i Y na grafikonima.

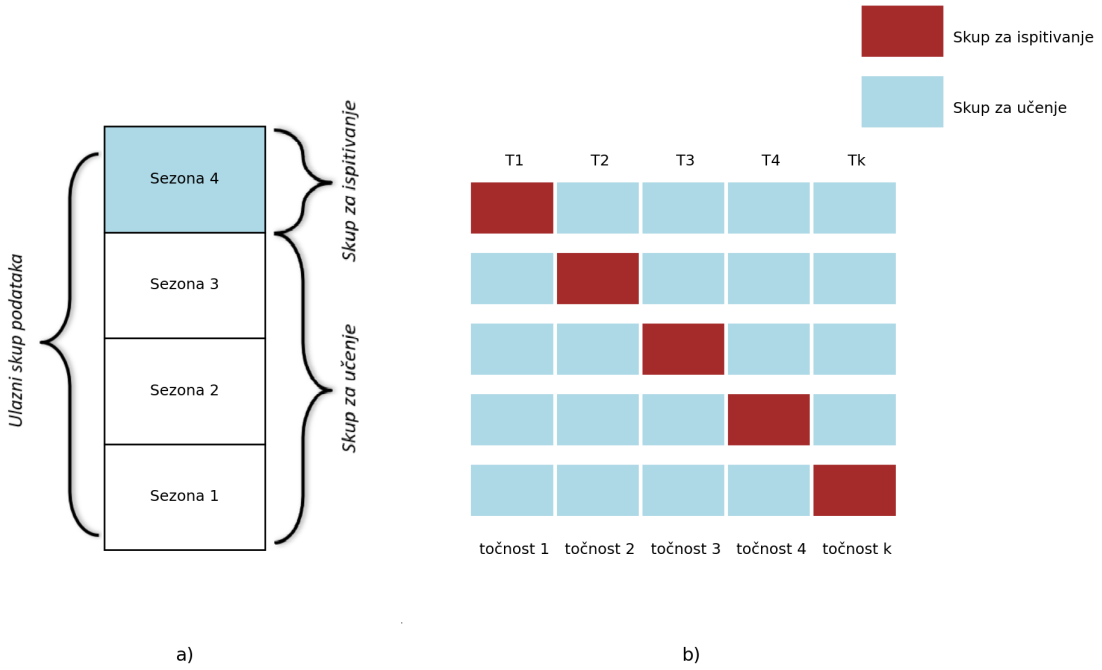
Odabir odgovarajućeg algoritma za regresiju zahtijeva promišljanje o osobitostima skupa podataka te o ciljevima analize. Proces istraživanja različitih modela i finog podešavanja parametara su ključni za postizanje željene točnosti predviđanja. Neprekidni napredak u području strojnog učenja otvara put za inovacije u tehnikama i algoritmima, što doprinosi razvoju regresijske analize za širok spektar primjena.

2.3. Metode evaluacije u strojnom učenju

Evaluacija modela ključan je korak u strojnom učenju kako bi se ocijenila njegova sposobnost generalizacije na novim, nepoznatim podacima. Različite metode evaluacije pomažu u procjeni performansi modela i identifikaciji potencijalnih problema. U nastavku, razmotrit ćemo ključne metode evaluacije kao što su podjelu skupa podataka, unakrsna validacija te druge relevantne parametre evaluacije. Slika 7 prikazuje dvije metode validacije podataka.

2.3.1. Podjela skupa podataka

Prilikom razvijanja modela strojnog učenja, esencijalno je podatke podijeliti na trening i testni skup. Trening skup poslužit će za obučavanje modela, gdje će algoritam naučiti obrasce inherentne podacima. Testni skup, koji mora biti odvojen i nekorišten tijekom faze treniranja, koristi se za evaluaciju modela, provjeravajući kako dobro naučeni obrasci funkcioniraju na novim, ranije neviđenim podacima. Značajno je postići balansirano zastupanje različitih klasa ili izlaznih vrijednosti u oba skupa podataka kako bi se osiguralo da evaluacija modela bude pravedna i reprezentativna. Ako je distribucija između treninga i testa neujednačena, može doći do pristranosti u modelu što vodi nevaljanim i netočnim predviđanjima [31].



Slika 7. Grafički prikaz metoda validacije. Slučaj a) prikazuje metoda podjele skupa, dok slučaj b) prikazuje metodu unakrsne provjere.

2.3.2. Unakrsna validacija

Tehnika unakrsne validacije (engl. *cross-validation*) je postupak verifikacije modela strojnog učenja koji osigurava pouzdanu ocjenu njegove sposobnosti generalizacije na nepoznatim podacima. Riječ je o standardiziranoj tehnici provjere točnosti modela, čija je svrha osigurati vjerodostojnu evaluaciju sposobnosti modela da generalizira zaključke na podacima koji nisu korišteni tijekom njegova učenja. Ovaj postupak je ključan kada se radi s ograničenim

količinama podataka, što može biti izazov za adekvatno ocjenjivanje performansi modela. Metoda tzv. *k*-folds unakrsne validacije je široko prihvaćena i često korištena varijanta ove tehnike, gdje se podaci razdvajaju na *k* podjednakih segmenata, ili „foldova“. Tijekom svake od *k* iteracija, jedan segment se određuje kao testni set, dok se svi ostali koriste za treniranje modela. Svaki segment se točno jedanput koristi kao testni skup, a proces se ponavlja *k* puta. Prosjek rezultata svih iteracija daje konačnu ocjenu učinkovitosti modela. Primjena unakrsne validacije smanjuje rizik od prekomjerne prilagodbe i neobjektivnosti modela, što omogućava pouzdaniju procjenu realnih performansi modela u praksi [31].

2.3.3. Drugi relevantni parametri evaluacije

Osim podjele skupa podataka i unakrsne validacije, postoje i drugi ključni parametri za procjenu performansi modela [31]:

1. Preciznost (engl. *Accuracy*) ─ Preciznost modela . prikazanom formulom (1) mjeri se omjer točno klasificiranih instanci prema ukupnom broju instanci.

$$Preciznost = \frac{Broj\ točno\ klasificiranih\ instanci}{Ukupan\ broj\ instanci} \qquad (1)$$

2. Osjetljivost (engl. *Recall*) ─ prikazanom formulom (2) mjeri se omjer prepoznatih stvarno pozitivnih instanci prema ukupnom broju stvarnih pozitivnih instanci.

$$Osjetljivost = \frac{Stvarno\ pozitivni}{Stvarno\ pozitivni + Lažno\ negativni} \qquad (2)$$

3. F1 Score ─ Prikazana formula (3) kombinira preciznost i osjetljivost te pruža izbalansiranu mjeru performansi modela.

$$F1 = 2\ x \frac{Preciznost\ x\ Osjetljivost}{Preciznost + Osjetljivost} \qquad (3)$$

4. Matrica konfuzije (engl. *Confusion matrix*) ─ Matrica konfuzije je tablica koja prikazuje stvarne i predviđene vrijednosti te pomaže u analizi grešaka modela.
5. AUC-ROC krivulja (engl. *Area Under the Receiver Operating Characteristic Curve*) ─ AUC-ROC krivulja koristi se za evaluaciju performansi klasifikacijskih modela, posebno u usporedbi s različitim pragovima odlučivanja.

Pravilno korištenje ovih metoda i parametara omogućuje precizniju procjenu performansi modela te olakšava identifikaciju i rješavanje problema poput prenaučenja (engl. *overtrained*) ili nedostatka generalizacije.

3. Prikupljanje podataka

Ovo poglavlje predstavlja ključnu fazu u procesu strojnog učenja, gdje se naglasak stavlja na identifikaciju izvora podataka i primjenu alata za njihovo prikupljanje. Kvalitetan izbor izvora podataka i efikasna obrada istih igraju presudnu ulogu u stvaranju uspješnih modela strojnog učenja. Istražuju se različiti izvori podataka, od web stranica i baza podataka do API-ja i lokalnih datoteka, naglašavajući važnost raznolikosti pristupa. Posebna pažnja posvećuje se etičkim i pravnim aspektima prikupljanja podataka, uz naglasak na poštivanje pravila web stranica i zaštite privatnosti. Također, detaljno se istražuju alati i tehnike za prikupljanje podataka s weba, uz naglasak na popularne biblioteke poput *Requests*, *BeautifulSoup* i *Pandas*. Kroz primjer prikupljanja podataka o utakmicama engleske Premier lige, ilustrira se praktična primjena web scrapinga i koraci potrebni za uspješno preuzimanje podataka s internetske platforme. Konačno, pruža se programski kod koji demonstrira proces prikupljanja i pohrane podataka, uz naglasak na korake i alate korištene u tom procesu. Stoga, ovo poglavlje omogućuje dublje razumijevanje važnosti prikupljanja podataka te pruža konkretne smjernice za njegovu uspješnu primjenu u kontekstu strojnog učenja.

3.1. Izvor podataka

Prikupljanje i obrada podataka predstavljaju ključne korake u procesu strojnog učenja, a identifikacija izvora podataka igra bitnu ulogu u formiranju modela. Raznolikost izvora omogućuje pristup različitim tipovima informacija, od strukturiranih tablica do nestrukturiranih tekstualnih dokumenata. Lokacija podataka označava mjesto na kojem su podaci fizički ili digitalno pohranjeni. Različiti izvori imaju svoje karakteristike [37]:

1. Web stranice često su izvor podataka putem web scrapinga, procesa ekstrakcije informacija s web stranica. Ovaj pristup omogućuje analizu širokog spektra podataka dostupnih na internetu.
2. Baze podataka često čuvaju strukturirane podatke, poput SQL baza ili NoSQL sustava, gdje se nalaze informacije poput transakcija, korisničkih podataka ili vremenskih serija.
3. API-ji (eng. *Application Programming Interface*) omogućuju pristup podacima putem programskog sučelja, što se često koristi za integraciju s vanjskim servisima poput društvenih mreža, financijskih usluga ili meteoroloških informacija.
4. Lokalne datoteke ponekad sadrže podatke pohranjene izravno na računalima ili serverima u obliku tekstualnih datoteka, Excel tablica ili CSV datoteka.

Razumijevanje lokacije podataka bitno je za uspostavljanje ispravnih metoda prikupljanja i obrade podataka. Svaki izvor može zahtijevati specifične tehnike kako bi se podaci učinkovito i točno ekstrahirali. Nužno je razmotriti i pravne aspekte prikupljanja podataka, posebno kada su u pitanju informacije koje su zaštićene pravilima privatnosti ili autorskim pravima.

3.2. Alati za prikupljanje podataka s weba

Prikupljanje podataka putem web scrapinga predstavlja moćan alat za ekstrakciju informacija s web stranica, no istovremeno zahtijeva pažljivu primjenu kako bi se poštovala etika i pravila. Ključni aspekti procesa scrapinga uz korištenje alata obuhvaćaju [37]:

1. Anaconda i JupyterLab ─ *Anaconda* distribucija i JupyterLab okruženje pružaju sveobuhvatno rješenje za analizu podataka. JupyterLab omogućuje interaktivno pisanje i izvođenje koda, dok Anaconda dolazi s predinstaliranim popularnim bibliotekama, uključujući *Pandas*, *BeautifulSoup* i *Requests*.
2. *Requests* ─ Python biblioteka koja se koristi za slanje HTTP zahtjeva prema web stranicama. *Requests* olakšava pristup stranicama, preuzimanje HTML koda i pripremu podataka za daljnju analizu.
3. *BeautifulSoup* ─ Python biblioteka koja se koristi za brzu i efikasnu analizu HTML i XML koda. *BeautifulSoup* olakšava proces traženja, izdvajanja i manipulacije podacima na web stranicama.
4. *Pandas* ─ *Pandas* je još jedna moćna Python biblioteka koja se koristi za analizu i manipulaciju podataka. Nakon što su podaci prikupljeni, *Pandas* omogućuje lako organiziranje i transformaciju podataka, uključujući pretvaranje u *DataFrame*, popularnu strukturu podataka za rad s tabelarnim podacima.
5. *Time* ─ Python biblioteka koja se koristi za manipulaciju vremenom. *Time* se može koristiti za postavljanje odgoda između zahtjeva prilikom scrapinga kako bi se smanjilo opterećenje na serverima.

3.3. Etika i pravila u prikupljanju podataka

Pri primjeni scrapinga, ključno je poštivati etičke smjernice i pravila kako bi se osigurao integritet informacija i zaštitila privatnost. Neki od ključnih aspekata uključuju [38] [39]:

1. Poštivanje Robots.txt Datoteke ─ Robots.txt datoteka često se nalazi na web stranicama i definira pravila o tome koji dijelovi stranice mogu biti indeksirani i prikupljeni, a koji ne. Pridržavanje ovih pravila ključno je za etičko prikupljanje podataka. Provjera Robots.txt datoteke pomaže osigurati da scraping proces ne krši postavljena ograničenja.
2. Pravne ograničenja ─ Prije početka scraping procesa važno je provjeriti uvjete korištenja stranice kako biste se uvjerili da nema zabrane scrapinga ili drugih ograničenja. Scraping podataka s web stranica može biti u suprotnosti s pravnim pravilima određenih web stranica. Nepoštivanje ovih pravila može rezultirati pravnim sankcijama.
3. Respektiranje privatnosti ─ Prikupljanje podataka povezanih s osobama ili osjetljivih informacija može ugroziti privatnost korisnika. U takvim slučajevima, važno je pridržavati se pravila o zaštiti podataka i osigurati zaštitu identiteta gdje god je to moguće. Posebna pažnja treba se posvetiti prikupljanju podataka o osobama, posebice u skladu s propisima o zaštiti podataka poput Opće uredbu o zaštiti podataka (GDPR) u Europskoj uniji.

Pravilna primjena scrapinga uz poštivanje etičkih smjernica ključna je za izgradnju održivog i odgovornog procesa prikupljanja podataka. Time se osigurava da analitičari mogu koristiti prikupljene podatke bez narušavanja integriteta informacija ili kršenja zakonskih normi.

3.4. Preuzimanje podataka o utakmicama engleske Premier lige

U svrhu dubinske analize performansi momčadi u engleskoj Premier ligi, odabrana je priznata internetska platforma - fbref.com. Ova platforma pruža obilje statističkih podataka o svakoj utakmici, omogućavajući analitičarima dublje proučavanje statistike, rezultata i ključnih parametara momčadi tijekom sezone [40]. Podaci potrebni za analizu dostupni su na stranici https://fbref.com/en/comps/9/Premier-League-Stats. Za sustavno prikupljanje ovih podataka, koristi se tehnika poznata kao web scraping. Web scraping omogućuje automatsko preuzimanje podataka s web stranica, što olakšava analitičarima pristup velikim količinama informacija. Za implementaciju ovog procesa koristi se Python, programski jezik široko korišten u analizi podataka. Dodatno, koriste se biblioteke poput *BeautifulSoup* za ekstrakciju HTML-a,

omogućujući precizno lociranje i izdvajanje željenih podataka, te *requests* za dohvaćanje web stranica. Ovaj pristup automatskom preuzimanju podataka omogućuje stvaranje konkretnog temelja za daljnje analize i istraživanja. Podaci o utakmicama, uključujući vremenske aspekte, rezultate, sastave momčadi, posjed lopte i mnoge druge parametre, bit će ključni u razumijevanju dinamike i trendova unutar engleske Premier lige. Time se dobiva dragocjen uvid u performanse momčadi.

3.5. Programski kod za preuzimanja i pohranu podataka s weba

Pri prikupljanju podataka s weba za analitičke svrhe, ključno je koristiti učinkovite metode automatizacije. U ovom poglavlju, predstavljen je postupak i programski kod koji se koristi za preuzimanje podataka s web stranica vezanih za englesku Premier ligu, te njihovo strukturiranje i spremanje za daljnju obradu. Korištene biblioteke jesu:

1. *Requests* ─ biblioteka koja se koristi za slanje HTTP zahtjeva i preuzimanje podataka s weba
2. *BeautifulSoup* iz bs4 ─ biblioteka koja se koristi za parsiranje i manipulaciju HTML sadržaja dobivenog iz web zahtjeva
3. *Pandas* ─ biblioteka koja se koristi za stvaranje i manipulaciju strukturiranim podacima (*DataFrames*) te njihovo spremanje u obliku CSV datoteka
4. *Time* ─ biblioteka koja se koristi za upravljanje vremenom između zahtjeva kako bismo izbjegli mogućnost preopterećenja web servera

Detaljan opis procesa i programskog koda:

1. Kod na slici prikazuje Python skriptu koja koristi module *requests*, *BeautifulSoup*, *pandas* i *time*. Skripta inicijalizira listu godina *years* koja sadrži godine u opadajućem redoslijedu od 2023. do 2019. te inicijalizira praznu listu *all_matches* i URL za podatke o Premier ligi *standings_url*. Pripadajući programski blok nalazi se u nastavku.

```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import time
years = list(range(2023, 2019, -1))
all_matches = []
standings_url = "https://fbref.com/en/comps/9/Premier-League-Stats"
```

2. Korištenje *requests* za slanje HTTP zahtjeva na definirani URL. Parsira se HTML sadržaj koristeći *BeautifulSoup*, izdvajaju se tablice s podacima te se generira puni URL za svaku momčad. Pripadajući programski blok nalazi se u nastavku.

```
for year in years:
    data = requests.get(standings_url)
    soup = BeautifulSoup(data.text, 'html.parser')
    standings_table = soup.select('table.stats_table')[0]
    links = [l.get("href") for l in standings_table.find_all('a') if
'/squads/' in l.get("href")]
    team_urls = [f"https://fbref.com{l}" for l in links]
    previous_season = soup.select("a.prev")[0].get("href")
    standings_url = f"https://fbref.com{previous_season}"
```

3. Preuzimanje i spajanje podataka: za svaku momčad, preuzimaju se podaci o utakmicama i statistikama igrača, te ih se spaja na osnovu datuma. Uključuje se i obrada grešaka kako bi se kod mogao nastaviti izvršavati i u slučaju problema s podacima. Pripadajući programski blok nalazi se u nastavku.

```
for team_url in team_urls:
    team_name = team_url.split("/")[-1].replace("-Stats", "").replace("-", "
")
    data = requests.get(team_url)
    matches = pd.read_html(data.text, match="Scores & Fixtures")[0]
```

4. Pohrana prikupljenih podataka; Nakon prikupljanja svih podataka, stvara se konsolidirana tablica i pohranjuje se u CSV datoteku za daljnju analizu. Pripadajući programski blok nalazi se u nastavku.

```
match_df = pd.concat(all_matches)
match_df.columns = [c.lower() for c in match_df.columns]
match_df.to_csv("matches.csv")
```

Kao rezultat, dobiva se opsežan skup podataka koji uključuje informacije o svakoj utakmici engleske Premier lige. Tijekom ovog procesa, prikupljaju se i analiziraju razni aspekti igre, uključujući datum i vrijeme održavanja, lokaciju, rezultate, broj postignutih i primljenih golova, kao i mnoge druge statistike koje su ključne za razumijevanje performansi momčadi i pojedinačnih igrača.

Primjer prikupljenih podataka koji omogućavaju stvaranje detaljnog pregleda lige uključuju:

1. Datum i vrijeme utakmice ─ ključni temporalni podaci o održavanju utakmica
2. Natjecanje i kolo ─ informacije o kontekstu u kojem se igra odvija
3. Rezultat ─ krajnji ishod utakmice koji predstavlja osnovu za analizu performansi

4. Očekivani golovi (engl. *expected goals*, kraće xG) ─ očekivani golovi postali su standard u analizi nogometa
5. Statistike igrača ─ detaljne informacije poput udaraca, udaraca u okvir gola i drugih parametara koji pružaju uvid u ofenzivnu i defenzivnu efikasnost

Ovaj način prikupljanja podataka ne samo da štedi vrijeme koje bi inače bilo potrebno za ručno prikupljanje informacija, već također povećava preciznost i pouzdanost podataka koji se koriste za analizu.

Programski kod je strukturiran na način da se izvršava efikasno i bez prekida. Strukturirani pristup je usvojen gdje je svaki korak jasno definiran i gdje se za svaku akciju provodi provjera ispravnosti kako bi se izbjegle greške koje bi mogle prekinuti proces ili dovesti do prikupljanja netočnih podataka. Konačni skup podataka pohranjuje se u CSV datoteku koja se može lako uvesti u alate za analizu podataka ili baze podataka za daljnje proučavanje. Takav metodološki pristup omogućava istraživačima i analitičarima fokusiranje na interpretaciju podataka, umjesto na njihovo prikupljanje.

Zaključno, razvijen je i implementiran programski kod koji omogućava automatizirano prikupljanje, obradu i pohranjivanje podataka o engleskoj Premier ligi. Kod predstavlja temelj za dublju analizu nogometnih performansi i može se prilagoditi te proširiti za prikupljanje podataka iz drugih liga ili sportskih događanja.

4. Predviđanje ishoda

Predviđanje ishoda sportskih događaja razvilo se u značajno područje istraživanja, posebice u svijetu najprestižnijih sportskih liga kao što je engleska Premier liga. U suvremenom dobu, obilje podataka prikupljenih s raznih sportskih događanja koristi se ne samo za detaljnu analizu već i za anticipiranje budućih rezultata.

Strojno učenje postaje sve važniji instrument u prognoziranju rezultata sportskih natjecanja. Ova grana istraživanja nudi alate trenerima i menadžerima u sportu za predviđanje ishoda nadolazećih utakmica, procjenu učinka pojedinih igrača ili cijele momčadi, otkrivanje mogućih ozljeda te identificiranje nadolazećih talenata. Osim toga, ovakvi podaci su neprocjenjivi za širu javnost, posebno u kontekstu sportskog klađenja, pružajući temelj za informiranije odluke.

Engleska Premier liga, poznata po svom visokom stupnju konkurentnosti, pruža obilje podataka koji su iznimno vrijedni za analitičare. Analiza faktora kao što su formacija momčadi, povjest međusobnih susreta, ozljede igrača i trenutna forma ključna je za primjenu metoda nadziranog strojnog učenja s ciljem predviđanja ishoda utakmica.

Dok su individualni sportovi s dva moguća ishoda, poput tenisa ili šaha, relativno jednostavniji za predviđanje, nogomet kao timski sport koji uključuje višestruke varijable, predstavlja znatno veći izazov. Ipak, proučavanje iznenađujućih sportskih događaja, kao što je neočekivano osvajanje Premier lige momčadi Leicester City-a u sezoni 2015./2016. godine, ukazuje na nepredvidivost sportskih natjecanja i potencijal za iznenađenja koja se mogu odigrati na terenu.

4.1. Alati za predviđanje ishoda

Scikit-learn predstavlja otvorenu biblioteku za strojno učenje u programskom jeziku Python, poznatu po svojoj jednostavnoj, ali učinkovitoj implementaciji raznovrsnih algoritama strojnog učenja te alata za procjenu modela. Zbog svoje široke funkcionalnosti u izradi, treniranju i evaluaciji modela, biblioteka je stekla popularnost među istraživačima i praktikantima u području strojnog učenja.

Algoritam logistička regresija (engl. *logistic regression)*, koji spada u modul *sklearn.linear_model*, koristi se za binarnu i višeklasnu klasifikaciju. Logistička regresija procjenjuje vjerojatnosti pripadnosti uzoraka određenim klasama pomoću logističke funkcije, pretvarajući ulazne značajke u izračun vjerojatnosti. Ovaj algoritam je posebno koristan zbog

svoje jednostavnosti i interpretabilnosti. Koristi se u raznim domenama poput financija, zdravstva i marketinga za predviđanje binarnih ishoda kao što su kreditni rizik, dijagnoza bolesti i uspješnost marketinških kampanja.

Algoritmi Naivni Bayes (engl. *Naive Bayes*), kao što su GaussianNB, MultinomialNB i BernoulliNB iz modula *sklearn.naive_bayes*, temelje se na Bayesovom teoremu s pretpostavkom o uvjetnoj neovisnosti značajki. Navedeni algoritmi su posebno učinkoviti za velike skupove podataka i često se koriste u klasifikacijskim problemima poput filtriranja neželjene pošte i analize sentimenta. Naivni Bayes se koristi zbog svoje brzine i učinkovitosti, posebno u situacijama kada je potrebno brzo obraditi velike količine podataka.

Algoritam stabla odluke (engl. *decision trees*), koji spada u modul *sklearn.tree*, koristi se za klasifikacijske zadatke. Ovaj algoritam dijeli podatke na podskupove na temelju atributa pomoću niza odlučujućih pravila, što rezultira strukturama koje se lako interpretiraju. Iako su stabla odluke sklona prenaučenosti, mogu biti vrlo učinkovita kada se koriste s pravilno podešenim parametrima. Stabla odluke su korisna u raznim područjima kao što su medicinska dijagnostika, gdje se koriste za donošenje odluka temeljenih na medicinskim kriterijima, te u financijama za procjenu kreditnog rizika.

Algoritam *k*-najbližih susjeda (engl. *k-nearest neighbours*) iz modula *sklearn.neighbors* temelji se na identifikaciji *k*-najbližih susjeda svakom podatkovnom primjerku za klasifikacijske zadatke. Ova metoda je poznata po svojoj jednostavnosti i intuitivnosti, te je vrlo učinkovita za probleme s više klasa i kompleksnim raspodjelama podataka. *K*-najbližih susjeda su korisni u raznim aplikacijama, uključujući preporučne sustave, prepoznavanje obrazaca i dijagnostiku.

Nadalje, u sklopu biblioteke, *sklearn.ensemble* obuhvaća implementacije različitih metoda ansambla, uključujući *bagging*, boosting i slučajne šume. Te metode unaprjeđuju sposobnosti predviđanja kombiniranjem prognoza višestrukih modela. Posebno, algoritam slučajnih šuma (engl. *random forest*), koji spada u kategoriju metoda ansambla modula *sklearn.ensemble*, temelji se na korištenju stabala odlučivanja. Ova metoda podatke razdvaja u podskupove kroz niz odlučujućih pravila, pri čemu *RandomForestClassifier* stvara višestruka stabla odluke na osnovi slučajno izabranih uzoraka značajki i podataka. Time se osigurava diverzifikacija među stablima i smanjuje rizik od prenaučenosti modela. Slučajne šume se široko koriste u područjima poput biomedicine, ekologije i ekonomije za rješavanje problema klasifikacije i regresije. Algoritam višeslojni perceptron (engl. *multilayer perceptron,* kraće MLP), koji se koristi za klasifikacijske zadatke, temelji se na boosting metodi, gdje se serija slabih klasifikatora kombinira kako bi se stvorio snažan klasifikator. Boosting metode,

uključujući LogitBoost, iterativno poboljšavaju model fokusiranjem na podatke koje su prethodni klasifikatori pogrešno klasificirali. Spomenuti algoritam je koristan u aplikacijama gdje je potrebno postići visoku točnost, kao što su predviđanje bolesti, financijske analize i detekcija prijevara.

Algoritam MLPClassifier iz modula *sklearn.neural_network* koristi strukturu višeslojnog perceptrona za klasifikacijske zadatke. Višeslojni perceptron (engl. *multilayer perceptron*, kraće MLP) je vrsta umjetne neuronske mreže koja se sastoji od ulaznog sloja, jednog ili više skrivenih slojeva te izlaznog sloja. MLP koristi metode nadziranog učenja za treniranje modela pomoću algoritma povratnog širenja pogreške, omogućujući efikasno rješavanje kompleksnih problema klasifikacije. Višeslojni perceptroni su primjenjivi u raznim domenama uključujući prepoznavanje rukopisa, analizu slika i prirodno jezično procesiranje. Za procjenu učinka modela, modul *sklearn.metrics* pruža niz metrika. Te metrike pružaju analizu točnosti, preciznosti, odziva, kao i drugih kritičnih mjera koje ocjenjuju efikasnost modela. Jedna od takvih metrika jest matrica konfuzije, koja se koristi za evaluaciju učinkovitosti klasifikacijskih modela. Matrica konfuzije nudi uvid u broj točnih i netočnih prognoza po klasama, omogućujući detaljno razumijevanje performansi modela, osobito u problemima višeklasne klasifikacije. Alati i moduli dostupni u biblioteci *Scikit-learn* olakšavaju proces razvoja i testiranja modela strojnog učenja u različitim primjenama, što istraživačima i praktikantima omogućuje usmjeravanje na inovacije i rješavanje konkretnih problema u području. Kroz korištenje ovih algoritama i alata, stručnjaci mogu razvijati sofisticirane modele za predviđanje ishoda, klasifikaciju i analizu podataka, čime doprinose napretku u znanosti, tehnologiji i industriji.

4.2. Koraci i programski kod

Ovo poglavlje detaljno razmatra postupak razvoja modela strojnog učenja koristeći Python biblioteke, s posebnim fokusom na biblioteku *Scikit-learn*. Kroz niz programskih primjera, objašnjava se cjelokupni proces razvoja modela, počevši od početne obrade podataka pa sve do konačne evaluacije modela. Svaki korak je sistematično prikazan kako bi se osiguralo jasno razumijevanje i mogućnost primjene naučenih metoda na različite analitičke izazove.

Potrebne biblioteke za manipulaciju podacima i različite modele za strojno učenje iz *Scikit-learn* se uvoze. Pripadajući programski blok nalazi se u nastavku.

```
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import accuracy_score
```

Podaci o utakmicama učitavaju se iz CSV datoteke i postavlja se prvi stupac kao indeks. Pripadajući programski blok nalazi se u nastavku.

```
# Učitavanje podataka
matches = pd.read_csv("MatchesSeason.csv", index_col=0)
```

Stupac *date* se pretvara u *datetime* format, kreira se ciljana varijabla *target* koja označava pobjedu rezultatom 1 ili poraz rezultatom 0. Kategorijalne varijable *venue* i *opponent* se kodiraju, izvlači se sat iz stupca *time*, te se dobiva dan u tjednu iz stupca *date*. Pripadajući programski blok nalazi se u nastavku.

```
# Predprocesiranje podataka
matches["date"] = pd.to_datetime(matches["date"])
matches["target"] = (matches["result"] == "W").astype("int")
matches["venue_code"] = matches["venue"].astype("category").cat.codes
matches["opp_code"] = matches["opponent"].astype("category").cat.codes
matches["hour"] = matches["time"].str.replace(":.+", "", regex=True).a-
stype("int")
matches["day_code"] = matches["date"].dt.dayofweek
```

Definiraju se značajke varijable koje će se koristiti za treniranje modela. Pripadajući programski blok nalazi se u nastavku.

```
# Definicija prediktora
predictors = ["venue_code", "opp_code", "hour", "day_code"]
```

Definiraju se stupci za koje će se izračunavati pomični prosjeci, dodajući sufiks *_rolling*. Funkcija *rolling_averages* sortira grupu podataka po datumu, računa pomične prosjeke za zadane stupce, dodaje nove stupce s pomičnim prosjecima i uklanja redove bez dovoljno podataka. Pripadajući programski blok nalazi se u nastavku.

```
# Definiranje stupaca za pomične prosjeke
cols = ["gf", "ga", "sh", "sot", "dist", "fk", "pk", "pkatt"]
new_cols = [f"{c}_rolling" for c in cols]

# Funkcija za izračunavanje pomičnih prosjeka
def rolling_averages(group, cols, new_cols):
    group = group.sort_values("date")
    rolling_stats = group[cols].rolling(3, closed='left').mean()
    group[new_cols] = rolling_stats
    group = group.dropna(subset=new_cols)
    return group
```

Funkcija *rolling_averages* se primjenjuje na svaki tim u podacima koristeći *groupby* i *apply* metode. Nakon toga, uklanja se dodatni indeks momčadi i resetiraju se indeksi redaka u *matches_rolling* okviru podataka. Pripadajući programski blok nalazi se u nastavku.

```
# Primjena pomičnih prosjeka na cijeli skup podataka
matches_rolling = matches.groupby("team").apply(lambda x: rolling_averages(x,
cols, new_cols))
matches_rolling = matches_rolling.droplevel('team')
matches_rolling.index = range(matches_rolling.shape[0])
```

Podaci se dijele na skupove za treniranje (utakmice prije 1. siječnja 2022.) i testiranje (utakmice nakon tog datuma). Pripadajući programski blok nalazi se u nastavku.

```
# Podjela podataka na skupove za treniranje i testiranje
train_data = matches_rolling[matches_rolling["date"] < '2022-01-01']
test_data = matches_rolling[matches_rolling["date"] >= '2022-01-01']
```

Definiraju se različiti modeli za treniranje, uključujući logističku regresiju, Naive Bayes, stabla odluke, k-najbližih susjeda, slučajnu šumu, AdaBoost s logističkom regresijom kao osnovnim modelom, i višeslojni perceptron. Pripadajući programski blok nalazi se u nastavku.

```
# Inicijalizacija modela
models = {
    "Logistic Regression": LogisticRegression(max_iter=1000),
    "Naive Bayes": GaussianNB(),
    "Decision Tree": DecisionTreeClassifier(),
    "K-Nearest Neighbors": KNeighborsClassifier(),
    "Random Forest": RandomForestClassifier(n_estimators=50, min_sam-
ples_split=10, random_state=1),
    "LogitBoost": AdaBoostClassifier(base_estimator=LogisticRegression(max_i-
ter=1000), n_estimators=50),
    "Multi-layer Perceptron": MLPClassifier(max_iter=1000)
}
```

Kreira se tablica za pohranu rezultata za svaki model i svaki tim. Nazivi stupaca su imena modela, a indeksi su jedinstvene momčadi iz testnog skupa podataka. Pripadajući programski blok nalazi se u nastavku.

```
# DataFrame za rezultate
results = pd.DataFrame(index=test_data["team"].unique(), columns=mo-
dels.keys())
```

Svaki model se trenira koristeći podatke za treniranje i predviđaju se ishodi za svaku momčad u testnom skupu podataka. Za svaku momčad se računa točnost predviđanja kao omjer ispravno predviđenih ishoda i ukupnog broja utakmica. Rezultati točnosti se pohranjuju u *team_results* tablicu. Na kraju, za svaki model se pohranjuje prosječna točnost po momčadi u *results* tablicu. Pripadajući programski blok nalazi se u nastavku.

```
# Iteriranje kroz svaki model
for model_name, model in models.items():
    # Treniranje modela
    model.fit(train_data[predictors + new_cols], train_data["target"])

    # DataFrame za rezultate svakog tima
    team_results = pd.DataFrame(index=test_data["team"].unique(), colu-
mns=["Accuracy", "Total Matches Tested", "Missed"])

    # Iteriranje kroz svaki tim radi testiranja
    for team in team_results.index:
        team_matches = test_data[test_data["team"] == team]

        if not team_matches.empty:
            preds = model.predict(team_matches[predictors + new_cols])
            total_matches = len(team_matches)
            missed = (preds != team_matches["target"]).sum()
            accuracy = (total_matches - missed) / total_matches

            team_results.loc[team, "Accuracy"] = accuracy
            team_results.loc[team, "Total Matches Tested"] = total_matches
            team_results.loc[team, "Missed"] = missed

    # Pohrana prosječne točnosti za svaki model
    results[model_name] = team_results["Accuracy"]
```

Na kraju, ispisuju se rezultati točnosti svakog modela za svaku momčad, te prosječna točnost po modelu koristeći *print* funkciju. Pripadajući programski blok nalazi se u nastavku.

```
# Ispis rezultata
print("\nResults:")
print(results)
print("\nAverage Accuracy per model:")
print(results.mean())
```

Kroz primjer izrade modela za predviđanje ishoda nogometnih utakmica pokazano je kako se tehnike strojnog učenja mogu primijeniti na stvarne podatke. Ovaj pristup omogućava razvoj efikasnih modela i dublje razumijevanje dinamike sportskih natjecanja. Naučene metode mogu se primijeniti na različite analitičke izazove u mnogim industrijama, otvarajući mogućnosti za inovacije i poboljšanje odlučivanja temeljenog na podacima.

5. Vizualizacija podataka

Vizualizacija je neophodna jer omogućava jasno i efikasno predstavljanje kompleksnih informacija, što je posebno važno u kontekstu sportskih analiza gdje brojne varijable i veliki volumeni podataka mogu otežati razumijevanje i interpretaciju. Korištenjem biblioteka kao što su *pandas* i *Matplotlib*, koje su standard u industriji za obradu i vizualizaciju podataka u Pythonu, transformiraju se suhoparne numeričke tablice u vizualno privlačne i intuitivne grafičke prikaze. Specifično, ovaj dio istraživanja fokusira se na predstavljanje rezultata analize utakmica nogometne Premier lige, koristeći podatke prikupljene iz 1160 utakmica tokom četiri sezone. Tablica 2 prikazuje osam značajki. Navedene značajke su ključne za analizu i predviđanja ishoda. Kroz ovaj segment, demonstrira se kako se podaci mogu efektivno koristiti ne samo za stvaranje modela strojnog učenja već i za vizualno prikazivanje performansi, trendova i potencijalnih anomalija unutar sezonskih performansi momčadi. Na temelju ovih vizualizacija, dobiva se dublji uvid u dinamiku lige, što dodatno pomaže u taktičkom planiranju i strategiji, kako za momčadi tako i za individualne igrače. Ovo poglavlje će uključivati primjere vizualizacija koje prikazuju točnost predviđanja, analizu momčadskih performansi i druge ključne statističke pokazatelje važnih za razumijevanje sportskih susreta. Tablica 2 prikazuje osam značajki korištenih za predviđanje ishoda utakmica.

Tablica 2. Korištene značajke

ZNAČAJKA	OPIS	ULOGA U MODELU
gf	Broj postignutih golova momčadi	Indikator ofenzivne snage momčadi.
ga	Broj primljenih golova momčadi	Mjera defenzivne stabilnosti momčadi.
sh	Broj udaraca na gol	Odraz agresivnosti momčadi na terenu.
sot	Broj udaraca u okvir gola	Pokazatelj preciznosti udarac na gol.
dist	Udaljenost pređena loptom	Mjera mobilnosti i dinamike igre.
fk	Broj slobodnih udaraca	Indikator prilika nastalih iz prekršaja.
pk	Broj kaznenih udaraca	Broj prilika za postizanje gola iz penala.
pkatt	Broj izvedenih kaznenih udaraca	Mjera učestalosti i uspješnosti penala.

5.1. Alati za vizualizaciju podataka

Vizualizacija podataka ključna je komponenta u analizi i prezentaciji podataka, omogućavajući bolje razumijevanje kompleksnih informacija kroz vizualno privlačne i

intuitivno razumljive grafičke prikaze. U kontekstu analize sportskih podataka, posebno rezultata nogometnih utakmica, vizualizacija igra nezamjenjivu ulogu u prikazivanju trendova, distribucija i odnosa među varijablama. Biblioteka *pandas* u Pythonu stoji u temelju obrade i vizualizacije podataka zahvaljujući svojoj sposobnosti da jednostavno manipulira podacima i generira različite vrste grafova i dijagrama iz tabličnih struktura. Ova biblioteka omogućuje stvaranje osnovnih grafikona kao što su histogrami, kružni dijagrami i linijski grafikoni, koji su instrumentalni za vizualnu analizu distribucija, udjela i trendova u podacima. Funkcionalnosti *pandas*-a koriste se kako za početnu analizu tako i za pripremu podataka koji će se dalje vizualizirati za dublje analize [41]. Za naprednije vizualizacije, biblioteka *Matplotlib* pruža obimne mogućnosti za prilagodbu izgleda i stila grafova. Ovaj alat omogućava kreiranje kompleksnijih vizualizacija kao što su raspršeni dijagram (engl. *scatter plot*), kutijasti dijagram i toplinske karte (engl. *heatmap*), koje su ključne za detaljnu analizu relacija između različitih varijabli. Biblioteka *Matplotlib* se često koristi za izradu grafičkih prikaza koji zahtijevaju detaljne prilagodbe, čime se omogućuje precizno prikazivanje specifičnih aspekata podataka [42]. Ove biblioteke su se pokazale kao neophodni alati u istraživanju i prezentaciji podataka relevantnih za analizu ishoda nogometnih utakmica. Korištenjem biblioteka *pandas*-a i *Matplotlib*-a, analitičari mogu jasno prikazati kako se različite varijable mijenjaju kroz vrijeme, kako međusobno koreliraju i kako utječu na ishode procesa, u ovom radu korištene za ishode nogometnih utakmica. Primjena ovih alata omogućuje ne samo bolje razumijevanje podataka, već i pruža osnovu za strateške odluke unutar sportskih organizacija.

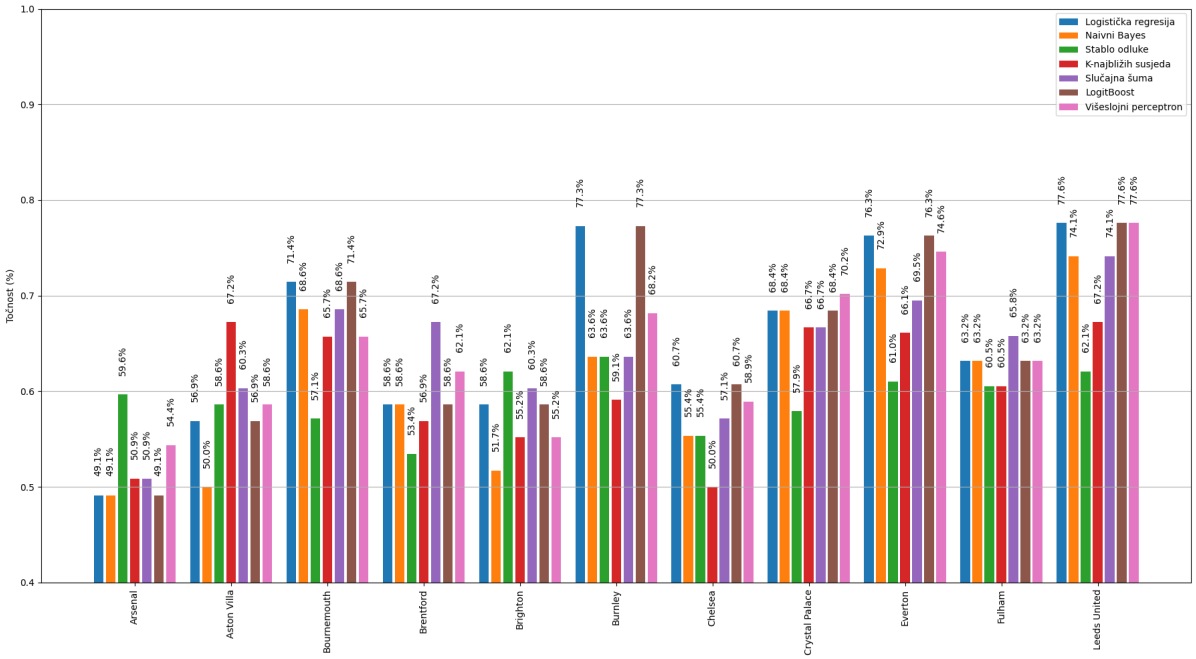
5.2. Izrada vizualizacija i programski kod

Vizualizacijom se transformiraju tzv. sirovi podaci u razumljive i zorno prikazane informacije, što je posebno važno prilikom analize rezultata nogometnih utakmica. U nastavku ovog poglavlja predstavit će se primjer kako se podaci o točnosti predviđanja ishoda utakmica momčadi engleske Premier lige mogu vizualno interpretirati koristeći programski jezik Python i biblioteku *Matplotlib*. Podaci koji se analiziraju obuhvaćaju četiri nogometne sezone, s ukupno 1160 utakmica. Proces započinje računanjem prosječne točnosti svakog modela. Zatim se definiraju parametri za vizualizaciju, uključujući širinu stupca, razmak između stupaca i razmak između grupa stupaca. Kreiraju se pozicije za sve stupce kako bi se osigurala pravilna raspodjela na grafu. Nakon toga, kreira se graf, gdje se za svaki model i svaku momčad crta odgovarajući stupac. Na vrh svakog stupca dodaju se postotci kako bi se olakšalo čitanje

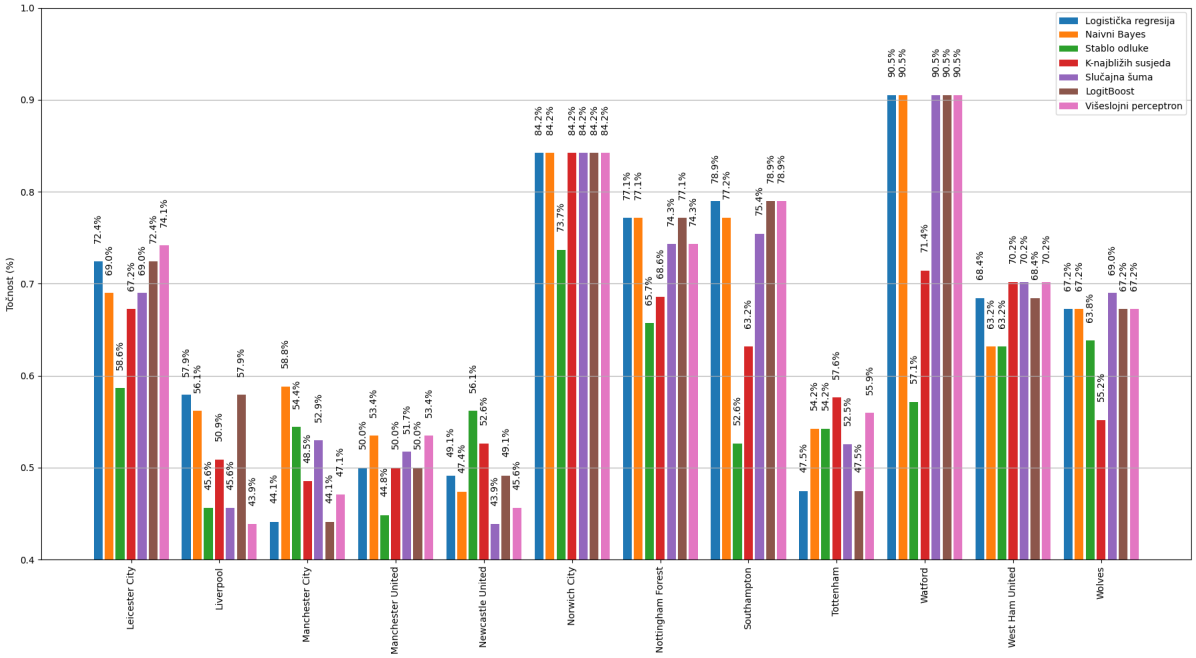
rezultata. Naslovi i oznake osi jasno su postavljeni, a rotacija oznaka na x-osi omogućava čitljivost naziva momčadi. Y-osi je postavljena od 0 do 1.1 kako bi svi podaci bili vidljivi, uključujući prostor za postotke iznad stupaca. Konačno, *plt.tight_layout()* osigurava pravilan raspored elemenata, a *plt.show()* prikazuje graf korisniku.

```
import matplotlib.pyplot as plt
import numpy as np

average_accuracy = results.mean()

bar_width = 0.4
gap = 0.1
group_gap = 0.5

indices = np.arange(len(results)) * (len(results.columns) * (bar_width + gap)
+ group_gap)

plt.figure(figsize=(18, 10))

for i, model_name in enumerate(results.columns):
    plt.bar(indices + i * (bar_width + gap), results[model_name], bar_width,
label=model_name)

for i, model_name in enumerate(results.columns):
    for j, value in enumerate(results[model_name]):
        plt.text(indices[j] + i * (bar_width + gap), value + 0.02,
f'{value:.1%}',
                 ha='center', va='bottom', fontsize=10, rotation=90)

plt.ylabel('Točnost (%)')
plt.xticks(indices + (len(results.columns) - 1) * (bar_width + gap) / 2,
results.index, rotation=90)
plt.ylim(0.4, 1)
plt.legend()
plt.grid(axis='y')
plt.tight_layout()
plt.savefig("Točnost predviđanja rezultata po momčadi")
plt.show()
```

Slika 8 i Slika 9 prikazuje stupčasti graf koji prikazuje točnost predviđanja za svaku momčad. Na grafu, svaki stupac predstavlja jednu momčad, dok visina stupca odgovara točnosti predviđanja ishoda njihovih utakmica. Kako bi imena momčadi bila jasnije vidljiva i lakša za čitanje, oznake na x-osi su rotirane. Ovaj graf ilustrira proces transformacije analitičkih podataka o nogometnim utakmicama u vizualne prikaze koji jasno komuniciraju rezultate istraživanja.



Slika 8. Točnost predviđanja ishoda po momčadi prvi dio



Slika 9. Točnost predviđanja ishoda po momčadi drugi dio

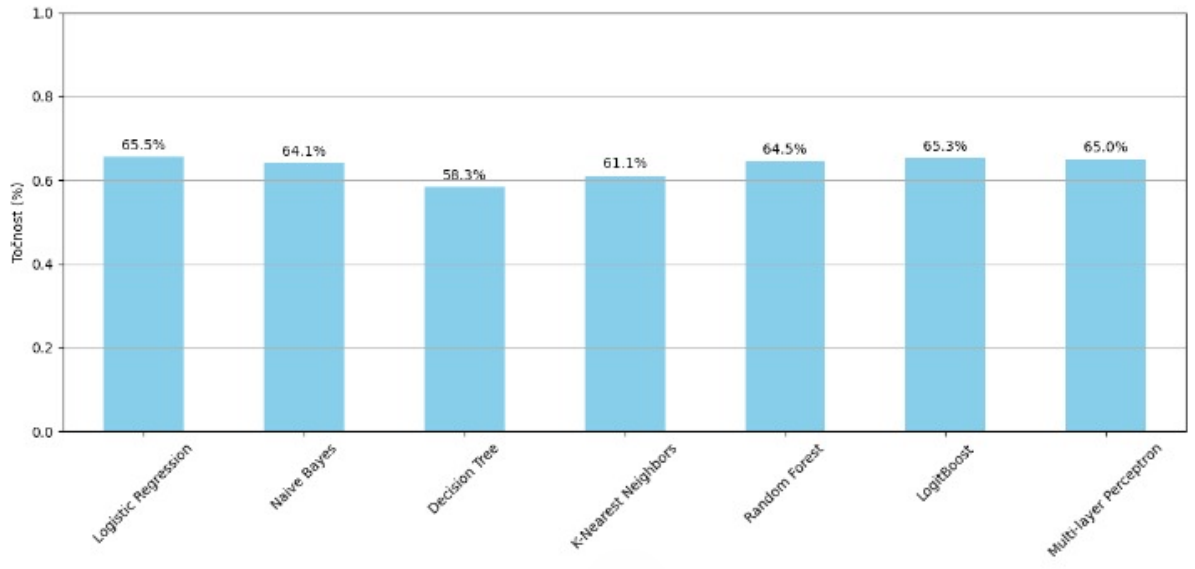
Kroz kombinaciju detaljnih analiza i efikasnih vizualnih prikaza, omogućeno je bolje razumijevanje ključnih pokazatelja performansi. Vizualizacije su olakšale interpretaciju točnosti predviđanja ishoda utakmica i pružile temelj za daljnje analitičko promišljanje. Ova metoda može poslužiti kao model za različite analitičke izazove i pruža osnovu za unapređenje strategija temeljenih na podacima unutar sportske industrije i šire.

5.3. Interpretacija rezultata

Naime, iako najbolji model postiže točnost od 65.50%, postoji značajan broj netočnih predviđanja. Ovo upućuje na potrebu za daljnjim analizama i potencijalnim poboljšanjima u modelima kako bi se povećala preciznost i smanjila pogreška u predviđanjima. Poboljšanja mogu uključivati reviziju značajki, optimizaciju parametara i testiranje alternativnih modeliranja, s ciljem ostvarivanja pouzdanijih i preciznijih rezultata.

Nadalje, prikazana je analiza rezultata dobivenih testiranjem modela strojnog učenja, koji je dizajniran za predviđanje ishoda nogometnih utakmica u Premier ligi. Tablica 3 detaljno prikazuje performanse modela za svaku od momčadi unutar lige, s posebnim osvrtom na točnost predviđanja. Ovakav format vizualizacije omogućava jasan uvid u efikasnost modela te identificiranje specifičnih trendova i anomalija u performansama različitih momčadi.

Slika 10 prikazuje ukupne performanse modela u obliku stupčastog dijagrama.



Slika 10. Detaljan prikaz točnosti algoritama

Vizualizacija jasno prikazuje razlike u točnosti različitih modela strojnog učenja. Postotci točnosti iznad stupaca omogućavaju jednostavnu usporedbu učinkovitosti svakog modela. Uočen je raspon točnosti od 58.3% do 65.5%, što sugerira da postoji prostor za unapređenje. Kroz daljnju analizu rezultata može se identificirati koji modeli imaju najviše potencijala za poboljšanje te koje su metode optimizacije najpogodnije za povećanje ukupne točnosti i smanjenje pogreške u predviđanjima.

Tablica 3. Detaljan prikaz točnosti algoritama

	Logistička regresija	**Naivni Bayes**	**Stablo odluke**	**k-NN**	**Slučajna šuma**	**LogitBoost**	**MLP**
Arsenal	49.12%	49.12%	54.39%	50.88%	50.88%	49.12%	45.61%
Aston Villa	56.90%	50.00%	56.90%	67.24%	60.34%	56.90%	55.17%
Bournemouth	71.43%	68.57%	57.14%	65.71%	68.57%	71.43%	71.43%
Brentford	58.62%	58.62%	58.62%	56.90%	67.24%	58.62%	65.52%
Brighton	58.62%	51.72%	53.45%	55.17%	60.34%	58.62%	67.24%
Burnley	77.27%	63.64%	63.64%	59.09%	63.64%	72.73%	72.73%
Chelsea	60.71%	55.36%	58.93%	50.00%	57.14%	58.93%	55.36%
Crystal Palace	68.42%	68.42%	54.39%	66.67%	66.67%	68.42%	71.93%
Everton	76.27%	72.88%	54.24%	66.10%	69.49%	76.27%	76.27%
Fulham	63.16%	63.16%	60.53%	60.53%	65.79%	60.53%	60.53%
Leeds United	77.59%	74.14%	70.69%	67.24%	74.14%	77.59%	77.59%
Leicester City	72.41%	68.97%	58.62%	67.24%	68.97%	72.41%	72.41%
Liverpool	57.89%	56.14%	43.86%	50.88%	45.61%	61.40%	50.88%
Manchester City	44.12%	58.82%	55.88%	48.53%	52.94%	44.12%	42.65%
Manchester United	50.00%	53.45%	50.00%	50.00%	51.72%	50.00%	58.62%
Newcastle United	49.12%	47.37%	54.39%	52.63%	43.86%	45.61%	43.86%
Norwich City	84.21%	84.21%	68.42%	84.21%	84.21%	84.21%	84.21%
Nottingham Forest	77.14%	77.14%	62.86%	68.57%	74.29%	77.14%	77.14%
Southampton	78.95%	77.19%	64.91%	63.16%	75.44%	78.95%	78.95%
Tottenham	47.46%	54.24%	50.85%	57.63%	52.54%	52.54%	50.85%
Watford	90.48%	90.48%	66.67%	71.43%	90.48%	90.48%	90.48%
West Ham United	68.42%	63.16%	57.89%	70.18%	70.18%	68.42%	68.42%
Wolverhampton	67.24%	67.24%	63.79%	55.17%	68.97%	67.24%	67.24%

Analizirajući rezultate predstavljene u tablici, uočene su značajne varijacije u točnosti predviđanja između različitih momčadi. Watford se ističe izuzetno visokom točnošću od 90.48%, što sugerira da su modeli vrlo uspješno predvidjeli njihovih ishoda, a to može

ukazivati na određenu predvidivost njihovih taktika ili stilova igre. Suprotno tome, najniža točnost od 42.65% zabilježena je za Manchester City, što upućuje na izazove modela u adaptaciji na varijabilnosti koje karakteriziraju njihove utakmice. Modeli su testiran na ukupno 1160 utakmica engleske Premier lige. Ove brojke dodatno ističu složenost predviđanja ishoda sportskih događaja, gdje različiti faktori poput forme momčadi, taktičkih odluka i slučajnih događaja tijekom igre mogu znatno utjecati na točnost predviđanja. Najveću točnost ostvaruje logička regresija od 65.50% što ukazuje na prilično zadovoljavajuću razinu učinkovitosti, za razliku od stabla odluke koji ostvaruje točnost od 58.30% što nam ostavlja prostor za poboljšanje, posebno u slučajevima momčadi s nižom točnošću predviđanja.

5.4. Rasprava

Analiza razlike između dva prikaza performansi modela za predviđanje ishoda nogometnih utakmica pokazuje da prvi prikaz obuhvaća ukupnu točnost za sve momčadi, dok se drugi fokusira na točnost za pojedine momčadi, s naglaskom na Watford. Istražuju se aspekti koji utječu na točnost modela i objašnjavaju razlike u učinkovitosti ovisno o karakteristikama i podacima svake momčadi. Ti aspekti jesu:

1. Maksimalna točnost modela (65.50%) ─ Ovaj postotak odražava maksimalnu točnost modela preko svih momčadi i svih testiranih utakmica. To znači da model, kad se primjenjuje na sve utakmice, u prosjeku točno predviđa ishod 65.50% vremena. Ova brojka uključuje sve moguće varijable i svu momčad, što može uključivati momčad koju je teže predvidjeti zbog različitih faktora kao što su nestabilnost performansi, česte promjene u sastavu momčadi ili nepredvidive taktike.
2. Specifična točnost za Watford (90.48%) ─ Ova visoka točnost se odnosi na predviđanja ishoda utakmica specifično za momčad Watford. Ovaj visoki postotak sugerira da su za Watford karakteristike koje modeli koriste osobito efikasne u predviđanju ishoda. Možda Watford igra na konzistentan način koji se dobro uklapa u parametre modela ili su podaci koji se koriste za treniranje modela posebno dobri u hvatanju faktora koji utječu na rezultate.

Razlika u točnosti postoji zbog:

1. Konzistencija u igranju ─ Neke momčadi poput Watforda mogu imati konzistentnije performanse ili manje varijacija u stilu igre što olakšava modelu točno predviđanje njihovih ishoda.

2. Varijabilnost ostalih momčadi ─ Momčadi s nižom točnošću predviđanja mogu imati više varijabilnosti u svojim performansama iz utakmice u utakmicu što otežava modelu da točno predvidi ishode.
3. Kvaliteta podataka ─ Razlike u kvaliteti i obujmu podataka dostupnih za različite momčadi također mogu utjecati na točnost modela.

Dva prikaza modela prikazuju različite aspekte performansi istog modela. Ukupna točnost daje općenit uvid u efikasnost modela preko svih momčadi i utakmica, dok specifična točnost za pojedine momčadi može pokazati kako modeli variraju u svojoj efikasnosti ovisno o specifičnim karakteristikama i podacima svake momčadi. Razumijevanje ovih razlika je ključno za daljnje prilagođavanje i poboljšanje modela.

Slika 10 prikazuje da modeli postižu maksimalnu točnost od 65.50% što sugerira relativno dobru sposobnost predviđanja ishoda utakmica. Ipak, značajan broj netočnih predviđanja ukazuje na potrebu za poboljšanjem preciznosti modela. To ukazuje na potrebu za daljnjim prilagođavanjem modela, prilagodbom hiperparametara ili promjenom korištenih značajki kako bi se postigla veća točnost.

Tablica 3 Prikazuje značajne varijacije u točnosti predviđanja između različitih momčadi. Na primjer, momčad Watford se ističe visokom točnošću od 90.48%, dok je točnost za Manchester City znatno niža (42.65%). Ukupno gledano, rezultati sugeriraju solidnu učinkovitost modela u predviđanju ishoda utakmica, ali ukazuju na potrebu za daljnjim analizama i poboljšanjima. Prilagođavanje modela, dodavanje novih značajki ili optimizacija parametara mogu pridonijeti poboljšanju točnosti predviđanja, čime se povećava vrijednost modela u praktičnim primjenama u sportskom klađenju ili analizi rezultata. Dobiveni rezultati također pružaju osnovu za daljnja istraživanja u primjeni strojnog učenja u sportskoj analizi.

Ovaj rad se oslanja na značajne doprinose i inovacije u polju strojnog učenja, posebno u analizi sportskih podataka. Prethodna istraživanja i analize, poput onih objavljenih od [41] i [42], pružaju temeljne koncepte i tehničke metode koje su korištene u ovom radu za obradu i vizualizaciju podataka. Biblioteke poput *pandas* i *Matplotlib*, koje su detaljno opisane u radovima [41] i [42], omogućile su efikasnu manipulaciju podacima i razvoj naprednih vizualizacijskih tehnika koje su ključne za analizu performansi nogometnih momčadi. Ove tehnike omogućuju bolje razumijevanje distribucija, udjela i trendova u podacima, čime se olakšava interpretacija složenih podatkovnih skupova te pomaže u donošenju strateških odluka unutar sportske industrije.

U konačnici, suvremeni nogomet karakterizira izuzetna dinamičnost i nepredvidivost što predstavlja izazov kako za trenere tako i za znanstvenike usmjerene na optimizaciju treninga i taktičkih odluka. U tom kontekstu, primjena strojnog učenja postaje ključna u transformaciji ogromnih količina podataka iz stvarni utakmica u korisne uvide koji mogu značajno utjecati na sportsku praksu. Kako bi se dublje razumjela trenutna primjena i potencijal strojnog učenja u nogometu, potrebno je sistematično razmotriti postojeća istraživanja koja koriste ove tehnologije. Prema tome, jedno od takvih istraživanja je studija [43] koja nudi sveobuhvatan pregled načina na koji se strojno učenje primjenjuje u analizi nogometnih podataka. Rad sistematizira rezultate različitih istraživačkih grupa koje su se fokusirale na tri ključna područja u sportu kao što su ozljede, performanse igrača i identifikaciju talenta. Autori su pregledali literaturu dostupnu do veljače 2021. godine, koristeći baze kao što su PubMed, SPORTDiscus i FECYT, uključujući različite znanstvene baze podataka kao što su Web of Sciences i MEDLINE. Ključni aspekti ovog istraživanja bile su ozljede, performanse i identifikacija talenta. Tako su autori identificirali sedam studija koje se bave predviđanjem ozljeda koristeći strojno učenje. Ove studije istražuju kako različiti algoritmi mogu efikasno predvidjeti rizik od ozljeda koristeći podatke poput prethodnih ozljeda, fizičke spremnosti i biomehaničkih testiranja. Cilj ovih istraživanja je bio omogućiti sportskim stručnjacima bolje alate za planiranje treninga i preventivnih mjera. Zatim je korištena dvadeset i jedna studija fokusirana na predviđanje sportskih performansi. U ovom segmentu, strojno se učenje koristi za analizu i predviđanje ishoda utakmica, fizioloških kapaciteta igrača te tehničkih i taktičkih aspekata igre. Na primjer, neke studije koriste modele za predviđanje ishoda utakmica na temelju prethodnih performansi momčadi, dok druge analiziraju kako individualne karakteristike igrača utječu na ishod utakmica, dok se pet studija bavilo primjenom strojnog učenja za identifikaciju i predviđanje sportskog talenta. Ova istraživanja koriste različite algoritme za analizu podataka prikupljenih tijekom natjecanja ili treninga kako bi identificirali igrače s najvećim potencijalom za uspjeh. To uključuje analizu tehničkih vještina, fizičkih atributa i psiholoških profila igrača. Ovo istraživanje zaključuje da strojno učenje nudi značajne mogućnosti za poboljšanje razumijevanja i predviđanja različitih aspekata nogometa. Studije pokazuju da, iako su trenutni modeli strojnog učenja vrlo obećavajući, postoji potreba za daljnjim istraživanjem koje će uključivati veće i detaljnije setove podataka. Također, ističe se važnost interdisciplinarnog pristupa koji kombinira znanja iz područja sportske znanosti, podatkovne analize i informatike [43].

Još jedno slično istraživanje imalo je za cilj odrediti kombinaciju značajki i klasifikatora koji će pružiti najbolju točnost predviđanja ishoda nogometnih utakmica engleske

Premier lige. Autori su predložili korištenje više klasifikatora strojnog učenja i kombinacija značajki kako bi odredili koja kombinacija daje najveću točnost predviđanja. Eksperimenti su obuhvatili različite kombinacije značajki i klasifikatora koristeći prethodne zapise utakmica za dvije sezone. Rezultati su pokazali kako kombinacija korištenja čimbenika domaće momčadi, gostujuće momčadi, prvih dvadeset dva igrača iz obje momčadi, i razlike u golovima zajedno s klasifikatorom *k*-NN strojnog učenja postiže najvišu točnost od 83,95%. Autori su usporedili najbolju kombinaciju značajki i klasifikatora s postojećim metodama predviđanja ishoda utakmica engleske Premier lige i otkrili su poboljšanje u odnosu na postojeće rezultate [44]. Uz navedeno, [45] istraživanje istražuje primjenu strojnog učenja za predviđanje ishoda profesionalnih nogometnih utakmica. Ovaj projekt se bavi nepredvidivom prirodom nogometnih utakmica, gdje ishodi nisu isključivo ovisni o broju postignutih golova zbog slučajnih elemenata u postizanju golova. Cilj je bio razviti modele strojnog učenja koji predviđaju rezultate i ishode nogometnih utakmica na temelju događaja u utakmici, a ne samo na temelju broja golova. Istraživanje je razvilo nove metrike poput očekivanih golova i kombiniralo ih s ocjenama ofenzivnih i defenzivnih sposobnosti momčadi za predviđanje budućih ishoda utakmica. Testirani su brojni modeli strojnog učenja, uključujući regresijske modele za bodovanje i klasifikacijske modele za predviđanje ishoda utakmica. Učinkovitost ovih modela uspoređena je s referentnim metodama, uključujući modele kladionica, kako bi se procijenila njihova točnost. Razvijeni modeli pružili su predviđanja s točnošću usporedivom s postojećim metodama koje koriste kladionice. Primjena strojnog učenja nudi značajan potencijal za poboljšanje točnosti predviđanja uključivanjem detaljnih podataka o događajima u utakmici umjesto oslanjanja isključivo na postignute golove. Studija je uspješno pokazala kako se strojno učenje može koristiti za pouzdanije predviđanje ishoda nogometnih utakmica fokusiranjem na detaljne događaje u utakmici i razvojem novih metričkih pokazatelja poput očekivanih golova. Budući radovi mogli bi proširiti ove tehnike, uključujući sveobuhvatnije skupove podataka i poboljšanje modela za još veću točnost [45]. Ovo istraživanje naglašava potencijal strojnog učenja u analitici sporta, posebno u poboljšanju predviđanja ishoda nogometnih utakmica prelaskom izvan tradicionalnih metrika.

Moderna primjena strojnog učenja u nogometu, kako je spomenuto u navedenim istraživanjima, pokazuje značajan napredak u predviđanju ishoda utakmica kroz sofisticirane analitičke modele. Studije kao što su [43] i [45] ističu kako napredne metrike poput očekivanih golova i detaljna analiza podataka mogu značajno poboljšati točnost predviđanja ishoda nogometnih utakmica. Takvi pristupi omogućuju dublje razumijevanje dinamike igre i pružaju ključne uvide koji mogu biti instrumentalni u taktičkom planiranju i strategiji igre.

6. Zaključak

Istraživanje usmjereno na predviđanje ishoda nogometnih utakmica u engleskoj Premier ligi pokazuje značajnu primjenu metoda nadziranog strojnog učenja i vizualizacije podataka. Kroz primjenu različitih modela strojnog učenja, demonstrirana je sposobnost preciznog prognoziranja ishoda utakmica što potvrđuje visok potencijal navedenih tehnologija u sportskoj analitici. Vizualizacije podataka su osigurale dodatni uvid u prikupljene podatke te omogućile usporedbu s prethodnim istraživanjima, ističući kako inovacije u metodama vizualizacije mogu dodatno obogatiti interpretaciju i prezentaciju rezultata.

Ovaj rad pruža novi pogled na integraciju strojnog učenja u predviđanju sportskih ishoda, pokazujući kako napredne analitičke tehnike mogu služiti kao alat za poboljšanje odluka u sportskom menadžmentu i taktičkom planiranju. Također, istraživanje ukazuje na važnost kontinuiranog prilagođavanja modela i metoda kako bi se osigurala njihova relevantnost i točnost u dinamičnom okruženju sportskih liga. Iako su rezultati obećavajući, istraživanje također otvara prostor za buduća poboljšanja, posebno u pogledu integracije većeg spektra podataka i daljnjeg rafiniranja algoritama strojnog učenja. Potencijal za proširenje ovog pristupa na druge sportske lige i različite tipove sportskih događaja naglašava univerzalnost i adaptabilnost primijenjenih metoda. Budući rad bi trebao fokusirati na testiranje različitih modela i algoritama u različitim kontekstima kako bi se unaprijedila točnost predviđanja i generalizacija modela.

U konačnici, ovaj rad postavlja čvrste temelje za daljnji razvoj u području sportske analitike i predviđanja ishoda, naglašavajući značaj multidisciplinarnog pristupa koji uključuje strojno učenje, analizu podataka, i domensko znanje sportskih disciplina. Kontinuirani razvoj i evaluacija primijenjenih metoda neizbježni su za postizanje maksimalne preciznosti i učinkovitosti, čime se osigurava da sportska analitika ostane relevantna i učinkovita u predviđanju ishoda u sve nepredvidivijem svijetu sporta.

7. Literatura

[1] A. McCabe i J. Travathan, »Artificial Intelligence in Sports Prediction,« u *Fifth International Conference on Information Technology: New Generations*, 2008.

[2] D. Buursma, »Predicting sports events from past res
ults: Towards effective betting on football matches,« u *14th Twente Student Conference on IT*, 2011.

[3] J. Hucaljuk i A. Rakipović, »Predicting football scores using machine learning techniques,« u *Proceedings of the 34th International Convention MIPRO*, 2011.

[4] F. Owramipur, P. Eskandarian i F. Mozneb Sadat, »Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team,« *International Journal of Computer Theory and Engineering,* svez. 5, br. 5, pp. 812-815, 2013.

[5] C. Igiri i E. Nwachukwu, »An Improved Prediction System for Football a Match Result,« *IOSR Journal of Engineering,* svez. 4, br. 12, pp. 12-20, 2014.

[6] N. Tax i Y. Joustra, »Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach,« *Transactions on Knowledge and Data Engineering,* svez. 10, br. 10, pp. 1-13, 2015.

[7] D. Prasetio i D. Harlili, »Predicting football match results with logistic regression,« u *International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, 2016.

[8] N. Zaveri, U. Shah, S. Tiwari, P. Shinde i L. K. Teli, »Prediction of Football Match Score and Decision Making Process,« *International Journal on Recent and Innovation Trends in Computing and Communication,* svez. 6, br. 2, pp. 162-165, 2018.

[9] J. Knoll i J. Stübinger, »Machine-learning-based statistical arbitrage football betting,« *KIKünstliche Intelligenz,* svez. 34, br. 1, pp. 69-80, 2020.

[10] J. Stübinger, B. Mangold i J. Knoll, »Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics,« *Applied Sciences,* svez. 10, br. 1, p. 46, 2020.

[11] R. Ievoli, L. Palazzo i G. Ragozini, »On the use of passing network indicators to predict football outcomes,« *Knowledge-Based Systems,* 2021.

[12] A. A. Azeman, A. Mustapha, N. Razali, A. Nanthaamomphong i M. H. A. Wahab, »Prediction of football matches results: Decision forest against neural networks,« u *8th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2021.

[13] N. Razali, A. Mustapha, N. Mustapha i F. M. Clemente, »A Bayesian approach for major European football league match prediction,« *International Journal of Nonlinear Analysis and Applications,* svez. 12, br. Special Issue, pp. 971-980, 2021.

[14] F. Rodrigues i Â. Pinto, »Prediction of football match results with machine learning,« *Procedia Computer Science,* svez. 204, pp. 463-470, 2022.

[15] Y. Ren i T. Susnjak, »Predicting Football Match Outcomes with eXplainable Machine Learning and the Kelly Index,« u *School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand*, 2022.

[16] R. M. Demartino, L. Egidi i N. Torelli, »Alternative ranking measures to predict international football results,« 2024.

[17] T. Horvat, J. Job, R. Logozar i Č. Livada, »A Data-Driven Machine Learning Algorithm for Predicting the Outcomes of NBA Games,« *Symmetry,* pp. 798-818, 2023.

[18] E. Alpaydin, Introduction to Machine Learning, Cambridge, MA, USA: MIT Press, 2020.

[19] D. Delen, D. Cogdell i N. Kasap, »A comparative analysis of data mining methods in predicting NCAA bowl outcomes,« *International Journal of Forecasting,* svez. 28, p. 543, 2012.

[20] C. Valero, »Predicting Win-Loss outcomes in MLB regular season games – A comparative study using data mining methods,« *International Journal of Computer Science in Sport,* svez. 15, pp. 91-112, 2016.

[21] T. Horvat i J. Job, »The use of machine learning in sport outcome prediction: A review,« *WIREs Data Mining and Knowledge Discovery,* svez. 10, p. e1380, 2020.

[22] D. Miljković, L. Gajić, A. Kovačević i Z. Konjović, »The use of data mining for basketball matches outcomes prediction,« u *Zbornik radova IEEE 8th International Symposium on Intelligent Systems and Informatics*, Subotica, Srbija, 10–11. rujna 2010.

[23] B. Loeffelholz, E. Bednar i K. Bauer, »Predicting NBA games using neural networks,« *Journal of Quantitative Analysis in Sports,* svez. 5, 2009.

[24] G. Zhang, »Neural networks for classification: A survey,« *IEEE Transactions on Systems, Man, and Cybernetics Part C,* svez. 30, pp. 451-462, 2000.

[25] J. Han i M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. The Morgan Kaufmann Series in Data Management Systems, Amsterdam, The Netherlands; Burlington, MA, USA; San Francisco, CA, USA: Elsevier, 2006.

[26] J. D. Kelleher i B. Tierney, Znanost o podacima, Zagreb: Mate d.o.o., 2021.

[27] E. Alpaydin, Strojno učenje, Zagreb: Mate d.o.o., 2021.

[28] F. Provost i T. Fawcett, Data Science for Business, O'Reilly Media Inc., 2013.

[29] E. E. Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, New Jersey: Wiley, 2015.

[30] A. Zulqarnain, »Data Science: Its Role and Importance,« *BS Data Science Islamia University Bahawalpur,* svez. 3, br. 3, pp. 1-16, 2024.

[31] G. James, D. Witten, T. Hastie i R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, New York: Springer, 2013.

[32] K. P. Murphy, Machine Learning: A Probabilistic Perspective, Cambridge: MIT Press, 2012.

[33] I. Goodfellow, Y. Bengio i A. Courville, Deep Learning, MIT Press: Cambridge, 2016.

[34] A. C. F. M. L. Q. Management, »Machine Learning Quality Management Guideline,« 2023.

[35] T. Hastie, R. Tibshirani i J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, New York: Springer, 2009.

[36] R. N. Mohalder, M. A. Hossain i N. Hossain, »Classifying The Supervised Machine Learning And Comparing The Performances Of The Algorithms,« *International Journal of Advanced Research,* svez. 12, br. 1, pp. 422-438, 2024.

[37] PMF, »6. Simpozij studenata doktorskih studija PMF-a,« Zagreb, 2022.

[38] *Opća uredba o zaštiti podataka,* 2016.

[39] J. Spinner, Web scraping with Python: Collecting more data from the modern web, Sebastopol: O'Reilly Media, 2018.

[40] J. McDaniel, Web Scraping with Python, Birmingham: Packt Publishing Ltd., 2020.

[41] W. McKinney, »Data Structures for Statistical Computing in Python,« u *Proceedings of the 9th Python in Science Conference*, 2010.

[42] J. D. Hunter, »Matplotlib: A 2D Graphics Environment,« *Computing in Science & Engineering,* svez. 9, br. 3, pp. 90-95, 2007.

[43] M. Rico-Gonzalez, J. Pino-Ortega, F. M. Clemente i A. L. Arcos, »Guidelines for performing systematic reviews in sports science,« *Biology of Sport,* svez. 39, br. 2, pp. 463-471, 2022.

[44] U. Haruna i J. Z. Maitama, »Predicting the Outcomes of Football Matches Using Machine Learning Approach,« *Informatics and Intelligent Applications,* svez. 1, br. 7, pp. 92-104, 2022.

[45] C. Herbinet, »Predicting Football Results Using Machine Learning Techniques,« London, 2018.

PRILOZI

Popis slika

Slika 1. Kutijasti dijagram analize rezultata ostalih istraživača. 4
Slika 2. Napredak algoritama strojnog učenja u nogometu po godinama. 7
Slika 3. Koncept nadziranog učenja..... 12
Slika 4. Koncept nenadziranog učenja..... 14
Slika 5. Koncept učenja uz podršku..... 16
Slika 6. Prikaz regresije i klasifikacije..... 19
Slika 7. Grafički prikaz metoda validacije. Slučaj a) prikazuje metoda podjele skupa, dok slučaj b) prikazuje metodu unakrsne provjere..... 20
Slika 8. Točnost predviđanja ishoda po momčadi prvi dio..... 37
Slika 9. Točnost predviđanja ishoda po momčadi drugi dio..... 37
Slika 10. Detaljan prikaz točnosti algoritama..... 38

Popis tablica

Tablica 1. Analizirani radovi poredani po godini objave. 3
Tablica 2. Korištene značajke..... 34
Tablica 3. Detaljan prikaz točnosti algoritama 39

Sveučilište Sjever

IZJAVA O AUTORSTVU

Završni/diplomski/specijalistički rad isključivo je autorsko djelo studenta koji je isti izradio te student odgovara za istinitost, izvornost i ispravnost teksta rada. U radu se ne smiju koristiti dijelovi tuđih radova (knjiga, članaka, doktorskih disertacija, magistarskih radova, izvora s interneta, i drugih izvora) bez navođenja izvora i autora navedenih radova. Svi dijelovi tuđih radova moraju biti pravilno navedeni i citirani. Dijelovi tuđih radova koji nisu pravilno citirani, smatraju se plagijatom, odnosno nezakonitim prisvajanjem tuđeg znanstvenog ili stručnoga rada. Sukladno navedenom studenti su dužni potpisati izjavu o autorstvu rada.

Ja, Mateo Vujčić *(ime i prezime)* pod punom moralnom, materijalnom i kaznenom odgovornošću, izjavljujem da sam isključivi autor/~~ica~~ završnog/~~diplomskog~~/~~specijalističkog~~ (*obrisati nepotrebno*) rada pod naslovom Upotreba metoda nadziranog strojnog učenja u predviđanju sportskih ishoda te vizualizacija podataka *(upisati naslov)* te da u navedenom radu nisu na nedozvoljeni način (bez pravilnog citiranja) korišteni dijelovi tuđih radova.

Student/~~ica~~:
(upisati ime i prezime)
Mateo Vujčić
(vlastoručni potpis)

Sukladno članku 58., 59. i 61. Zakona o visokom obrazovanju i znanstvenoj djelatnosti završne/diplomske/specijalističke radove sveučilišta su dužna objaviti u roku od 30 dana od dana obrane na nacionalnom repozitoriju odnosno repozitoriju visokog učilišta.

Sukladno članku 111. Zakona o autorskom pravu i srodnim pravima student se ne može protiviti da se njegov završni rad stvoren na bilo kojem studiju na visokom učilištu učini dostupnim javnosti na odgovarajućoj javnoj mrežnoj bazi sveučilišne knjižnice, knjižnice sastavnice sveučilišta, knjižnice veleučilišta ili visoke škole i/ili na javnoj mrežnoj bazi završnih radova Nacionalne i sveučilišne knjižnice, sukladno zakonu kojim se uređuje umjetnička djelatnost i visoko obrazovanje.