

UNIVERZA V LJUBLJANI
EKONOMSKA FAKULTETA

MAGISTRSKO DELO

**ISKANJE TALENTIRANIH IGRALCEV ZA NOGOMETNI KLUB S
POMOČJO STROJNEGA UČENJA**

Ljubljana, september 2024

ŽIGA OCVIRK

IZJAVA O AVTORSTVU

Podpisani Žiga Ocvirk, študent Ekonomske fakultete Univerze v Ljubljani, avtor predloženega dela z naslovom Iskanje talentiranih igralcev za nogometni klub s pomočjo strojnega učenja, pripravljenega v sodelovanju s svetovalcem doc. dr. Urošem Godnovom

IZJAVLJAM

1. da sem predloženo delo pripravil samostojno;
2. da je tiskana oblika predloženega dela istovetna njegovi elektronski obliki;
3. da je besedilo predloženega dela jezikovno korektno in tehnično pripravljeno v skladu z Navodili za izdelavo zaključnih nalog Ekonomske fakultete Univerze v Ljubljani, kar pomeni, da sem poskrbel, da so dela in mnenja drugih avtorjev oziroma avtoric, ki jih uporabljam oziroma navajam v besedilu, citirana oziroma povzeta v skladu z Navodili za izdelavo zaključnih nalog Ekonomske fakultete Univerze v Ljubljani;
4. da se zavedam, da je plagiatorstvo – predstavljanje tujih del (v pisni ali grafični obliki) kot mojih lastnih – kaznivo po Kazenskem zakoniku Republike Slovenije;
5. da se zavedam posledic, ki bi jih na osnovi predloženega dela dokazano plagiatorstvo lahko predstavljalo za moj status na Ekonomski fakulteti Univerze v Ljubljani v skladu z relevantnim pravilnikom;
6. da sem pridobil vsa potrebna dovoljenja za uporabo podatkov in avtorskih del v predloženem delu in jih v njem jasno označil;
7. da sem pri pripravi predloženega dela ravnal v skladu z etičnimi načeli in, kjer je to potrebno, za raziskavo pridobil soglasje etične komisije;
8. da soglašam, da se elektronska oblika predloženega dela uporabi za preverjanje podobnosti vsebine z drugimi deli s programsko opremo za preverjanje podobnosti vsebine, ki je povezana s študijskim informacijskim sistemom članice;
9. da na Univerzo v Ljubljani neodplačno, neizključno, prostorsko in časovno neomejeno prenašam pravico shranitve predloženega dela v elektronski obliki, pravico reproduciranja ter pravico dajanja predloženega dela na voljo javnosti na svetovnem spletu preko Repozitorija Univerze v Ljubljani;
10. da hkrati z objavo predloženega dela dovoljujem objavo svojih osebnih podatkov, ki so navedeni v njem in v tej izjavi;
11. da sem preveril verodostojnost informacij, ki izhajajo iz zapisov na podlagi uporabe orodij umetne inteligence.

V Ljubljani, dne _____

Podpis študenta: _____

KAZALO

1	UVOD	1
2	PREGLED LITERATURE	5
2.1	Uspeh nogometnega kluba	5
2.2	Pregled prepoznavanja in zaposlovanja nadarjenih igralcev v nogometu	6
2.3	Tradicionalne metode skavtiranja in njihove omejitve	8
2.4	Novi trendi in izzivi na področju nogometne analitike	10
2.5	Strojno učenje in statistika pri prepoznavanju talentov v nogometu	13
3	METODOLOGIJA	15
3.1	Teoretični okvir in zasnova raziskave	15
3.2	Zbiranje in priprava podatkov	16
3.3	Čiščenje in preoblikovanje podatkov	19
3.4	Raziskovalna analiza podatkov in izbira atributov	23
3.5	Razvoj in vrednotenje modela strojnega učenja	29
3.6	Multivariatna analiza	32
4	REZULTATI	33
4.1	Pregled podatkov in ugotovitev	33
4.1.1	Branilci	33
4.1.2	Napadalci	33
4.1.3	Vezisti	34
4.1.4	Vratarji	35
4.1.5	Podrobna analiza vpliva ključnih atributov	35
4.2	Ocena uspešnosti modela strojnega učenja	36
4.2.1	Branilci	36
4.2.2	Napadalci	41
4.2.3	Vezisti	45
4.2.4	Vratarji	49
4.2.5	Izboljšave	49
4.3	Rezultati multivariatne analize	50
5	RAZPRAVA	52
6	SKLEP	55

LITERATURA IN VIRI	57
---------------------------------	-----------

KAZALO TABEL

Tabela 1: Študije nogometne analitike	13
Tabela 2: Izbrani atributi za branilce.....	33
Tabela 3: Izbrani atributi za napadalce	34
Tabela 4: Izbrani atributi za veziste	34
Tabela 5: Izbrani atributi za vratarje	35
Tabela 6: Najboljša točnost modela branilcev	36
Tabela 7: Zmogljivost razreda 0 branilcev.....	37
Tabela 8: Zmogljivost razreda 1 branilcev.....	37
Tabela 9: Najboljši parametri branilcev	38
Tabela 10: Najboljša točnost modela napadalcev	41
Tabela 11: Zmogljivost razreda 0 napadalcev.....	41
Tabela 12: Zmogljivost razreda 1 napadalcev.....	42
Tabela 13: Najboljši parametri napadalcev	42
Tabela 14: Najboljša točnost modela vezistov	45
Tabela 15: Zmogljivosti razreda 0 vezistov	45
Tabela 16: Zmogljivosti razreda 1 vezistov	46
Tabela 17: Najboljši parametri vezistov.....	46

KAZALO SLIK

Slika 1: Zaslonski prikaz datotek s podatki o igralcih na platformi Kaggle	17
Slika 2: Zaslonski prikaz uporabniškega vmesnika Snowsight	18
Slika 3: Položaji na nogometnem igrišču.....	20
Slika 4: Zaslonski prikaz uporabe SQL window funkcij za izračun TEAM_SCORE.....	22
Slika 5: Zaslonski prikaz poročila Pandas Profiling po čiščenju podatkov	22
Slika 6: Zaslonski prikaz knjižnice Pandas v beležnici Jupyter.....	23
Slika 7: Zaslonski prikaz poročila Pandas Profiling pred čiščenjem podatkov	24
Slika 8: Opozorila v Pandas Profiling poročilu pred čiščenjem podatkov.....	25
Slika 9: Distribucija atributa po imputaciji z n-sosedi	25
Slika 10: Primer porazdelitve atributa odvisne spremenljivke.....	26
Slika 11: Izbrani atributi z Lassoovo regresijo in prikaz optimalne vrednosti alpha	27
Slika 12: Toplotni zemljevid korelacij med pomembnimi atributi	28
Slika 13: Sintetična generacija manjšinskega razreda.....	30
Slika 14: Primeri uporabljenih modelov in njihovih parametrov.....	30
Slika 15: Zaslonska slika matrike zmede za branilce.....	39
Slika 16: Zaslonska slika krivulje sprejemnika delovanja za branilce.....	40

Slika 17: Zaslonska slika matrike zmede za napadalce	43
Slika 18: Zaslonska slika krivulje sprejemnika delovanja za napadalce	44
Slika 19: Zaslonska slika matrike zmede za veziste	47
Slika 20: Zaslonska slika krivulje sprejemnika delovanja za veziste	48
Slika 21: Dunnov test za minute	51
Slika 22: Dunnov test za končane igre	52

SEZNAM KRATIC

angl. – angleško

fr. – francosko

ANOVA – (angl. Analysis of Variance); analiza variance

AUC – (angl. Area Under the Curve); območje pod krivuljo

CSV – (angl. Comma-Separated Values, v nadaljevanju); vrednosti so ločene z vejico

EU – (angl. European Union); Evropska unija

FIFA – (fr. Fédération Internationale de Football Association); Mednarodna nogometna zveza

FP – Napačno Pozitivni (angl. False Positives); Napačno Pozitivni

FN – Napačno Negativni (angl. False Negatives); Napačno Negativni

FPR – (angl. False Positive Rate) delež napačno prepoznanih negativnih rezultatov

KNN – (angl. K-Nearest Neighbors); k-najbližjih sosedov

OLAP – (angl. Online Analytical Processing); analitična baza

ROC – (angl. Receiver Operating Characteristic); Krivulja sprejemnika delovanja

SMOTE – (angl. Synthetic Minority Over-sampling Technique, v nadaljevanju); tehnika sintetičnega nadzorčnega manjšinskih primerov.

SQL – (angl. Structured Query Language); Strukturirani poizvedbeni jeziki

UEFA – (angl. Union of European Football Associations); Zveza evropskih nogometnih zvez

UI – (angl. User Interface); uporabniški vmesnik

TP – (angl. True Positives); Pravilno Pozitivni

TN – (angl. True Negatives); Pravilno Negativni

TPR – (angl. True Positive Rate) delež pravilno prepoznanih pozitivnih rezultatov

1 UVOD

Vrednost podatkov v današnjem svetu je neprecenljiva. Podatki so temelj procesov odločanja, bogatijo naše razumevanje in spodbujajo inovacije, ki oblikujejo našo prihodnost. Z natančnim zbiranjem in analiziranjem podatkov pridobivamo ključne vpoglede, ki spodbujajo kulturo samoizpopolnjevanja in učinkovitosti. Ta analitični pristop ne le personalizira naše izkušnje, temveč tudi širi znanje, kar omogoča družbo, ki je bolj opremljena za reševanje zapletenih izzivov. Tako so podatki na stičišču tehnologije in človeštva ter služijo kot temeljni vir, ki presega osebno izboljšanje in postaja temelj družbenega napredka (Wolf, 2010).

V dinamičnem svetu mednarodnega nogometa je svetovno prvenstvo v nogometu leta 2022 pokazalo več kot le atletske spretnosti. Služilo je kot dokaz tihe revolucije, ki jo poganja analitika velikih podatkov. Ta preobrazba, subtilna, a globoka, na novo opredeljuje pridobivanje talentov, razvoj strategij in optimizacijo uspešnosti v nogometu. Turnir je postal več kot le globalni športni dogodek, je združitev podatkov, tehnologije in človeških spretnosti. Vključevanje velikih podatkov v nogomet pomeni premik paradigme od zanašanja predvsem na instinkt in izkušnje k uporabi podatkovno podprtih strategij, ki z matematično natančnostjo analizirajo vsako podajo, sprint in taktično potezo. Ta premik vključuje velike naložbe v analizo podatkov, ki podpirajo pridobivanje igralcev, sestavo ekipe in taktiko igre. Vendar pa ta preobrazba zahteva tudi skrbno proučitev veljavnosti, pomembnosti in etičnosti podatkov. Svetovno prvenstvo 2022 je primer, kako lahko podatkovna analitika ob preišljeni uporabi preseže tradicionalne meje skavtstva (Adam, 2022).

Nastajajoča disciplina strojnega učenja, podmnžica umetne inteligence, uvaja transformativni pristop k analizi podatkov, ki sistemom omogoča samostojno učenje iz podatkov in napovedovanje. Sprejetje strojnega učenja v nogometu predstavlja pomemben paradigmatški premik, saj zagotavlja analitično globino, ki presega tradicionalne metode. Ta tehnologija omogoča analizo obsežnih zbirk podatkov o uspešnosti igralcev, kar vpliva ne le na izide tekem, temveč tudi na odločitve, povezane z zaposlovanjem igralcev, strategijo ekipe in taktičnim načrtovanjem (García-Aliaga in drugi, 2021).

Vključevanje strojnega učenja v nogomet se razširi tudi na analizo tekem v realnem času, kjer služijo podrobni statistični podatki o igralcih kot ključni kazalniki uspešnosti za trenerje in športne fiziologe. Združevanje velikih količin podatkov z zapletenimi algoritmi klubom omogoča, da pridobijo niansiran vpogled v attribute igralcev, vključno s prednostmi, slabostmi in vedenjskimi nagnjenji, kar olajša strateško načrtovanje in razvoj igralcev na ravni, kakršne še ni bilo. Ta paradigma temelji na kritičnih razmišljanjih Lee in Taylor (2004) v zvezi z oceno gospodarskega učinka svetovnega prvenstva v nogometu leta 2002, ki poudarjajo zapletenost in pomembnost takšnih dogodkov.

Medtem ko so tradicionalne metode skavtiranja že dolgo koristile nogometni skupnosti, pa uvedba strojnega učenja ponuja pomembne prednosti. Napredni algoritmi lahko zaznajo zapletene vedenjske vzorce in soodvisnosti v dejanjih, ki bi jih človeški analitiki lahko spregledali. To izpopolnjeno razumevanje dinamike igralcev in njihovo vključevanje v strategije ekip je neprecenljivo. Poleg tega so napovedne zmožnosti strojnega učenja še posebej transformativne na področju prepoznavanja talentov, saj ekipam omogočajo, da z novo stopnjo natančnosti prepoznavajo mlade talente in vlagajo vanje (Bergkamp in drugi, 2022).

Bistvo učinkovitega prepoznavanja talentov je v njegovem celostnem pristopu, ki presega zgolj tehnične attribute in fizično zmogljivost ter vključuje tudi nogometno inteligenco in psihološko odpornost. Izziv je razviti sistematično in objektivno metodo, ki se oddaljuje od tradicionalnih stereotipov in razširja iskanje talentov na obetavne igralce, ki morda ne ustrezajo običajnim profilom. V tem pogledu se strojno učenje s svojo nepristransko analizo, ki temelji na podatkih, kaže kot ključno orodje, ki bi lahko spremenilo način, kako nogometne ekipe iščejo in razvijajo prihodnje zvezde.

Področje prepoznavanja talentov doživlja velike spremembe, ki jih poganja napredek na področju strojnega učenja in podatkovne analitike. Tradicionalno skavtstvo sicer ponuja dragocene izkustvene vpogleda, vendar se spopada s subjektivnostjo ter pogosto spregleda skrite talente, ki se izmikajo standardnim pričakovanjem. Strojno učenje ponuja obetavno alternativo s svojo zmožnostjo sistematične analize velikih podatkovnih nizov in napovedovanja potencialnih talentov z uporabo napovednih modelov. Vendar ostajajo izzivi, kot so opredelitev "talenta" v merljivem smislu, izbira ustreznih atributov za analizo in vključevanje modelov strojnega učenja v obstoječe okvire skavtiranja. Poleg tega je prepoznavanje talentov povezano z velikimi ekonomskimi tveganji. Uspešno prepoznavanje in vzgajanje talentov lahko pomembno vpliva na finančno zdravje kluba in njegovo konkurenčno prednost.

Namen magistrskega dela je odgovoriti na več ključnih raziskovalnih vprašanj, da bi pospešili uporabo strojnega učenja pri prepoznavanju talentov v nogometu:

- Kaj opredeljuje "talent" v profesionalnem nogometu in kako lahko strojno učenje prispeva k njegovi merljivi oceni?
- Kateri atributi najbolj napovedujejo potencialni uspeh nogometaša in kako jih je mogoče sistematično izbrati in uporabiti v modelih strojnega učenja?
- Katere so trenutne omejitve uporabe strojnega učenja pri iskanju talentov v nogometu in katere strategije lahko te izzive premagajo?
- Na kakšen način lahko izvajanje strojnega učenja v procesih prepoznavanja talentov finančno koristi nogometnim klubom?
- Kako skavti prepoznavajo talente in kateri subjektivni dejavniki vplivajo na njihove odločitve?

- Kako lahko nogometni klubi izkoristijo strojno učenje pri pridobivanju talentov in na drugih področjih, da bi optimizirali svoje delovanje in povečali splošno konkurenčnost?

Cilj teh vprašanj je opredeliti podatke, ki lahko pomagajo razviti modele strojnega učenja za napovedovanje nogometnih talentov, s poudarkom na ključni vlogi kazalnikov uspešnosti na turnirjih z visokimi vložki. Konec koncev se raziskuje širšo uporabo podatkov, ki presega prepoznavanje talentov in vključuje tudi operativno učinkovitost in strateško odločanje.

Glavni cilj magistrskega dela je izkoristiti strojno učenje za opredelitev prepoznavanja talentov v profesionalnem nogometu. S poglobljeno analizo definicije "talenta" in opredelitvijo merljivih atributov, ki jih je mogoče oceniti s strojnim učenjem, ta raziskava uvaja raven natančnosti in objektivnosti, ki je s tradicionalnimi metodami skavtstva prej ni bilo mogoče doseči. Izbira in potrditev napovednih atributov, ki so ključne za ocenjevanje potenciala in uspeha igralcev, sta bistvenega pomena za razvoj učinkovitih modelov strojnega učenja. Ti modeli obljublajo, da bodo spremenili proces skavtstva, saj bodo z vpogledom v podatke razkrili skrite talente, namesto da bi se zanašali zgolj na subjektivno presojo. Uporaba strojnega učenja pri prepoznavanju talentov je pomembna priložnost za nogometne klube, da povečajo svoje naložbe v skavtsko dejavnost. Večja učinkovitost in natančnost prepoznavanja obetavnih igralcev klubom omogoča, da dosežejo boljši donos na svoje naložbe, zmanjšajo stroške skavtstva in zmanjšajo tveganje dragih napačnih presoj. Ta pristop ne povečuje le konkurenčne prednosti kluba na igrišču, temveč tudi njegovo finančno vzdržnost in potencial rasti zunaj igrišča.

Vključitev strojnega učenja v iskanje talentov je ključni korak k učinkovitejšim in uspešnejšim procesom. Sposobnost modelov strojnega učenja, da obdelajo in analizirajo velike nabore podatkov, daleč presega človeške zmogljivosti, kar omogoča širše in bolj raznoliko iskanje talentov. Ta operativna izboljšava omogoča klubom, da vstopijo na prej neraziskane trge in demografske skupine, kar demokratizira iskanje nogometnih talentov in razširja koristi na razvoj igralcev, pri čemer strojno učenje zagotavlja personaliziran vpogled v usposabljanje igralcev, upravljanje zdravja in optimizacijo zmogljivosti. Sprejetje strojnega učenja pri prepoznavanju talentov omogoča bolj informirano sprejemanje odločitev. S kvantifikacijo talenta in potenciala lahko klubi sprejemajo objektivne, na podatkih temelječe odločitve, ki so v skladu z dolgoročnimi strateškimi cilji. Ta metoda zmanjšuje tveganja, povezana z velikimi finančnimi naložbami v nakupe igralcev, in omogoča bolj strateško razporejanje virov. Poleg tega lahko napovedne zmogljivosti modelov strojnega učenja klubom zagotovijo konkurenčno prednost, saj prepoznajo nastajajoče trende in talente, še preden postanejo splošno prepoznavni.

Učinkovito prepoznavanje in razvoj nogometnih talentov s pomočjo strojnega učenja imata globoke gospodarske učinke. Neposredne koristi za klube vključujejo nižje stroške skavtiranja, večjo donosnost naložb v pridobivanje igralcev in boljše vrednotenje igralcev s strategijami razvoja, ki temeljijo na podatkih. Posredno pa uspešno prepoznavanje talentov prispeva k športnim uspehom, ki povečujejo prihodke zaradi denarnih nagrad, prodaje

trgovskega blaga in prestopnih pravic. Gospodarski učinek se širi tudi na lokalna gospodarstva, razvoj mladih in svetovni športni trg, kar poudarja ključno vlogo nogometa v kulturnem in gospodarskem kontekstu po vsem svetu.

Končno si ta raziskava prizadeva utreti pot, na kateri lahko sinergija strojnega učenja, statistične analize in tradicionalnih skavtskih paradigem revolucionarno spremeni prepoznavanje in zaposlovanje nogometnih talentov. Zavzema se za integriran pristop, ki ne le izpopolnjuje postopke izbire igralcev, temveč tudi povečuje konkurenčno prednost kluba, operativno učinkovitost in ekonomsko vzdržnost v sodobni dobi nogometa.

Naše magistrsko delo temelji na podatkovno vodenem izboru atributov, kar pomeni, da predhodno ne moremo določiti, kateri atributi bodo vključeni v končni model. Ta pristop zahteva, da so naše hipoteze oblikovane dovolj splošno, da ostanejo relevantne ne glede na izbrane attribute. Zato smo se odločili za oblikovanje hipotez, ki preizkušajo variabilnost pomena in porazdelitve atributov med različnimi položaji igralcev, ne da bi se osredotočili na specifične attribute. Ta metodologija nam omogoča prilagodljivost pri testiranju hipotez in zagotavlja, da so naše ugotovitve robustne in zanesljive. Hipotezi sta naslednji:

Hipoteza 1: "Statistični pomen atributov se znatno razlikuje med različnimi položaji nogometnih igralcev." Ta hipoteza predpostavlja, da bo pomen (ali statistična pomembnost) določenih atributov različen glede na položaj igralca. To omogoča prilagodljivost, saj ne določa, kateri atributi bodo pomembni, ampak navaja, da bo prišlo do variabilnosti v pomembnosti atributov med položaji. Hipotezo lahko preizkusimo tako, da proučimo, katere attribute ohrani model Lasso za različne položaje, in kako ti atributi vplivajo na merila uspešnosti modela, kot sta točnost ali območje pod krivuljo (angl. Area Under the Curve, v nadaljevanju AUC).

Hipoteza 2: "Izbrani atributi bodo pokazali različne statistične porazdelitve med položaji igralcev, kar odraža edinstvene vloge in prispevke posameznih položajev." Ta hipoteza predvideva, da se atributi, ki jih ohrani model Lasso, ne bodo le razlikovali po pomembnosti, ampak tudi po svoji porazdelitvi ali vedenju med različnimi položaji. Na primer, atributi, povezani z ofenzivnimi dejanji, bi lahko bili bolj spremenljivi ali imeli višje vrednosti za napadalce v primerjavi z branilci. Hipotezo lahko preizkusimo z uporabo raziskovalnih metod analize podatkov, kot so grafikon kvantilov ali primerjave povprečij med skupinami, in tudi z bolj formalnimi statističnimi testi, kot sta analiza variance (angl. Analysis of Variance, v nadaljevanju ANOVA) ali Kruskal-Wallis, odvisno od porazdelitve podatkov. Vsako hipotezo smo jasno opredelili in navedli specifične metode, s katerimi jo nameravamo preizkusiti. Za preverjanje hipoteze o variabilnosti pomena atributov bomo uporabili Lasso model za izbor atributov za različne igralne položaje in analizirali, kako izbrani atributi vplivajo na merila uspešnosti modela, kot sta točnost in AUC. Drugo hipotezo, ki predvideva različne statistične porazdelitve atributov, bomo preizkusili z uporabo raziskovalnih tehnik analize podatkov, kot so grafikon kvantilov in statistični testi ANOVA ali Kruskal-Wallis, odvisno od normalnosti porazdelitev. Te hipoteze neposredno podpirajo glavne cilje naše

študije, ki želijo razumeti, kako se različni atributi igralcev odražajo v njihovih igralnih položajih in vplivajo na njihovo uspešnost. Z jasnim testiranjem teh hipotez ne le odgovarjamo na osnovna raziskovalna vprašanja, ampak tudi prispevamo k širšemu razumevanju vloge podatkovne analize in strojnega učenja v športni znanosti. Ta pristop zagotavlja, da so naše ugotovitve pomembne in aplicirane, kar omogoča boljše strategije v skavtstvu in analizi igralcev na različnih položajih.

2 PREGLED LITERATURE

2.1 Uspeh nogometnega kluba

Nogomet je v 20. stoletju postal najpomembnejši šport na svetu in se je razvil v zelo donosno industrijo, v katero se vlaga veliko denarja. Samo v sezoni 2014/15 je 20 najboljših klubov ustvarilo več kot 6,6 milijarde evrov, kar je pritegnilo pozornost navijačev in znanstvenikov, ki šport analizirajo kvantitativno. Glavni cilj korporacij, vključno z nogometnimi klubi, je finančni dobiček. Vendar se ustvarjanje prihodkov po vsem svetu razlikuje zaradi različnih dejavnikov. V nekaterih ligah klubi veliko vlagajo v svoja moštva in izkoriščajo komercialne podvige, kot so pravice za televizijski prenos in prodajo blaga. Vendar ta model ni izvedljiv za klube v manj bogatih ligah, kjer je razvoj igralcev z visokim potencialom in njihov prenos v bogatejše lige bolj izvedljiva strategija za ustvarjanje dobička. Uspeh klubske strategije je odvisen od njegove skrbne umestitve v prestopni market, zlasti glede na pomembnost in attribute. Poleg igranja se uspeh profesionalnih klubov prepleta z dejavniki, kot so naložbe, družbena omrežja, usklajenost ekip in celo z navidezno nepomembnimi podrobnostmi, kot so barve ekip. V osnovi je uspeh kluba v veliki meri odvisen od sestavljanja seznama zelo nadarjenih igralcev, ki lahko uresničijo rezultate znatnih naložb. Pridobivanje talentov, pogosto s prestopi ali posojili, ki vključujejo visoke pristojbine, je pogosto, in čeprav so prestopi igralcev predmet akademskega zanimanja, je sistematičnih raziskav, ki bi povezovale dejavnosti na trgu prestopov z uspešnostjo kluba, malo (Liu in drugi, 2016).

Za razliko od številnih ameriških športov v evropskem nogometu ni sistema zaposlovanja, ki bi kvalificirane igralce enakomerno porazdelil med moštva. Namesto tega so igralci v Evropi zaposleni v klubih z možnostjo prestopa v drug klub za odškodnino pred iztekom pogodbe. Takšna ureditev omogoča dinamičen sistem prestopov igralcev, ki temelji na medsebojnih dogovorih in finančnih pogajanjih. Kljub temu se je evropski nogometni prostor v času pandemije soočil z velikimi gospodarskimi izzivi, kar je povečalo finančno ranljivost klubov in nevarnost stečaja. Ti izzivi so povzročili tudi splošno zmanjšanje prestopnih provizij v panogi, kar je pomenilo velik pritisk na klube, da znova ocenijo in prilagodijo svoje finančne strategije (Wand, 2022).

V angleški premier ligi, vrhu angleškega nogometa, so glavni stroški plače igralcev in prestopni stroški. To predstavlja edinstveno priložnost za športne ekonomiste, da neposredno ocenijo mejni produkt igralčeve uspešnosti. Čeprav so začetne raziskave temeljile na

nemških podatkih, je vključitev statistike v angleški nogomet spodbudila vplivne empirične študije. Te študije so za modeliranje prestopov uporabile podatke Premier League in ugotovile, da so starost, odigrane tekme v prejšnji sezoni, doseženi goli in odigrane mednarodne tekme pomembni dejavniki, ki določajo zneske prestopov. Sodba Bosman iz leta 1995, ki je povečala prožnost prestopov v državah članicah Evropske unije (angl. European Union- EU), je pomembno vplivala na zaposlovanje tujih igralcev v Premier League. Da bi preprečila ta trend, sta Zveza evropskih nogometnih zvez (angl. Union of European Football Associations - UEFA) in kasneje Premier League leta 2008 oziroma 2010 uvedli politiko, ki je od ekip zahtevala vključitev minimalnega števila domačih igralcev, da bi okrepili uspešnost nacionalnih reprezentanc. Čeprav teoretično delo kaže, da prožnejši trgi dela vodijo k daljšim pogodbam in višjim povprečnim tržnim vrednostim, je treba to še temeljito preveriti pri nogometnih plačah. Študije o vplivu prožnosti trga dela na plače nogometašev so dale mešane rezultate, ki kažejo, da je vpliv "domačih" igralcev na plače po prilagoditvi za stroške plač minimalen. Vendar pa bi lahko vključitev ekip, ki niso iz Premier League, v te študije izkrivljala rezultate, saj je odločitev o Bosmanu bolj vplivala na Premier League kot na druge lige. Kljub temu vloge domačih igralcev ne bi smeli podcenjevati v študijah, v katerih se naslov obravnava kot odvisna spremenljivka pri vrednotenju prestopnega marketa. Omejitve sedanje raziskave so predvsem posledica uporabe statističnih podatkov, ki posredno odražajo igralčevo produktivnost in ne njegove dejanske vrednosti za ekipo. Vendar je napredek v tehnologiji sledenja igralcev v zadnjem času omogočil objavo podrobnejših statističnih podatkov (Preston, 2012).

2.2 Pregled prepoznavanja in zaposlovanja nadarjenih igralcev v nogometu

Trenutna akademska prizadevanja za opredelitev in merjenje nadarjenosti v športu običajno iščejo splošno sprejemljivo opredelitev, vendar pogosto ne morejo obravnavati zapletenosti. Običajno se nadarjenost kategorizira kot fiziološka, na primer vzdržljivost ali moč, ali psihološka, na primer odločnost ali tekmovalnost. Te vrste nadarjenosti se merijo z metodami, ki lahko količinsko opredelijo, koliko teh atributov ima športnik, na primer, kako hitro lahko teče ali koliko lahko fizično zdrži. Vendar je uporaba teh meritev za ugotavljanje nadarjenosti zahtevna. Takšne fizične in psihološke attribute se namreč lahko med športnikovim razvojem, zlasti v obdobju odraščanja, ki ne poteka po ravni poti in nanj vplivajo njegove pretekle dejavnosti ter številni drugi dejavniki, značilni za vsakega posameznika, zelo razlikujejo (Miller in drugi, 2015).

Nekateri raziskovalci menijo, da so psihološki atributi morda boljši kazalniki nadarjenosti, saj vrhunski športniki pogosto kažejo večjo stopnjo predanosti in tekmovalnosti. Vendar je razumevanje povezave med temi atributi in športnim uspehom zapleteno in je odvisno od zrelosti športnika, zahtev športa in njegovih duševnih sposobnosti. Ključno spoznanje zadnjih študij je, da nadarjenost ni statičen atribut, temveč dinamična in razvijajoči se atribut, ki se spreminja od osebe do osebe in skozi čas. V nasprotju s tradicionalnim stališčem, da je nadarjenost prirojen atribut, raziskave poudarjajo, da ugotavljanja nadarjenosti ne smemo

obravnavati ločeno od razvoja nadarjenosti. Nasprotno, ta dva procesa sta medsebojno povezana in nenehno vplivata drug na drugega. Prepoznavanje in razvijanje nadarjenosti sta stalen, razvijajoč se proces, ki ni odvisen le od genetskih ali prirojenih sposobnosti, temveč od tega, kako se športniki razvijajo in rastejo skozi čas (Miller in drugi, 2015).

V akademskih razpravah se nadarjenost v športu pogosto obravnava kot posebna, naravna sposobnost, ki bistveno poveča možnosti za uspeh. Raziskovalci pa menijo, da so za doseganje izjemnih ravni spretnosti v športu ključnega pomena resnični ustvarjalci razlik, to so prirojeni talenti ali darovi. Za uvrstitev med 10 % najboljših športnikov na primer običajno ni potrebna le prirojena nadarjenost, temveč tudi ustrezen trening in ugodni pogoji. Proces prepoznavanja nadarjenosti pomeni zgodnje odkrivanje teh naravnih sposobnosti pri športnikih in napovedovanje njihovih možnosti za uspeh v prihodnosti. Pri tem ne gre le za prepoznavanje trenutnih sposobnosti, temveč za predvidevanje, kdo bi lahko postal elitni odrasli športnik. Izziv v svetu športa, zlasti v nogometu, je pomanjkanje univerzalne opredelitve "talenta", kar otežuje postopek. Izrazi, ki se uporabljajo za opisovanje ravni nadarjenosti, kot sta "elita" in "super elita", pogosto niso jasno opredeljeni, kar še povečuje zmedo. Razlike v tem, kako šport prepozna in neguje nadarjenost, lahko prav tako povzročijo razlike. Na primer, mladega nogometaša lahko opazujejo tako na ravni akademije kot na ravni nacionalne ekipe, medtem ko je atlet lahko deležen pozornosti le na nacionalni ravni. Ta neskladja poudarjajo potrebo po natančnejši taksonomiji talenta v nogometu, da bi bolje upravljali postopke prepoznavanja in razvoja. Takšen strukturiran pristop bi pomagal razjasniti zapletenost nadarjenosti v nogometu in si prizadeval izboljšati način odkrivanja in vzgoje mladih talentov (Larkin in Reeves, 2018).

Trenutna spoznanja v športni znanosti kažejo, da bi lahko osredotočanje samo na fizične attribute zgodnjih zrelih športnikov pri prepoznavanju talentov omejilo ocenjevanje na trenutne ravni uspešnosti in nehote spregledalo športnike, ki dozorijo pozneje. Študije so proučevale, kako lahko agilnost in sprinterski čas pomagata pri razlikovanju med bolj in manj spretnimi mladostniškimi nogometaši. Vendar se zdaj zagovarja širši pristop, ki priznava, da je uspešnost v nogometu posledica medsebojnega delovanja več sistemov. Pomembna študija, ki je ugotovila zmerno povezanost med tehničnimi spretnostmi in uspešnostjo igre, nakazuje, da povezovanje različnih meril uspešnosti omogoča celovitejši pogled na igralčeve sposobnosti. Tradicionalno je bilo prepoznavanje nadarjenosti v veliki meri odvisno od subjektivnih ocen trenerjev ali skavtov, kar je pogosto vodilo do nedoslednih in pristranskih rezultatov. Prehod na znanstveno utemeljeno metodologijo, ki vključuje fiziološke, antropometrične, psihološke, sociološke in tehnične ocene, naj bi zagotovil objektivnejši in celovitejši pogled na nadarjenost. Znanstveni pristop je široko podprt, vendar je njegovo izvajanje v praksi še vedno izziv zaradi dinamične narave razvoja nadarjenosti, ki se razvija z rastjo in učenjem športnikov. Ta proces zahteva nenehno ocenjevanje, pri čemer se poudarja dolgoročni razvoj športnika in ne kratkoročni. Različni nogometni klubi lahko dajejo prednost določenim atributom, zato so potrebni prilagojeni modeli prepoznavanja talentov, ki odražajo te preference in so občutljivi na razvojne stopnje

mladih igralcev. Strokovnjaki opozarjajo, da so zgodnje statične meritve potenciala nezanesljive v primerjavi s stalnim opazovanjem in interakcijo. Trenerji imajo v tem procesu ključno vlogo, saj njihovo dožemanje nadarjenosti in razvoja vpliva na to, kako vzgajajo mlade športnike. Pri prepoznavanju obetavnih talentov se pogosto zanašajo na svojo izkustveno presojo, ki jo ubeseduje fraza "Ko vidim, vem, da je". Ta izkustveni pristop je v nasprotju z bolj togimi, k atributom usmerjenimi metodami, ki prevladujejo v nekaterih akademskih krogih. Trenerji poudarjajo pomembnost igralčeve prilagodljivosti in odzivnosti na treninge, ki lahko bistveno vpliva na njegovo razvojno pot. Če povzamemo, sta prepoznavanje in razvoj športnih talentov vse bolj prepoznana kot kompleksna, dinamična procesa, ki jima koristi celosten, večplasten pristop. Ta perspektiva izpodbija tradicionalne metode, ki se osredotočajo zgolj na fizične ali psihološke attribute, in namesto tega zagovarja model, ki združuje različne vidike športnikovih zmogljivosti in razvojnega potenciala (Unnithan in drugi, 2012; Miller in drugi, 2015).

2.3 Tradicionalne metode skavtiranja in njihove omejitve

Prepoznavanje talentov v nogometu je zapleten in strateški proces, v katerem imajo skavti ključno vlogo pri odkrivanju prihodnjih zvezd. V nasprotju s trenerji, ki se osredotočajo na spodbujanje dolgoročnega razvoja, je naloga skavtov opazovanje in ocenjevanje igralcev, predvsem tistih, ki še niso vključeni v vrhunske akademije, da bi napovedali njihov potencial za prihodnjo odličnost. Skavti ocenjujejo igralce med preizkušnjami ali tekmami in se osredotočajo na tiste, ki imajo obetavne attribute. Kljub njihovim prizadevanjem je natančno napovedovanje prihodnje uspešnosti zahtevno in odvisno od več dejavnikov. Skavti morajo določiti ustrezno starostno skupino za ocenjevanje, opredeliti ključne nogometne attribute, ki lahko napovedujejo uspeh, in celostno presoditi o prihodnji uspešnosti igralca. Raziskave kažejo na neskladje med starostnimi skupinami, na katere se osredotočajo skavti, in njihovim zaupanjem v točne napovedi. Skavti menijo, da lahko najbolj zanesljivo napovedujejo uspešnost od starosti približno 13,6 leta naprej. Vendar pogosto začnejo raziskovati igralce, ki so mlajši od 12 let, saj se zavedajo, da zgodnji znaki prihodnjega uspeha pri tako mladih letih morda niso očitni ali jih ni mogoče zlahka oceniti. Pri ocenjevanju potenciala igralcev skavti poudarjajo različne attribute, ki jih razvrščajo v tehnične spretnosti, igralno inteligenco, fiziološke sposobnosti in psihično trdnost. Tehnične spretnosti veljajo za najbolj ključne, vendar se cenijo tudi taktično razumevanje in kognitivne sposobnosti, sledijo pa fizični atributi in osebni atributi. Kljub strukturiranemu pristopu se lahko metode skavtov za ocenjevanje teh atributov zelo razlikujejo. Številni se pri ocenjevanju igralčevega potenciala zanašajo na splošni vtis, namesto da bi uporabili sistematičen postopek odločanja, ki temelji na pravilih. Ta subjektivni pristop pogosto privede do različnih napovedi glede igralčevega prihodnjega uspeha. Da bi povečali natančnost in zanesljivost prepoznavanja talentov, je bistvenega pomena, da bi skavti koristili usposabljanje, ki poudarja uporabo strukturiranih ocen. Poleg tega bi razvoj posebnih meril za ocenjevanje prihodnje uspešnosti in sistematična uporaba teh meril lahko izboljšala postopek napovedovanja. Na splošno je za izboljšanje tega, kako skavti prepoznavajo in ocenjujejo mlade talente, potreben bolj

standardiziran pristop, ki jim omogoča ne le učinkovitejše prepoznavanje potenciala, temveč tudi napovedovanje, ki je dosledno in zanesljivo v različnih skavtih in kontekstih (Bergkamp in drugi, 2022).

Metode za ocenjevanje stanja zrelosti lahko bistveno izboljšajo razlago podatkov telesnih, fizioloških in tehničnih testov v mladinskih športnih programih. Te metode pomagajo zagotoviti, da tehnično spretni, a telesno manj zreli igralci ne bodo spregledani zaradi predsodkov, povezanih z zrelostjo, ki bi sicer lahko ovirali prepoznavanje nadarjenih igralcev z velikim potencialom. Trenerje spodbujamo, da taktičnih odločitev, kot je dodeljevanje igralnih mest, ne sprejemajo le na podlagi telesnih atributov, temveč upoštevajo širši nabor atributov, saj samo ti telesni atributi morda ne napovedujejo uspeha na višjih ravneh tekmovanja. Dolgoročni pristop k razvoju igralcev je ključnega pomena. Njegov cilj je z zagotavljanjem enakih možnosti in kakovostnega usposabljanja zadržati širok nabor nadarjenih igralcev. Ta pristop priznava, da lahko mlajši ali telesno manj zreli igralci dosežejo uspešne rezultate, če dobijo priložnost za razvoj v okviru podpornega programa. Nacionalne nogometne zveze, kot so tiste na Škotskem, Irskem in v Kanadi, so na primer sprejele model dolgoročnega razvoja igralcev, ki spodbuja celovit pristop k vzgoji mladih talentov. Poleg tega lahko na način izbire igralcev vpliva porazdelitev datumov rojstva, pojav, znan kot "učinek relativne starosti". To pristranskost je treba skrbno upravljati, da bi spodbudili sodelovanje brez pretiranega pritiska in zagotovili pravičnost. Kar zadeva fiziološke in tehnične ocene, je ključnega pomena, da so ti testi specifični za posamezne športe ter ustrezajo starosti in znanju mladih igralcev. Pri časovnem razporedu teh ocenjevanj je treba upoštevati tudi razlike znotraj športnih sezon in med njimi, da se zagotovita natančnost in ustreznost. Učinkoviti testi za razlikovanje med bolj in manj spretnimi mladimi igralci in napovedovanje prihodnjih uspehov vključujejo meritve sprinta, moči nog, agilnosti, maksimalne anaerobne moči, vzdržljivosti in sposobnosti driblinga z žogo. Poleg tega lahko z vključitvijo večdimenzionalnega pristopa, ki vključuje psihološke vprašalnike, kognitivno-percepcijske ocene in ocene v odprtem okolju, zagotovimo celovitejše razumevanje igralčevih sposobnosti. Na skavtskih dnevih bi razvrščanje igralcev po velikosti, datumu rojstva ali stopnji zrelosti lahko omogočilo pravičnejše primerjave in natančnejše ocene. Nazadnje je priporočljivo, da se odločitve glede akademskega zadržanja med puberteto in izbire za profesionalne pogodbe odložijo na čas po tej kritični razvojni fazi, da se omogočijo pravične možnosti in ustrezno ocenjevanje potencialov (Meylan in drugi, 2010).

Iskanje talentov je ključni vidik zaposlovanja vrhunskih športnikov v profesionalnem športu, pri čemer se precej razlikujejo pristopi v različnih športih in regijah. V športih, kot je nogomet, se iskanje talentov v veliki meri opira na človeško strokovno znanje in mreže, ki vključujejo skavte, agente Mednarodne nogometne zveze (fr. Fédération Internationale de Football Association, v nadaljevanju FIFA) in agente igralcev, medtem ko imajo športi, kot sta bejzbol in košarka, daljšo tradicijo uporabe statističnih podatkov pri sprejemanju odločitev. Nedavni napredek je uvedel nove tehnologije, kot so programska oprema za

analizo igralcev in večpredstavnostne spletne zbirke podatkov, ki napovedujejo uspešnost igralcev in izide tekem v različnih pogojih. Vendar se kljub potencialu teh novih medijskih orodij za izboljšanje procesa skavtstva njihova uporaba zelo razlikuje. Medtem ko na primer ameriški športi vse bolj vključujejo te tehnologije, italijanski nogometni sistem ostaja neodločen in se bolj zanaša na tradicionalne metode ter strokovno znanje izkušenih strokovnjakov iz igrskega in trenerskega okolja. Ta zadržanost deloma izhaja iz neskladja med poznavanjem tehnologije s strani starejše generacije in spretnostjo mlajših športnikov pri uporabi družbenih medijev za prepoznavnost. Čeprav nove tehnologije ponujajo možnosti za izpostavljanje igralcev in krepitev blagovne znamke, številni v nogometni industriji počasi prepoznavajo njihov potencial pri pridobivanju talentov. Takšno stanje poudarja, da morajo nogometni klubi razviti vodstvene sposobnosti, ki vključujejo nove tehnologije in izboljšujejo procese odločanja pri zaposlovanju. Vendar pa ta digitalna in večpredstavnostna orodja veljajo za dopolnilo in ne za nadomestilo tradicionalnih metod iskanja. Zdi se, da je združevanje vodstvenih, statističnih in igralnih izkušenj ključno za ohranjanje zanesljivega sistema iskanja talentov (Radicchi in Mozzachiodi, 2016).

Poleg tega se nogometne akademije, ki so bistvenega pomena pri vzgoji mladih talentov, pogosto obravnavajo kot prestižna učna okolja, ki pomagajo utrditi igralčevo zavezanost k nadaljevanju nogometne kariere. Vendar pa to okolje ustvarja tudi pritisk, zaradi katerega je samopodoba igralcev krhka zaradi nenehnega pričakovanja izboljšav. Če igralci ne izpolnijo teh standardov, to lahko ogrozi njihovo identiteto nadarjenih posameznikov. Obravnava teh identitetnih ranljivosti, obvladovanje primerjav z vrstniki, ohranjanje identitete po izločitvi in spodbujanje več identitet so ključna področja za prihodnje raziskave v programih razvoja nogometnih talentov (Clarke in drugi, 2018).

2.4 Novi trendi in izzivi na področju nogometne analitike

Analiza ekipnih športov, kot je nogomet, je zapletena in zahteva kombinacijo različnih vrst podatkov in analitičnih metod. Vključuje razumevanje učinkov strategij in taktik na gibanje igralcev in uspešnost ekipe, pri čemer je treba upoštevati tako tekmovalnost kot sodelovanje v okviru sklopa pravil igre. Da bi to opravili učinkovito, analitiki zbirajo različne podatke, od fizičnih gibov, zajetih z optičnim prepoznavanjem in lokalnimi sistemi za določanje položaja, do video in senzoričnih podatkov. Ti podatki vključujejo tudi okoljske attribute, kot sta lokacija tekme in vreme. Vključevanje tehnoloških orodij se je v zadnjih letih močno razširilo. Ta orodja združujejo podatke iz različnih virov, kar zagotavlja celovit pristop k analizi ekipnih športov. Ti podatki so pogosto zelo razsežni in vključujejo besedilo, slike in videoposnetke, pri čemer imajo vsak svoje specifikacije glede velikosti, natančnosti in ločljivosti. Pri tehnološko naprednejših pristopih se gibanje igralcev zajame s senzorji, pritrjenimi na igralce ali opremo, kar omogoča prenos podatkov v realnem času. Poleg podatkov o gibanju in dogodkih imajo pri športni analitiki ključno vlogo tudi opisne statistike. Te statistike, ki vključujejo attribute, kot sta pogostost podaj in hitrost igralcev, se pogosto še vedno zbirajo ročno, kljub napredku tehnologij za samodejno beleženje podatkov.

Na splošno je povezovanje različnih vrst podatkov in analitičnih metod temelj učinkovite analize ekipnih športov, ki pomaga izboljšati trenerske strategije in uspešnost igralcev. Nenehni razvoj tehnologij za analizo podatkov obljublja nadaljnji napredek pri razumevanju in napovedovanju rezultatov v športu (Stein in drugi, 2017).

Strojno učenje ponuja velik potencial za preoblikovanje tradicionalnih metod opazovanja pri analizi nogometa z zagotavljanjem avtomatiziranih, kvantitativnih vpogledov. Ta tehnologija obdeluje predvsem dve vrsti vhodnih podatkov: podatke o dogodkih in podatke o sledenju. Analiza podatkov o dogodkih se osredotoča na prepoznavanje vzorcev moštva, določanje ključnih kazalnikov uspešnosti in napovedovanje rezultatov. Z vrednotenjem diskretnih dogodkov med tekmo, kot so podaje, preigravanja in streli, lahko algoritmi strojnega učenja odkrijejo trende in dejavnike uspeha, ki vplivajo na uspešnost ekipe. Za razliko od podatkov o dogodkih so podatki o sledenju procesno usmerjeni in obravnavajo podatke o neprekinjenem gibanju igralcev med igro. Ta vrsta analize vključuje ocenjevanje učinkovitosti podaj in napovedovanje verjetnosti doseganja golov. Vendar hitro in nepredvidljivo gibanje igralcev skupaj z prekinitvami v igri predstavlja izziv pri ohranjanju natančnosti podatkov o sledenju. Ti zapleteni sistemi, ki pogosto delujejo kot "črne škatle", ponujajo le malo vpogleda v to, kako so sklepi oblikovani, kar otežuje delo trenerjem in praktikom, ki imajo morda raje bolj preprosto analitiko. Opazno je tudi pomanjkanje sodelovanja med ključnimi zainteresiranimi stranmi, kot so podatkovni znanstveniki, športni znanstveniki, analitiki tekem in trenerji, kar omejuje praktično uporabo rezultatov raziskav. Večina študij se osredotoča na napovedovanje tekem, kar se ne odraža vedno učinkovito v uporabnih strategijah za ekipe. Učinkovito sodelovanje med analitiki strojnega učenja, športnimi znanstveniki in nogometnimi trenerji je ključnega pomena za izpopolnjevanje algoritmov, ki bolje obdelujejo položajne podatke in zagotavljajo vpogleda, ki presegajo sposobnosti človeškega opazovanja. Povezovanje multidisciplinarnih ekip je bistvenega pomena za pretvorbo zapletenih rezultatov strojnega učenja v praktične strategije, ki jih trenersko osebje zlahka razume in izvaja. Ta pristop ne le povečuje natančnost in uporabnost nogometne analitike, temveč tudi zagotavlja, da je tehnološki napredek utemeljen v praktični realnosti športnega treninga in upravljanja (Herold in drugi, 2019).

Zbirke podatkov FIFA so se izkazale za obetavne pri napovedovanju izidov nogometnih tekem, saj so pogosto enako uspešne ali boljše od drugih virov podatkov o nogometu. Zlasti so bili ti podatkovni nizi učinkovito uporabljeni v eksperimentalnih postavitvah za prikaz prednosti nelinearnih metod pred tradicionalnimi pristopi za napovedovanje tržne vrednosti nogometašev. Ta raziskava ima posledice, ki presegajo področje videoiger, saj kaže na širšo uporabo v športni analitiki (Al-Asadi in Tasdemir, 2022).

V okviru videoigre FIFA, zlasti v načinu Ultimate Team, igralci upravljajo svoje ekipe s pomočjo različnih kart za tekme. Raziskava poudarja pomen določanja objektivnih cen v igri za igralce, saj trdi, da je trgovanje v igri boljša alternativa kot poraba pravega denarja za virtualna sredstva. Te ugotovitve so pomembne tudi za dejanska pogajanja med nogometnimi klubi in igralskimi agenti, saj zagotavljajo podatkovno podprto podlago za

racionalizacijo in objektivizacijo ocen tržne vrednosti igralcev. Inovativne so metode za ocenjevanje vrednosti igralcev z uporabo nabora podatkov igre FIFA 20. To vključuje uporabo algoritma za grozdenje za segmentacijo nabora podatkov v kategorije, kot so vratarji, napadalci, branilci in vezisti. Za vsako kategorijo se nato opravi regresijska analiza, da se ugotovijo in uporabijo najpomembnejši atributi pri napovedovanju tržnih vrednosti igralcev. Ta metoda poudarja novost vključevanja samodejnega grozdenja v fazi usposabljanja za kategorizacijo igralcev in izpopolnjevanje ocen vrednosti, ki je dodatno okrepljena s hevristično izbiro atributov za odpravo odvečnih atributov (Behravan in Razavi, 2021).

Zanimivo je poudariti raziskavo, ki razširja širšo uporabo strojnega učenja v športu, pri čemer se osredotoča na razumevanje in napovedovanje izidov tekem. Različne tehnike strojnega učenja ponujajo različne vpogleds v odnose med atributi igre in njihov vpliv na rezultate tekem. Na primer, naivni algoritem Bayes in pristop k-najbližjih sosedov (angl. K-Nearest Neighbors, v nadaljevanju KNN) zagotavljata temeljno razumevanje, vendar se razlikujeta po izgradnji modela in prispevku k razumevanju. Nasprotno pa bayesovska omrežja, zgrajena s strani strokovnjakov, kljub morebitnim pristranskostim ponujajo podrobne modele medsebojnih povezav atributov in so učinkovita tudi pri omejenih podatkih za usposabljanje (Joseph in drugi, 2006).

Druga raziskava obravnava izzive napovedovanja izidov nogometnih tekem z razvojem specializirane programske rešitve. Ta sistem, preizkušen za optimizacijo kombinacij atributov in klasifikatorjev, se je izkazal za zelo učinkovitega, saj je preseegal svoje cilje glede točnosti in pokazal potencial podrobnega modeliranja v sezoni za povečanje točnosti napovedi. Priporočilo za razširitev nabora podatkov za usposabljanje poudarja potencial za nadaljnje izpopolnjevanje te napovedne zmogljivosti (Hucaljuk in Rakipović, 2011).

Na splošno te študije poudarjajo pomemben potencial uporabe naprednih tehnik analize podatkov in strojnega učenja za izboljšanje razumevanja in napovedovanja dinamike nogometne igre in tržnih vrednosti igralcev, kar ima praktične posledice za virtualna in realna nogometna okolja. Strojno učenje spreminja način razumevanja kolektivnega vedenja v ekipnih športih, pri čemer se osredotoča na interaktivno dinamiko igralcev, kakršna je značilna za nogomet. Ta pristop presega tradicionalne individualne attribute, kot so fiziološki podatki, in se ukvarja s tem, kako se igralci kolektivno odzivajo na različne situacije v igri, vključno z različnimi igralnimi položaji in fazami tekme. Z uporabo metodologije, ki jo je navdihnila fizika, za analizo podatkov o položaju, raziskovalci odkrivajo zapletene vzorce obnašanja ekipe, ki so močno povezani z uspehom ekipe. Ta pretanjena analiza razkriva vloge, ki jih imajo posamezni igralci v kolektivni dinamiki, in kaže, kako so njihovi gibi in odločitve na igrišču medsebojno povezani. Ugotovitve teh študij ne izboljšujejo le taktičnega razumevanja igre, temveč se praktično uporabljajo tudi na področjih, kot so iskanje talentov, preprečevanje poškodb in optimizacija zmogljivosti. S količinsko opredelitvijo podobnosti in razlik v vzorcih gibanja med soigralci in nasprotniki razviti atributi ponujajo nova spoznanja, ki so lahko podlaga za strateške odločitve trenerjev, skavtov in osebja ekip med

tekami in na trgu prestopov. Poleg tega se raziskava povezuje s širšimi študijami o kolektivnem vedenju, kot so tiste, ki jih opažamo pri živalskih skupinah, in bogati naše razumevanje temeljnih pravil, ki urejajo interakcije v kateri koli skupini ali ekipi. Ta interdisciplinarni vpogled je neprecenljiv ne le v športu, temveč v vsaki analizi kolektivnega vedenja. Povezovanje športne znanosti in računalništva je bilo ključnega pomena za razvoj tega področja. Vsako področje prinaša edinstvene poglede in metodologije, kar kaže na potrebo po cikličnem sodelovanju, ki krepi obe področji. Športna znanost prispeva z oblikovanjem ustreznih raziskovalnih vprašanj in preizkušanjem teoretičnih konstruktov, računalništvo pa ponuja izpopolnjena analitična orodja in zmogljivosti za obdelavo podatkov, ki so potrebni za reševanje teh zapletenih problemov. Vendar pa se pri tem interdisciplinarnem sodelovanju pojavljajo pomembni izzivi, ki so povezani predvsem z obravnavo in obdelavo velikih nizov podatkov, upravljanjem časovnega združevanja in pridobivanjem uporabnih vpogledov iz velike količine prostorskih podatkov, zbranih med igrami. Kljub tem izzivom sinergija med športno znanostjo in računalništvom spodbuja inovativne pristope v športni analitiki. Ta skupna prizadevanja ne pospešujejo le akademskih raziskav, temveč se kažejo tudi v praktičnih strategijah, ki lahko pomembno vplivajo na športno uspešnost v resničnem svetu (Marcelino in drugi, 2020; Goes in drugi, 2021).

2.5 Strojno učenje in statistika pri prepoznavanju talentov v nogometu

Vloga strojnega učenja in napredne analitike v športu se je eksponentno povečala, kar omogoča poglobljene vpoglede v zmogljivost igralcev, taktično odločanje in strateško načrtovanje. Tabela 1 povzema več ključnih študij, ki poudarjajo raznolike aplikacije in preobrazbeni potencial teh tehnologij v športu. Vsaka študija prikazuje premik proti bolj znanstveno rigoroznemu in pristopu na podlagi podatkov (angl. data-driven), ki izboljšuje različne vidike športnih disciplin z uporabo zapletnih analitičnih metod.

Tabela 1: Študije nogometne analitike

Avtor študije	Glavne ugotovitve
Almulla in Alam (2020)	Uporabili so logistično regresijo za napovedovanje izidov nogometnih tekem s točnostjo nad 80 %, osnovano na kazalnikih uspešnosti, kot so število strellov na gol, uspešne podaje in tekaške razdalje. Ta študija poudarja vpliv podatkovno vodenih strategij na upravljanje in uspešnost športnih ekip.

se nadaljuje

Tabela 1: Študije nogometne analitike (nad.)

Baboota in Kaur (2019)	Pregledali so napovedne modele v angleški Premier ligi, ki kažejo, kako modeli, kot so Gradientno povečanje, Naključni gozd in metoda Podpornih vektorjev, optimizirajo napovedi izidov tekem z učinkovitim izborom primernih atributov in modelov, vključno z ocenami FIFAIndex.com in gol razliko.
García-Aliaga in drugi (2021)	Uporabili so strojno učenje za klasifikacijo položajev igralcev na podlagi zgolj njihovega tehničnega in taktičnega obnašanja, kar dokazuje sposobnost umetne inteligence za izboljšanje strateškega pozicioniranja in razvoja igralcev brez uporabe prostorsko-časovnih podatkov.
Hughes in Franks (2008)	Študija poudarja pomembnost povratnih informacij pri športni zmogljivosti in zagovarja napredne metode zbiranja podatkov, kot so razsevni diagram in frekvenčne tabele. Ta pristop izboljšuje analizo in razumevanje uspešnosti športnikov, kar omogoča natančnejšo prilagoditev treningov in taktik.
Jamil in drugi (2021)	Analizirali so več kot 14.000 vratarjev za prepoznavanje atributov, ki ločujejo elitne od subelitnih izvajalcev, kot sta kratka distribucija in uspešna podaja, kar premika fokus k širši oceni vratarskih atributov v nogometu, ki temeljijo na podatkih.
Kelly in drugi (2022)	Osredotoča se na razvoj mladih igralcev z uporabo linearne regresije Lasso za identifikacijo ključnih vplivnih atributov, kot so psihološka odpornost, tehnične sposobnosti in starost. Ugotovitve zagovarjajo celosten pristop k treningu, pri čemer opozarjajo na omejitve zaradi retrospektivne narave študije in majhnosti podatkovnega niza.
McGarry (2013)	Primerjal je tradicionalne metode z modernimi pristopi, kot so napovedno modeliranje in nevronske mreže, poudarjajoč nujnost vključevanja napredne analitike za izboljšanje strateškega odločanja v ekipnih in individualnih športih, kar omogoča boljše razumevanje in izboljšanje taktik in uspešnosti.
Waldron in Worsfold (2010)	Uvaja učinkovito analizo uspešnosti v realnih igralnih okoljih v primerjavi s tradicionalnimi testi spretnosti in normativno profiliranje za razlikovanje med elitnimi in subelitnimi igralci, kar izboljšuje strategije prepoznavanja talentov z uporabo igralnih podatkov v živo za natančnejše ocenjevanje.

Vir: lastno delo.

Te študije prikazane v tabeli 1 skupaj poudarjajo pomembno vlogo, ki jo imata analiza podatkov in strojno učenje pri izboljšanju razumevanja športne uspešnosti in strateškega načrtovanja. Integracija teh tehnologij ne samo izboljšuje ocene igralcev in taktične odločitve, ampak tudi nadgrajuje metodologije treninga, spodbuja bolj dinamičen in znanstveno utemeljen pristop k športni analitiki. Ta trend je ključen za neprestano izboljševanje konkurenčnih rezultatov in strateško evolucijo različnih športnih disciplin.

3 METODOLOGIJA

3.1 Teoretični okvir in zasnova raziskave

Ker se tehnologije, kot so sistemi za sledenje igralcev s kamerami in tehnologije, ki temeljijo na senzorjih, še naprej razvijajo, ustvarjajo velike količine podatkov, ki zagotavljajo podroben vpogled v položaj, gibanje in fizično stanje igralcev v realnem času. Ta podatkovna revolucija je več kot le kopičenje statističnih podatkov, kar pomeni spremembo paradigme v smeri cikla nenehne analize in uporabe, ki izboljšuje strategije tekmovanja, posodablja prakse zaposlovanja in bogati sodelovanje z navijači. Če pogledamo v prihodnost, je prihodnost športne analitike v premikanju meja zbiranja podatkov v smeri napovedne analitike, ki bo še bolj izpopolnila protokole usposabljanja, razvila nove arhetipe igralcev in inovirala strategije ekip v virtualnih okoljih usposabljanja. Možnosti, ki jih omogoča računalništvo v oblaku, bodo te zmožnosti močno razširile, kar bi lahko spremenilo izkušnje navijačev in odprlo nove poti za analizo in izboljšanje športne uspešnosti na svetovni ravni (Chase, 2020).

Sprejemanje podatkovne revolucije nas postavlja na ključno stičišče tehnologije in tradicije ter napoveduje prihodnost, v kateri šport, ki ga poganjajo podatki, umetna inteligenca in strojno učenje, odpira možnosti, ki so bile nekoč predstavljene v znanstveno fantastiko. Ta raziskava priznava preobrazbeno moč teh tehnologij in jih poskuša izkoristiti za točnost prepoznavanje in razvoj nogometnih talentov. Magistrsko delo temelji na teoretični podlagi, ki združuje športno znanost, kognitivno psihologijo, podatkovno analitiko in tehnološke študije ter si prizadeva za celovito razumevanje talenta, ki presega zgolj fizične ali prirojene sposobnosti in vključuje psihološke, okoljske, genetske in tehnične dejavnike.

Tradicionalne metode prepoznavanja talentov se pogosto močno zanašajo na subjektivno presojo skavtov, kar lahko vodi do pristranskosti pri ocenjevanju igralcev. To delo z vključevanjem strojnega učenja in podatkovne znanosti predlaga objektivna orodja, ki izboljšujejo, izpodbijajo in izpopolnjujejo konvencionalno skavtsko modrost. Ta pristop ne zavrača vrednosti izkušenih skavtov, temveč njihova spoznanja dopolnjuje z zanesljivo analizo, ki temelji na podatkih, kar zagotavlja bolj uravnoteženo in natančno oceno potenciala igralcev.

Temeljit pregled literature postavlja pod vprašaj obstoječe paradigme o nadarjenosti in skuša na novo opredeliti, kaj naredi nogometaša uspešnega v okviru najbolj gledanega športa na svetu. Ta pregled se razširi na proučevanje, kako igralci, trenerji, skavti in športni znanstveniki dojemajo in opredeljujejo nadarjenost, kar zagotavlja celosten pristop k študiji. Poleg tega se delo pogloblja v psihološko literaturo, ki se osredotoča na procese odločanja in kognitivne pristranskosti, kar zagotavlja vpogled v to, kako lahko ti dejavniki vplivajo na iskanje talentov. Ta vidik je ključen za razumevanje subtilnih vplivov, ki lahko vplivajo na odločitve skavtov in ocenjevanje igralcev.

Raziskava ocenjuje obstoječe modele strojnega učenja, ki se uporabljajo pri ocenjevanju igralcev, ter obravnava njihovo točnost napovedi in primerjavo s tradicionalnimi metodami skavtiranja. Analiziranih je več študij primerov, ki uspešno izvajajo podatkovno podprte pristope, da bi prikazali praktično uporabo in rezultate teh tehnologij.

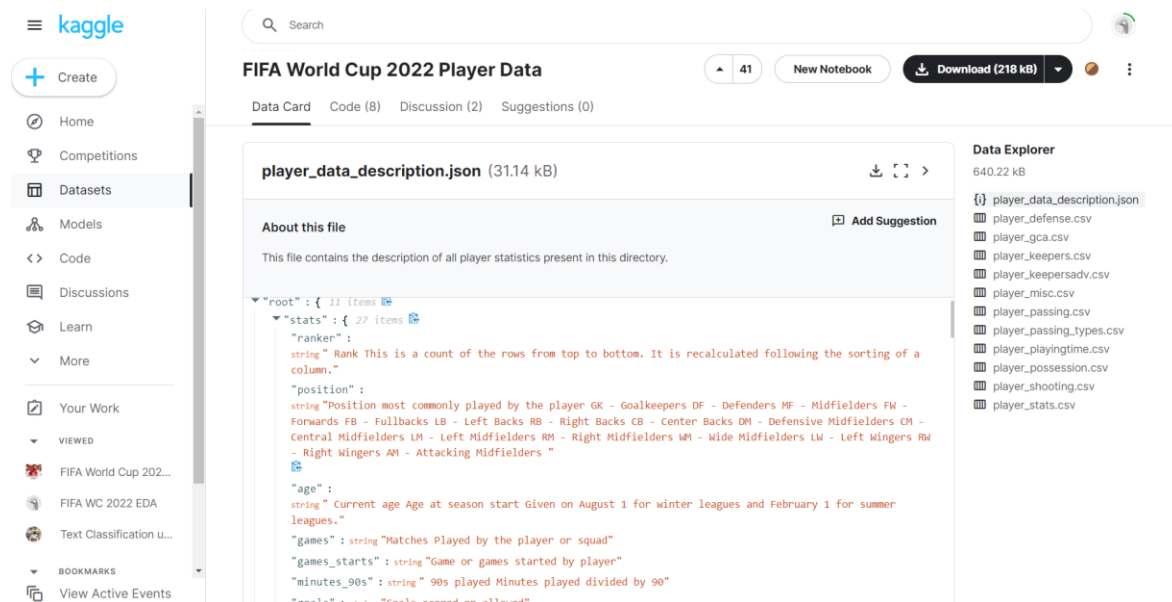
Učinkovitost povezovanja športne znanosti, tehnologije in tradicionalnih skavtskih metod je dodatno ovrednotena s primerjalnimi raziskavami in analizo študij primerov. Ta multidisciplinarni pristop ne le povečuje zanesljivost raziskave, temveč tudi zagotavlja, da so ugotovitve uporabne in pomembne za realna okolja nogometnih klubov. Z zagovarjanjem multidisciplinarnega okvira magistrsko delo poudarja pomen uporabe podatkov za sprejemanje upravljavskih odločitev, ki spodbujajo uspeh na igrišču in finančno blaginjo nogometnih klubov. Ta celovita, s podatki podprta perspektiva je pripravljena revolucionirati nogometni menedžment in skavtiranje, saj združuje bogato tradicijo tega športa z najsodobnejšimi dosežki na področju tehnologije in analitike.

3.2 Zbiranje in priprava podatkov

Za empirično raziskavo vloge strojnega učenja pri ocenjevanju nogometnega talenta smo uporabili obsežen podatkovni niz z naslovom "FIFA World Cup 2022 Player Data", ki je na voljo na portalu Kaggle (Swaptr, 2022). Ta podatkovni niz vključuje podrobne statistične podatke in attribute s svetovnega prvenstva v nogometu 2022, kar zagotavlja trdno kvantitativno podlago za analizo. Zajema široko paleto atributov igralcev, od rumenih kartonov do dejanj, ki prispevajo k zadetkom, zato je neprecenljiv vir za podatkovne znanstvenike. Kaggle prikazan na sliki 1, ki je znan po organizaciji tekmovanj v podatkovni znanosti, na katerih se rešujejo problemi iz resničnega sveta, je odlična platforma za dostop do takšnih zbirk podatkov. Kljub njenemu potencialu pa so rezultati teh tekmovanj v napovedovanju poslov pogosto premalo uporabljeni v skupnosti napovedovalcev (Bojer in Meldgaard, 2021). Odločitev za uporabo tega nabora podatkov je bila pogojena z njegovo globino in širino, saj ponuja edinstven objektiv, skozi katerega lahko proučimo uspešnost igralcev na enem od najprestižnejših nogometnih turnirjev. Njegov globalni obseg omogoča analizo talentov v različnih nogometnih kulturah in ligah, kar omogoča vpogled tako v univerzalne kazalnike uspešnosti kot tudi v kazalnike uspešnosti, specifične za posamezne regije. Dobro opredeljene atributr podatkovnega niza so idealne za uporabo modelov

strojnega učenja za raziskovanje napovednih modelov prepoznavanja talentov. Omejitev pa je, da se podatkovni niz osredotoča le na podatke o posameznih turnirjih, kar morda ne zajema vzdolžne uspešnosti in razvoja igralcev v različnih sezonah in tekmovanjih. Prihodnje študije bi lahko to izboljšale z vključitvijo longitudinalnih podatkov o igralcih, s čimer bi ustvarili celovitejši model nogometnega talenta.

Slika 1: Zaslonski prikaz datotek s podatki o igralcih na platformi Kaggle



Vir: Swaptr (2022).

Podatkovni niz "FIFA World Cup 2022 Player Data" iz Kaggle služi kot strukturiran in obsežen vir statističnih podatkov o igralcih za našo raziskavo. Vključuje več Excelovih datotek s podrobnimi podatki o atributih, kot so položaji igralcev, starost, nastopi na tekmah, goli, asistencije, pričakovani goli in akcije ustvarjanja zadetkov. Vsaka datoteka se osredotoča na posebne kazalnike uspešnosti in ponuja celosten pogled na prispevke na turnirju. Vključitev tega podatkovnega niza v našo raziskavo vključuje obdelavo in analizo različnih podatkov, da bi prepoznali vzorce in vpoglede, ki so ključni za prepoznavanje nogometnih talentov. Z vrednotenjem atributov na različnih igralnih položajih in kazalnikov uspešnosti je cilj naše študije opredeliti ključne attribute, ki prispevajo k uspehu in potencialu igralcev.

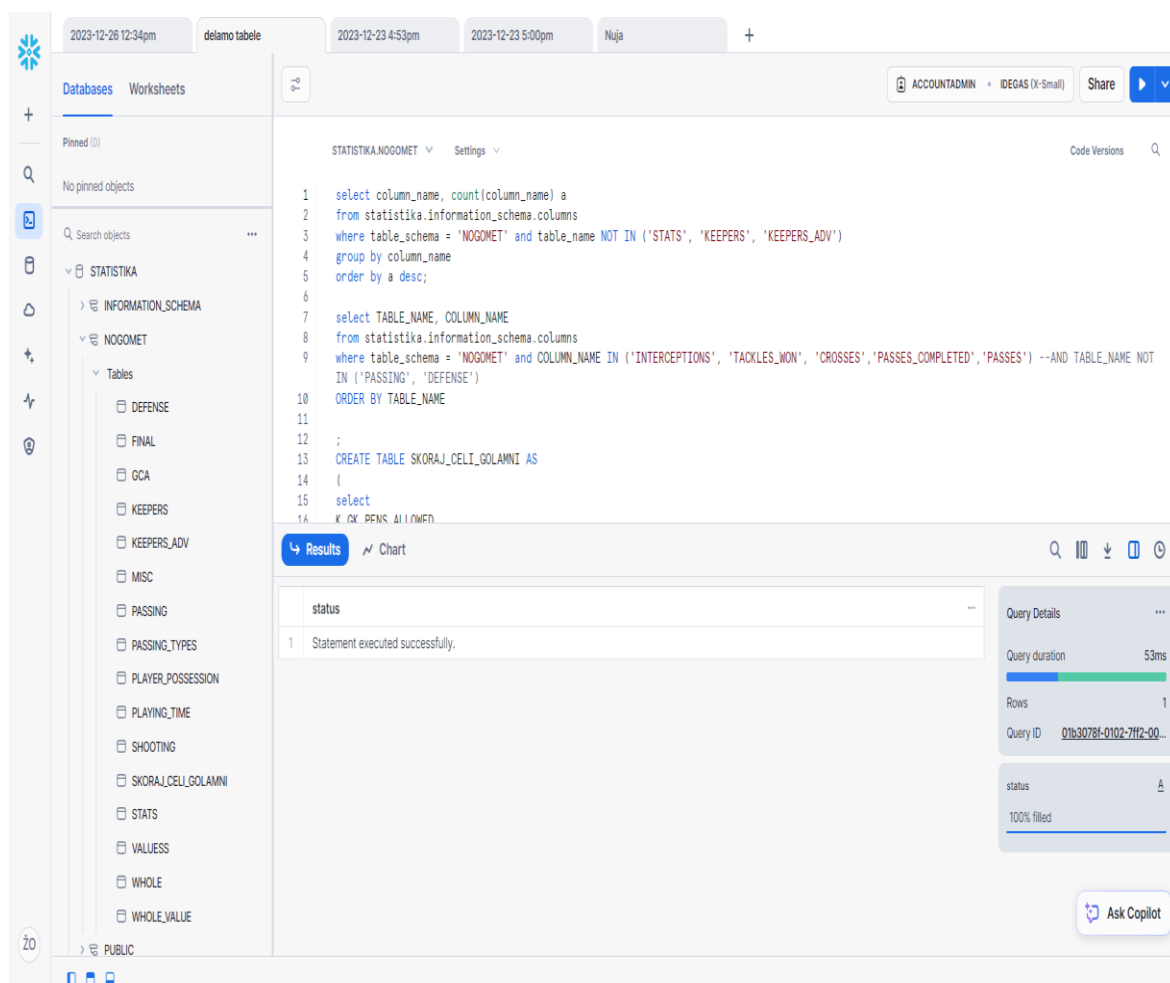
Za tiste, ki jih zanimajo nadaljnje raziskave, je nabor podatkov na voljo neposredno na portalu Kaggle pod "FIFA World Cup 2022 Player Data". Naložili smo datoteko `player_data_description.json`, ki vključuje podroben opis vseh statističnih podatkov o igralcih v 11 datotekah, katerih vrednosti so ločene z vejico (angl. Comma-Separated Values, v nadaljevanju CSV). V tej datoteki so na voljo naslednje datoteke CSV (Swaptr, 2022):

- `player_defense.csv` – obrambne statistike
- `player_gca.csv` – statistični podatki o akcijah za doseganje golov
- `player_keepers.csv` – statistika vratarjev

- player_keepersadv.csv – razširjena statistika o vratarjih
- player_misc.csv – različna statistika
- player_passing.csv – statistika podaj
- player_passing_types.csv – statistika o vrstah podaj
- player_playingtime.csv – statistika igralnega časa
- player_possession.csv – statistični podatki o posesti žoge
- player_shooting.csv – statistika streljanja
- player_stats.csv – splošna statistika igralcev

Vsaka datoteka CSV vsebuje vsaj 20 različnih atributov. Za vsako datoteko smo v analitični bazi (angl. Online Analytical Processing - OLAP) Snowflake preko uporabniškega vmesnika (angl. User Interface - UI) Snowsight, ki je prikazan na sliki 2 z uporabo Strukturirano poizvedbenega jezika (angl. Structured Query Language, v nadaljevanju SQL) pripravili ustrezne tabele.

Slika 2: Zaslonski prikaz uporabniškega vmesnika Snowsight



Vir: lastno delo.

V današnji digitalni dobi, ko so spletne storitve sestavni del naših vsakodnevnih dejavnosti, so tradicionalne metode shranjevanja in upravljanja podatkov pogosto pomanjkljive, saj se ne morejo prilagoditi hitro spreminjajoči se naravi in različnim oblikam podatkov. Snowflake se pojavlja kot najsodobnejša rešitev, zasnovana posebej za okolja v oblaku. Predstavljajte si Snowflake kot zelo prilagodljivo digitalno skladišče, ki ne le hrani velike količine podatkov, temveč uporabnikom omogoča tudi izjemno učinkovito upravljanje in analiziranje teh podatkov. Sistem Snowflake, ki je na voljo v oblaku Amazon Cloud, na novo opredeljuje podatkovno skladišče s sistemom, ki je več-grozdni, analitičen, varen, skalabilen in elastičen. Podpira tako strukturirane (angl. structured) kot polstrukturirane (angl. semi structured) podatke, zato je vsestransko orodje za različne poslovne potrebe. Njegova edinstvena arhitektura z več grozdi in skupnimi podatki ne omogoča le neprimerljive skalabilnosti in stroškovne učinkovitosti, temveč je usklajena tudi z modelom programske opreme kot storitve (angl. Software as a Service), zaradi česar so napredni sistemi za upravljanje podatkov dostopni širšemu krogu uporabnikov (Dageville in drugi, 2016).

Datoteke CSV so bile nato prenesene v te tabele z uporabo Snowsighta, ki podpira prenos posameznih datotek do velikosti 250 megabajtov. Po podrobnem pregledu strukture podatkovnega niza in zagotovitvi njegove združljivosti s sistemom za upravljanje podatkov v Snowflake smo začeli postopek integracije podatkov. Z uporabo čarovnika za uvoz podatkov v sistemu Snowflake smo metodično ustvarili posebne tabele za različne vidike podatkovnega niza, kot so statistični podatki o igralcih in kazalniki uspešnosti tekem. S poizvedbami SQL smo obdelali in analizirali podatke, da bi pridobili dragocene vpogleds, pomembne za prepoznavanje nogometnih talentov. Ta strukturiran pristop je omogočil izčrpno analizo, pri čemer smo izkoristili zanesljive zmogljivosti obdelave v orodju Snowflake za učinkovito obdelavo velikih naborov podatkov.

3.3 Čiščenje in preoblikovanje podatkov

Po vzpostavitvi potrebnih podatkovnih nizov v sistemu Snowflake je bil poudarek na integraciji teh podatkovnih zbirk. Ta korak je bil nujen, da bi se izognili neučinkovitosti pri izvajanju ločenih analiz za vsak niz podatkov ter zagotovili celovitost in natančnost združenega podatkovnega niza. Postopek integracije je vključeval združevanje podatkov iz 11 različnih datotek CSV s poudarkom na skupnih stolpcih podatkov. Ta metoda je podobna sestavljanju sestavljanke, pri kateri se mora vsak delček natančno prilegati brez odvečnih elementov. Med integriranimi podatki so bili podrobni statistični podatki o nogometaših s svetovnega prvenstva v nogometu leta 2022, razvrščeni po vlogah igralcev na igrišču, kot prikazana na sliki 3.

Slika 3: Položaji na nogometnem igrišču



vir: lastno delo na podlagi Inkwina (2008).

Vsak igralni položaj na nogometnem igrišču ima specifično vlogo in odgovornosti, ki so ključne za uspeh ekipe. Od vratarjev, ki preprečujejo gole, do napadalcev, ki so osredotočeni na doseganje zadetkov, je prispevek vsakega igralca bistven. Pregled glavnih posameznih igralnih vlog je sledeč:

- Vratarji so zadnja obrambna linija, ki je odgovorna za preprečevanje golov in sprožanje ekipnih napadov. Edinstveni so po tem, da lahko v svojem kazenskem prostoru uporabljajo roke.
- Branilci: so ključni pri onemogočanju nasprotnikovih napadov, odstranjujejo žogo z nevarnih območij in se razlikujejo od osrednjih branilcev do polnih branilcev, ki pomagajo pri napadih.
- Vezisti so motor ekipe, ki povezuje obrambo in napad, nadzoruje žogo in določa tempo igre. Njihova vloga je ključna pri vzpostavljanju ravnotežja na igrišču, saj skrbijo za

pretok igre med različnimi igralnimi področji. Poleg tega pogosto delujejo kot prvi kreatorji napadov in pomagajo obrambi pri povrnitvi posesti.

- Napadalci: osredotočeni na napad, pri doseganju golov uporabljajo hitrost, spretnost in sposobnost zaključevanja, vključno s krilnimi igralci, ki premagujejo nasprotnike po straneh.

Cilj tega analitičnega pristopa ni bil le združiti in poenostaviti podatke, temveč je želja te podatke izkoristiti za razvrstitev igralcev na podlagi potencialne prihodnje uspešnosti. Igralcem smo dodali binarno vrednost ("1" za velik potencial in "0" za manjši potencial) na podlagi meril, ki so vključevala tehnične spretnosti, fizično pripravljenost, preteklo uspešnost in trenutno formo. Pri ocenjevanju so bili uporabljeni podrobni pregledi literature in ocene igralcev iz uglednih virov, dodatno sta podatkovni niz popestrila lestvica Ballon d'Or in spletna stran Transfermarkt. Bistveni vidik te ocene je njena zgodovinska natančnost, ki zagotavlja, da potencial, prepoznan med svetovnim prvenstvom, ni anomalija, temveč kaže na stalno uspešnost. Ballon d'Or, ki se podeljuje najboljšemu nogometašu sezone na podlagi glasov kapetanov reprezentanc, trenerjev in novinarjev, pomaga potrditi uspešnost igralcev skozi čas (Morsinkhof, 2014). Poleg tega Transfermarkt zagotavlja tržne vrednosti igralcev, ki se pogosto uporabljajo v športni ekonomiji in v nogometnih klubih pri pogajanjih o pogodbah. Te vrednosti se redno posodablja na podlagi ocen, ki jih na spletnem mestu posredujejo uporabniki, in odražajo razpoloženje na trgu in vrednotenje igralcev v realnem času.

Širši cilj te analize je zagotoviti trdno, na podatkih temelječo podlago za prepoznavanje in vzgojo nogometnih talentov. S sistematično analizo podatkov o igralcih želimo ponuditi jasnejšo in manj subjektivno oceno potenciala igralcev ter tako pomagati klubom in nacionalnim reprezentancam pri sprejemanju odločitev. To delo ponazarja praktično uporabo naprednega upravljanja in analize podatkov v športu ter poudarja preobrazbeni vpliv tehnologije na tradicionalne prakse v nogometnem skavtstvu in upravljanju. Transfermarkt je splošno priznan kot ponudnik tržnih vrednosti za velik vzorec profesionalnih nogometašev, na katere se pogosto sklicujejo raziskovalci v literaturi o športni ekonomiji in upravljanju ter v popularnem tisku. Tržne vrednosti so pomembne, ker vplivajo na pogajanja o pogodbah med zastopniki igralcev in klubi. Na platformi Transfermarkt registrirani uporabniki prispevajo k vrednotenju igralcev tako, da razpravljajo in spreminjajo svoja mnenja na forumih, namenjenih posameznim igralcem. Ta participativni pristop zagotavlja, da vrednotenje odraža široko soglasje, ki se med nogometno sezono redno posodablja in odraža trenutni položaj igralcev na trgu (Peeters, 2018).

Razvrstitev Ballon d'Or, na kateri je navedenih 50 najboljših igralcev, od katerih jih je veliko sodelovalo na nedavnem svetovnem prvenstvu v Katarju, ponuja še en pogled. V naši analizi smo se osredotočili na 100 najdonosnejših igralcev januarja 2024 po podatkih Transfermarkta, da bi sistematično ocenili njihov potencial. Ta pristop nam je omogočil, da smo poleg elitnih prepoznali tudi subelitne igralce, ki izstopajo zaradi svoje uspešnosti in potencialne zaželenosti za klube, kar dokazujejo različni viri in ne le spekulativne novice.

Za učinkovito upravljanje in analizo teh podatkov smo uporabili zmogljivosti programa Snowflake in ustvarili celovit podatkovni niz. Različni atributi, kot so oblike starosti igralcev iz "let-dni", smo pretvorili v enotno številčno obliko in potencialne attribute igralcev združili v stolpec "točke ekipe" z uporabo SQL window funkcij, vidne na sliki 4.

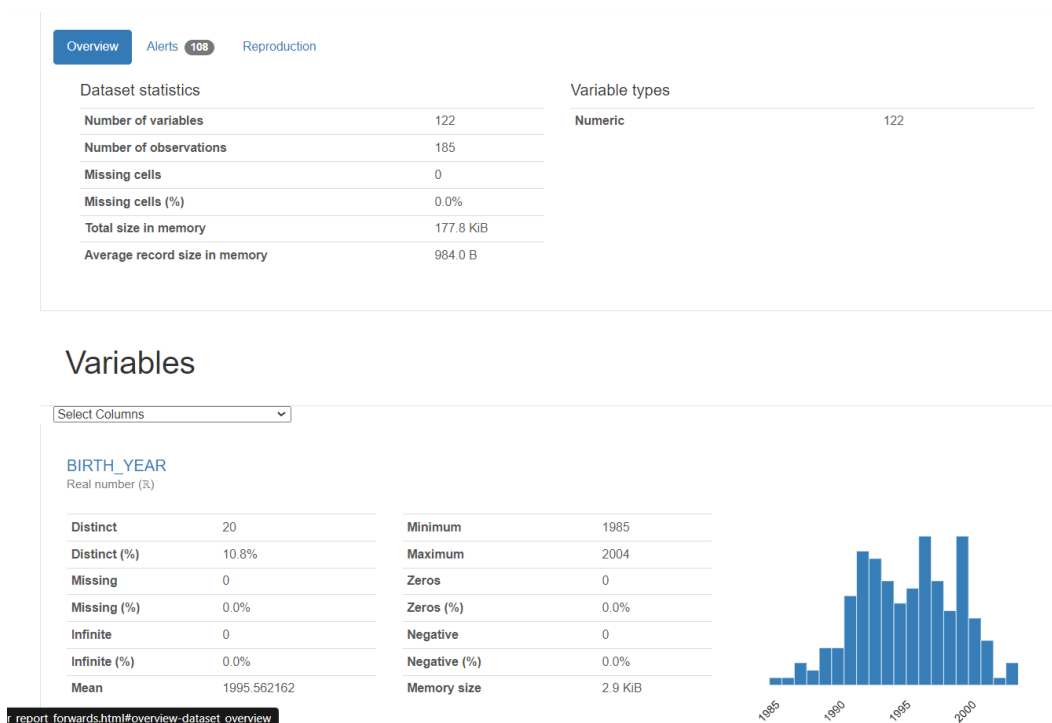
Slika 4: Zaslonski prikaz uporabe SQL window funkcij za izračun TEAM_SCORE

```
select *, split_part(wv.age, '-',1) + split_part(wv.age, '-',2)/365 as tage, sum(wv.value) over(partition by wv.team order by wv.team) as
team_score
from whole_value wv
```

Vir: lastno delo.

To je olajšalo poglobljeno analizo, kako lahko koncentracija talentov v ekipi vpliva na individualne in kolektivne attribute uspešnosti. Poleg tega smo za zagotovitev poštenih primerjav podatke normalizirali na attribute na 90 minut, ki nam omogočajo pravično primerjavo igralcev, ne glede na čas, ki so ga preživeli na igrišču. Ta prilagoditev je ključnega pomena, zlasti pri analizi igralcev z različno dolgim igralnim časom. Prečiščene podatke prikazani na sliki 5, smo na koncu zbrali v datoteko CSV, ki je pripravljena za nadaljnjo analizo ali deljenje, s čimer smo zagotovili, da so pridobljena spoznanja utemeljena na zanesljivih, sistematično zbranih podatkih, ki odražajo tako trenutno uspešnost kot morebitne prihodnje uspehe.

Slika 5: Zaslonski prikaz poročila Pandas Profiling po čiščenju podatkov



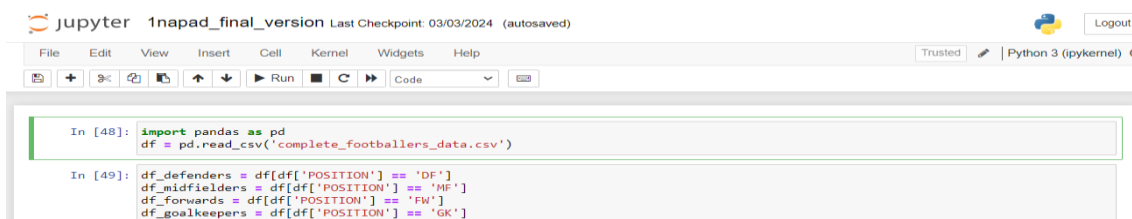
Vir: lastno delo.

3.4 Raziskovalna analiza podatkov in izbira atributov

Beležnica Jupyter je postala sestavni del našega raziskovalnega procesa, ki podpira ne le beleženje delovnih postopkov, kode in podatkov, temveč tudi izboljšuje vizualno predstavitev naših ugotovitev. To orodje se odlično ujema z načeli FAIR (angl. Findable, Accessible, Interoperable, Reusable), katerih cilj je izboljšati upravljanje in izmenjavo digitalnih raziskovalnih podatkov. Tako naša študija uporablja Jupyter Notebook, da bi zagotovila, da je naša raziskava o ocenjevanju nogometnih talentov odprta in dostopna, kar spodbuja večje znanstveno sodelovanje in komunikacijo (Randles in drugi, 2017).

Začetna faza naše analize podatkov, ki je prikazana na sliki 6 je vključevala uporabo Pandas, robustne knjižnice Python, ki je ključna za manipulacijo in analizo podatkov na različnih področjih, vključno s podatkovno znanostjo in financami. To orodje nam je omogočilo učinkovit uvoz in razčlenitev našega nabora podatkov, pri čemer smo podatke o igralcih razdelili v štiri posebne pozicijske kategorije: branilce, napadalce, srednje igralce in vratarje, pri čemer je vsaka imela edinstvene metrike uspešnosti, pomembne za njihovo vlogo na igrišču. Takšna segmentacija je bistvenega pomena, saj priznava, da različni položaji zahtevajo različne sklope spretnosti in kazalnike uspešnosti za točno ocenjevanje talentov.

Slika 6: Zaslonski prikaz knjižnice Pandas v beležnici Jupyter



```
In [48]: import pandas as pd
df = pd.read_csv('complete_footballers_data.csv')

In [49]: df_defenders = df[df['POSITION'] == 'DF']
df_midfielders = df[df['POSITION'] == 'MF']
df_forwards = df[df['POSITION'] == 'FW']
df_goalkeepers = df[df['POSITION'] == 'GK']
```

Vir: lastno delo.

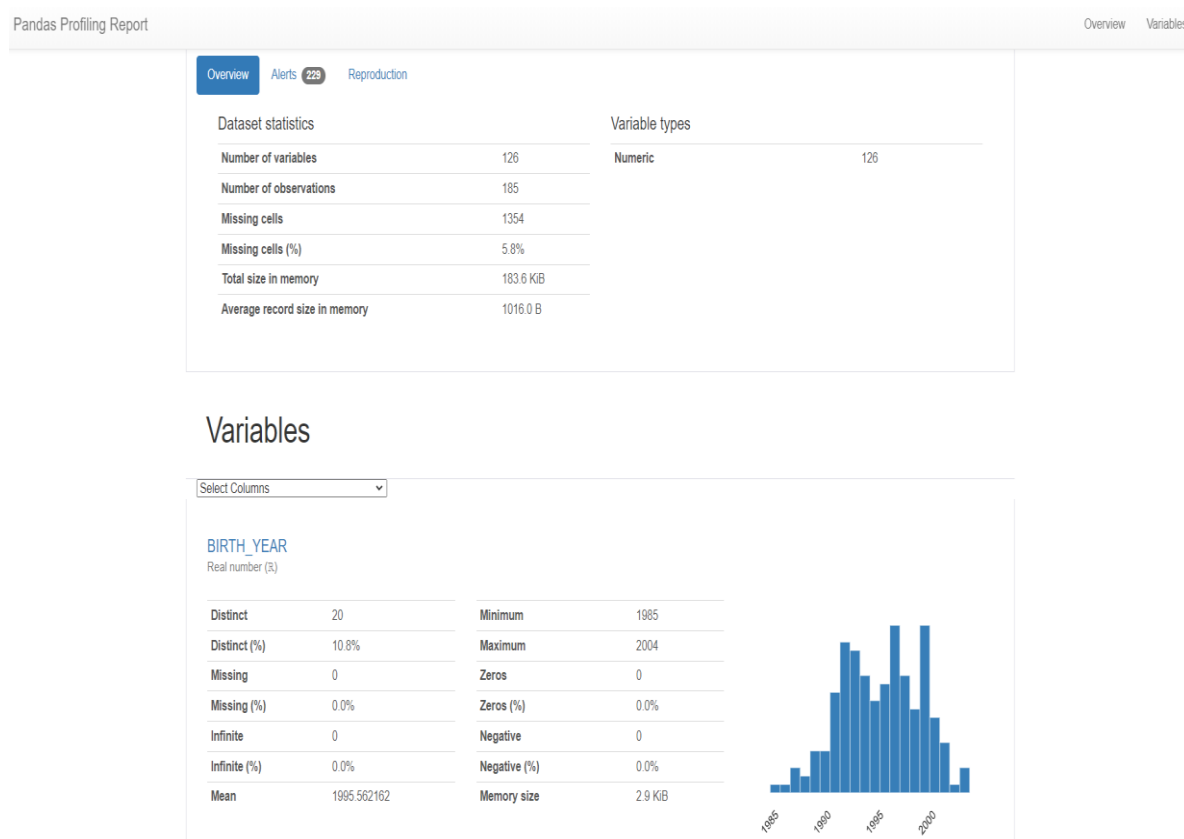
Po segmentaciji podatkov smo izvedli raziskovalno analizo podatkov z uporabo knjižnice `pandas_profiling`. To zmogljivo orodje avtomatizira večino začetnega raziskovanja podatkov in zagotavlja izčrpna poročila, ki vključujejo statistiko, korelacije in analize porazdelitve atributov. S tem nismo prihranili le veliko časa, temveč smo tudi izpostavili ključna vprašanja glede kakovosti podatkov, ki jih je bilo treba odpraviti, da smo lahko izpopolnili naše napovedne modele. Natančneje, ugotovili in obravnavali smo težave z ničelnimi vrednostmi, odvečnimi podatki in manjkajočimi vrednostmi:

- Ničelne vrednosti: veliko igralcev je zabeležilo ničle v kategorijah, kot so doseženi goli ali prejeti rumeni kartoni, kar je glede na naravo športa pričakovano. Vendar smo morali zaradi analitične jasnosti razlikovati med pravimi ničlami, ki kažejo na dejanske dogodke v igri, in manjkajočimi podatki.
- Odvečni podatki: zaradi poenostavitve naših modelov smo odstranili stolpce, ki vsebujejo konstante in se ne spreminjajo in tako nimajo napovedne moči.

- Ravnanje z manjkajočimi podatki: pri stolpcih, v katerih manjka precej podatkov (od 10 % do 50 %), smo uporabili algoritem KNN za pripisovanje vrednosti na podlagi podobnosti z drugimi podatki, s čimer smo povečali robustnost našega nabora podatkov. Stolpci z manj kot 10 % manjkajočih podatkov so bili prilagojeni z uporabo mediane vrednosti stolpca, da bi se ohranila celovitost porazdelitve podatkov. Ostale stolpce smo odstranili iz analize odstranili (več kot 50% manjkajočih podatkov).

Prinazorjenih na sliki 7 manjkajočih podatki v našem podatkovnem nizu, katerih vrednost zajema od 10% do 50% smo se lotili uporabili računalniško metodo KNNImputer, ki ocenjuje manjkajoče vrednosti z uporabo povprečja najbližjih sosedov. Ta metoda je za našo rešitev še posebej ugodna, ker je sposobna ohraniti notranjo strukturo podatkov, zaradi česar je idealna za podatkovne nize, kjer je relacijska kontinuiteta med atributi ključnega pomena.

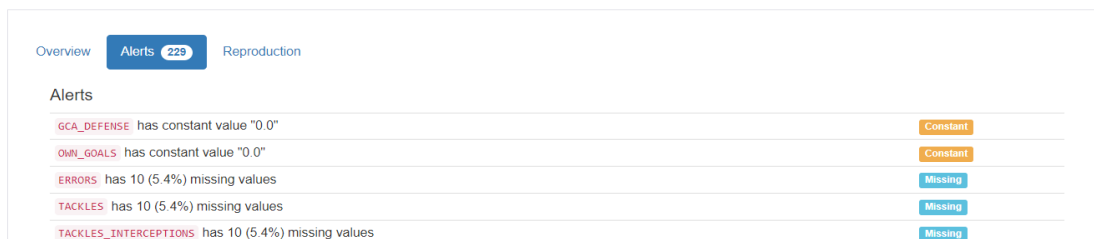
Slika 7: Zaslonski prikaz poročila Pandas Profiling pred čiščenjem podatkov



Vir: lastno delo.

Na sliki 8 je prikazano poročilo o napakah in težavah s podatki, ki se pojavljajo pred čiščenjem podatkov. Prikazana so opozorila, ki opozarjajo na konstantne vrednosti in manjkajoče podatke v določenih stolpcih. Ti problemi, kot so manjkajoče vrednosti ali konstantne vrednosti brez variacije, so običajno identificirani pred začetkom analiz in zahtevajo dodatno pozornost, da se zagotovi ustrezna kakovost podatkov za nadaljnjo obdelavo.

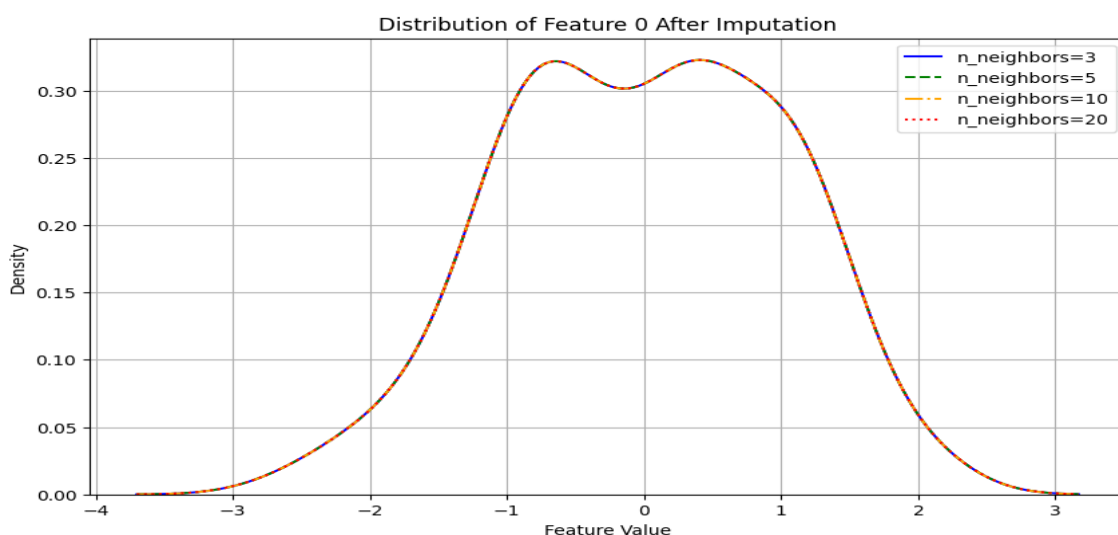
Slika 8: Opozorila v Pandas Profiling poročilu pred čiščenjem podatkov



Vir: lastno delo.

Da bi optimizirali postopek imputiranja, smo izvedli vrsto poskusov prikazanih na sliki 9, da bi ugotovili učinek spreminjanja števila najbližjih sosedov, uporabljenih pri izračunu. Preizkusili smo vrednosti `n_neighbors` 3, 5, 10 in 20, da bi zajeli razpon od manjše do večje velikosti sosedstva. Naš cilj je bil določiti ravnovesje med vključitvijo zadostnih informacij o soseskah in izogibanjem pretirani razvojenitvi ustreznih atributov.

Slika 9: Distribucija atributa po imputaciji z *n*-sosedi



Vir: lastno delo.

Ta skrbna priprava podatkov ni le izboljšala kakovosti našega podatkovnega niza, temveč je postavila tudi trdne temelje za razvoj napovednih modelov za posamezne položaje, ki natančno odražajo različne attribute, bistvene za različne nogometne vloge. Ta pristop zagotavlja, da naše ugotovitve niso le znanstveno stroge, temveč tudi praktično pomembne za iskalce nogometnih talentov in analitike, ki želijo izkoristiti podatkovno podprta spoznanja za ocenjevanje in zaposlovanje igralcev.

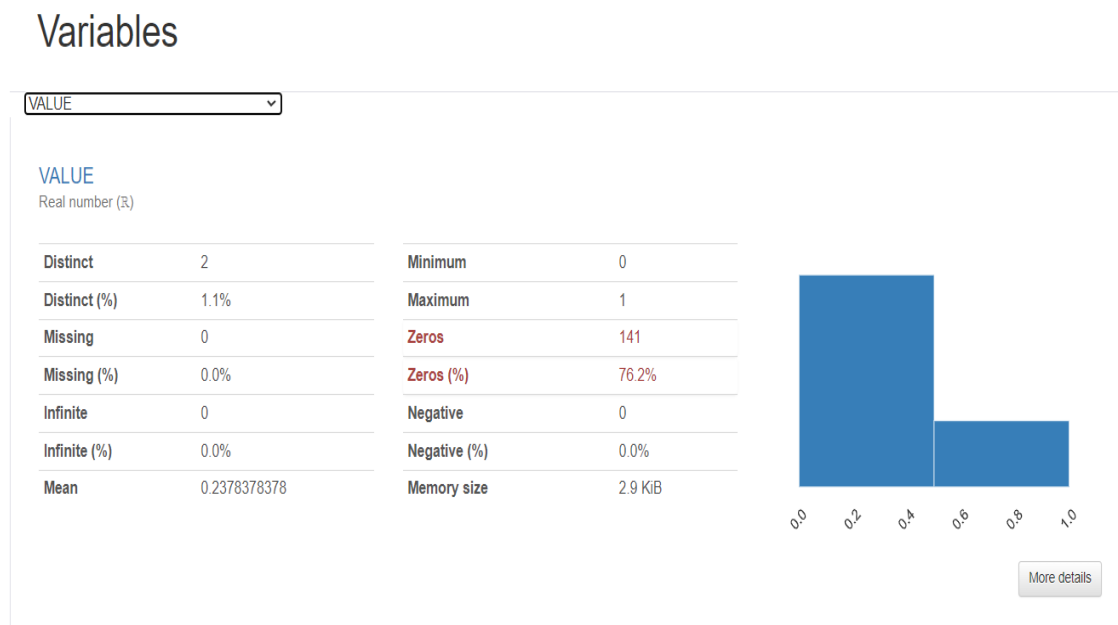
V naši študiji smo ugotovili, da številčni atributi bistveno poenostavijo izdelavo klasifikacijskih modelov v primerjavi z besedilnimi atributi, ki lahko otežijo napovedne analize. Za optimizacijo našega modela smo se osredotočili na pretvorbo vseh pomembnih

besedilnih podatkov v številčne oblike in izločitev atributov, ki ne bi učinkovito prispevale k našim napovednim ciljem:

- Imena igralcev in pripadnost ekipi: ti atributi so sicer edinstveni, vendar nimajo napovedne vrednosti za prepoznavanje talentov, zato smo jih odstranili iz nabora podatkov.
- Zastopanost držav: ta atribut je bil nadomeščen s številčnim atributom TEAM_SCORE, ki bolje ustreza kvantitativni analizi.
- Leta igralca: zaradi lažjega izvajanja statističnih operacij je bil pretvorjen iz besedilne v številčno obliko.
- Poleg tega so bile opuščeni tudi atributi s konstantnimi vrednostmi ali veliko variabilnostjo (več kot 90-odstotna edinstvenost), da bi se povečala učinkovitost modela in točnost napovedovanja.

Po čiščenju podatkov smo izvedli ponovno preverjanje z uporabo `pandas_profiling`, da bi zagotovili celovitost podatkov v vseh štirih podatkovnih nizih. Ta korak je bil ključen za potrditev čistosti podatkov po spremembi. Pri izbiri atributov smo razlikovali med odvisnimi in neodvisnimi spremenljivkami, pri čemer smo kot odvisno spremenljivko (y) določili "vrednost", ki kategorizira igralce kot nadarjene (1) ali ne (0), kar je tudi prikazano na sliki 10. Vse druge številčne attribute pa kot neodvisne (x). Te neodvisne attribute smo standardizirali, da smo izenačili njihovo lestvico, kar je potreben korak za smiselno primerjavo in učinkovito usposabljanje modelov.

Slika 10: Primer porazdelitve atributa odvisne spremenljivke

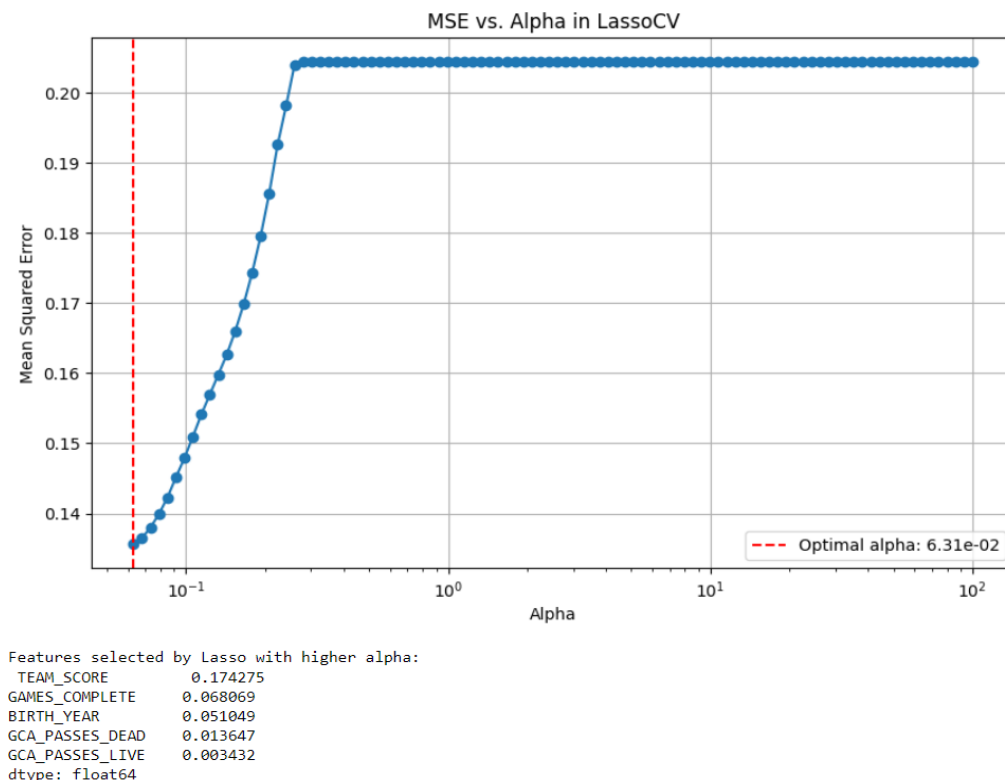


Vir: lastno delo.

Med različnimi tehnikami izbire atributov se je kot najučinkovitejša izkazala Lasso regresija. Ta metoda ne le poenostavi model z zmanjšanjem manj pomembnih koeficientov atributov na nič, temveč tudi odpravi prekletstvo večrazsežnosti z izbiro samo najpomembnejših atributov. Sposobnost Lasso regresije, da obvladuje multikolinearnost, je še posebej koristna v primerjavi z drugimi metodami, kot sta ExtraTreesClassifier in SelectKBest, ki ne obvladujejo multikolinearnosti, kar vodi v dolgotrajnejše in manj učinkovite postopke izbire atributov. Vključitev pristopa navzkrižnega preverjanja (angl. cross-validation) pri izbiri parametra regularizacije (alfa) z metodo Lasso pomaga preprečiti tako preprileganja (angl. overfit) kot tudi podprileganja (angl. underfit).

Za določitev optimalne vrednosti alfa smo uporabili LassoCV, ki sistematično preizkuša različne vrednosti alfa in izbere tisto, ki minimizira napako prek navzkrižnega preverjanja. To zagotavlja, da je model dobro generaliziran in učinkovit na novih podatkih. V naši analizi je optimalna vrednost alfa prikazana na sliki 11, ki je bila izbrana z metodo LassoCV, poudarjena na grafičnem prikazu, ki vizualno predstavi, kako različne vrednosti alfa vplivajo na srednje kvadratno napako. Ta pristop omogoča jasno in empirično utemeljeno izbiro najboljših vrednosti alfa. V naši analizi je Lasso zahteval odstranitev največ enega atributa na podatkovni niz, da bi obvladal multikolinearnost, kar je povečalo točnost modela. Nazadnje smo korelacije med izbranimi atributi vizualizirali s toplotnim zemljevidom, s čimer smo potrdili usklajenost in učinkovitost izbire atributov Lasso regresije.

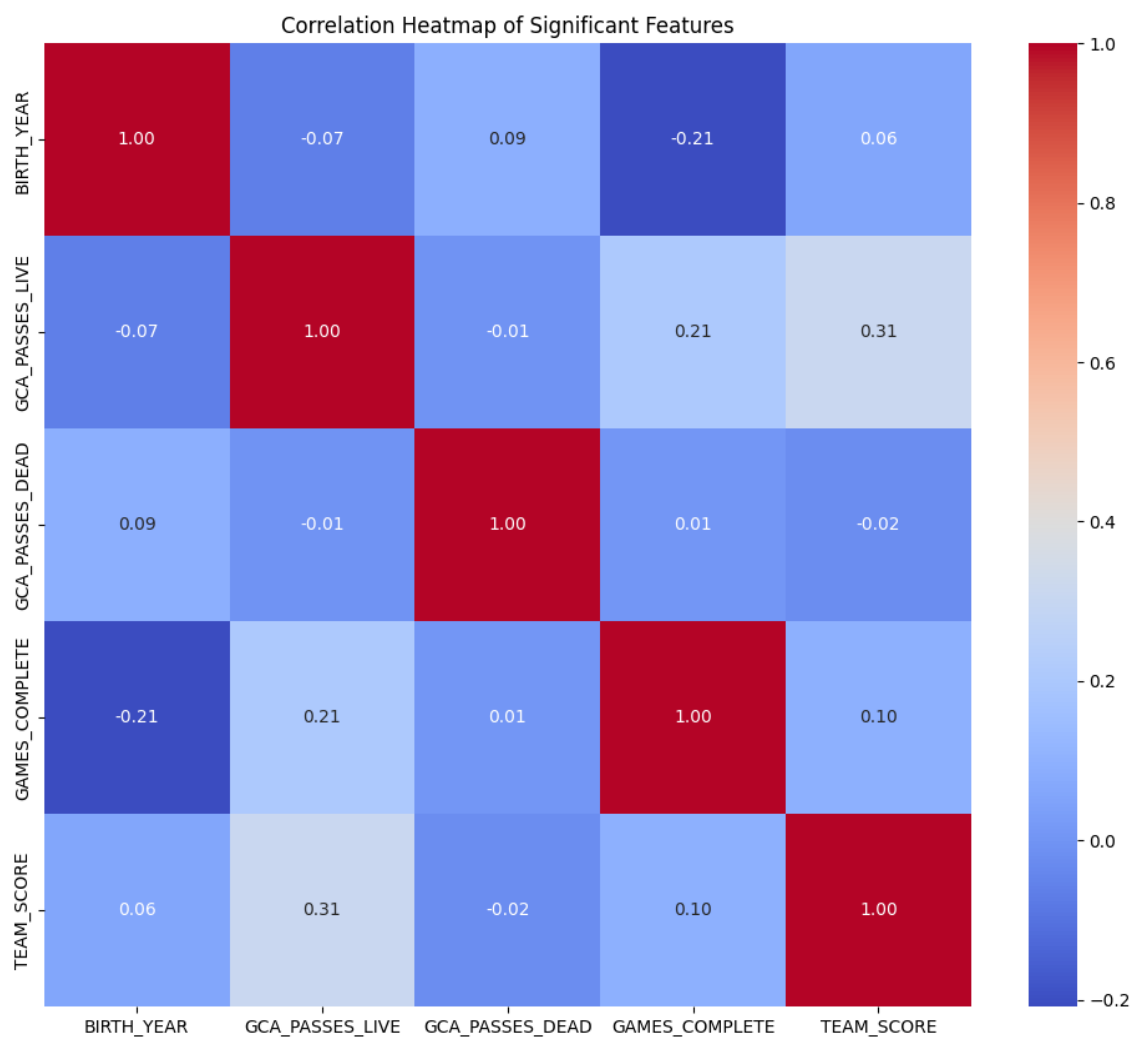
Slika 11: Izbrani atributi z Lasso regresijo in prikaz optimalne vrednosti alfa



Vir: lastno delo.

Prikazan Toplotni zemljevid na sliki 12 ni potrdil multikolinearnosti med atributi, ki jih je izbral Lasso, kar zagotavlja zanesljivost napovednika na različnih položajih, vključno z vezisti, napadalci in branilci. Da bi dodatno potrdili pomembnost teh atributov glede na binarno odvisno spremenljivko, smo izvedli Mann-Whitneyjeve U-teste.

Slika 12: Toplotni zemljevid korelacij med pomembnimi atributi



Vir: lastno delo.

Pri vezistih so atributi TEAM_SCORE, MINUTE, BIRTH_YEAR in XG_PLUS_MINUS pokazali pomembne razlike s pomembnimi velikostmi učinkov, kar kaže na bistvene razlike med skupinami. Čeprav je tudi PENS_MADE pokazala statistično pomembnost, je bila njena velikost učinka majhna, kar kaže na omejen praktični pomen; zato je njena vključitev odvisna od premislekov, specifičnih za posamezen igralni položaj. Pri napadalcih so TEAM_SCORE, GAMES_COMPLETE, BIRTH_YEAR in GCA_PASSES_LIVE pokazali pomembne razlike s pomembnimi velikostmi učinkov, kar potrjuje njihovo pomembnost. GCA_PASSES_DEAD je imel kljub statistično pomembnemu učinku majhno velikost učinka, kar upravičuje razmislek na podlagi domenskega znanja. Pri branilcih so lastnosti

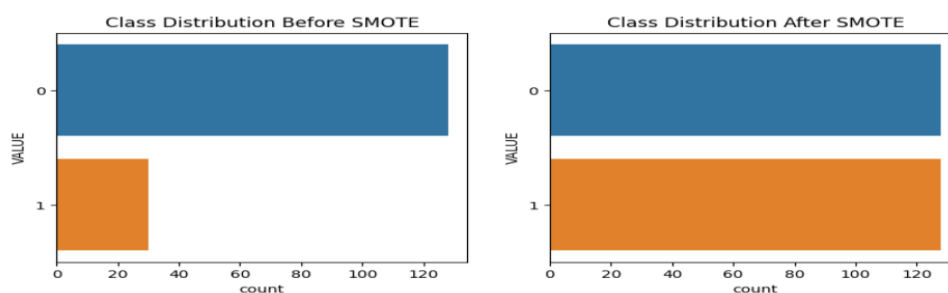
TEAM_SCORE, BIRTH_YEAR in MINUTES pokazale pomembne razlike s smiselno velikostjo učinka, kar potrjuje njihov pomen pri razlikovanju med skupinami.

Atributi TEAM_SCORE, MINUTE, BIRTH_YEAR, XG_PLUS_MINUS in GAMES_COMPLETE so se na splošno izkazale za pomembne napovedne attribute s precejšnjo velikostjo učinka na različnih položajih. Značilnosti, kot sta PENS_MADE in GCA_PASSES_DEAD, so imele kljub statistični pomembnosti manjše velikosti učinkov, zato bi bilo treba njihovo vključitev oceniti na podlagi nadaljnjih spoznanj, specifičnih za posamez igralni položaj. Ta celovita analiza zagotavlja, da je naš model statistično zanesljiv in praktično ustrezen, saj vključuje tako statistično pomembne attribute kot domensko znanje za izboljšanje napovedne uspešnosti.

3.5 Razvoj in vrednotenje modela strojnega učenja

Za pripravo in obdelavo podatkov smo uporabili Python, ki je zaradi svoje robustnosti in obsežne knjižnice idealen za analizo podatkov. Najprej smo podatke ustrezno pripravili in uporabili napredne tehnike za zagotovitev kakovosti podatkov, vključno s tehniko Lassoove regresije za izbor atributov. Lassoova regresija je bila ključna za redukcijo večrazsežnosti, saj omogoča izločanje manj pomembnih atributov, kar osredotoča model na najbolj vplivne attribute. Za izdelavo in optimizacijo napovednih modelov smo implementirali več tehnik strojnega učenja, med katerimi je logistična regresija služila kot osnova za razumevanje vpliva posameznih atributov. Poleg logistične regresije smo uporabili tudi bolj kompleksne modele, kot so drevo odločanja, naključni gozdovi in klasifikatorji z gradientnim povečevanjem. Integrirali smo tehniko sintetičnega nadzorčnega manjšinskih primerov. (angl. Synthetic Minority Over-sampling Technique, v nadaljevanju SMOTE) v naš pipeline za obdelavo podatkov, kar nam je omogočilo uravnoteženje razredov v našem učnem naboru. Ta metoda, ki je prikazana na sliki 13 sintetično generira primerke manjšinskih razredov, s čimer preprečuje pristranskost modela zaradi neuravnoteženosti razredov. Uporaba SMOTE je bila ključna za izboljšanje zanesljivosti in točnosti naših modelov pri generalizaciji na novih podatkih. Za vsak model smo uporabili GridSearchCV za avtomatizirano iskanje optimalnih hiperparametrov, kar je vključevalo regularizacijsko moč, globino drevesa, število ocenjevalcev in druge pomembne nastavitve. GridSearchCV, skupaj z navzkrižnim preverjanjem, je zagotovil, da so izbrane konfiguracije najbolj učinkovite na več delih podatkovnega niza. Faza priprave podatkov je poleg Lassoove regresije zajemala tudi delitev na učno in testno množico (70 % učna množica, 30 % testna množica), kar omogoča objektivno oceno uspešnosti modelov na nevidenih podatkih. Ta metodološki pristop zagotavlja, da so modeli zmožni dobro generalizirati, kar je ključnega pomena za njihovo praktično uporabo. Uspešnost vsakega modela smo ocenili z uporabo natančnosti, priklica, F1-ocene in obsežnega poročila o klasifikaciji, kar nam je omogočilo poglobljen vpogled v sposobnost vsakega modela napovedati nove podatke.

Slika 13: Sintetična generacija manjšinskega razreda



Vir: lastno delo.

Po usposabljanju in optimizaciji vsakega prikazanega modela na sliki 14 smo jih ocenili z uporabo različnih atributov, kot so natančnost, priklic, točnost in F1-ocena. Podrobno smo opisali uspešnost vsakega modela v poročilu o klasifikaciji, ki nam je omogočilo vpogled v to, kako učinkovito vsak model napoveduje nove podatke. Primerjali smo tudi modele, da bi razumeli njihove relativne prednosti in slabosti ter določili, kateri model se najbolje obnese za naš specifičen podatkovni niz. S tem celovitim pristopom smo našli model, ki najbolje ustreza našim podatkom, in pridobili vpogled v zapleteno dinamiko napovedovanja uspešnosti nogometašev. To nam je omogočilo sprejemanje utemeljenih odločitev o ocenjevanju igralcev in strategiji ekipe ter zagotavljanje, da so naše napovedi robustne in zanesljive.

Slika 14: Primeri uporabljenih modelov in njihovih parametrov

```
models = {
  "Logistic Regression": {
    "model": LogisticRegression(random_state=42, solver='lbfgs', max_iter=5000, tol=0.01),
    "param_grid": {
      'classifier__C': np.logspace(-4, 4, 20),
      'classifier__penalty': ['l2']
    }
  },
  "Decision Tree": {
    "model": DecisionTreeClassifier(random_state=42),
    "param_grid": {
      'classifier__max_depth': [None, 10, 20, 30, 40, 50],
      'classifier__min_samples_split': [2, 5, 10, 20],
      'classifier__min_samples_leaf': [1, 2, 5, 10]
    }
  },
  "Random Forest": {
    "model": RandomForestClassifier(random_state=42),
    "param_grid": {
      'classifier__n_estimators': [100, 200, 300],
      'classifier__max_depth': [None, 10, 20, 30],
      'classifier__min_samples_split': [2, 5, 10],
      'classifier__min_samples_leaf': [1, 2, 4],
      'classifier__bootstrap': [True, False]
    }
  },
  "Gradient Boosting": {
    "model": GradientBoostingClassifier(random_state=42),
    "param_grid": {
      'classifier__n_estimators': [100, 200, 300],
      'classifier__learning_rate': [0.01, 0.1, 0.2],
      'classifier__max_depth': [3, 5, 7],
      'classifier__min_samples_split': [2, 5, 10],
      'classifier__min_samples_leaf': [1, 2, 5]
    }
  }
}
```

Vir: lastno delo.

Pri ocenjevanju uspešnosti napovednih modelov za različne nogometne položaje smo dodatno vsak model ovrednotili z dvema ključnima statističnima orodjema: krivuljo sprejemnika delovanja (angl. Receiver Operating Characteristic, v nadaljevanju ROC) z AUC in matriko zmede. Ta orodja so bistvena za razumevanje, kako dobro vsak model razlikuje med različnimi razredi igralcev in za kvantificiranje točnosti napovedi modela.

Krivulja ROC je grafični prikaz, ki ilustrira diagnostično sposobnost sistema binarnega klasifikatorja, ko se spreminja njegov prag diskriminacije. Na krivulji so prikazani delež pravilno prepoznanih pozitivnih rezultatov (angl. True Positive Rate, v nadaljevanju TPR) proti deležu napačno prepoznanih negativnih rezultatov (angl. False Positive Rate, v nadaljevanju FPR) pri različnih nastavitvah praga. Delež pravilno prepoznanih pozitivnih rezultatov je delež dejanskih pozitivov, ki jih model pravilno identificira, medtem ko je delež napačno prepoznanih negativnih rezultatov delež negativov, ki so napačno klasificirani kot pozitivni.

AUC meri celotno dvodimenzionalno območje pod celotno krivuljo ROC od (0,0) do (1,1). To območje predstavlja sposobnost modela za razlikovanje med razredi. AUC v vrednosti 0,5 kaže na ničelno diskriminacijsko sposobnost (enako naključju), medtem ko AUC v vrednosti 1,0 predstavlja popolno diskriminacijo. Višja vrednost AUC kaže na bolje delujoč model.

Matrika zmede je posebna tabela, ki omogoča vizualizacijo delovanja algoritma. Še posebej uporabna je za merjenje priklica, točnosti in natančnosti klasifikacijskega modela ter bolj široko njegovo sposobnost pravilnega napovedovanja razredov. Vsaka vrstica matrike predstavlja primere v dejanskem razredu, medtem ko vsak stolpec predstavlja primere v napovedanem razredu. Matrika je razdeljena na štiri dele:

- Pravilno Pozitivni (angl. True Positives, v nadaljevanju TP): Pravilno napovedana pozitivna opazovanja.
- Pravilno Negativni (angl. True Negatives, v nadaljevanju TN): Pravilno napovedana negativna opazovanja.
- Napačno Pozitivni (angl. False Positives, v nadaljevanju FP): Napačno napovedani pozitivni rezultati (napaka tipa I).
- Napačno Negativni (angl. False Negatives, v nadaljevanju FN): Napačno napovedani negativni rezultati (napaka tipa II).

Razumevanje teh matrik pomaga ne le pri ocenjevanju točnosti, ampak tudi pri robustnosti modela glede na njegov priklic in natančnost.

Zaključna faza našega projekta je vključevala primerjavo modelov, da bi identificirali tistega z najboljšo uspešnostjo za naš specifičen podatkovni niz. Ta celovit pristop nam je omogočil izbiro najučinkovitejšega modela, kar je zagotovilo vpogled v dinamiko napovedovanja

uspešnosti in omogočilo utemeljene odločitve pri ocenjevanju in strategiji izbora igralcev. Te korake smo nato ponovili za vse štiri podatkovne nize.

3.6 Multivariatna analiza

V tem razdelku podrobno obravnavamo proces multivariatne analize, ki je ključna za razumevanje razlik med igralnimi položaji in za prepoznavanje ključnih atributov, ki vplivajo na uspešnost igralcev. Ta analiza temelji na našem strokovnem pristopu k obdelavi in analizi podatkov, ki zagotavlja, da so naše ugotovitve zanesljive in podkrepljene z robustnimi statističnimi metodami. Priprava podatkov je prvi korak, ki zagotavlja kakovostno osnovo za vse nadaljnje analize. Izbrali smo samo numerične stolpce iz naše podatkovne zbirke, kar je bilo nujno za izvedbo statističnih testov. Da bi obravnavali manjkajoče podatke, smo uporabili KNN imputacijo, ki izkoristi prostorsko podobnost med primeri za natančno oceno manjkajočih vrednosti. Prav tako smo uporabili tehniko winsorizacije za omejevanje vpliva izstopajočih vrednosti, kar je ključno za preprečevanje potencialnih izkrivljanj v analizi.

Izbor atributov po položajih je bil opravljen z namenom identifikacije relevantnih atributov za vsak položaj. To nam omogoča specifično analizo, ki bolje odraža razlike med igralnimi položaji. Atributi, kot so 'TEAM_SCORE', 'MINUTES', 'BIRTH_YEAR', med drugimi, so bili izbrani na podlagi njihove predpostavljene pomembnosti za posamezne igralne položaje. Statistična analiza je vključevala uporabo testov Kruskal-Wallis in Dunnov post-hoc test, ki nam omogoča natančnejše primerjave med pari skupin po identifikaciji statistično pomembnih razlik med položaji.

Test Kruskal-Wallis, uporabljen za primerjave med položaji, je bil izbran zaradi svoje učinkovitosti pri analizi podatkov brez predpostavke normalne porazdelitve. Ko smo odkrili statistično pomembne razlike, smo izvedli post-hoc test Dunnovih parnih primerjav, ki nudi podrobne analize med pari skupin in določa, ali so opažene razlike med položaji statistično značilne. Za vizualizacijo rezultatov smo uporabili Dunnov post-hoc test, ki je bil prikazan z grafičnimi predstavitvami povprečnih razlik in intervalov zaupanja. Te vizualizacije so ključne za intuitivno razumevanje obsega in narave statističnih razlik med položaji. Dokumentacija in shranjevanje rezultatov je bila izvedena z uporabo PdfPages, kar omogoča enostavno arhiviranje in pregledovanje rezultatov v profesionalno oblikovanem formatu. To ne samo, da zagotavlja transparentnost našega analitičnega procesa, ampak tudi omogoča lažjo komunikacijo in predstavitev ugotovitev zainteresiranim stranem.

Ta celostni pristop k multivariatni analizi zagotavlja, da so naše metode in ugotovitve natančne, zanesljive in neposredno uporabne za nadaljnje raziskovanje in praktično aplikacijo v športni analitiki.

4 REZULTATI

4.1 Pregled podatkov in ugotovitev

Naš podatkovni niz obsega bogato zbirko atributov nogometašev na različnih položajih, vključno z branilci, vezisti, napadalci in vratarji. Poleg vratarjev pregled podatkovnega niza razkrije uravnoteženo zastopanost položajev, kar omogoča zanesljivo primerjavo med različnimi vlogami na igrišču. Vratarji imajo zaradi narave svojega položaja poseben nabor atributov, prilagojen njihovim specifičnim odgovornostim. Hkrati so rezultati pri vratarjih najmanj relevantni, saj ima ta položaj bistveno manj instanc glede na druge položaje, ki jih je večje število na igrišču. V raziskovalni analizi smo opazili tudi veliko število korelacij med atributi, saj so nekateri atributi podmnožice drugih atributov. Dober primer je, da je rdeči karton sestavljen iz dveh rumenih. Zanimivo, da v analizi vidimo, tudi korelacije atributov, ki nimajo logične povezave, saj dotiki v tretji četrtini, ne pomenijo, da bodo srednje dolge podaje uspešne. Končna izbira podatkov kaže visoko stopnjo popolnosti, brez osamljencev in z ustreznim soočanjem manjkajočih vrednostmi, kar je posledica strogega zbiranja in priprave podatkov. Ta robustnost je ključna za zanesljivost nadaljnjih modelov strojnega učenja. Izbrane attribute lahko razdelimo na igralne položaje.

4.1.1 Branilci

Lasso je izbral attribute pri podatkovnem nizu o branilcih, ki so vidni v tabeli 2. Višja TEAM_SCORE znatno poveča rezultat, kar kaže na pomembnost timske uspešnosti. Na rezultat pozitivno vplivajo tudi izkušeni igralci (BIRTH_YEAR) in tisti, ki so odigrali več minut (MINUTES), kar poudarja pomen kvalitete soigralcev, izkušenj in zaupanje trenerja.

Tabela 2: Izbrani atributi za branilce

Atribut	Koeficient
TEAM_SCORE	0,175703
BIRTH_YEAR	0,063825
MINUTES	0,038965

Vir: lastno delo.

4.1.2 Napadalci

Pri analizi podatkovnega niza o napadalcih je izbor funkcij Lasso identificiral ključne attribute, ki so vidni v tabeli 3. Višje vrednosti v atributih TEAM_SCORE, BIRTH_YEAR, in GAMES_COMPLETE izrazito pozitivno vplivajo na napovedane izide, kar poudarja ključne aspekte igralčeve uspešnosti in ekipne dinamike. Manjši koeficienti za GCA_PASSES_DEAD (atribut odraža igralčev prispevek h golom iz mrtvih situacij, kot so

prosti streli in koti, kar je manjši, a pomemben faktor v napovedovanju uspešnosti) in GCA_PASSES_LIVE (koeficient poudarja igralčev vpliv na igro skozi prispevke v živi igri, čeprav s še manjšim vplivom) prav tako poudarjajo, kako specifične igralčeve veščine vplivajo na skupni uspeh in strategije na igrišču. To nam ponuja vpogled v raznolike načine, kako igralčeva vloga in interakcije na igrišču prispevajo k celotni učinkovitosti ekipe.

Tabela 3: Izbrani atributi za napadalce

Atributi	koeficienti
TEAM_SCORE	0,174275
GAMES_COMPLETE	0,068069
BIRTH_YEAR	0,051049
GCA_PASSES_DEAD	0,013647
GCA_PASSES_LIVE	0,003432

Vir: lastno delo.

4.1.3 Vezisti

Pri naboru podatkov o vezistih igrišča je Lasso natančno določil ključne attribute, ki so vidni v tabeli 4. Višji koeficienti za TEAM_SCORE, MINUTES in BIRTH_YEAR poudarjajo, da so ekipni uspehi, igralni čas in izkušnje ključni dejavniki za napovedovanje uspešnosti vezistov. Manjši, a pozitivni koeficienti za PENS_MADE in XG_PLUS_MINUS nakazujejo, da imajo tudi specifični prispevki, kot so izvedba kazenskih strellov in razmerje med pričakovanimi in dejanskimi goli, pomembno vlogo, čeprav so ti vplivi manj izraziti. Za vsak atribut smo skrbno proučili njegov vpliv na model, kar nam je omogočilo globlji vpogled v to, kako posamezni atributi igralcev prispevajo k skupnemu uspehu na igrišču. Ta natančna analiza je ključna za razumevanje dinamike sredine igrišča in strategij, ki najbolj vplivajo na uspešnost vezistov.

Tabela 4: Izbrani atributi za veziste

Atributi	koeficient
TEAM_SCORE	0,138288
MINUTES	0,075809
BIRTH_YEAR	0,027152
PENS_MADE	0,008574
XG_PLUS_MINUS	0,001818

Vir: lastno delo.

4.1.4 Vratarji

Pri naboru podatkov o vratarjih je Lasso natančno določil ključne attribute, ki so vidni v tabeli 5. V začetnih fazah te študije so bili za napovedni model upoštevani vratarji. Vendar smo se zaradi nezadostnega števila podatkov za ta položaj in edinstvene narave vratarjev, ki v primerjavi z igralci na zunanjem igrišču vključuje drugačne sklope atributov, odločili, da se vratarji izključijo iz končne analize. Ta omejitev je priznana in je bila upoštevana v obsegu te raziskave. Analiza se bo zato osredotočila na branilce, napadalce in srednje igralce.

Tabela 5: Izbrani atributi za vratarje

Atributi	Koeficient
PASSES_BLOCKED	0,164066
SCA_PASSES_DEAD	0,110589
DRIBBLES_COMPLETED	0,062834
TACKLES	0,007699

Vir: lastno delo.

4.1.5 Podrobna analiza vpliva ključnih atributov

Naša študija je pokazala, kako različni ključni atributi vplivajo na uspešnost igralcev na nogometnem igrišču, razen pri vratarjih, kjer je bila analiza omejena zaradi manjšega števila instanc. Kljub temu smo identificirali nekatere pomembne attribute za vratarje, kot so PASSES_BLOCKED, SCA_PASSES_DEAD, DRIBBLES_COMPLETED, in TACKLES, ki poudarjajo specifične vloge vratarjev, ki so precej različne od drugih igralcev, zato nima pomena primerjava z drugimi položaji na igrišču ali podrobna analiza njihovih atributov.

TEAM_SCORE, BIRTH_YEAR, in MINUTES so se izkazali kot pomembni napovedniki za večino položajev. To poudarja, da so ekipni uspehi, izkušnje in igralni čas ključni dejavniki za uspešnost vseh igralcev. Leto rojstva kaže, da starejši igralci, ki imajo več izkušenj, verjetno igrajo celotne tekme in prispevajo k zanesljivosti na igrišču.

Zanimivo je, da pri napadalcih ni tako pomembno število doseženih golov kot število pričakovanih golov, kar kaže na njihovo sposobnost vplivanja na igro tudi v kritičnih trenutkih. Dodatno, atributi, kot so GCA_PASSES_DEAD in GCA_PASSES_LIVE, nakazujejo, kako napadalci prispevajo k dinamiki in strategijam igre s svojimi spretnostmi in taktičnimi odločitvami.

Za branilce so izredno pomembni izključno atributi TEAM_SCORE, BIRTH_YEAR in MINUTES, ki odražajo vpliv ekipnih dosežkov, izkušenj in zanesljivosti igralca, ki se kažejo v igralnem času. Višje vrednosti teh atributov neposredno prispevajo k njihovi sposobnosti učinkovitega branjenja in vzdrževanja obrambne strukture. Stabilnost in zanesljivost, ki jo prinašajo starejši, izkušenejši igralci, so ključne za celovito obrambno delovanje, kjer sta na

igrišču odločilni kontinuiteta in zbranost. Zanimivo je, da tradicionalni obrambni statistični podatki, kot so tackles ali drugi neposredno obrambni atributi, niso bili izbrani. To lahko pomeni, da so obrambne sposobnosti, kot so predvidevanje igre in pozicioniranje, ki se ne merijo neposredno preko teh statistik, prav tako ali celo bolj pomembne za uspešnost branilcev.

Vezisti, ki predstavljajo mešanico napadalnih in obrambnih atributov, poudarjajo svoj prispevek s streli, pričakovanimi goli in kazenskimi poskusi. Njihova vloga pri povezovanju obrambnih in napadalnih priložnosti je ključna, s čimer poudarjajo svojo sposobnost vplivanja na različne faze igre. V naslednjih razdelkih je podrobno analiziran vpliv atributov na uspešnost branilcev, napadalcev in vezistov. Vratarji so iz te analize izključeni zaradi prej omenjenega pomanjkanja podatkov in njihove posebne vloge v nogometu.

Ta izčrpna analiza različnih igralnih položajev omogoča globok vpogled v dinamiko nogometne uspešnosti in dokazuje, da tradicionalne statistike morda ne zajamejo vedno celotnega bistva igralčevega vpliva na igrišču. Tako podrobno razumevanje pomaga selektorjem in trenerjem pri sprejemanju informiranih odločitev na podlagi zanesljivih podatkovnih vpogledov ter zagotavlja, da so ocene igralčevega potenciala čim bolj natančne in predvidljive.

4.2 Ocena uspešnosti modela strojnega učenja

4.2.1 Branilci

Tabela 6 omogoča enostavno primerjavo različnih modelov strojnega učenja in njihovih atributov točnosti. Stolpec Model navaja imena uporabljenih modelov strojnega učenja, kot sta logistična regresija in klasifikator z gradientnim povečevanjem. Stolpec točnosti prikazuje skupni odstotek pravih napovedi, ki jih je pripravil vsak model. Ta preglednica služi kot začetno merilo uspešnosti modelov in omogoča hitro primerjavo, kako natančno vsak model napoveduje rezultate v podatkovni zbirki.

Tabela 6: Najboljša točnost modela branilcev

Model	Točnost (angl. Accuracy)
Logistična regresija	0,91
Drevo odločanja	0,78
Naključni gozd	0,87
Klasifikator z gradientnim povečevanjem	0,87

Vir: lastno delo.

Tabela 7 podrobno prikazuje uspešnost modelov pri napovedovanju primerov razreda 0. Vsebuje stolpce za model, natančnost (angl. Precision), prepoznavnost (angl. Recall), F1

rezultat (angl. F1 Score) in podporo (angl. Support). Natančnost prikazuje delež pravih pozitivnih napovedi od vseh pozitivnih napovedi za razred 0. Prepoznavnost prikazuje delež dejanskih pozitivnih primerov, pravilno opredeljenih za razred 0. F1 rezultat je uravnoteženo merilo, ki združuje natančnost in prepoznavnost. Podpora prikazuje število primerov razreda 0 v testnem podatkovnem nizu.

Tabela 7: Zmogljivost razreda 0 branilcev

Model	Natančnost (angl. Precision)	Prepoznavnost (angl. Recall)	F1 rezultat (angl. F1 Score)	Podpora (angl. Support)
Logistična regresija	0,98	0,92	0,95	59
Drevo odločanja	0,91	0,83	0,87	59
Naključni gozd	0,93	0,92	0,92	59
Gradientno povečevanje	0,93	0,92	0,92	59

Vir: lastno delo.

Tabela 8 se osredotoča na uspešnost modelov pri napovedovanju primerov razreda 1. Struktura tabele je enaka kot pri atributih za razred 0, vendar se osredotoča na natančnost (angl. Precision), prepoznavnost (angl. Recall), F1 rezultat (angl. F1 Score) in podporo (angl. Support) za razred 1.

Tabela 8: Zmogljivost razreda 1 branilcev

Model	Natančnost (angl. Precision)	Prepoznavnost (angl. Recall)	F1 rezultat (angl. F1 Score)	Podpora (angl. Support)
Logistična regresija	0,64	0,90	0,75	10
Drevo odločanja	0,33	0,50	0,40	10
Naključni gozd	0,55	0,60	0,57	10
Gradientno povečevanje	0,55	0,60	0,57	10

Vir: lastno delo.

Tabela 9 podrobno opisuje optimalne konfiguracije za vsak model, ki so bile določene z optimizacijo. Vsebuje stolpca Model in Najboljši parametri. Stolpec Najboljši parametri

podrobno opisuje konfiguracije, ki so se izkazale za najučinkovitejše za doseganje najboljše učinkovitosti vsakega modela.

Tabela 9: Najboljši parametri branilcev

Model	Najboljši parametri
Logistična regresija	{'C': 0,0006951927961775605, 'penalty': 'l2'}
Drevo odločanja	{'max_depth': None, 'min_samples_leaf': 10, 'min_samples_split': 2}
Naključni gozd	{'bootstrap': True, 'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 200}
Gradientno povečevanje	{'learning_rate': 0,01, 'max_depth': 3, 'min_samples_leaf': 5, 'min_samples_split': 2, 'n_estimators': 200}

Vir: lastno delo.

Interpretacije so naslednje:

- Pri logistični regresiji je bila točnost 91,30 % z visoko natančnostjo (0,98) za neuspešne igralce in natančnostjo (0,64) za uspešne igralce. Ta model je uporaben za prepoznavanje neuspešnih igralcev z visoko natančnostjo, vendar je potrebno izboljšanje pri prepoznavanju uspešnih igralcev.
- Drevesu odločanja je bila točnost 78,26 % z uravnoteženimi stopnjami prepoznavnosti (0,83 za neuspešne in 0,50 za uspešne igralce). Ta model zagotavlja dobro ravnoesje med natančnostjo in prepoznavnostjo, vendar njegova nižja prepoznavnost uspešnih igralcev kaže potrebo po izboljšavah.
- Pri naključnem gozdu je bila točnost 86,96 % z močno prepoznavnostjo (0,92) za neuspešne igralce, vendar nižjo prepoznavnostjo (0,60) za uspešne igralce. Medtem ko je ta model zanesljiv za prepoznavanje neuspešnih igralcev, je treba njegovo delovanje pri prepoznavanju uspešnih igralcev izboljšati. Koristen je v primerih, kjer je izogibanje lažnim pozitivnim rezultatom ključnega pomena.
- Pri gradientnem povečevanju je bila točnost 86,96 % z uravnoteženimi F1-rezultati, kar kaže na močno splošno uspešnost. Ta model izstopa kot natančen in zagotavlja uravnotežene stopnje natančnosti in prepoznavnosti. Posebej je učinkovit za celovito napovedovanje uspešnosti igralcev in bi moral biti primarni model za strateške odločitve.

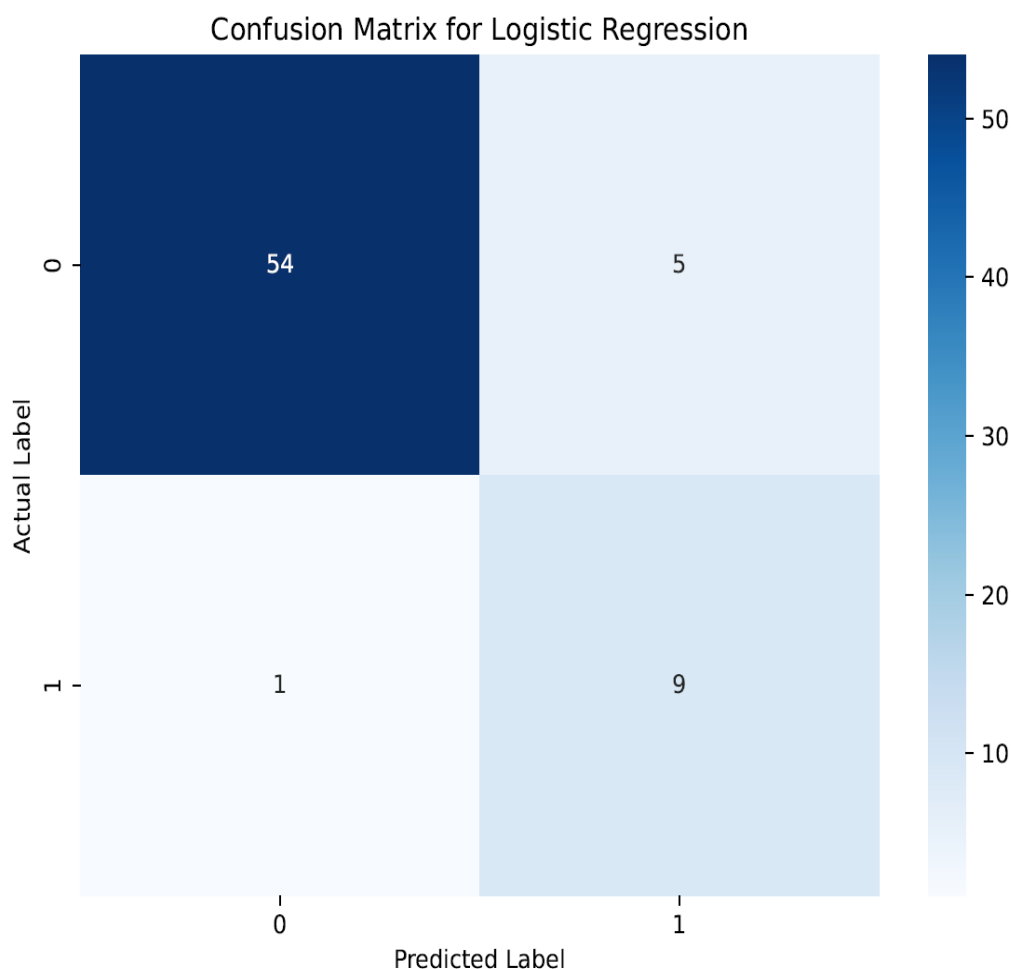
Matrika zmede na sliki 15 zagotavlja vizualni povzetek razlik v uspešnosti med dejanskimi in napovedanimi klasifikacijami. Tukaj je pomen posameznih delov matrike v našem kontekstu:

- TP: 9 – Model je pravilno napovedal 9 branilcev kot talentirane igralce.
- TN: 54 – Model je pravilno napovedal 54 branilcev kot netalementirane igralce.
- FP: 5 – Model je napačno napovedal 5 netalementiranih branilcev kot talentirane igralce.

- FN: 1 – Model ni prepoznal 1 talentiranega branilca, označil ga je kot netaletriranega.

Ta matrika nam pomaga oceniti sposobnost modela za pravilno identifikacijo talentiranih in netaletriranih igralcev, pri čemer izpostavlja njegove moči pri identifikaciji pravilno negativnih in področja, kjer bi lahko izboljšal, zlasti pri zmanjševanju napačno pozitivnih.

Slika 15: Zaslonska slika matrike zmede za branilce



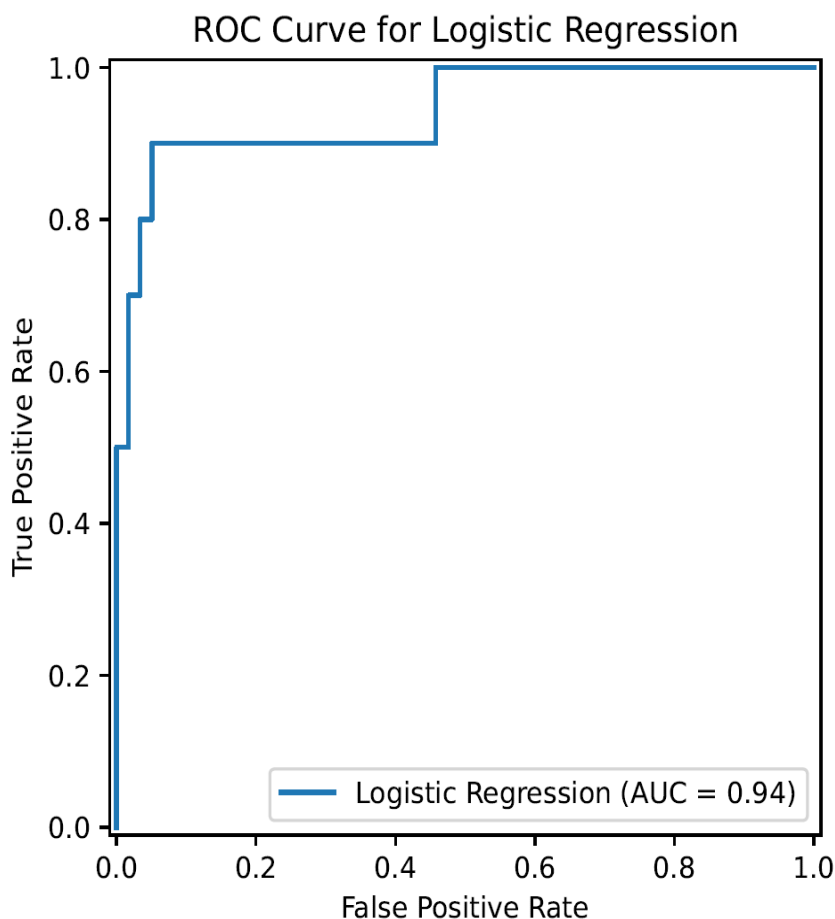
Vir: lastno delo.

Krivulja ROC je prikazana na sliki 16 in ilustrira diagnostično sposobnost klasifikatorja, ko se spreminja njegov prag diskriminacije. Za naš model logistične regresije so ključni vidiki naslednji:

- TPR, narisani na y-osi, ki predstavlja občutljivost modela ali njegovo sposobnost za pravilno zaznavanje talentiranih igralcev.
- FPR na x-osi, ki prikazuje delež talentiranih igralcev, ki so napačno klasificirani kot talentirani igralci.

Krivulja ROC zagotavlja celovit pogled na kompromis med priklicom in specifičnostjo ($1 - \text{FPR}$) pri različnih pragovih. Višja krivulja proti zgornjemu levemu kotu kaže na boljšo uspešnost.

Slika 16: Zaslonska slika krivulje sprejemnika delovanja za branilce



Vir: lastno delo.

Površina Pod Krivuljo $\text{AUC} = 0,94$ je merilo celotne sposobnosti modela za razlikovanje med razredi (talentirani in ne talentirani branilci). $\text{AUC} 0,94$ nakazuje, da je model zelo učinkovit pri razlikovanju med obema skupinama, pri čemer višja ocena kaže na boljši model. Model je precej zanesljiv pri identifikaciji pravilno negativnih in ima dobro AUC , kar kaže na močno diskriminativno moč. Zmanjšanje števila napačno pozitivnih bi lahko naredilo model še bolj učinkovit, saj bi bilo manj netalementirani branilcev napačno klasificiranih kot talentirani igralci.

Razumevanje teh grafov pomaga pri optimizaciji našega modela z nastavljanjem pragov in morebitno prefinjenostjo izbire značilnosti za izboljšanje natančnosti napovedi in zmanjšanje napačnih klasifikacij. Model logistične regresije s SMOTE je dosegel najvišjo točnost, vendar model gradientnega povečevanja s SMOTE kaže uravnotežene rezultate in

visoko natančnost. Te ugotovitve bodo vodile naše prihodnje analitične pristope in strateške odločitve pri razvoju igralcev in pripravi tekem.

4.2.2 Napadalci

Tabela 10 omogoča enostavno primerjavo različnih modelov strojnega učenja in njihovih splošnih atributov točnosti. Stolpec Model navaja imena uporabljenih modelov strojnega učenja, kot sta logistična regresija in klasifikator z gradientnim povečevanjem. Stolpec točnosti prikazuje skupni odstotek pravilnih napovedi, ki jih je pripravil vsak model. Ta preglednica služi kot začetno merilo uspešnosti modelov in omogoča hitro primerjavo, kako natančno vsak model napoveduje rezultate v podatkovnem nizu.

Tabela 10: Najboljša točnost modela napadalcev

Model	Točnost (Accuracy)
Logistična regresija	0,82
Drevo odločanja	0,89
Naključni gozd	0,86
Klasifikator z gradientnim povečevanjem	0,86

Vir: lastno delo.

Tabela 11 podrobno prikazuje uspešnost modelov pri napovedovanju primerov razreda 0. Vsebuje stolpce za model, natančnost (angl. Precision), prepoznavnost (angl. Recall), F1 rezultat (angl. F1 Score) in podporo (angl. Support). Natančnost prikazuje delež pravilnih pozitivnih napovedi od vseh pozitivnih napovedi za razred 0. Prepoznavnost prikazuje delež dejanskih pozitivnih primerov, pravilno opredeljenih za razred 0. F1 rezultat je uravnoteženo merilo, ki združuje natančnost in prepoznavnost. Podpora prikazuje število primerov razreda 0 v preskusni zbirki podatkov.

Tabela 11: Zmožljivost razreda 0 napadalcev

Model	Natančnost (angl. Precision)	Prepoznavnost (angl. Recall)	F1 rezultat (angl. F1 Score)	Podpora (angl. Support)
Logistična regresija	0,95	0,83	0,89	48
Drevo odločanja	0,98	0,90	0,93	48
Naključni gozd	0,95	0,88	0,91	48
Gradientno povečevanje	0,95	0,88	0,91	48

Vir: lastno delo.

Tabela 12 se osredotoča na uspešnost modelov pri napovedovanju primerov razreda 1. Struktura tabele je enaka kot pri atributih za razred 0, vendar se osredotoča na natančnost

(angl. Precision), prepoznavnost (angl. Recall), F1 rezultat (angl. F1 Score) in podporo (angl. Support) za razred 1.

Tabela 12: Zmogljivost razreda 1 napadalcev

Model	Natančnost (angl. Precision)	Prepoznavnost (angl. Recall)	F1 rezultat (angl. F1 Score)	Podpora (angl. Support)
Logistična regresija	0,43	0,75	0,55	8
Drevo odločanja	0,58	0,88	0,70	8
Naključni gozd	0,50	0,75	0,60	8
Gradientno povečevanje	0,50	0,75	0,60	8

Vir: lastno delo.

Tabela 13 podrobno opisuje optimalne konfiguracije za vsak model, ki so bile določene z optimizacijo. Vsebuje stolpca Model in Najboljši parametri. Stolpec Najboljši parametri podrobno opisuje konfiguracije, ki so se izkazale za najučinkovitejše za doseganje najboljše učinkovitosti vsakega modela.

Tabela 13: Najboljši parametri napadalcev

Model	Najboljši parametri
Logistična regresija	{'C': 0,23357214690901212, 'penalty': 'l2'}
Drevo Odločanja	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 20}
Naključni gozd	{'bootstrap': True, 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 100}
Gradientno povečevanje	{'learning_rate': 0,01, 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}

Vir: lastno delo.

Interpretacija je naslednja:

- Pri logistični regresiji je točnost 82,14 % z visoko natančnostjo (0,95) za neuspešne igralce in natančnostjo (0,43) za uspešne igralce: Ta model je uporaben za prepoznavanje neuspešnih igralcev z visoko natančnostjo, vendar je potrebno izboljšanje pri prepoznavanju uspešnih igralcev.
- Pri odločitvenem drevesu je bila točnost 89,29 % z uravnoteženimi stopnjami prepoznavnosti (0,90 za neuspešne in 0,88 za uspešne igralce). Ta model zagotavlja dobro ravnovesje med natančnostjo in prepoznavnostjo, kar ga naredi za zanesljivo

izbiro za prepoznavanje obeh skupin igralcev. Njegova interpretativnost je prednost pri praktičnem odločanju.

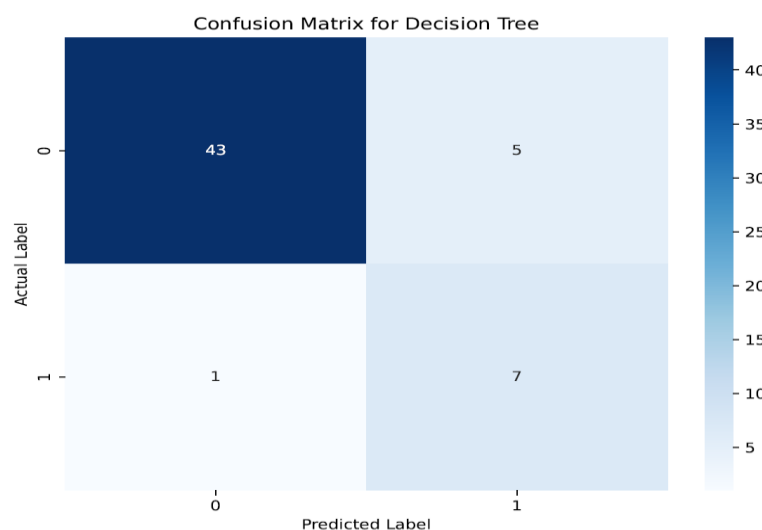
- Pri Drevesu odločanja je bila točnost 85,71 % z močno prepoznavnostjo (0,88) za neuspešne igralce, vendar nižjo prepoznavnostjo (0,75) za uspešne igralce. Medtem ko je ta model zanesljiv za prepoznavanje neuspešnih igralcev, je treba njegovo delovanje pri prepoznavanju uspešnih igralcev izboljšati. Koristen je v primerih, kjer je izogibanje lažnim pozitivnim rezultatom ključnega pomena.
- Pri Gradientnem povečevanju je bila točnost 85,71 % z uravnoteženimi F1-rezultati, kar kaže na močno splošno uspešnost. Ta model izstopa kot natančen in zagotavlja uravnotežene stopnje natančnosti in prepoznavnosti. Posebej je učinkovit za celovito napovedovanje uspešnosti igralcev in bi moral biti primarni model za strateške odločitve.

Matrika zmede na sliki 17 za model odločitvenega drevesa, ki se uporablja za analizo napadalcev, prikazuje naslednje:

- TP: 7 – Model je pravilno napovedal 7 napadalcev kot talentirane igralce.
- TN: 43 – Model je pravilno napovedal 43 napadalcev kot netalementirane igralce.
- FP: 5 – Model je napačno napovedal 5 netalementiranih napadalcev kot ključne igralce.
- FN: 1 – Model ni prepoznal 1 talentiranega napadalca, označil ga je kot netalementiranega.

Ta matrika nudi vpogled v natančnost modela pri razlikovanju med talentiranimi in netalementiranimi igralci, izpostavlja njegove moči pri identifikaciji pravilno negativnih in področja, kjer bi se lahko izboljšal, zlasti pri zmanjševanju napačno pozitivnih.

Slika 17: Zaslonska slika matrike zmede za napadalce



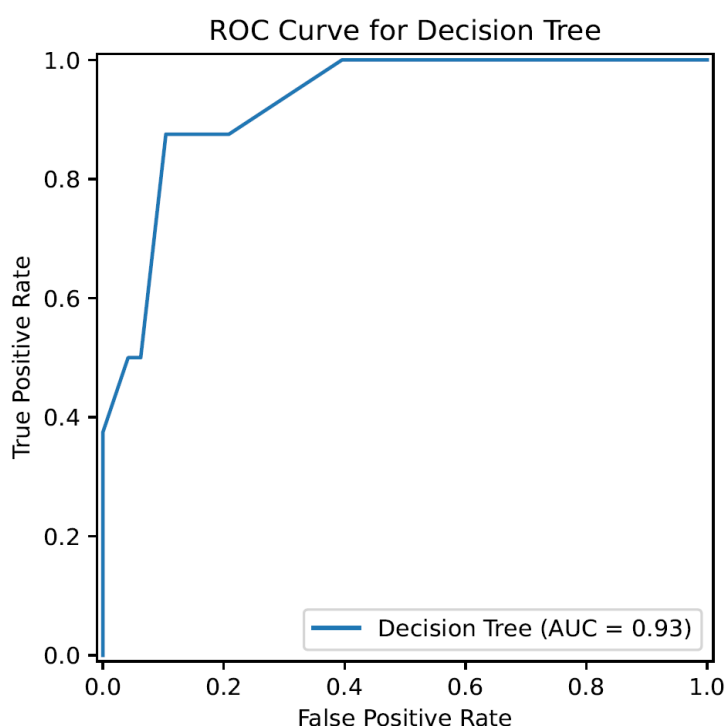
Vir: lastno delo.

Krivulja ROC prikazana na sliki 18 za model Drevesa odločanja prikazuje:

- TPR, narisani na y-osi, ki predstavlja občutljivost modela ali njegovo sposobnost za pravilno zaznavanje talentiranih igralcev.
- FPR na x-osi, ki prikazuje delež netalementiranih igralcev, ki so napačno klasificirani kot talentiranimi igralci.

Krivulja ROC zagotavlja celovit pogled na kompromis med priklicom in specifičnostjo ($1 - \text{FPR}$) pri različnih pragovih. Višja krivulja proti zgornjemu levemu kotu kaže na boljšo uspešnost.

Slika 18: Zaslonska slika krivulje sprejemnika delovanja za napadalce



Vir: lastno delo.

Površina Pod Krivuljo $\text{AUC} = 0,93$ je merilo celotne sposobnosti modela za razlikovanje med razredi (talentirani in netalementirani napadalci). $\text{AUC} 0,93$ nakazuje, da je model zelo učinkovit pri razlikovanju med obema skupinama, pri čemer višja ocena kaže na boljši model.

Analiza z matriko zmede in krivuljo ROC kaže, da ima model Drevesa odločanja dobro osnovo za razlikovanje med talentiranimi in netalementiranimi napadalnimi igralci. Z nadaljnjimi prilagoditvami in izboljšanjem modela, zlasti pri zmanjšanju napačno pozitivnih in povečanju občutljivosti za pravilno pozitivne, bi lahko še izboljšali njegovo natančnost in uporabnost.

Model odločitvenega drevesa s SMOTE je dosegel najvišjo točnost in uravnotežene rezultate, kar ga naredi zelo učinkovitega za napovedovanje uspešnosti napadalcev. Te ugotovitve bodo vodile naše prihodnje analitične pristope in strateške odločitve pri razvoju igralcev in pripravi tekem.

4.2.3 Vezisti

Tabela 14 omogoča enostavno primerjavo različnih modelov strojnega učenja in njihovih splošnih atributov točnosti. Stolpec Model navaja imena uporabljenih modelov strojnega učenja, kot sta logistična regresija in klasifikator z gradientnim povečevanjem. Stolpec točnosti prikazuje skupni odstotek pravih napovedi, ki jih je pripravil vsak model. Ta preglednica služi kot začetno merilo uspešnosti modelov in omogoča hitro primerjavo, kako natančno vsak model napoveduje rezultate v podatkovnem nizu.

Tabela 14: Najboljša točnost modela vezistov

Model	Točnost (angl. Accuracy)
Logistična regresija	0,82
Drevo odločanja	0,84
Naključni gozd	0,81
Klasifikator z gradientnim povečevanjem	0,85

Vir: lastno delo.

Tabela 15 podrobno prikazuje uspešnost modelov pri napovedovanju primerov razreda 0. Vsebuje stolpce za model, natančnost (angl. Precision), prepoznavnost (angl. Recall), F1 rezultat (angl. F1 Score) in podporo (angl. Support). Natančnost prikazuje delež pravih pozitivnih napovedi od vseh pozitivnih napovedi za razred 0. Prepoznavnost prikazuje delež dejanskih pozitivnih primerov, pravilno opredeljenih za razred 0. F1 rezultat je uravnoteženo merilo, ki združuje natančnost in prepoznavnost. Podpora prikazuje število primerov razreda 0 v testnem podatkovnem nizu.

Tabela 15: Zmožljivosti razreda 0 vezistov

Model	Natančnost (angl. Precision)	Prepoznavnost (angl. Recall)	F1 rezultat (angl. F1 Score)	Podpora (angl. Support)
Logistična regresija	0,90	0,87	0,88	53
Drevo odločanja	0,94	0,85	0,89	53
Naključni gozd	0,87	0,89	0,88	53
Gradientno povečevanje	0,89	0,92	0,91	53

Vir: lastno delo.

Tabela 16 se osredotoča na uspešnost modelov pri napovedovanju primerov razreda 1. Struktura tabele je enaka kot pri atributih za razred 0, vendar se osredotoča na natančnost (angl. Precision), prepoznavnost (angl. Recall), F1 rezultat (angl. F1 Score) in podporo (angl. Support). za razred 1.

Tabela 16: Zmogljivosti razreda 1 vezistov

Model	Natančnost (angl. Precision)	Prepoznavnost (angl. Recall)	F1 rezultat (angl. F1 Score)	Podpora (angl. Support)
Logistična regresija	0,59	0,67	0,62	15
Drevo odločanja	0,60	0,80	0,69	15
Naključni gozd	0,57	0,53	0,55	15
Gradientno povečevanje	0,69	0,60	0,64	15

Vir: lastno delo.

Tabela 17 podrobno opisuje optimalne konfiguracije za vsak model, ki so bile določene z optimizacijo. Vsebuje stolpca Model in Najboljši parametri. Stolpec Najboljši parametri podrobno opisuje konfiguracije, ki so se izkazale za najučinkovitejše za doseganje najboljše učinkovitosti vsakega modela.

Tabela 17: Najboljši parametri vezistov

Model	Najboljši parametri
Logistična regresija	{'C': 0,0886, 'penalty': 'l2'}
Drevo odločanja	{'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 10}
Naključni gozd	{'bootstrap': True, 'max_depth': None, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 200}

Vir: lastno delo.

Interpretacija je naslednja:

- Pri Logistični regresiji je dosežena točnost 82,35 % z visoko natančnostjo (0,90) za neuspešne igralce, vendar nižjo natančnostjo (0,59) za uspešne igralce. Ta model je uporaben za prepoznavanje neuspešnih igralcev z visoko natančnostjo, vendar je potrebno izboljšanje pri prepoznavanju uspešnih igralcev.
- Pri odločitvenem drevesu je bila nekoliko višja točnost 83,82 % z uravnoteženimi stopnjami prepoznavnosti (0,85 za neuspešne in 0,80 za uspešne igralce). Ta model zagotavlja dobro ravnovesje med natančnostjo in prepoznavnostjo, kar ga naredi

zanesljivo izbiro za prepoznavanje obeh skupin igralcev. Njegova interpretabilnost je prednost pri praktičnem odločanju.

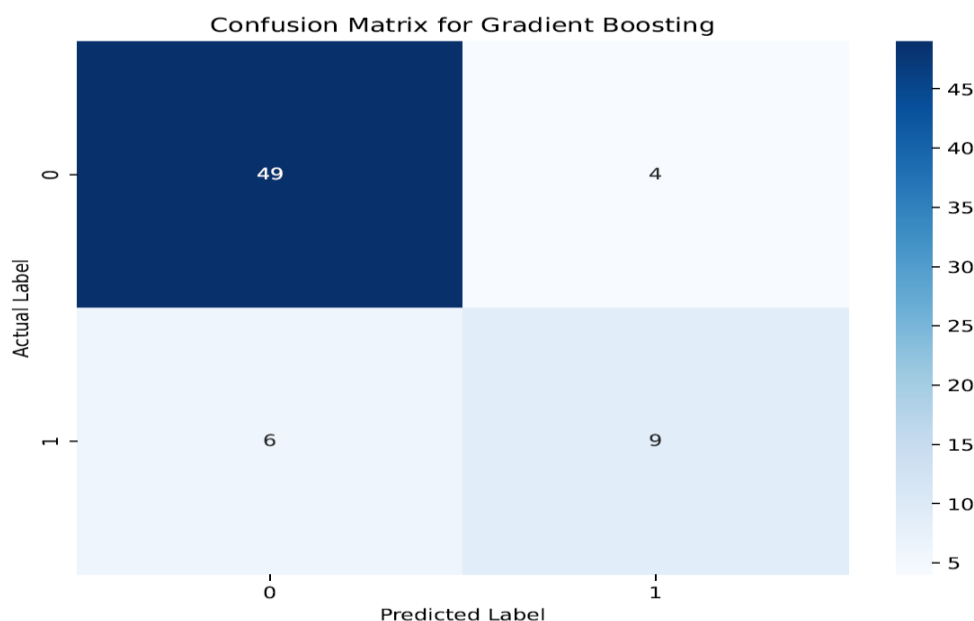
- Pri Naključnem gozdu je bila točnost 80,88 % z močno prepoznavnostjo (0,89) za neuspešne igralce, vendar nižjo prepoznavnostjo (0,53) za uspešne igralce. Medtem ko je ta model zanesljiv za prepoznavanje neuspešnih igralcev, je njegovo delovanje pri prepoznavanju uspešnih igralcev potrebno izboljšati. Koristen je v primerih, kjer je izogibanje lažnim pozitivnim rezultatom ključnega pomena.
- Pri Gradientnem povečevanju Najvišja točnost pri 85,29 % z uravnoteženimi F1-rezultati, kar kaže na močno splošno uspešnost. Ta model izstopa kot najnatančnejši, saj zagotavlja uravnotežene stopnje natančnosti in prepoznavnosti. Posebej je učinkovit za celovito napovedovanje uspešnosti igralcev in bi moral biti primarni model za strateške odločitve.

Matrika zmede na sliki 19 za model gradientnega spodbujanja, uporabljenega za analizo veznih igralcev, prikazuje naslednje:

- TP: 9 – Model je pravilno napovedal 9 veznih igralcev kot talentirane igralce.
- TN: 49 – Model je pravilno napovedal 49 veznih igralcev kot netalementirane igralce.
- FP: 4 – Model je napačno napovedal 4 netalementirane veziste kot talentirane igralce.
- FN: 6 – Model ni prepoznal 6 talentiranih vezistov, označil jih je kot netalementirane.

Zgornja matrika nudi vpogled v natančnost modela pri razlikovanju med talentiranimi in netalementiranimi igralci, izpostavlja njegove moči pri identifikaciji pravilno negativnih in področja, kjer bi lahko izboljšal, zlasti pri zmanjševanju napačno pozitivnih.

Slika 19: Zaslonka slika matrike zmede za veziste



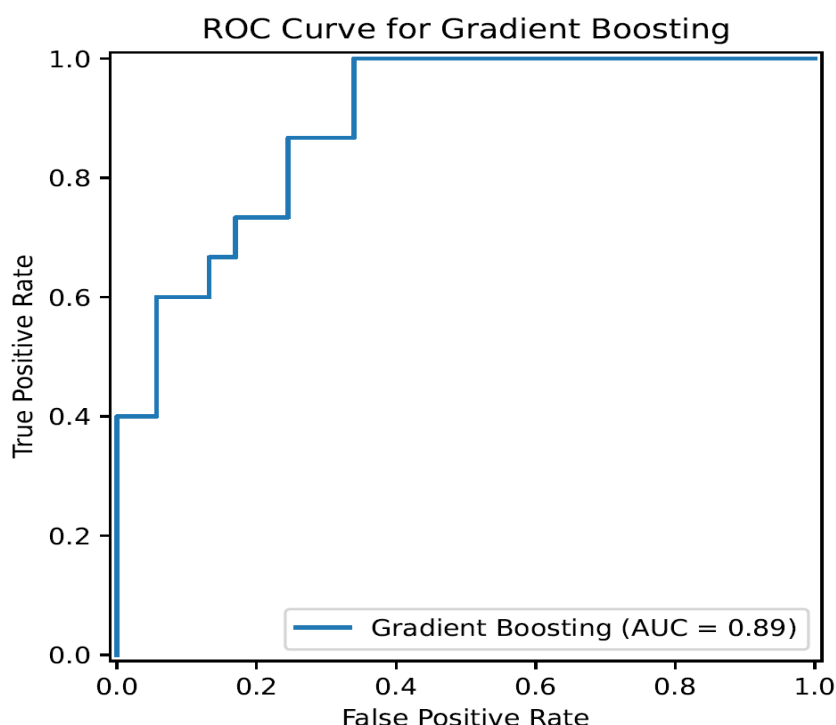
Vir: lastno delo.

Krivulja ROC prikazan na sliki 20 za model gradientnega spodbujanja prikazuje:

- TPR, narisanih na y-osi, ki predstavlja občutljivost modela ali njegovo sposobnost za pravilno zaznavanje talentiranih igralcev.
- FPR na x-osi, ki prikazuje delež netalementiranih igralcev, ki so napačno klasificirani kot talentirani igralci.

Krivulja ROC zagotavlja celovit pogled na kompromis med priklicom in specifičnostjo ($1 - \text{FPR}$) pri različnih pragovih. Višja krivulja proti zgornjemu levemu kotu kaže na boljšo uspešnost.

Slika 20: Zaslonska slika krivulje sprejemnika delovanja za veziste



Vir: lastno delo.

Površina Pod Krivuljo $\text{AUC} = 0,89$ je merilo celotne sposobnosti modela za razlikovanje med razredi (talentirani in netalementirani vezisti). $\text{AUC} 0,89$ nakazuje, da je model precej učinkovit pri razlikovanju med obema skupinama, pri čemer višja ocena kaže na boljši model.

Analiza z matriko zmede in krivuljo ROC kaže, da ima model gradientnega spodbujanja solidno osnovo za razlikovanje med talentiranimi in netalementiranimi veznimi igralci. Z nadaljnjimi prilagoditvami in izboljšanjem modela, zlasti pri zmanjšanju napačno pozitivnih in povečanju občutljivosti za pravilno pozitivne, bi lahko še izboljšali njegovo natančnost in uporabnost.

Model gradientnega povečevanja s SMOTE, z najvišjo točnostjo in uravnoteženimi metričnimi rezultati, je priporočljiv za napovedovanje uspešnosti vezistov. Vendar pa model odločitvenega drevesa prav tako kaže močno uspešnost in ponuja dodatno prednost interpretativnosti. Te ugotovitve bodo vodile naše prihodnje analitične pristope in strateške odločitve pri razvoju igralcev in pripravi tekem.

4.2.4 Vratarji

Zaradi omejenih podatkov, ki so na voljo za vratarje, in posebne narave njihove vloge, ki vključuje edinstvene attribute, ki niso neposredno primerljivi s tistimi, ki jih imajo igralci na drugih igralnih položajih, so bili izključeni iz napovednega modela. Ta odločitev je bila sprejeta, da bi zagotovili robustnost in zanesljivost modela. Ocena uspešnosti modela strojnega učenja bo tako temeljila na podatkih za branilce, napadalce in srednje igralce.

4.2.5 Izboljšave

Za izboljšanje teh modelov bi bilo treba razširiti in diverzificirati podatkovni niz. To vključuje zbiranje podatkov iz različnih lig in tekmovanj, ne samo iz svetovnih prvenstev, temveč tudi iz domačih lig, pokalov in celo iz mladinskih in rezervnih lig. Razširitev podatkovnega niza ne poveča samo količine podatkov, ki so na voljo za učenje, temveč omogoča tudi modelom, da se naučijo iz širšega spektra igralnih situacij in stilov igre, kar izboljša njihovo splošno robustnost in zmožnost generalizacije. Poleg tega bi bilo koristno nadalje razvijati in prilagajati attribute, ki so specifične za posamezne igralne položaje. Napredne tehnike ansambla, kot sta XGBoost ali LightGBM, ki so znane po svoji sposobnosti obvladovanja kompleksnih interakcij in nelinearnih odnosov, bi prav tako lahko izboljšale uspešnost modelov.

Sistematično nastavljanje parametrov modela prek metod, kot sta mrežno iskanje ali naključno iskanje, lahko pomaga pri iskanju optimalnih nastavitvev, ki uravnotežijo priklic in natančnost, zlasti za manjšinske razrede. Razvoj prilagojenih funkcij izgube, ki bolj kaznujejo napačno klasifikacijo manjšinskih razredov, bi lahko prav tako usmeril učenje modelov k boljšemu prepoznavanju talentiranih igralcev. Ta pristop bi bil še posebej učinkovit v kombinaciji z uporabo SMOTE, saj bi skupaj zagotovili boljše ravnotežje med zaznavanjem manjšinskih in večinskih razredov.

V sklepu, medtem ko SMOTE naslavlja težave z neuravnoteženostjo razredov, analiza jasno kaže, da obstajajo nadaljnji izzivi, ki jih je treba rešiti za izboljšanje napovedne moči in zanesljivosti modelov. Večplastni pristop, ki vključuje izboljšanje podatkovnih nizov, prilagajanje modelov in optimizacijo njihovih parametrov, bi lahko v veliki meri izboljšal uspešnost modelov v realnih aplikacijah. Ta pristop ne samo da izboljšuje statistično uspešnost modelov, ampak tudi zagotavlja, da modeli resnično prispevajo k boljšemu razumevanju in napovedovanju igralnih učinkov v različnih nogometnih scenarijih.

Z izboljšanjem zmogljivosti modela s poglobljenim inženiringom atributov odkrijemo več napovednih atributov, ki so še posebej koristne za manjšinske razrede. Dosledna učinkovitost klasifikatorja z gradientnim povečevanjem v različnih zbirkah podatkov poudarja njegovo vrednost kot robustnega in prilagodljivega orodja za različne izzive strojnega učenja. Nadaljnja avtomatizacija postopkov usposabljanja, uglaševanja in uvajanja modelov bi lahko te postopke poenostavila, kar bi zagotovilo, da modeli ostanejo na tekočem z razvijajočimi se podatkovnimi trendi, hkrati pa zmanjšalo potrebo po ročnem nadzoru.

Povečanje podatkovnega niza, zlasti z več predstavitvami iz manjšinskega razreda, bi lahko bistveno izboljšala osnove za usposabljanje modelov, kar bi povečalo njihove sposobnosti posploševanja in natančnost pri napovedovanju manjšinskega razreda. Poleg tega bi obogatitev podatkovnega niza lahko pomagala pri izboljšanju atributov in natančnem prilagajanju parametrov, da bi bolje zajeli posebnosti podatkovnega niza. Ta analiza ne izpostavlja le skupnih izzivov in prednosti pri napovednem modeliranju, temveč ponazarja tudi možnosti za avtomatizacijo modelov in pomemben vpliv, ki bi ga lahko imeli dodatni podatki na izboljšanje učinkovitosti modelov. Okrepljena prizadevanja za zbiranje podatkov, zlasti za manj zastopane razrede, bi lahko privedla do bolj uravnotežene uspešnosti modelov med razredi, kar bi zagotovilo pravičnost in povečalo praktično uporabnost teh modelov v realnih okoljih.

4.3 Rezultati multivariatne analize

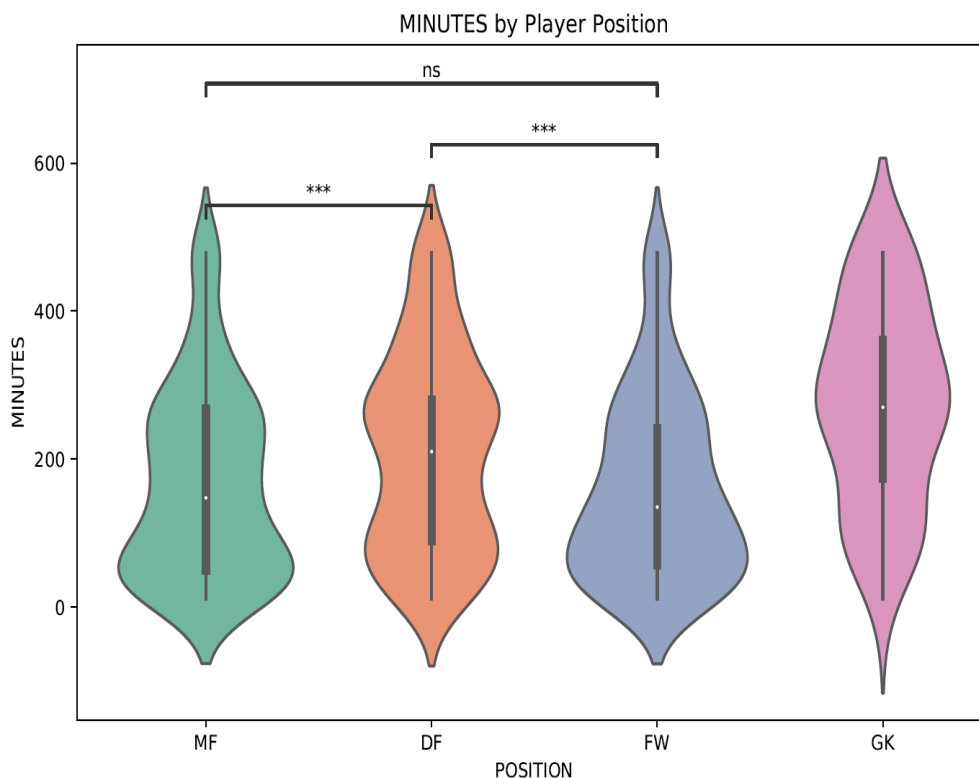
Analiza je osredotočena na oceno, kako se ključni atributi razlikujejo med položaji in kako te razlike vplivajo na razumevanje uspešnosti igralcev ter njihovih vlog:

- Statistična analiza za branilce je pokazala značilne razlike v času igranja ('MINUTES'), kjer je Kruskal-Wallisov test pokazal visoko statistično pomembnost (p -vrednost = $5,1358374433086104e-05$). Dunnov test je dodatno identificiral, da branilci igrajo bistveno več minut kot napadalci in vezisti, kar odraža njihovo ključno vlogo v obrambnih nalogah igre. Razlike v 'TEAM_SCORE' niso bile statistično pomembne, kar kaže na možno homogenost vpliva branilcev na ekipne rezultate ne glede na položaj.
- Za napadalce je bilo pomembno odkritje v 'GAMES_COMPLETE', kjer je Kruskal-Wallisov test pokazal izjemno statistično pomembnost (p -vrednost = $4,9475939324629496e-24$). Dunnov test je razkril, da napadalci dokončajo manj iger kot branilci, kar lahko odraža taktične zamenjave ali razlike v fizičnih zahtevah položaja. Te ugotovitve so ključne za razumevanje, kako upravljanje igralnega časa vpliva na zmogljivost in strategijo ekip.
- Vezisti so pokazali statistično pomembne razlike v 'MINUTES' (p -vrednost = $5,1358374433086104e-05$) podobno kot branilci, kar potrjuje njihovo stalno vlogo v dinamiki igre. Dunnov test ni pokazal pomembnih razlik med napadalci in vezisti, kar lahko nakazuje na podobnosti v zahtevah teh dveh položajev v sodobnem nogometu.

V celotni analizi je bilo opaziti, da so nekateri atributi, kot je 'BIRTH_YEAR', pokazali konstantnost čez položaje, kar nakazuje, da starostna dinamika ne vpliva bistveno na igralne razlike. Poleg tega so bili nekateri atributi, kot 'PENS_MADE' in 'GCA_PASSES_DEAD', izključeni iz analize zaradi konstantnih podatkov, kar poudarja potrebo po nadaljnji preiskavi teh variabel v kontekstu bolj specifičnih raziskovalnih nastavitvev.

Slika 21 grafično predstavlja Dunnov test za 'MINUTES' in jasno kaže, da obstajajo statistično pomembne razlike v številu odigranih minut med branilci in napadalci ter branilci in vezisti. Branilci igrajo bistveno več minut, kar potrjuje njihovo stalno in ključno vlogo v obrambi. Razlika med napadalci in vezisti ni statistično pomembna, kar lahko kaže na podobne taktične zahteve ali vzdržljivost teh dveh položajev.

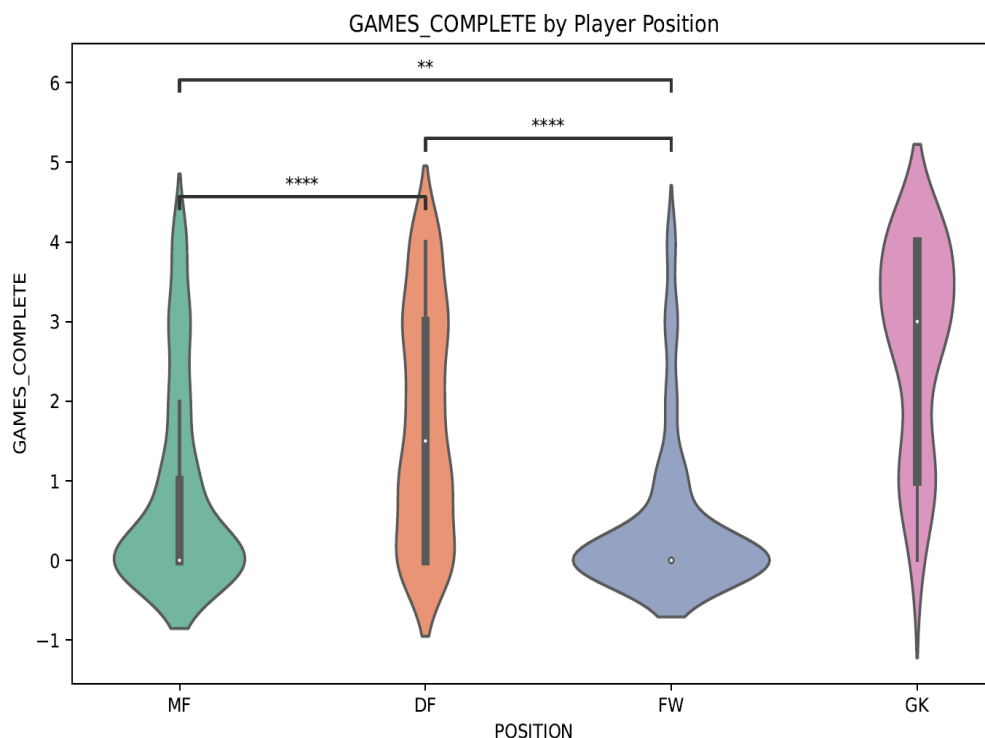
Slika 21: Dunnov test za minute



Vir: lastno delo.

Slika 22 grafično prikazuje rezultate Dunnovega testa za 'GAMES_COMPLETE', kjer so razlike med branilci in napadalci ter branilci in vezisti prav tako statistično pomembne. Podatki kažejo, da branilci dokončajo več iger v primerjavi z napadalci, kar lahko odraža manjšo rotacijo na obrambnih položajih ali večjo fizično zanesljivost teh igralcev. Zanimivo je, da so razlike med napadalci in vezisti minimalne, kar nakazuje na morebitne podobnosti v igralnem času med tema dvema skupinama igralcev.

Slika 22: Dunnov test za končane igre



Vir: lastno delo.

Samo dva atributa, 'MINUTES' in 'GAMES_COMPLETE', sta bila predmet grafičnih analiz zaradi njune statistične značilnosti in pomembnosti. Drugi atributi, kot sta 'BIRTH_YEAR' in 'XG_PLUS_MINUS', niso prikazali statistično pomembnih rezultatov ali so bili izključeni zaradi konstantnosti podatkov, kar omejuje njihovo uporabnost za vizualno analizo. V primerih, kjer podatki niso omogočali jasne razlage ali interpretacije, smo se osredotočili na tiste, ki ponujajo največjo vrednost za naše raziskovalne cilje.

5 RAZPRAVA

Rezultati poudarjajo naraščajoči trend nogometnih klubov, da pri sprejemanju odločitev uporabljajo pristop, ki temelji na podatkih. Kljub izzivom, povezanim z neuravnoteženostjo podatkov, so naši modeli z izjemno natančnostjo in točnostjo uspeli prepoznati zelo nadarjene igralce, podprte z obsežnimi statističnimi podatki. Ta zmožnost bistveno poenostavi postopek skavtstva, kar vodstvu kluba omogoča hitrejše in bolj utemeljene odločitve. Izboljšana kakovost podatkov in strokovni vpogledi lahko potencialno spremenijo nogometni klub v zelo dobičkonosno in uspešno podjetje. Kakovost podatkov je ključnega pomena, saj zagotavlja potrjevanje vpogledov skavtov in zmanjšuje tveganje slabega odločanja, kar je pozitivno povezano z uspešnostjo na igrišču in ustvarjanjem prihodkov.

Hipotezo 1 (Statistični pomen atributov se znatno razlikuje med različnimi položaji nogometnih igralcev) so naši rezultati močno podprli. Uporabljeni statistični testi, zlasti Kruskal-Wallisov test, so razkrili pomembne razlike v atributih, kot sta 'MINUTES' in 'GAMES_COMPLETE', med različnimi igralnimi položaji. Na primer, branilci so pokazali bistveno večje število 'MINUTES' v primerjavi z napadalci in vezisti, kar poudarja kritično in konstantno vlogo, ki jo branilci igrajo skozi celotno trajanje tekme. Ta razlika v času igre med položaji poudarja raznolike taktične in fizične zahteve, ki so postavljene na igralce v različnih vlogah. Poleg tega so pomembne ugotovitve pri 'GAMES_COMPLETE' za napadalce nakazale, da se vzorci zamenjav in stopnje dokončanja iger močno razlikujejo od branilcev. To smo podrobno analizirali z Dunnovim post-hoc testom, ki je pokazal, da napadalci dokončajo manj iger kot branilci, kar bi lahko pripisali strateškim odločitvam ali zahtevam po fizični vzdržljivosti, specifičnih za te vloge.

Te ugotovitve so ključne, saj potrjujejo, da določeni atributi imajo večji pomen za določene položaje, kar vpliva na to, kako bi klubi lahko pristopili k treningom, strategiji iger in oceni igralcev. Hipoteza 2 (Izbrani atributi bodo pokazali različne statistične porazdelitve med položaji igralcev, kar odraža edinstvene vloge in prispevke posameznih položajev) je bila delno podprta. Čeprav so bile opažene pomembne razlike v tem, kako so atributi, kot sta 'MINUTES' in 'GAMES_COMPLETE', porazdeljene med položaji, druga atributa, kot sta 'BIRTH_YEAR' in 'XG_PLUS_MINUS', nista pokazala pomembnih razlik. To kaže, da čeprav nekateri atributi igrajo ključno vlogo pri razlikovanju med vlogami igralcev, druge ne prispevajo razločno k razumevanju razlik med različnimi položaji. Pomanjkanje pomembnih razlik pri 'BIRTH_YEAR' nakazuje, da starost sama po sebi ne vpliva tako močno na vloge igralcev kot metrike igre, kot so odigrane minute ali dokončane igre.

Rezultati kažejo, da bi lahko športni analitiki in trenerji imeli koristi od osredotočanja na attribute, specifične za položaje, pri razvijanju modelov uspešnosti igralcev ali poročil o skavtih. Z razumevanjem, katere značilnosti so statistično pomembne za vsako vlogo, lahko ekipe bolj učinkovito prilagodijo svoje programe treningov in strategije rekrutiranja. Poleg tega poudarimo potrebo po nadaljnjih raziskavah glede atributov, ki niso pokazale pomembnih razlik. Raziskovanje, ali te atributi vplivajo na druge vidike uspešnosti igralcev ali dinamike ekipe, bi lahko razkrilo nove vpogled. Prav tako bi bilo koristno vključiti bolj granularne podatke, kot so fiziološke ali psihološke metrike, ki bi lahko razkrile globlje vpogled v medsebojno delovanje med položaji igralcev in njihovimi metrikami uspešnosti. Na splošno naše ugotovitve zagotavljajo niansiran pogled na to, kako različni atributi vplivajo na uspešnost igralcev čez različne nogometne položaje. S potrditvijo pomembnih razlik pri nekaterih ključnih atributih, medtem ko opažamo enotnost pri drugih, ta študija prispeva k bolj rafiniranemu razumevanju analize igralcev v nogometu. Prihodnje študije bi morale razmisliti o razširitvi podatkovnega nabora in vključitvi bolj raznolikih metrik, da bi nadgradile temelje, ki so bili vzpostavljeni tukaj.

Naša analiza je trenutno omejena na podatke o igralcih s svetovnega prvenstva 2022 v Katarju. Omejeni podatki vsebujejo 680 igralcev iz 32 držav na samo 64 tekmah v enem

mesecu, kar omejujejo naše modele. Če bi naš nabor podatkov razširili na več igralcev, narodnosti in tekem v daljšem obdobju, bi izboljšali točnosti in učinkovitost napovedovanja naših modelov. Naši modeli so trenutno osredotočeni le na tehnične spretnosti, vendar bi jim koristil širši, multidisciplinarni pristop, ki vključuje tehnične, fizične, socialne in psihološke vidike razvoja igralcev. Perspektivna uporaba strojnega učenja osvetljuje nove poti v razvoju igralcev in poudarja njegov potencial močnega orodja v športni znanosti. Vendar retrospektivna narava naših podatkov in njihov omejen obseg poudarjata potrebo po nadaljnjih raziskavah z uporabo naprednih tehnik analize podatkov.

Pri izvajanju te raziskave je bilo treba zaradi pomanjkanja podatkov iz napovednega modela izključiti vratarje. Vratarji predstavljajo edinstven položaj v nogometu, ki v primerjavi z drugimi igralnimi položaji zahteva drugačne attribute in metrike uspešnosti. Zaradi njihove specializirane narave atributov, ki prispevajo k njihovi uspešnosti, niso neposredno primerljivi z atributi branilcev, napadalcev in srednjih igralcev. Ta omejitev poudarja potrebo po obsežnejšem zbiranju podatkov o vratarjih v prihodnjih študijah. Ločen model, prilagojen vratarjem, bi bil primernejši za natančen zajem dejavnikov, ki vplivajo na njihovo uspešnost.

Raziskovanje alternativnih opredelitev odvisne spremenljivke, ki je trenutno binarna, bi lahko odprlo možnosti, ki presegajo logistično regresijo in klasifikacijske napovedne modele. Na primer, ocenjevanje uspešnosti igralcev na lestvici (npr. 1–100) bi omogočilo uporabo različnih modelov za ocenjevanje uspešnosti igralcev, kar bi dodalo element subjektivnosti, vendar bi lahko povečalo učinkovitost rezultatov. Druga možnost je, da bi združevanje igralcev v grozde na podlagi njihovih statističnih podatkov odkrilo vzorce ali tipe igralcev, ki s standardno statistično analizo niso takoj razvidni, kar bi pomagalo pri prepoznavanju podcenjenih talentov ali strateškem oblikovanju sestave ekipe. Vendar pri takšnih poenostavitvah obstaja tveganje, da bodo zanemarjene edinstvene spretnosti igralcev ali ključne igralne sposobnosti, če ne bodo dovolj statistično zajete. Brez strokovne interpretacije takšna združevanja morda ne bodo prinesla uporabnih spoznanj, kar poudarja pomen povezovanja strokovnega znanja s področja s statističnimi analizami za učinkovito uporabo pri upravljanju ekipe ali skavtiranju.

Prihodnje raziskave bi morale raziskati alternativne opredelitve odvisne spremenljivke in različne pristope modeliranja, ki bi lahko zajeli širši spekter igralčevih sposobnosti. Raziskovanje metod, kot je grozdenje, bi lahko razkrilo nove vzorce in strateška spoznanja, ki jih z običajnimi tehnikami ni lahko odkriti. V idealnem primeru bi bilo mogoče razmisliti o podrobnejši kategorizaciji igralnih položajev in ne le štirih osnovnih. Vendar je potrebna previdnost, da ne bi preveč poenostavili edinstvenih talentov igralcev, kar poudarja potrebo po združitvi statističnih vpogledov z globokim področnim znanjem. To tekoče raziskovanje uporabe strojnega učenja v športni analitiki poudarja potrebo po nenehnih raziskavah in inovacijah na tem področju ter nakazuje, da bo kombinacija natančnih podatkov, metodološkega napredka in celostnega pristopa še izboljšala naše razumevanje in vzgojo nogometnih talentov.

6 SKLEP

Vključevanje strojnega učenja v kvantitativno ocenjevanje nogometnih talentov je obetavna priložnost za izboljšanje postopkov identifikacije in razvoja. Z uporabo vpogledov, ki temeljijo na podatkih, lahko nogometni klubi natančno določijo in vzgajajo obetavne talente ter uskladijo operativne strategije s celostnim razumevanjem sposobnosti igralcev, ki vključujejo fizične, tehnične, psihološke in socialne razsežnosti. Vendar pa izzivi ostajajo, kot so spremenljivost atributov igralcev skozi čas, nelinearne razvojne faze športnikov in različne spremenljivke, ki edinstveno zaznamujejo posamezne talente. Poleg tega dinamična narava nadarjenosti, kot jo opisujejo trenerji in praktiki, nakazuje, da sta prepoznavanje in razvoj nadarjenosti reflektivni, stalni proces in ne statična ocena, ki temelji na takojšnjih opazanjih.

Nabor podatkov zagotavlja bogato paleto atributov za nogometaše na različnih položajih, pri čemer so uravnoteženo zastopani vsi, razen vratarjev, pri katerih so atributi posebej prilagojeni njihovi vlogi. Analiza razkriva pomembne korelacije med spremenljivkami, kar kaže na to, da so nekateri atributi podskupine drugih. Presenetljivo so korelacije tudi med navidezno nepovezanimi spremenljivkami, kar dodatno zaplete analizo. Zanesljiva priprava podatkov zagotavlja popolnost in ustrezno ravnanje z manjkajočimi vrednostmi, kar je ključno za zanesljivost poznejših modelov strojnega učenja. Za položaje, kot so vezisti, napadalci, branilci, se ključni atributi, kot so leto rojstva, odigrane tekme in ekipni rezultat (ki označuje število nadarjenih igralcev v isti ekipi), pojavljajo kot napovedniki uspešnosti. Atribut leta rojstva kaže, da imajo prednost starejši in izkušenejši igralci, kar se ujema z atributom odigranih tekem in poudarja, da uspešni igralci pogosto odigrajo celotne tekme. Zanimivo je, da atribut TEAM_SCORE, čeprav ni intuitiven, morda nakazuje, da nekatere države proizvajajo večje število uspešnih igralcev. Med raziskavo je bilo ugotovljeno, da razpoložljivi podatki za vratarje ne zadostujejo za razvoj zanesljivega napovednega modela. Poleg tega je treba vratarje zaradi njihovih edinstvenih spretnosti in kazalnikov uspešnosti analizirati posebej. Zato so bili vratarji izključeni iz analize, ki se je osredotočila na branilce, napadalce in veziste. Ta omejitev poudarja pomen zbiranja podrobnejših podatkov za vratarje in upoštevanja njihove edinstvene vloge v prihodnjih študijah, da se zagotovi celovito modeliranje.

V zaključku tega magistrskega dela smo potrdili, da se statistični pomen in porazdelitve določenih atributov, kot sta 'MINUTES' in 'GAMES_COMPLETE', zelo razlikujejo med različnimi igralnimi položaji. Branilci so pokazali večjo količino odigranih minut in večje število dokončanih tekem, kar odraža njihovo neprekinjeno in ključno vlogo skozi celotno tekmo. To potrjuje našo hipotezo in ponuja pomembne vpogleds za športno analitiko, še posebej v smislu izboljšanja ocenjevanja igralcev, prilagajanja programov treninga in taktičnih odločitev. Kljub temu so nekateri atributi kazali, da ne prispevajo tako izrazito k razumevanju razlik med položaji, kar poudarja potrebo po nadaljnjem raziskovanju teh atributov v prihodnosti. Ta ugotovitev nas spodbuja k razširitvi podatkovnega nabora, da bi

vključili bolj raznolike metrike, kot so fiziološki in psihološki podatki, ki bi lahko razkrili globlje vpogled v uspešnost igralcev in dinamiko med različnimi položaji.

Priporočamo tudi nadaljnjo uporabo naprednejših analitičnih modelov ali tehnik, ki bi lahko še dodatno izpilile razumevanje podatkov in vodile do bolj niansiranih interpretacij uspešnosti igralcev v različnih kontekstih in igralnih ravneh. Takšne študije bi lahko še dodatno podprle strateške odločitve v upravljanju igralcev in oblikovanju igralne strategije. Omejitve magistrskega dela izpostavljajo potrebo po nadaljnjih raziskavah in razširitvi podatkovnega nabora, kar bi omogočilo večjo natančnost in učinkovitost modelov za analizo nogometnih igralcev. Te omejitve vključujejo naslednje:

- Omejen obseg podatkov: Raziskava temelji na podatkih z enega turnirja, svetovnega prvenstva 2022, ki vključuje 680 igralcev iz 32 držav na samo 64 tekmah v enem mesecu. To omejuje možnosti generalizacije in natančnosti modelov. Povečanje števila igralcev, narodnosti in tekem bi izboljšalo rezultate.
- Izključitev vratarjev: Zaradi pomanjkanja podatkov o specifičnih atributih za vratarje so bili ti izključeni iz modelov. Vratarji predstavljajo edinstven položaj, ki zahteva drugačne metrike uspešnosti. V prihodnjih raziskavah bi bilo treba razviti ločen model, prilagojen vratarjem.
- Binarna opredelitev odvisne spremenljivke: Uporaba binarne odvisne spremenljivke omejuje uporabo naprednejših metod. Ocena uspešnosti na lestvici ali uporaba grozdenja bi lahko prinesla bolj niansirane rezultate.
- Pomanjkanje multidisciplinarnega pristopa: Trenutni modeli ne vključujejo celostnega pristopa, ki bi vključeval več disciplin (tehnične, fizične, psihološke metrike), kar omejuje globlje razumevanje vpliva atributov na igralčevo uspešnost.

Naša raziskava prispeva k bolj rafiniranemu in empirično podprtemu razumevanju analize igralcev v nogometu, kar je še posebej pomembno v času hitrih sprememb v športni znanosti in analitiki. Kljub omejitvam naše študije poudarjamo pomen nadaljnjih raziskav na tem področju, ki bi zagotovile zanesljivost in praktično uporabnost ugotovitev v resničnih okoliščinah. S tem zaključujemo našo raziskavo, ki ne le da je potrdila določene hipoteze, ampak je tudi odprla vrata za nove priložnosti za poglobljeno razumevanje športne analitike in izboljšanje športne prakse. Analiza različnih položajev ne razkriva le skupnih izzivov, temveč tudi možnosti za avtomatizacijo uporabe modelov, kar povečuje učinkovitost uvajanja modelov. Poleg tega bi lahko dodajanje več podatkov, zlasti za premalo zastopane razrede, znatno izboljšalo učinkovitost modelov, kar bi omogočilo točnejše napovedi in boljše posplošitev na nove podatkovne nize.

Skratka, uporaba strojnega učenja pri prepoznavanju nogometnih talentov lahko bistveno spremeni ekonomsko dinamiko v klubih. Optimizacija procesov skavtstva in izboljšanje točnosti napovedi talentov omogočata bolj premišljene naložbene odločitve, kar zmanjšuje finančna tveganja, povezana z nakupi igralcev. Ta tehnološki napredek klubom omogoča, da prepoznajo podcenjene talente, povečajo prizadevanja skavtov in potencialno odkrijejo

skrite dragulje, ki bi lahko ponudili konkurenčno prednost. Poleg tega lahko pristop, ki temelji na podatkih, spodbuja strateški operativni etos, osredotočen na dolgoročni razvoj igralcev in ustvarjanje vrednosti, kar ne le krepi položaj kluba na trgu s pametnejšim zaposlovanjem, temveč tudi povečuje potencialno donosnost naložb z razvojem in/ali prodajo igralcev po višjih vrednostih.

LITERATURA IN VIRI

1. Adam, D. (2022) Science and the World Cup: how big data is transforming football. *Nature*, 611(7936), 444–446.
2. Al-Asadi, M. A. in Tasdemir, S. (2022). Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques. *IEEE Access*, 10, 22631–22645.
3. Almulla, J., in Alam, T. (2020). Machine Learning Models Reveal Key Performance Metrics of Football Players to Win Matches in Qatar Stars League. *IEEE Access*, 8, 213695–213705.
4. Baboota, R., in Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, 35(2), 741–755.
5. Behravan, I., in Razavi, S. M. (2021). A novel machine learning method for estimating football players' value in the transfer market. *Soft Computing*, 25(3), 2499–2511.
6. Bergkamp, T. L. G., Frencken, W. G. P., Niessen, A. S. M., Meijer, R. R., in Den Hartigh, Ruud. J. R. (2022). How soccer scouts identify talented players. *European Journal of Sport Science*, 22(7), 994–1004.
7. Bojer, C. S., in Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, 37(2), 587–603.
8. Chase, C. (2020). *The Data Revolution: Cloud Computing, Artificial Intelligence, and Machine Learning in the Future of Sports*. Springer
9. Clarke, N. J., Cushion, C. J., in Harwood, C. G. (2018). Players' understanding of talent identification in early specialization youth football. *Soccer in Society*, (str. 1–15).
10. Dageville, B., Cruanes, T., Zukowski, M., Antonov, V., Avanes, A., Bock, J., Claybaugh, J., Engovatov, D., Hentschel, M., Huang, J., Lee, A. W., Motivala, A., Munir, A. Q., Pelley, S., Povinec, P., Rahn, G., Triantafyllis, S., in Unterbrunner, P. (2016). The Snowflake Elastic Data Warehouse.V *Proceedings of the 2016 International Conference on Management of Data*. Association for Computing Machinery.
11. García-Aliaga, A., Marquina, M., Coterón, J., Rodríguez-González, A., in Luengo-Sánchez, S. (2021). In-game behaviour analysis of football players using machine

- learning techniques based on player statistics. *International Journal of Sports Science in Coaching*, 16(1), 148–157.
12. Goes, F. R., Meerhoff, L. A., Bueno, M. J. O., Rodrigues, D. M., Moura, F. A., Brink, M. S., Elferink-Gemser, M. T., Knobbe, A. J., Cunha, S. A., Torres, R. S., in Lemmink, K. A. P. M. (2021). Unlocking the potential of big data to Podpora tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science*, 21(4), 481–496.
 13. Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C., in Meyer, T. (2019). Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science in Coaching*, 14(6), 798–817.
 14. Hucaljuk, J., in Rakipović, A. (2011). Predicting football scores using machine learning techniques. V *Zbornik 34. Mednarodne konvencije MIPRO* (str. 1623–1627). IEEE.
 15. Hughes, M., in Franks, I. M. (2008). *The essentials of performance analysis: An introduction*. Routledge.
 16. Inkwina. (2008). *Soccer field transparent*.
https://commons.wikimedia.org/wiki/File:Soccer_Field_Transparant.svg
 17. Jamil, M., Phatak, A., Mehta, S., Beato, M., Memmert, D., in Connor, M. (2021). Using multiple machine learning algorithms to classify elite and sub-elite goalkeepers in professional men's football. *Scientific Reports*, 11(1), 22703.
 18. Joseph, A., Fenton, N. E., in Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7), 544–553.
 19. Kelly, A. L., Williams, C. A., Cook, R., Sáiz, S. L. J., in Wilson, M. R. (2022). A Multidisciplinary Investigation into the Talent Development Processes at an English Football Academy: A Machine Learning Approach. *Sports*, 10(10), 159.
 20. Larkin, P., in Reeves, M. J. (2018). Junior-elite football: Time to re-position talent identification? *Soccer in Society*, (str. 1–10).
 21. Lee, C.-K., in Taylor, T. (2005). Critical reflections on the economic impact assessment of a mega-event: the case of 2002 FIFA World Cup. *Tourism Management*, 26(4), 595–603.
 22. Liu, X. F., Liu, Y.-L., Lu, X.-H., Wang, Q.-X., in Wang, T.-X. (2016). The Anatomy of the Global Football Player Transfer Network: Club Functionalities versus Network Properties. *PLOS ONE*, 11(6), e0156504.
 23. Marcelino, R., Sampaio, J., Amichay, G., Gonçalves, B., Couzin, I. D., in Nagy, M. (2020). Collective movement analysis reveals coordination tactics of team players in football matches. *Chaos, Solitons in Fractals*, 138, 109831.
 24. McGarry, T. (Ur.). (2013). *Routledge handbook of sports performance analysis*. Routledge.

25. Meylan, C., Cronin, J., Oliver, J., in Hughes, M. (2010). Talent Identification in Soccer: The Role of Maturity Status on Physical, Physiological and Technical Characteristics. *International Journal of Sports Science in Coaching*, 5(4), 571–592.
26. Miller, P. K., Cronin, C., in Baker, G. (2015). Nurture, nature and some very dubious social skills: An interpretative phenomenological analysis of talent identification practices in elite English youth soccer. *Qualitative Research in Sport, Exercise and Health*, 7(5), 642–662.
27. Morsinkhof, J. (2014). *The influence of social identity theory on voting behavior in the FIFA Ballon d'Or competition*. Diplomsko delo, Univerza Twente.
28. Peeters, T. (2018). Testing the Wisdom of Crowds in the field: Transfermarkt valuations and international soccer results. *International Journal of Forecasting*, 34(1), 17–29.
29. Preston, T. (2012). The determinants of football transfer market values: an age of financial restraint. *Diplomsko delo*, Univerza Warwick.
30. Radicchi, E., in Mozzachiodi, M. (2016). Social Talent Scouting: A New Opportunity for the Identification of Football Players? *Physical Culture and Sport. Studies and Research*, 70(1), 28–43.
31. Randles, B. M., Pasquetto, I. V., Golshan, M. S., in Borgman, C. L. (2017). Using the Jupyter Notebook as a Tool for Open Science: An Empirical Study. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, (str. 1–2).
32. Stein, M., Janetzko, H., Seebacher, D., Jäger, A., Nagel, M., Hölsch, J., Kosub, S., Schreck, T., Keim, D., in Grossniklaus, M. (2017). How to Make Sense of Team Sport Data: From Acquisition to Data Modeling and Research Aspects. *Data*, 2(1), 2.
33. Swaptr. (2022). *FIFA World Cup 2022 Player Data*. <https://www.kaggle.com/swaptr/fifa-world-cup-2022-player-data>
34. Unnithan, V., White, J., Georgiou, A., Iga, J., in Drust, B. (2012). Talent identification in youth soccer. *Journal of Sports Sciences*, 30(15), 1719–1726.
35. Waldron, M., in Worsfold, P. (2010). Differences in the Game Specific Skills of Elite and Sub-Elite Youth Football Players: Implications for Talent Identification. *International Journal of Performance Analysis in Sport*, 10(1), 9–24.
36. Wand, T. (2022). Analysis of the Football Transfer Market Network. *Journal of Statistical Physics*, 187(3), 27.
37. Wolf, G. (2010, april 28). *The Data-Driven Life*. <https://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html>