Fakultet tehničkih nauka

Univerzitet u Novom Sadu

Prijedlog predmetnog projekta

PREDVIĐANJE NAJBOLJE POZICIJE FUDBALSKOG IGRAČA

1. Uvod

Ovim dokumentom dat je prijedlog projekta iz predmeta Sistemi za istraživanje i analizu podataka. Data je definicija problema i objašnjen je doprinos i značaj njegovog rješavanja. Navedeni su i ukratko opisani radovi koji se bave rješavanjem srodnih problema. Opisani su: skup podataka, metodologija rješavanja problema, kao i metod evaluacije. Naveden je softver koji će biti korišćen.

Prilikom implementacije biće oslanjanja na ranije radove na sličnu temu, više o tome će biti u stavki “literatura”. Osim toga govoriće se o korišćenim softverima, metodama procjene tačnosti rezultata i planu rada.

Tema projekta je predviđanje najbolje pozicije igrača na osnovu njegovih vještina. Na raspolaganju je više hiljada igrača sa njihovom fizičkim sposobnostima, a u obzir je uzeto 6 različitih pozicija.

2. Definicija projekta

Ovaj projekat se bavi predviđanjem najbolje pozicije igrača na osnovu vještina. Ideja je da pozicija igrača u velikoj mjeri zavisi od njegovih fizičkih, tehničkih osobina. Za potrebe ovoga projekta u obzir su uzete sledeće pozicije:

- GK - Goalkeeper (Golman)
- CB - Centre-back (Štoper)
- WB - Wing-back (Bek)
- CM – Centre midfield (Centralni vezni)
- W – Winger (Krilo)
- ST – Striker (Napadač)

3. Motivacija

S obzirom da je fudbal najvažnija sporedna stvar na svijetu, o popularnosti fudbala nećemo mnogo govoriti. Osnovni motiv ovoga projekta jeste olakšavanje posla fubalskim stručnjacima koji se profesionalno bave svojim poslom. Rješenje ovoga projekta bi u velikoj mjeri olakšalo posao trenerima čiji klubovi se susreću sa mnoštvom povreda, pa je potrebno često vršiti promjenu formacije tima. Takođe to bi pomoglo da se u maksimalnoj mjeri iskoristi potencijal svakog igrača i time postignu što bolji rezultati. Posao skauta bi bio znatno olakšan, što bi omogućilo efikasniji izbor novih igrača koji bi popunili postojeće "rupe" u timu.

4. Pregled vladajućih stavova i shvatanja u literaturi

- [1] Predicting Player Position for Talent Identification in Association Football, Nazim Razali and Aida Mustapha (https://iopscience.iop.org/article/10.1088/1757-899X/226/1/012087)

  Rad se bavi predviđanjem i prepoznavanjem talenta u fudbalu na osnovu individualnih kvaliteta igrača, koji su klasifikovani u 3 grupe: fizičke, mentalne i tehničke. Tu spadaju: brzina, agilnost, skok, visina, snaga, sposobnost čitanja igre, smirenost, kreativnost, samouvjerenost, sposobnost u predvođenju tima, pas igra, šut, igra glavom, završnica, uklizavanje, ubačaj, odbrana. Kombinacija kvaliteta ocijenjena od strane trenera je zatim korišćena za predviđanje pozicije igrača koja mu najbolje odgovara u određenoj formaciji tima.

  Evaluacija predloženog okvira je dvostruka; kvantitativno putem klasifikacijskih eksperimenata za predviđanje položaja igrača i kvalitativno putem stranice za identifikaciju talenata koja je razvijena za postizanje istog cilja. Rezultati eksperimentiranja klasifikacije pomoću Bayesian Networks, Decision Trees, i K-Nearest Neighbor pokazali su prosječnost od 98% tačnosti, što će promovisati dosljednost u odlučivanju i eliminisati subjektivnu pristrasnost u odabiru tima. Pozitivne kritike na fudbalskoj stranici takođe pokazuju da je rad dovoljan da posluži kao osnova razvoja inteligentnog sistema upravljanja timom u različitim sportovima, gdje se mogu nadgledati i identifikovati napredak i razvoj sportskih igrača. Primjera radi, defanzivnom veznom igraču potreban je visoki kvalitet u pas igri, visoke mogućnosti šuta i povratnog pasa, visok potencijal "čitanja" igre, kao i velika fizička spremnost. Najveća tačnost klasifikacije postignuta tokom eksperimenata iznosila je 99% za Bayesian Networks, nakon čega slijedi 98% za Decision Trees i 97% za K-Nearest Neighbor.

- [2] Football Player Position Prediction, Neil Chandarana
  (https://towardsdatascience.com/classification-football-player-position-prediction-part-2-f5cc15163f8d)

  Cilj projekta jeste implementacija modela za preporučivanje pozicije igrača na osnovu fizičkih atributa. Rješenje je predstavljeno uz pomoć linearnih tehnika. Naime, linearna regresija, linearna diskriminatorna analiza (LDA) i kvadratna diskriminatorna analiza (QDA) i multinomalna logistička regresija (MLR).

  MLR je višeklasna ekstenzija logističke regresije koja modeluje odnos između ulaznih promjenljivih i binarne izlazne promjenljive procjenom vjerovatnoće koristeći temeljnu logističku funkciju. Ulazne promjenljive su snaga i izdržljivost i imaju vrijednosti između 0 i 100. Izlazna promjenljiva je pozicija i uzima vrijednosti {"CB": srednji bek, "CM": srednji vezni, "ST": napadač} koji predstavljaju 3 ključne pozicije. Odbrambeni igrači su klasifikovani tako da su u prosjeku jači i imaju manje izdržljivosti. Igrači centralnog veznog reda su u prosjeku prepoznati kao slabiji. Karakteristike napadača su dobra kontrola lopte. Prema podacima, prosječna snaga i izdržljivost napadača su ustvari između prosjeka za odbrambene i centralne vezne igrače.

- [3] Predicting player positions, David Schoch
  (http://blog.schochastics.net/post/predicting-player-positions)

  Ovaj rad se bavi istraživačkom analizom uz pomoć koje se predviđa položaj odnosno pozicija igrača koristeći različite algoritme mašinskog učenja. Podaci koju su korišćeni sadrže oko 18.000 igrača sa 75 karakteristika po igraču.

  Za potrebe ovog projekta kao obučavajući skup uzeto je 80% podataka, a za testni preostalih 20% podataka. Prilikom izrade modela korišteni su sledeći algoritmi:
    - K-Nearest Neighbors
    - Random Forest
    - Support Vector Machine
    - Neural Network

  Tačnost pomenutih algoritama iznosila je: KNN-0.8197; RF-0.8306; SVM-0.8289; NNET 0.8303. Bitno je napomenuti da je preciznost prilikom izbora golmana iznosila 100%. Ovo nije iznenađenje jer se njihove tehničke osobine znatno razlikuju od ostalih igrača.

- [4] Clustering Based Multi-Label Classification for Image Annotation and Retrieval (https://www.researchgate.net/publication/224087097_Clustering_Based_Multi-Label_Classification_for_Image_Annotation_and_Retrieval)

  Ovaj rad se bavi multi-label klasifikacijom za domene sa velikim brojem ciljnih labela. Metod se sastoji iz dvije faze: prva faza – podrazumjeva inicijalno klasterovanje podataka u cilju razbijanja inicijalnog data set-a na disjunktne cjeline, druga faza – podrazumjeva izgradnju multi-label klasifikacionih modela za svaki klaster. U ovom radu korišteni su: Binary Relevance (BR), Label Powerset(LP), Random k-labelset(RAkEL), Multi label k nearest neighbors(MLkNN), koji u stvari predstavlja prilagođeno proširenje kNN algoritma.

5. Skup podataka

Skup podataka nad kojim će se vršiti analize preuzet je sa interneta, tačnije sa web sajta (https://public.tableau.com/s/sites/default/files/media/fifa18_clean.csv) na osnovu analize iz 2018-te godine. Za potrebu ovoga projekta oduzete su kolone koje su smatrane suvišnim za rješenje ovog problema (kao što su: plata, tržišna vrijednost, nacionalnost, trenutni klub, id igrača. Predviđanje se vrši na osnovu dvadesetak parametara kao što su: brzina, snaga, pregled igre, pas igra, šut, skok, pas u prostor, uklizavanje itd. S obzirom na ogroman broj igrača za koje su dostupni podaci, obučavanje ćemo vršiti nad nekoliko hiljada igrača, a ostale podatke koristimo za testiranje.

6. Metodologija

Na osnovu relevantnih radova sa sličnom problematikom za potrebe ovog projekta koristićemo K-Nearest Neighbors Classifier, Random Forest Classifier i Gaussian Naive Bayes kao iz projekta [1]. Na pomenutom radu su ulazni podaci klasifikovani u 3 grupe, što u našem radu neće biti slučaj. Na osnovu rada [2] smatramo da je moguće iskoristiti multinomalnu logističku regresiju(MLR), linearnu diskriminatornu analizu (LDA) i kvadratnu diskriminatornu analizu (QDA) . Izlazna promjenljiva je imala 3 vrijednosti, odn. moguće pozicije, što ćemo mi u našem slučaju nadograditi, kao što smo naveli u specifikaciji. Takođe, smatramo da je za nas slučaj povoljno iskoristiti Suport Vector Mashine (SVM) kao u radu [3]. Ulazne podatke cemo klasterizovati i nad njima primjeniti klasifikacijske algoritme kao u radu [4]. S obzirom da pojedini igrači mogu igrati na više od jedne pozicije (koje su međusobno slične), pojavljuje se multi-label classification problem, koji cemo transformisati u single-label problem uz pomoć Classifier Chains algoritma. Pošto je povezanost između pozicija velika, Classifier Chains uzima u obzir odnose između ciljanih labela što će doprinijeti boljem ishodu. Koristićemo metode

eksplorativne analize da bismo dobili bolji uvid u skup podataka, ispitali odnose između atributa i pronašli veze između ciljnog i ostalih atributa.

7. Softver

U fazi izrade projekta planirano je korišćenje alata RapidMiner. Ukoliko pomenuti alat ne bude nudio dovoljno fleksibilan skup procesa za realizaciju, projekat će biti realizovan u programskom jeziku Python, korišćenjem biblioteka scipy i sckikit-learn.

8. Metod evaluacije

Tačnost korišćenih podataka algoritma u ovom radu procjenjivaće se upoređivanjem realnih rezulata i rezultata dobijenih primjenom modela nad testnim skupom podataka. Kao testni skup podataka uzećemo 20% slučajno selektovanih igrača, dok će se kao obučavajući skup koristiti preostalih 80%.

Za evaluaciju tačnosti koristićemo *Hamming loss* koji se bazira na *loss* funkciji, što znači da je njegova optimalna vrijednost 0. Takođe, kao dodatnu metodu koristićemo *Jaccard index* koji predstavlja odnos unije i presjeka.

9. Plan razvoja projekta

Plan projekta sadrži sledeće korake:

- Priprema podataka,
- Primjena algoritama za istraživanje i analizu nad podacima – kreiranje modela,
- Procjena kvaliteta dobijenih modela,
- Testiranje sa novim podacima i evaluacija dobijenih rezultata.

10. Tim

Tim čine : Dragan Škiljević (E2 93/2019), Nikola Slijepčević (E2 110/2019) i Dejan Doder (RA 220 /2015).