Fakultet Tehničkih Nauka

Trg Dositeja Obradovića 6, Novi Sad 21000

Projekat iz predmeta: Računarska Inteligencija

Predikcija Pobednika na UEFA Euro 2024

Student: Dušica Trbović SV 42/2021

Novi Sad, jul 2024. god

Sadržaj

1. Uvod

1.1. Pregled projekta

Cilj ovog projekta je predviđanje ishoda međunarodnih fudbalskih utakmica koristeći istorijske podatke o utakmicama i tehnike mašinskog učenja. Projektom se koristi Elo rangiranje timova i različiti algoritmi mašinskog učenja kako bi se izgradio robustan prediktivni model.

1.2. Ciljevi

1. **Prikupljanje i čišćenje podataka**: Prikupljanje podataka o rezultatima međunarodnih fudbalskih utakmica od prve zvanične utakmice 1872. godine do 2024. godine. Čišćenje podataka kako bi se osigurala tačnost i konzistentnost.
2. **Razumevanje Elo rangiranja**: Primena Elo rangiranja za ocenjivanje relativnih veština timova na osnovu njihovih rezultata.
3. **Inženjering karakteristika**: Kreiranje novih karakteristika (features) koje mogu biti korisne za predikciju ishoda utakmica.
4. **Građenje modela**: Odabir i obuka modela mašinskog učenja, kao što je Random Forest Classifier, za predviđanje ishoda utakmica.
5. **Evaluacija modela**: Korišćenje različitih metrika za ocenjivanje performansi modela, kao što su tačnost (accuracy), preciznost (precision), ROC kriva i ROC AUC skor.
6. **Tumačenje rezultata**: Analiza važnosti karakteristika i tumačenje odluka modela kako bi se razumelo koje karakteristike najviše doprinose tačnosti predikcija.

1.3. Značaj mašinskog učenja

Mašinsko učenje omogućava analizu velikih količina podataka i otkrivanje obrazaca koji nisu očigledni tradicionalnim metodama analize. U ovom projektu, mašinsko učenje se koristi za:

- **Predikciju ishoda utakmica**: Na osnovu istorijskih podataka i trenutnih rangiranja, mašinski modeli mogu da predvide verovatnoću pobede, poraza ili nerešenog ishoda.
- **Unapređenje strategija timova**: Treneri i analitičari mogu koristiti predikcije za bolje planiranje strategija i donošenje odluka pre i tokom utakmica.
- **Evaluaciju performansi timova**: Analiza modela može pomoći u identifikaciji ključnih faktora koji utiču na performanse timova, što može biti korisno za razvoj i unapređenje timova.

U daljem tekstu, detaljno ćemo objasniti sve korake preduzete u ovom projektu, od prikupljanja i čišćenja podataka do građenja i evaluacije modela mašinskog učenja.

2. Pregled Podataka

2.1. Izvor Podataka

Podaci korišćeni u ovom projektu sadrže ishode međunarodnih fudbalskih utakmica od prve zvanične utakmice 1872. godine do 2024. godine. Ovi podaci uključuju sledeće informacije:

- **Datum utakmice**: Datum kada je utakmica odigrana.
- **Domaći tim**: Naziv tima koji je bio domaćin utakmice.
- **Gostujući tim**: Naziv tima koji je bio gost na utakmici.
- **Rezultat domaćeg tima**: Broj golova koje je postigao domaći tim.
- **Rezultat gostujućeg tima**: Broj golova koje je postigao gostujući tim.
- **Turnir**: Naziv takmičenja u okviru kojeg je utakmica odigrana (npr. prijateljska utakmica, Svetsko prvenstvo, Evropsko prvenstvo).
- **Grad**: Grad u kojem je utakmica odigrana.
- **Država**: Država u kojoj je utakmica odigrana.
- **Neutralni teren**: Da li je utakmica odigrana na neutralnom terenu.

2.2. Učitavanje Podataka

Prvi korak u analizi je učitavanje podataka iz CSV datoteke i pregled sadržaja. Koristimo biblioteku Pandas za manipulaciju podacima i njihovo učitavanje u DataFrame.

2.3. Čišćenje Podataka

Nakon učitavanja podataka, neophodno je očistiti ih kako bi bili spremni za analizu. Proces čišćenja podataka obuhvata sledeće korake:

1. **Uklanjanje redova sa nedostajućim vrednostima**: Redovi koji sadrže nedostajuće vrednosti mogu negativno uticati na analizu i performanse modela. U ovom slučaju, uklanjamo redove koji imaju nedostajuće rezultate utakmica (nedostajući golovi).

Kao poslednji korak u ovoj sekciji, pregledamo osnovne statistike i informacije o očišćenim podacima kako bismo se uverili da su spremni za upotrebu.

Ovime smo završili proces čitanja i čišćenja podataka, što je ključni korak za uspešnu analizu i izgradnju modela mašinskog učenja.

2. Razumevanje Elo Rangiranja

3.1. Elo Rang Sistem

Elo rang sistem je metod za ocenjivanje relativnih veština timova na osnovu njihovih rezultata u utakmicama. Ovaj sistem se često koristi u sportu, a razvijen je prvobitno za ocenjivanje šahista. Glavne karakteristike Elo rang sistema su:

1. **Dinamičnost**: Rangiranje timova se menja nakon svake utakmice, uzimajući u obzir rezultate utakmica i jačinu protivnika.
2. **Jednostavnost**: Promene u rangiranju se izračunavaju pomoću relativno jednostavne formule koja uzima u obzir očekivani rezultat i stvarni rezultat utakmice.
3. **Komparativnost**: Sistem omogućava poređenje veština timova na osnovu njihove trenutne forme i istorijskih rezultata.

U ovom projektu, Elo rangiranje je korišćeno za procenu trenutne forme timova i njihovih šansi za pobedu u budućim utakmicama.

3.2. Postavljanje Parametara

Za konstrukciju našeg skupa podataka i primenu Elo rang sistema, potrebno je postaviti nekoliko ključnih parametara:

1. **Početni Elo rejting**: Svi timovi započinju sa početnim Elo rejtingom, koji je postavljen na 1500.
2. **K faktor**: K faktor određuje koliko će se rejting tima promeniti nakon svake utakmice. Veći K faktor znači veće promene u rejtingu, dok manji K faktor znači stabilniji rejting. Tipično, K faktor se postavlja između 20 i 40.
3. **Težinski faktor turnira**: Različiti turniri imaju različite težine koje utiču na promenu Elo rejtinga. Na primer, pobeda na Svetskom prvenstvu ima veći uticaj na Elo rejting nego pobeda u prijateljskoj utakmici. Težinski faktori se postavljaju na osnovu važnosti turnira.

3.3. Primena u Projektu

U projektu, primenjen je Elo rang sistem kako bismo ocenjivali timove na osnovu njihovih istorijskih rezultata i kreirali dinamičan skup podataka koji odražava trenutnu formu timova. Na ovaj način, moguće je bolje da modeliramo ishode budućih utakmica, uzimajući u obzir kako trenutnu formu, tako i istorijske performanse timova.

Elo rangiranje nam je omogućilo da:

- **Identifikujemo favorite**: Timovi sa višim Elo rejtingom su favoriti u utakmicama.
- **Procena performansi**: Praćenje promene rejtinga tokom vremena omogućava procenu kako se timovi poboljšavaju ili pogoršavaju.
- **Predikcija rezultata**: Koristili smo Elo rangiranje kao jednu od ključnih karakteristika u našem modelu mašinskog učenja za predikciju ishoda utakmica.

4. Preprocesiranje Podataka

Preprocesiranje podataka je ključni korak u pripremi podataka za modeliranje mašinskog učenja. Ovaj proces obuhvata inženjering karakteristika i normalizaciju/skaliranje podataka kako bi se poboljšale performanse modela.

4.1. Inženjering Karakteristika

Inženjering karakteristika uključuje kreiranje novih karakteristika iz postojećih podataka koje mogu poboljšati performanse modela. U ovom projektu, kreirali smo nekoliko ključnih karakteristika na osnovu Elo rangiranja i istorijskih podataka o utakmicama.

1. **Elo Rejting Timova**: Za svaku utakmicu, izračunali smo Elo rejting za domaći i gostujući tim pre utakmice. Ove vrednosti koriste se kao ulazne karakteristike za model.
2. **Razlika u Elo Rejtingu**: Kreirali smo novu karakteristiku koja predstavlja razliku u Elo rejtingu između domaćeg i gostujućeg tima. Ova karakteristika pomaže modelu da bolje razume relativnu snagu timova.
3. **Istorijski Performans Timova**: Dodali smo karakteristike koje odražavaju istorijski performans timova, kao što su broj pobeda, poraza i nerešenih utakmica u poslednjih nekoliko godina. Ove informacije pomažu modelu da uzme u obzir trenutnu formu timova.
4. **Težina Turnira**: Za svaku utakmicu, dodali smo karakteristiku koja predstavlja težinu turnira. Na primer, pobeda na Svetskom prvenstvu ima veću težinu nego pobeda u prijateljskoj utakmici.
5. **Lokacija Utakmice**: Dodali smo binarnu karakteristiku koja pokazuje da li je utakmica odigrana na neutralnom terenu ili ne. Utakmice na domaćem terenu obično daju prednost domaćem timu.
6. **Datum Utakmice**: Datum utakmice smo transformisali u različite vremenske karakteristike kao što su godina, mesec i dan u nedelji. Ovo pomaže modelu da uzme u obzir sezonske efekte.

4.2. Normalizacija/Skaliranje

Normalizacija i skaliranje podataka su važni koraci u preprocesiranju, posebno kada se koriste algoritmi koji su osetljivi na skalu karakteristika. U ovom projektu, koristili smo nekoliko tehnika normalizacije i skaliranja.

1. **Normalizacija numeričkih atributa:** Numerički atributi `home_score` i `away_score` su normalizovani kako bi se osigurala ujednačenost skale vrednosti. Koristila se standardizacija, pri čemu su vrednosti prebačene na z-skalu (oduzimanjem proseka i deljenjem sa standardnom devijacijom).
2. **Kodiranje kategoričkih atributa:** Kategorički atributi (`home_team`, `away_team`, `tournament`, `city`, `country`) su konvertovani u numeričke vrednosti koristeći One-Hot Encoding. Na ovaj način je svaka kategorička vrednost predstavljena kao binarna (dummy) varijabla, što omogućava lakšu upotrebu u mašinskim modelima.
3. **Skaliranje podataka:** Svi atributi su skalirani kako bi se osiguralo da su u istom opsegu, čime se poboljšava performansa modela mašinskog učenja.

4.3. Procesiranje Podataka

1. **Izbor Karakteristika**: Odabrali smo ključne karakteristike koje će se koristiti kao ulazi za model mašinskog učenja. Ovo uključuje karakteristike kreirane u procesu inženjeringa karakteristika i postojeće karakteristike iz skupa podataka.
2. **Podela Podataka**: Podelili smo podatke na trening i test skupove kako bismo mogli da obučimo model i procenimo njegove performanse. Koristili smo funkciju `train_test_split` iz biblioteke Scikit-learn.
3. **Balansiranje Podataka**: U slučaju da imamo nesrazmerne klase (npr. više pobeda nego poraza), primenili smo tehnike balansiranja kao što su oversampling ili undersampling kako bismo obezbedili bolje performanse modela.

Preprocesiranje podataka je ključni korak u osiguravanju da model mašinskog učenja može efikasno učiti iz podataka i praviti tačne predikcije. Pravilno inženjering karakteristika i skaliranje podataka značajno poboljšavaju performanse modela i njegovo generalizovanje na nove podatke.

5. Model – Random Forest

Za rešavanje problema predikcije pobednika UEFA Euro 2024, izabran je algoritam **Random Forest**. Ovaj algoritam je odabran zbog svoje robusnosti i sposobnosti da rukuje velikim brojem ulaznih atributa, kao i zbog svoje otpornosti na prekomerno prilagođavanje (overfitting). Random Forest algoritam kombinuje rezultate više stabala odluka kako bi poboljšao tačnost i stabilnost predikcija.

5.1. Podela podataka i treniranje modela

Podaci su podeljeni na trening set i test set tako što je glavni parametar bio datum (pre 13. juna 2024. i nakon). Trening set je korišćen za obuku modela, dok je test set korišćen za evaluaciju performansi modela.

Model je treniran na trening setu koristeći sledeće parametre:

- Broj stabala (estimators): 100
- Maksimalna dubina stabala: Nije ograničena

Nakon treniranja modela, postignuta je tačnost od 0.75 na testnom setu, što ukazuje da je model uspešno predvideo 75% ishoda mečeva.

- Od ukupno 23 stvarnih negativnih ishoda, model je tačno predvideo 20, dok je 6 puta pogrešno predvideo kao pozitivne ishode.
- Od ukupno 10 stvarnih pozitivnih ishoda, model je tačno predvideo 7, dok je 3 puta pogrešno predvideo kao negativne ishode.

	actual	predicted	date	home_team	opponent
4954	1	1	2024-06-14	Germany	108
6307	1	1	2024-06-15	Italy	0
5605	0	0	2024-06-15	Hungary	118
10681	1	1	2024-06-15	Spain	28
10229	0	0	2024-06-16	Slovenia	35
8291	0	0	2024-06-16	Poland	85
9923	0	0	2024-06-16	Serbia	39
9184	1	0	2024-06-17	Romania	126
1395	0	1	2024-06-17	Belgium	112
848	0	0	2024-06-17	Austria	44
11704	1	1	2024-06-18	Turkey	47

Slika 1 - Očekivani i Stvarni rezultati

Na osnovu rezultata evaluacije, možemo zaključiti da Random Forest model ima solidne performanse sa tačnošću od 75% i preciznošću od 53.8%. Međutim, postoji prostor za poboljšanje, posebno u povećanju preciznosti i smanjenju lažno pozitivnih predikcija. Dodatne metode za poboljšanje performansi mogu uključivati optimizaciju hiperparametara, korišćenje naprednih tehnika za balansiranje dataset-a ili kombinovanje više modela.

6. Rezultati

6.1. Proces predikcije

Nakon što je model Random Forest treniran i evaluiran, pristupilo se procesu predikcije pobednika UEFA Euro 2024. Korišćeni su istorijski podaci o utakmicama kako bi se izvršile predikcije za buduće mečeve. Model je bio zadužen za analizu i predikciju ishoda svakog meča na osnovu prethodno definisanih atributa kao što su timovi koji igraju, rezultati prethodnih utakmica, turnir i lokacija.

6.2. Predikcija po koracima

1. **Unos podataka za početne utakmice po grupama:**
   - Na početku su uneti podaci za sve timove i njihove početne utakmice po grupama. Ovi podaci su uključivali informacije o svakom timu i njihovim prethodnim performansama.

	Group	Home_Team	Away_Team	Home_att	Home_def	Away_att	Away_def	XGhome	XGaway	draw_prob	away_prob	home_prob
0	A	Germany	Scotland	2.202630	0.849587	1.122092	1.008867	2.222160	0.953315	0.189761	0.148860	0.653462
1	A	Hungary	Switzerland	1.164442	1.208869	1.525448	0.890410	1.036831	1.844067	0.226869	0.559763	0.210333
2	A	Germany	Hungary	2.202630	0.849587	1.164442	1.208869	2.662692	0.989295	0.154851	0.116052	0.709769
3	A	Scotland	Switzerland	1.122092	1.008867	1.525448	0.890410	0.999122	1.538974	0.256014	0.497139	0.245692
4	A	Switzerland	Germany	1.525448	0.890410	2.202630	0.849587	1.296001	1.961243	0.219165	0.525030	0.251325
5	A	Scotland	Hungary	1.122092	1.008867	1.164442	1.208869	1.356463	1.174767	0.266745	0.323400	0.409117
6	B	Spain	Croatia	2.164003	0.542828	1.610074	0.854836	1.849868	0.873993	0.223132	0.172996	0.600854
7	B	Italy	Albania	1.499667	0.593307	0.857377	1.201844	1.802366	0.508687	0.213154	0.097694	0.686562
8	B	Croatia	Albania	1.610074	0.854836	0.857377	1.201844	1.935058	0.732917	0.208717	0.133378	0.654092
9	B	Spain	Italy	2.164003	0.542828	1.499667	0.593307	1.283917	0.814061	0.288621	0.236295	0.474686
10	B	Albania	Spain	0.857377	1.201844	2.164003	0.542828	0.465408	2.600794	0.122547	0.813919	0.046338
11	B	Croatia	Italy	1.610074	0.854836	1.499667	0.593307	0.955267	1.281969	0.283394	0.439180	0.276992
12	C	Slovenia	Denmark	1.047052	1.101898	1.416472	0.780355	0.817072	1.560808	0.254529	0.548268	0.196018
13	C	Serbia	England	1.488930	1.107900	1.863344	0.600768	0.894502	2.064399	0.202096	0.639012	0.153490

Slika 2 - Spemnosti timova u trenutku igranja utakmica u grupnoj fazi

2. **Statistička analiza grupne faze:**
   - Uz pomoć statističke analize, dobili smo predikcije kako bi grupna faza trebala da izgleda. Model je predvideo koji tim će završiti na kom mestu u grupi, kao i sa koliko bodova.

```
Group A: [('Germany', 9), ('Switzerland', 4), ('Scotland', 2), ('Hungary', 1)]
Group B: [('Spain', 9), ('Italy', 4), ('Albania', 3), ('Croatia', 1)]
Group C: [('England', 7), ('Denmark', 7), ('Serbia', 1), ('Slovenia', 1)]
Group D: [('Netherlands', 9), ('France', 4), ('Poland', 4), ('Austria', 0)]
Group E: [('Ukraine', 7), ('Slovakia', 6), ('Belgium', 3), ('Romania', 1)]
Group F: [('Portugal', 7), ('Czech Republic', 4), ('Georgia', 2), ('Turkey', 2)]
```

Slika 3 - Rezultati grupne faze takmičenja

3. **Određivanje timova za narednu fazu:**
   - U sledeću fazu prošli su svi timovi koji su bili na prvom i drugom mestu u svojim grupama, kao i top 4 trećeplasirana tima. Ukupno 16 timova je nastavilo takmičenje.
4. **Predikcije za fazu osmine finala:**

- Za svaku utakmicu u fazi 16-tine finala, model je ponovo korišćen za predikciju ishoda. Analizirani su mečevi i na osnovu predikcija, određeno je koji timovi prolaze u sledeću fazu.

```
Round of 16 Results:
--------------------
Germany vs Denmark -> Winner: Germany
Spain vs Scotland -> Winner: Scotland
Portugal vs Serbia -> Winner: Serbia
France vs Slovakia -> Winner: Slovakia
Switzerland vs Italy -> Winner: Italy
England vs Georgia -> Winner: England
Ukraine vs Albania -> Winner: Ukraine
Netherlands vs Czech Republic -> Winner: Netherlands
```

Slika 4 - Utakmice u osmii finala i rezultati

5. **Predikcije za četvrtfinale finala:**
   - Nakon što su određeni timovi koji prolaze u osminu finala, ponovljene su predikcije za svaki meč u ovoj fazi. Timovi koji su prošli dalje su oni koji su imali najbolje šanse za pobedu prema modelu.

```
Quarterfinal Results:
--------------------
Germany vs Scotland -> Winner: Germany
Serbia vs Slovakia -> Winner: Slovakia
Italy vs England -> Winner: England
Ukraine vs Netherlands -> Winner: Ukraine
```

Slika 5 - Utakmice u četvrtini finala i rezultati

6. **Predikcije za polufinale:**
   - Proces predikcije je nastavljen i za faze polufinala. Svaki meč je detaljno analiziran, a predikcije su korišćene za određivanje timova koji će se takmičiti u finalu.
   -

```
Semifinal Results:
--------------------
Germany vs Slovakia -> Winner: Germany
England vs Ukraine -> Winner: England
```

Slika 6 - Utakmice u polufinalu i rezultati

7. **Finale:**
   - U finalnom meču, prema predikcijama modela, sastali su se Engleska i Nemačka. Analiza je pokazala da Nemačka ima najveće šanse za pobedu.

6.3. Konačni rezultati

Na osnovu predikcija modela, konačni pobednik UEFA Euro 2024 je **Nemačka**. Model je pokazao da Nemačka ima najviše šansi za pobedu, uzimajući u obzir sve relevantne faktore i

istorijske podatke. Ovaj rezultat se zasniva na detaljnoj analizi i predikcijama svakog meča, što omogućava pouzdanost u konačnu prognozu.

Tokom ovog procesa, model je pružio uvid u potencijalne performanse svakog tima i omogućio detaljnu analizu njihovih snaga i slabosti. Ovi rezultati mogu biti korišćeni za dalju strategiju i planiranje od strane trenera i analitičara timova učesnika.

7. Zaključak

Ovaj projekat je demonstrirao kako se mašinsko učenje može primeniti na sportske analize za predikciju ishoda fudbalskih turnira. Korišćenjem istorijskih podataka o utakmicama i naprednih algoritama kao što je Random Forest, razvijen je model koji predviđa ishode mečeva na UEFA Euro 2024.

7.1. Ključni koraci i postignuća

1. **Pregled i analiza podataka:** Detaljno su analizirani podaci kako bi se razumela njihova struktura i identifikovali relevantni atributi za predikciju pobednika.
2. **Pretprocesiranje podataka:** Obradjeni su nedostajući podaci, normalizovani numerički atributi i kodirani kategorički atributi kako bi podaci bili spremni za treniranje modela.
3. **Izbor i treniranje modela:** Izabran je Random Forest algoritam zbog svoje efikasnosti i robusnosti. Model je treniran i postigao zadovoljavajuće rezultate na testnom setu.
4. **Evaluacija modela:** Model je evaluiran koristeći različite metrike performansi, pri čemu je postignuta tačnost od 75% i preciznost od 53.8%.
5. **Predikcija ishoda turnira:** Korišćenjem treniranog modela, izvršene su predikcije za sve faze turnira, uključujući grupnu fazu, 16-tinu finala, osminu finala, četvrtfinale, polufinale i finale.

7.2. Konačni rezultat

Model je predvideo da će Nemačka biti pobednik UEFA Euro 2024, pobedivši Englesku u finalu. Ova predikcija je zasnovana na detaljnoj analizi i evaluaciji performansi svih timova tokom turnira.

7.3. Zaključci

- **Robusnost modela:** Random Forest algoritam se pokazao kao efikasan za predikciju ishoda sportskih događaja.
- **Značaj kvalitetnih podataka:** Kvalitet i obim ulaznih podataka značajno utiču na performanse modela.
- **Primena analize:** Rezultati ove analize mogu biti korisni za trenerske timove i analitičare kako bi optimizovali strategije i pripremu timova.

Ovaj projekat pruža osnovu za dalja istraživanja i poboljšanja u oblasti sportske analitike i predikcije. Napredne tehnike mašinskog učenja i veštačke inteligencije mogu se dalje razvijati i primenjivati za sveobuhvatniju analizu i tačnije predikcije u budućnosti.