MINISTERUL EDUCAȚIEI

UNIVERSITATEA TEHNICĂ
DIN CLUJ-NAPOCA

FACULTATEA DE AUTOMATICĂ ŞI CALCULATOARE
DEPARTAMENTUL CALCULATOARE

Predicția rezultatelor meciurilor de fotbal internaționale folosind metode de învățare automată.

Student: **Viman Mario Bogdan**

2024

Cuprins

Capitolul 1. Introducere

Fotbalul, ca o activitate sportivă de masă răspandită la nivel mondial, este atât o comunitate complexă și interconectată, cât și un mijloc de divertisment. În anul 2004 forul FIFA (Fédération Internationale de Football Association) a recunoscut China ca locul de naștere a fotbalului. În jurul anului 200 î.e.n. chinezii jucau un sport asemănător numit cuju.

Legile jocului au fost formate in Anglia de The Football Association în 1863, primind denumirea de Association Football, pentru a nu se confunda cu alte forme de fotbal existente în acele timpuri, însă primul meci internațional a avut loc doar în anul 1872 la data de 30 Noiembrie, între Scoția si Anglia terminându-se cu scorul de 0 – 0.

Recent, aplicații tot mai complexe ale inteligenței artificiale au fost dezvoltate în multe industrii, inclusiv în sport, acesta devenind un domeniu tot mai receptiv la integrarea tehnologiilor avansate. Printre acestea, predicția rezultatelor meciurilor de fotbal cu ajutorul tehnicilor de învățare automată este un topic la mare căutare.

Am ales această temă datorită pasiunii purtate pentru sportul rege, dar și unor experimente pe care mi le derulam în minte în cadrul anumitor turnee finale (Ex: “Cine ar fi câștigat dacă s-ar fi întâlnit respectivele două?”), astfel că motivația personală privind acest sport va juca un rol important în realizarea proiectului, pasiunea pentru fotbal împingându-mă să dedic timp cercetării și analizei datelor, și să îmi ajustez modelul pentru rezultate satisfăcătoare.

De altfel, pe lângă motivația personală, vor fi și alte elemente importante care vor fi luate în considerare în cadrul proiectului, precum analiza datelor statistice, evaluarea modelului (evaluarea performanței prin intermediul testelor și a comparațiilor cu alte modele existente), și interpretarea rezultatelor obținute pentru a identifica care factori au cel mai mare impact asupra predicțiilor.

Capacitatea de a putea prezice rezultatele meciurilor nu doar că ar putea alimenta pasiunea și entuziasmul fanilor, ci poate afecta semnificativ și alte domenii, precum cel al pariurilor sportive, predicțiile precise putând influența strategia de investiție și deciziile de pariere. De asemenea, aceste statistici asupra șanselor de victorie a unei echipe ar putea ajuta cluburile de fotbal în îmbunătățirea performanțelor sportive.

În cadrul realizării acestui proiect, am ales o bază de date a rezultatelor internaționale de fotbal (1870 – 2024), aceasta fiind formată din tabela *results*.

În concluzie, sunt entuziasmat de potențialul acestui proiect și cred că, alături de o abordare corectă, demersul va fi unul de succes.

Capitolul 2. Context

2.1. Descrierea bazei de date

Am ales această bază de date datorită faptului că aceasta conține informații cuprinzătoare despre fiecare meci, inclusiv datele acestuia, echipele implicate, scorul final, turneul în cadrul căruia s-a desfășurat și alte detalii relevante precum orașul și țara gazdă.

De asemenea, datele sunt structurate în ordine cronologică, permițându-ne să evaluăm performanța de-a lungul timpului și să identificăm tendințe sau modele care ar putea influența rezultatele meciurilor. Acest lucru ne ajută să avem o viziune mai clară asupra datelor și să obținem o predicție mai precisă.

În concluzie, alegerea acestei baze de date este justificată de structura clară, ordinea cronologică și relevanța datelor pentru obiectivele noastre de predicție.

- results (date, home_team, away_team, home_score, away_score, tournament, city, country, neutral).

Link-ul bazei de date: https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017

2.2. Cerințe

Cerințele funcționale includ dezvoltarea și implementarea unui algoritm de învățare automată utilizat în scopul predicției rezultatelor meciurilor, integrarea și prelucrarea eficientă a datelor și furnizarea de rezultate interpretabile.

2.3. Obiectivele proiectului

În mod clar, obiectul principal îl reprezintă dezvoltarea și antrenarea modelului de predicție, astfel încât să avem predicții precise și fiabile. Acest lucru implică evaluarea și antrenarea algoritmilor de învățare automată folosind datele disponibile, iar ulterior testarea performanței modelului pe date noi sau de test.

În varianta finală, aș dori ca modelul să citească dintr-un dicționar date de forma: dată, echipa gazdă, echipa oaspete, țară, turneu, neutru; și să poată să facă o predicție cât mai exactă pe baza acestora.

2.4. Analiză

În cadrul bazei noastre de date avem un total 46 289 de meciuri internaționale disputate, împărțite între un total de 336 de echipe.

2.4.1. Distribuția scorurilor meciurilor

În figura de mai jos, este prezentată distribuția scorurilor într-o ordine descrescătoare, oferind o imagine mai clară despre dinamica rezultatelor. Cele mai întâlnite rezultate internaționale au fost, de departe, "1-0" (4767 de rezultate) și "1-1" (4616 de rezultate), urmate de "0-0" (3711 rezultate), "2-0" (3603 rezultate), "2-1" (3534 rezultate) etc.

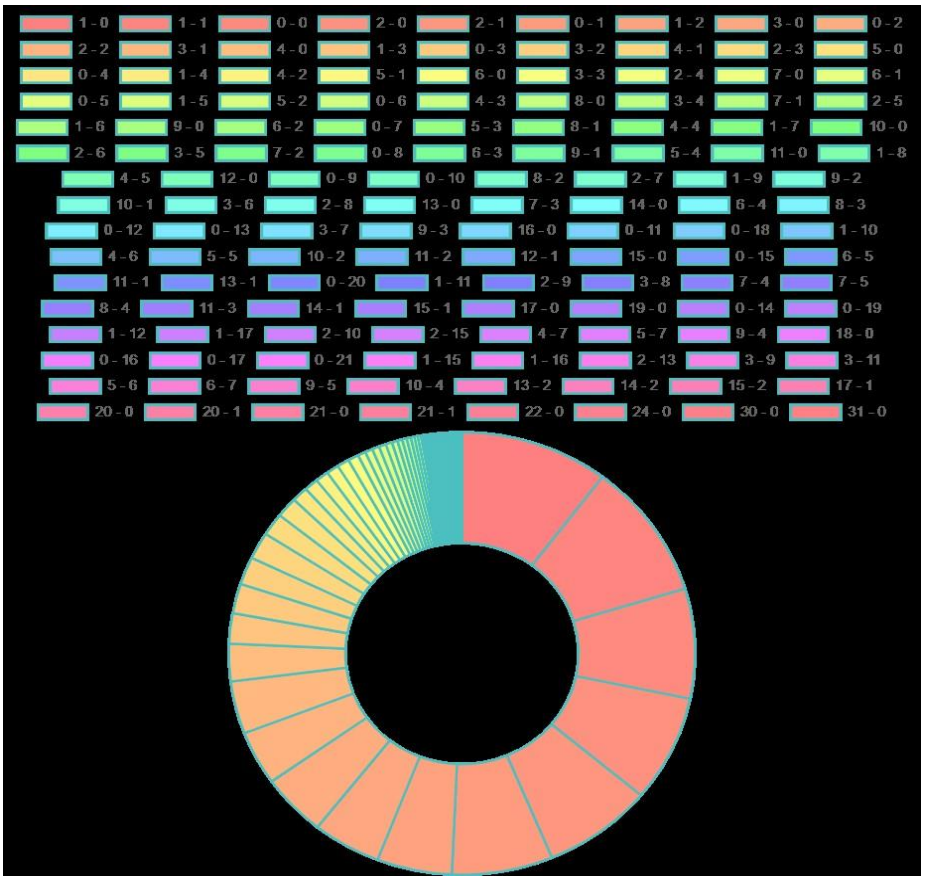


Figure 1

2.4.2. Distribuția pe turnee

Fotbalul cuprinde o varietate de competiții la nivel global, chiar și când vorbim doar de competițiile internaționale. De-a lungul timpului, au existat o diversitate deosebită de turnee internaționale, mai exact 163, însă majoritatea dintre ele și-au încetat existența din diverse motive. Distribuirea pe turnee ne ajută în analizarea performanței unei echipe în anumite condiții, putând evidenția chiar o anumită dominație a unor echipe asupra unor turnee specifice.

După cum putem vedea în figura de mai jos, cele mai multe meciuri din istorie sau disputant sub statutul de “amicale”, nefăcând parte dintr-o competiție anume, deși official, spre deosebire de meciurile amicale din cadrul cluburilor, in contextul echipelor naționale acestea contribuie la coefficient și statistici fiind considerat cel putin la nivel teoretic un meci *official*. După amicale urmatoarele competiți ar fi reprezentate

de: Calificările la Cupa Mondială (8013), Calificările la Campionatul European (2815), Calificările la Cupa Africii (2116) etc.

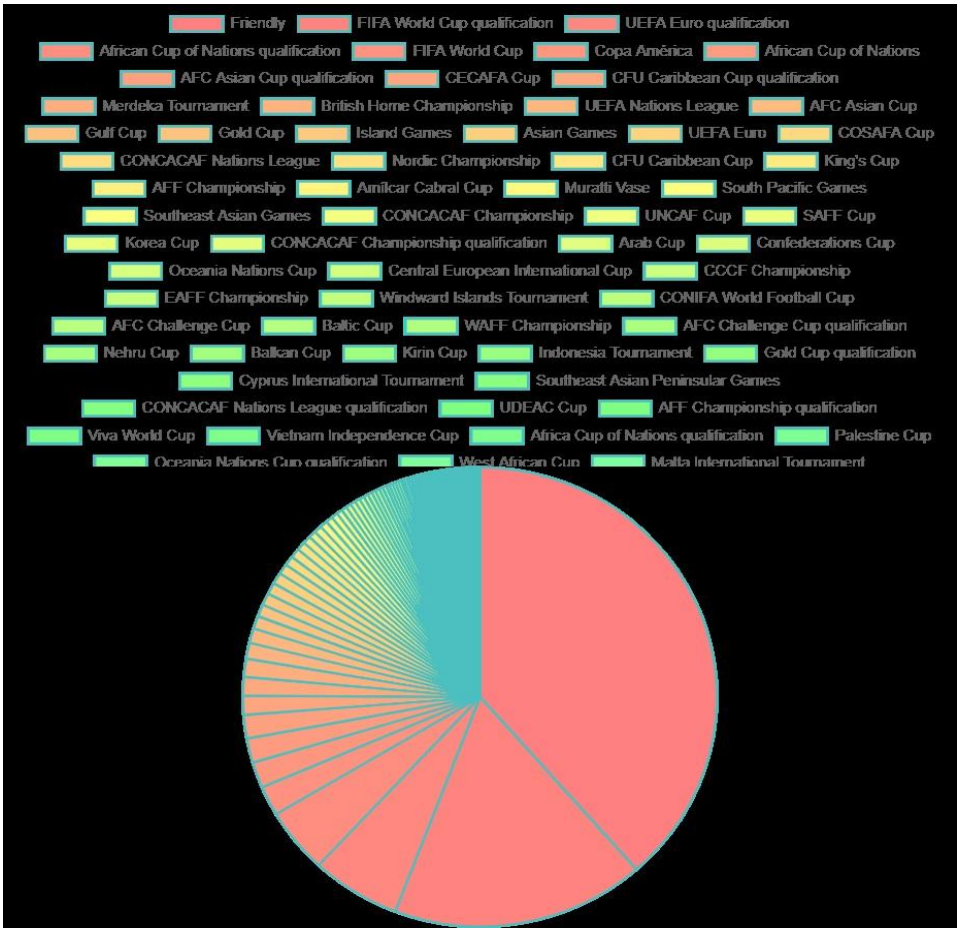


Figure 2

2.5. Modificări asupra datelor

Au fost făcute diferite modificări asupra datelor originale datorită diverselor anomalii observate.

Pentru eliminarea unei astfel de anomalii, am adoptat divizarea coloanei date în alte 3 coloane datorită faptului că aceasta nu era echivalentă pentru toate datele din tabelă. Aceasta era sub formatul an-zi-lună până în anul 1899 inclusiv, având o lungime fixă de 8 caractere, însă problema se complică de la anul 1900, fiind sub forma lună-zi-an, și având o lungime variabilă între 6 și 8 caractere (fiindcă dacă era de exemplu luna Iunie apărea doar 6 în loc de 06). Astfel că, prin divizarea în coloane separate pentru fiecare dată în parte, și convertirea datelor după 1899 folosind un regex cu dimensiune variabilă, problema a fost rezolvată.

O altă modificare necesară a fost descoperită în momentul calculării corelației, fiindcă nu o puteam face datorită faptului că toate coloanele relevante conțineau șiruri de caractere, așa că am apelat la atribuirea unui cod unic fiecărui element nou găsit în momentul iterării, iar când dădea de un element găsit anterior, îi atribuia codul respectiv.

Capitolul 3. Aspecte teoretice

3.1. Corelații

Există mai multe tipuri de corelații, printre cele mai comune fiind corelația Pearson, Spearman și corelația Kendall. Corelația Pearson măsoară relația liniară între două variabile continue, fiind cea utilizată în cadrul proiectului, spre deosebire de corelația Spearman și Kendall, care sunt mai potrivite pentru date ordonate sau categorice.

Valoarea corelației este întotdeauna între -1 și 1. Cu cât valoarea este mai aproape de 1 respectiv -1, cu atât mai puternică este corelația. O corelație de 0 indică lipsa unei corelații liniare între variabile.

În matricea de corelație prezentată mai jos pentru tabela results, se poate observa cum pe diagonala principală avem o corelație de 1, deoarece se întâmplă între aceeași coloană (are o corelație puternică cu ea însăși). De altfel, singurele corelații puternice pe care le mai avem în afară de acestea sunt între home_team_code (echipa care joacă acasă) și country_code (țara în care se dispută meciul), fiind de circa 0.81. Aceasta ar putea fi de 1.00 dacă nu luăm în calcul situațiile când există un turneu final, pentru că atunci practic ambele echipe sunt în deplasare (fiind pe teren neutru). Mai jos se pot observa toate aceste aspecte:

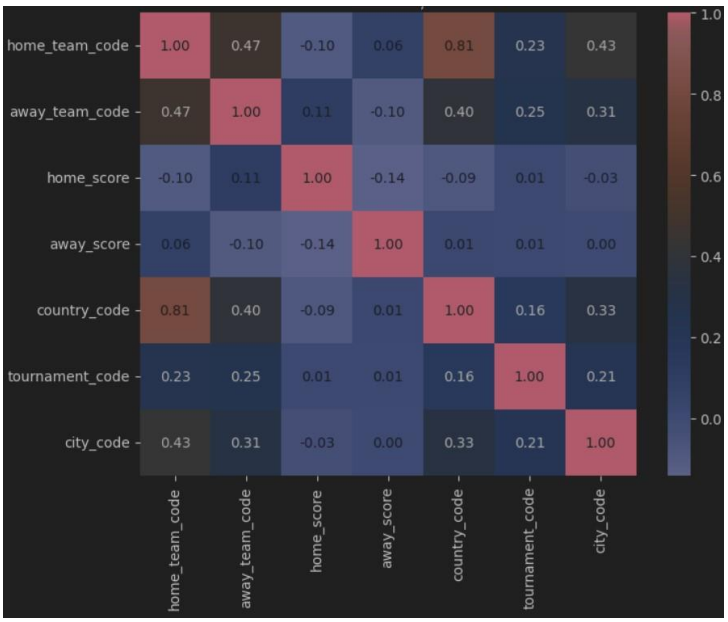


Figure 3

De asemenea, se poate observa în figură cum există o mulțime de corelații într-o foarte slabă legatură („country_code – city_code”, „home_team_code – city_code” etc) sau chiar lipsa unei corelații („away-score – city_code” , „tournament_code – city_code” etc).

Corelația nu implică cauzalitate, o corelație pozitivă sau negativă între două variabile nu înseamnă întotdeauna că una o cauzează pe cealaltă, fiind posibil să existe un alt factor extern care influențează ambele variabile.

De asemenea, [1], Mark A. Hall spune în teza lui că o caracteristică ar fi utilă doar dacă este corelată cu sau predictivă a clasei, practic acesta insumează faptul că orice obiect sau eveniment poate fi utilizat pentru a-l descrie sau clasifica. Ca să fie corelație ar trebui să fie o asociere statistică semnificativă între o caracteristică și o clasă, iar ca sa fie predictivă, ar trebui să aibă abilitatea de a prezice cu exactitate la ce clasă aparține un obiect sau eveniment.

O altă caracteristică a corelațiilor este reprezentată de simetrie, deși în cadrul corelațiilor noastre acest lucru nu este valabil, acestea fiind absolut simetrice. În mod normal, corelația dintre coloana 1 și coloana 2 poate fi diferită de cea dintre coloana 2 și coloana 1.

3.2. Gini Index

Gini index este o măsură a inegalității sau repartizării unei distribuții, lucru amintit si de [2] *Stefano Nembrini et al* în articolul lor. În contextul actual îl folosim pentru a evalua distribuția golurilor marcate de echipele de fotbal într-un meci (cele de acasa vs cele din deplasare).

În cadrul proiectului interpretarea este urmatoarea :

- pentru echipele gazdă (home_team), indicele este de aproximativ 0.576 ceea ce reprezintă o distribuție inegală a golurilor marcate pe teren propriu. Cu cât indicele este mai aproape de 1, cu atât distribuția este mai inegala, însemnând că există o variație mare între golurile marcate acasă.
- pentru echipa oaspete (away_team), indicele este de aproximativ 0.441, acest lucru indicând o distribuție mai egală (sau mai constantă am putea spune) a golurilor marcate de echipa oaspete în comparație cu echipa gazdă, totuși existând în continuare o anumită inegalitate.

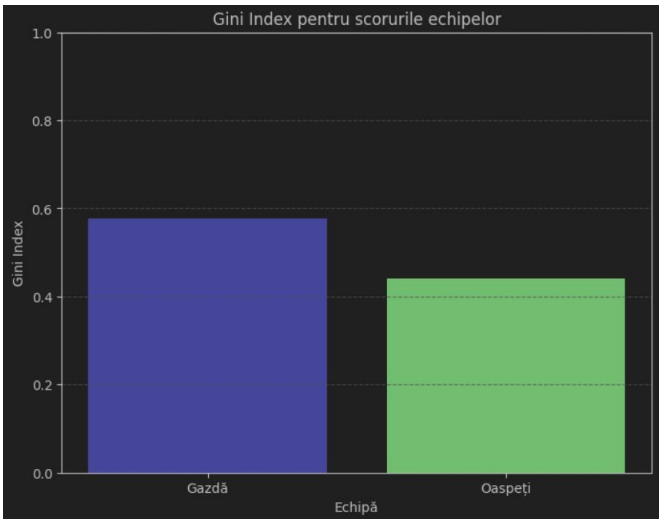


Figure 4

În cadrul calculării gini indexului pentru variabile categorice îl vom folosi pe ‚country', ‚tournament' si ‚city'.

Interpretarea indexului:

- pentru variabila „country", avem un gini index de aproximativ 0.98 care indica o inegalitate extrem de mare in găzduirea meciurilor intre țări. Acest lucru ne spune faptul

că un numar mic de echipe găzduiesc majoritatea meciurilor, lucru posibil din cauza găzduirii a mai multe turnee finale in cadrul aceloraşi tari.
- pentru variabila „tournament” avem un indice de aproximativ 0.63 indică si el o inegalitate, consemnând faptul ca un număr relativ mic de turnee organizează majoritatea meciurilor internaționale, lucru care este normal dat fiind că cele mai multe meciuri sunt ori amicale, in cadrul turneelor finale (European Championship, CONCAF, World Cup, Asia Cup, Gold Cup) sau a calificărilor la aceste turnee.
- pentru variabila „city”, avem o inegalitate de 0.993, fiind aproape perfecte, indicând faptul că un numar extrem de mic de orașe găzduiesc majoritatea meciurilor, deoarece din nou turneele finale joacă un rol extrem de important.

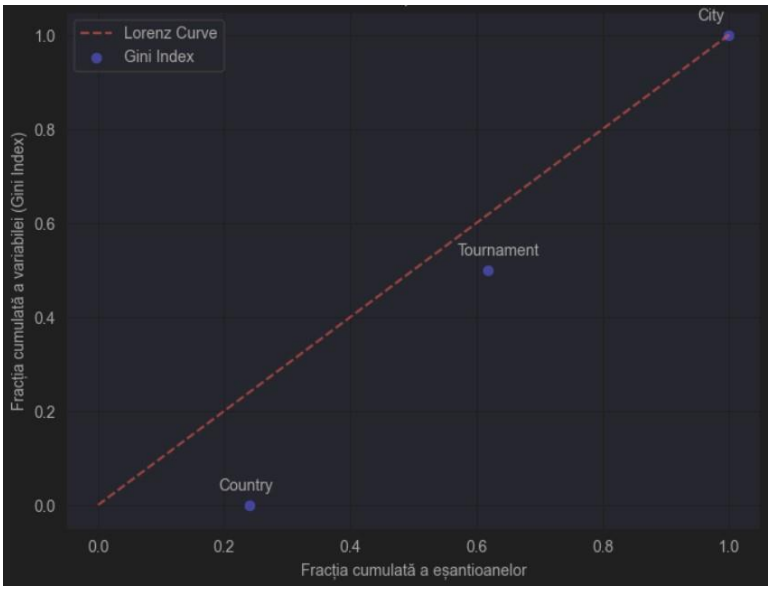


Figure 5

3.3. Entropie

Entropia reprezinta o masura a dezordonarii sau incertitudinii intr-un set de date. Cu cat probabilitatile sunt mai uniform distribuite, cu atat entropia este mai mare, in schimb daca aceasta este cat mai mica atunci setul de date este bine organizat.

Ca si un exemplu am calculat entropia golurilor de acasa, care a rezultat intr-o entropie de 2.516 care ne indica faptul ca distributia scorurilor este destul de variabila. Nefiind o distributie uniforma a golurilor marcate.

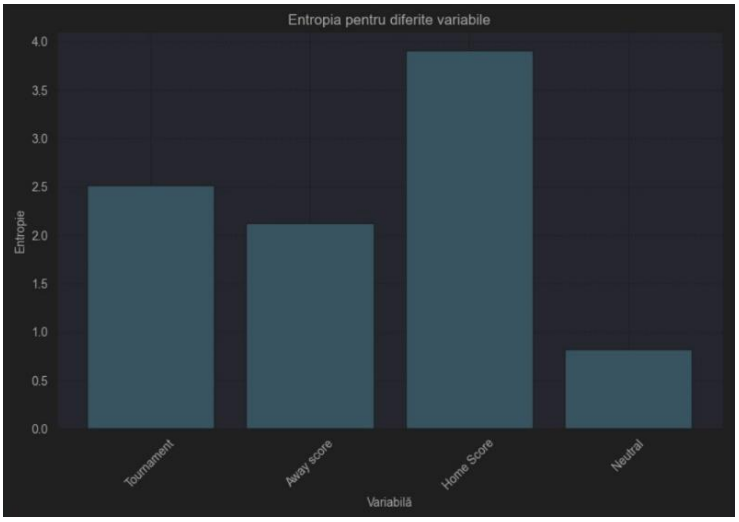


Figure 6

3.4. Starea actuală a domeniului

[3] Thomas Reilly & David Gilbourne, elaborează în lucrarea lor cum că fotbalul in general era văzut ca inadecvat cercetăriilor științifice, și ca o bună parte a instituțiilor care se ocupau de bunăstarea acestuia, vedeau oamenii de știință cu scepticism. Primul Congres Mondial al Științei si Fotbalului , ce a avut loc in anul 1987 a reprezentat un pas uriaș în evoluția dintre teorie si practică, de altfel fiind pentru prima dată când reprezentanții tuturor forurilor fotbalistice s-au reunit pentru un scop comun [4].

Începând cu acel moment, echipele de fotbal profesioniste in mod special, au primit sprijin sistematic din știință, de asemenea fiind creat “Body of knowledge in science and football” [3], pentru a se putea adresa problemelor specifice domeniului, deoarece competivitatea și lupta pentru obținerea unui avantaj semnificativ in acest sport a generat o mulțime de intrebări de cercetare cărora oamenilor de știință să li se poată adresa.

In primele 4 congresse ale fotbalului [5], s-au stabilit diverse caracteristici care ar urma să fie îmbunătățite cu ajutorul științei precum:

- analiza meciurilor,
- medicina și aspectele de mediu;
- testarea fitness-ului jucătorilor;
- psihologia antrenamentului;
- psihologia jucătorilor;
- management-ul echipei și coaching-ul;
- biomecanica;
- știința exercițiului pedriatric;
- metabolism și nutriție;
- psihologia meciurilor;
- sociologia.

Odată cu realizarea unei înțelegeri mai bune a jocurilor din punct de vedere disciplinar, au apărut diverse tehnici adaptate utilizate în investigațiile experimentale, Gleeson et al [6] , pentru a investiga integritatea musculo-scheletică au folosit un

protocol modificat, de altfel o utilizare a abilităților cu mingea în invățarea motrica [7] pentru luarea unor decizii optime la loviturile de pedeapsă [8], precum și multe altele.

Odată cu dezvoltarea științei, au apărut noi domenii de exercitare a acesteia in fotbal, si anume incercarea de a prezice rezultatele între echipe. Predicțiile sportive se bazează pe ideea de a colecta un număr mare de date, de la performanța istorice până la rezultatele meciurilor și date despre jucători, pe baza acestora incercând a se înțelege șansele de a câștiga sau a pierde a unei echipe [9]. Scopul cercetariilor în acest domeniu este de a prezice meciuri de fotbal folosind algoritmi de invățare automate precum ANN (Artificial Neural Network) si DNN (Deep Neural Network).

Pentru a prezice rezultatul unui meci , o persoană va trebui să țină cont de multe lucruri, precum performanța echipei în ultimul timp, dacă meciul va fi jucat acasă sau în deplasare, transferurile sau lotul de jucători, staff-ul curent, etc. Problema atunci când cineva face o predicție asupra rezultatului este că multe lucruri, cum ar fi preferintele personale pentru o anumita echipă, percepția omului asupra anumitor jucători ai echipei și altele, pot influența decizia, studiile arătând că, chiar și culoarea echipamentului poate influența o astfel de decizie.

Piața pariurilor de fotbal a realizat cea mai rapidă creștere dintre toate jocurile de noroc in ultimii ani, astfel ca au fost propuse diverse strategii econometrice pentru predicția rezultatelor de la primele cercetări [10] [11]. Datorită acestor aspecte, a fost introdusă o arhitectura ierarhică a ANN pentru a identifica modelele tactice, prin utilizarea hartilor auto-organizate, fiindca acest tipic de arhitectura poate identifica diverse modele tactice si variații ale acestora.

Metodologia începe de la preprocesarea seturilor de date, care are mai multe straturi. Modelul citește prima dată setul de date pentru a se asigura că este valid, putand-ul returna in caz contrar sau dacă există vreo problemă cu acesta. Focus-ul principal in multe cazuri se face asupra jucatorilor, fiindcă dacă un jucator important lipsește într-un anumit meci jucat la un moment specific, ar putea fi o fluctuație in rezultat [9]. De asemenea se tine cont de ordinea meciurilor si evenimentelor, prin urmare este nevoie de exploatarea unei retele de memorie pe termen lung (LSTM).

In concluzie, cercetarea in continuă dezvoltare privind predicția meciurilor de fotbal folosind algortimi de inteligentă artificială și invățare profundă va continua să revoluționeze industria pariurilor sportive, si chiar de a oferi informații utile fanilor, antrenorilor si managerilor de echipe. Predicțiile ar trebui sa devină din ce in ce mai utile și precise pe masură ce modelel AI se îmbunătățesc.

Ca și direcții viitoare, pe lângă creșterea complexității și eficienței modelelor, putem conștientiza integrarea de noi surse de date, precum cele de pe rețele sociale pentru a contura mai bine profilul psihologic al unui jucator, si de asemenea integrarea datelor de la senzorii purtati de jucatori.

Capitolul 4. Implementare

Pentru atingerea obiectivelor propuse în acest proiect, am ales explorarea a 3 modele de învățare automată diferite: KNN (K-Nearest Neighbors), Random Forest și Neural Network Regression. Am selectat aceste 3 modele datorită diversității și performanței lor în alegerea diferitelor aspecte ale datelor, sperând că această abordare va permite obținerea de rezultate cât mai precise în cadrul prezicerii meciurilor de fotbal, cu scopul ca în final să alegem cel mai bun model din cele 3.

4.1. Tehnologii folosite

- Jupyter Notebook
- Python 3.12
- SQL Lite Database

4.2. Compararea modelelor

4.2.1. Arhitectură

KNN (K Nearest Neighbors) este un model care functioneaza pe baza similaritatii intre exemplele de antrenare. Pentru efectuarea unei predictii acesta va identifica cei mai apropiati K vecini ai acestuia si le va atribui clasa cea mai frecventa intre acestia.

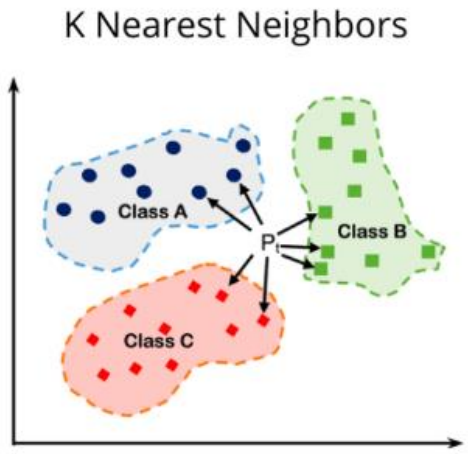


Figure 7

Random Forest reprezintă un grup de arbori de decizie, fiecare dintre ei fiind antrenat la un nivel mai mic, predicția finală fiind făcută prin votare sau o medie a predicțiilor fiecărui arbore.

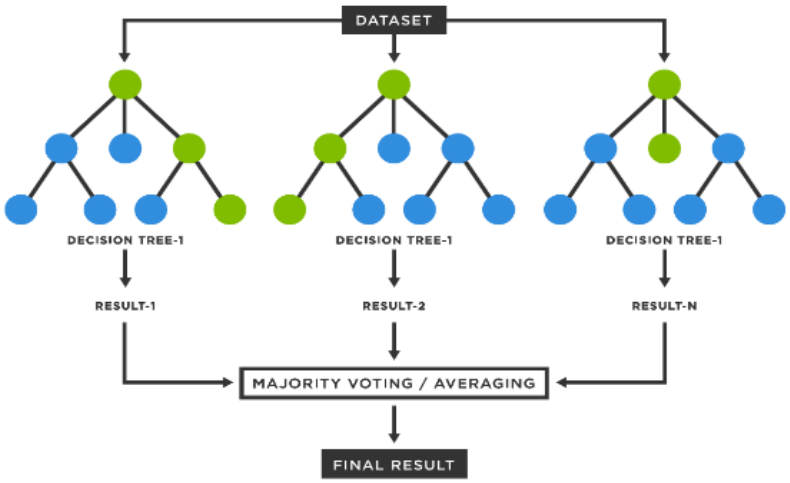


Figure 8

Neural Network Regression, este un model de regresie care se bazează pe rețele neurale artificiale ce utilizează mai multe straturi de neuroni interconecțati pentru a invăța relațiile complexe dintre caracteristicile de intrare si ieșire.

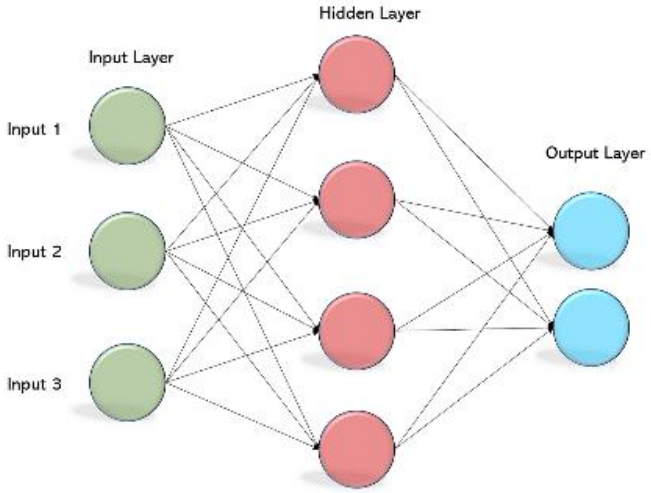


Figure 9

4.2.2. Performanță

Performanța KNN este una aparte, acesta este un algoritm simplu și ușor de înțeles, însă a cărui capacitate poate fi afectată de dimensiunea setului de date, devenind ineficient. Pentru setul nostru de date, acesta poate să aibă o marjă de eroare foarte mare dacă datele de care avem nevoie nu sunt printre acei K vecini.

Spre deosebire de KNN, Random Forest este cunoscut pentru capacitatea de a gestiona seturi mari de date cu diverse caracteristici, având tendința de a obține performanțe bune și fără prelucrare intensivă și slefuire a parametrilor (lucru care va putea fi observat în cele ce vor urma).

Neural Network Regression, conține retele neurale care pot invăța relații complexe între date și pot fi foarte flexible in ceea ce privește modelarea diferitelor

tipuri de date, insă antrenarea necesită multe resurse si o prelucrare intensivă a datelor si arhitecturii retelei.

4.2.3. Unde se folosesc?

KNN este folosit de obicei pentru probleme cu date etichetate si o distributie locală a acestora, Random Forest folosit in probleme cu date multiple si variabile complexe, precum și în selecția caracteristicilor, iar Neural Network Regression este folosit într-o gamă largă de probleme precum predicția prețurilor, estimarea stocurilor, modelarea seriilor de timp etc.

4.3. Construirea modelului

Aceasta constă in definirea caracteristicilor X care sunt *home_team, away_team, day, month, year, tournament, country, neutral*, în timp ce vom avea doua tipuri de variabile tinta y, si anume *home_score* si *away_score*, astfel că au urmat două antrenari cu același (X_train, X_test), însă cu y diferiți (y_train1, y_test1) si (y_train2, y_test2).

După importarea modelului și antrenarea corespunzătoare a acestuia, urmează predicția separată pe fiecare set de date, ca pe urmă crearea unei rotunjiri custom pentru a rotunji partea zecimală (<0.5 la 0 și >=0.5 la intreg).

4.4. Analiză și optimizare

In cadrul analizei inițiale, am evaluat performanța fiecărui model folosind diverse metode precum MSE (Mean Squarred Error) și MAE (Mean Absolute Error), rezultatele obținute oferindu-ne o perspectivă detaliată asupra eficienței modelelor in anticiparea rezultatelor.

Inițial modelele au fost antrenante pe coloanele 'day', 'month', 'year', 'home_team_code', 'away_team_code', 'home_score', 'away_score', 'tournament_code', 'country_code' , si încercând să facă predicția ca un total_score, făcându-se o singură antrenare , abia in modelul MLP ulterior modificat am început revizuirea acestui aspect, ca pe urmă în modelul final să avem modificările complete, lucru pe care îl vom discuta mai încolo.

Interpretarea rezultatelor a devenit un pas esențial pentru înțelegerea capacității fiecărui model astfel că:

- modelul KNN a obținut un MSE de 4.94 și un MAE de 1.56, având o eroare colosală in numărul de goluri (aproximativ radical din 5).
- modelul Random Forest a obținut un scor MSE de aproximativ 0.004 și MAE de 0.002 fiind exact genul de precizie pe care îl cauți în cadrul unui astfel de model.
- de asemenea, și modelul Neural Network Regression emite niște scoruri atractive, având MSE de aproximativ 0.18.

Mai departe am început optimizarea hiperparametrilor pentru cele 3 modele, pentru a vedea daca va exista o imbunătățire in rezultate, astfel că, concluziile sunt urmatoarele:

- pentru modelul KNN am ales optimizare prin Grid Search, aceasta implementându-se prin definirea unui grid de valori posibile pentru hiperparametrii modelului, ca mai apoi să evaluăm performanța modelului pentru fiecare set de valori. Rezultatele deși nu sunt

mulțumitoare, prezintă o imbunătățire semnificativă față de atunci când nu există o optimizare, astfel că avem un MSE de aproximativ 3.82 si MAE de 1.29. Totuși avem și un avertisment din partea programului "The least populated class in y has only 1 member" care ne indică faptul ca în setul de date există cel puțin o clasa care apare foarte rar sau are un număr foarte mic de instanțe in setul de antrenare. (vezi Fig. 9). În acest caz, judecând că avem scoruri, este posibil să existe scoruri rare care apar doar o singură dată în setul de date, ceea ce aduce avertismentul respectiv, lucru de altfel normal dacă ne uităm în extremitatea superioara a scorurilor, de-a lungul instoriei.

```
C:\Users\Viman Mario\Desktop\Programare\Python\Machine
Learning\PredictieMeciuriInternationale\venv\lib\site-packages\sklearn\model_selection\_split.py:737: UserWarning:
 The least populated class in y has only 1 members, which is less than n_splits=5.
  warnings.warn(

Cea mai bună combinație de hiperparametrii: {'n_neighbors': 9, 'p': 1, 'weights': 'distance'}
Acuratețea modelului KNN cu cei mai buni hiperparametrii: 0.31129833657377404
Mean Squared Error: 3.828904731043422
Mean Absolute Error: 1.2929358392741412
```

Figure 10

- în cazul modelului Random Forest, am ales tot o optimizare Grid Search in prima fază, dar din cauza timpului mare de rulare fără vreun rezultat (circa peste 30 de minute) am renunțat la această, înlocuind-o cu Randomized Search. Deși cu parametrii impliciți deja aveam un rezultat extrem de bun, am vrut sa vad daca printr-o optimizare am observa vreo imbunatățire, lucru care nu s-a intamplat, având un MSE de 0.08 față de 0.004 cât era inainte, și un MAE de 0.05 față de 0.002.
- pentru modelul Neural Network Regression nu am mai efectuat o optimizare având deja experiența modelului precedent.

Lucrând la modele, în final am observat o greșeală, sau mai bine zis o abordare imposibil de realizat, în momentul în care am încercat predicția unui meci. Am luat ca și meci pentru predicție African Cup of Nations din 11 Februarie 2024 dintre Nigeria și Coasta de Fildeș scor 1-2.

Ideea care nu poate fi aplicată în cadrul modelului meu este de a include în datele de antrenare home_score și away_score ca modelul să învețe pe baza acestora, fiindcă în momentul creării dicționarului input_data nu pot include în el și aceste scoruri. Astfel, voi fi atenționat de editor că există o discrepanță între setul de antrenare (9 coloane) și data_input (7 coloane). Astfel, în final, am renunțat la antrenarea pe baza scorurilor, și pe lângă asta am adăugat coloana neutral (exclusiv în modelul final).

Acest lucru ne-a dus la pierderea acelor erori minuscule (MSE și MAE) de la modelul Random Forest și Neural Network Regression și aducându-ne la niște erori mult mai mari: MSE de 4.33, respectiv 4.92, pe întregul algoritm NNR. Totuși, trebuie menționat că începând de la MLP, am calculat eroarea MSE și separat pentru scorurile de acasă și cele de deplasare (practic separat pentru fiecare echipă din acel meci), obținând 4.18 pentru 'home_score prediction' și 1.99 pentru 'away_score prediction'. Eroarea mai mare pentru 'home_score prediction' poate rezulta din cauza turneelor finale, în care echipa de acasă practic nu joacă pe teren propriu, aceasta fiind tot în deplasare, ca și cealaltă, putând implica factorul de outlier care indică o posibilă eroare mai mare în determinarea scorului, putând sugera acel element surpriză.

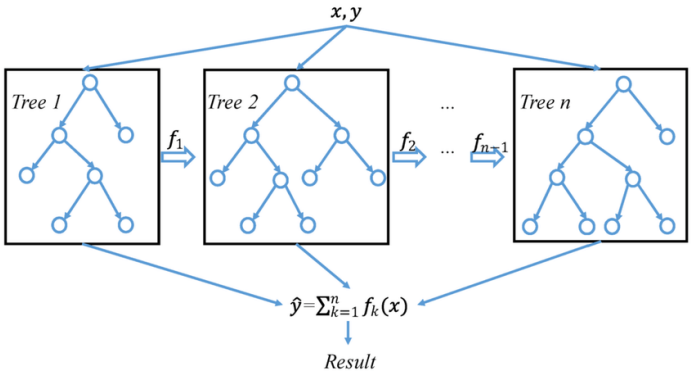
De aceea, pentru a ajunge la aceste rezultate s-au făcut două antrenări (y_train1, y_test1) și (y_train2, y_test2), însă ambele cu același (X_train, X_test), calculându-se astfel și acuratețea, constatând-o la aproximativ 30% pentru modelul MLP. Acestea s-au făcut de asemenea pentru fiecare scor separat.

Predicția scorului acelui meci a fost următoarea: [0.55 goluri] pentru echipa de acasă și [1.42 goluri] pentru echipa din deplasare, care după rotunjire ne aduce la un scor de 1-1, care nu este foarte departe de adevăr, chiar și cu o acuratețe atât de mică.

Capitolul 5. Model final

Deși anterior am exersat antrenarea/predicția asupra 3 modele: KNN, Random Forest și Neural Network Regression, am ales ca și model final XGBoost, pe care îl vom aborda în cele ce urmează.

Funcționarea algoritmului XGBoost constă în combinarea porțiunilor care învață ineficient, cum ar fi arborii de decizie, pentru a forma un model mai precis și robust. Această abordare valorifică atributele cheie ale fiecărei părți, având grijă în același timp de punctele slabe ale acestora



Figură 11

Spre deosebire de antrenările precedente, modelul final va lucra pe 8 coloane: day, month, year, home_team, away_team, country, tournament și neutral, care a fost adăugat exclusiv pentru acesta (toate acestea reprezentând caracteristicile X). Data are scop orientativ, fiindcă un meci jucat în 1905 între două echipe nu este același cu altul jucat în 2005 între aceleași două echipe.

Modelul va avea 2 seturi de antrenare cu aceleași caracteristici x (X_train, X_test), însă cu y diferit (y_train1, y_test1) folosit pentru predicția home_score, în timp ce (y_train2, y_test2) este folosit pentru predicția away_score. După inițializarea modelului, îl antrenăm pe datele y aferente și facem predicția folosind metoda nn_model.predict(X_input), care va returna un array conținând setul de valori din X (în cazul nostru o singură valoare în X fiindcă facem doar câte o predicție pe rând).

Datele pe baza cărora dorim să facem predicția vor fi introduse prin intermediul unui dicționar.

Ca și erori avem pentru home_score MSE de 2.71 și o acuratețe de 27%, iar pentru away_score un MSE de 1.71 și o acuratețe de 33%.

Ca și un lucru extra, am încercat să introduc importanța turneului în modelul final, fiecare având un număr atribuit în funcție de importanța într-un dicționar, ca să văd dacă ar putea influența cumva predicția. Din păcate sau din fericire, (depinde cum dorim să o luăm), nu a influențat în vreun fel rezultatul, lucru despre care am putea spune că este posibil că modelul conștientizează importanța turneului respectiv.

Capitolul 6. Rezultate

6.1. Teste

Pentru a putea verifica eficiența modelului de predicție, am ales simularea cupei mondiale din 2022, iar în cele ce urmează sunt prezentate rezultatele grupelor:

Group A

		Pts	MP	W	D	L	GF	GA	+/-
ADVANCE	Netherlands	9	3	3	0	0	6	3	3
ADVANCE	Qatar	2	3	0	2	1	3	4	-1
	Ecuador	2	3	0	2	1	3	4	-1
	Senegal	2	3	0	2	1	3	4	-1

Auto / Clear

1	November 22, 2022 11:00 AM ET	SENEGAL	1 : 2	NETHERLANDS
	November 21, 2022 11:00 AM ET	QATAR	1 : 1	ECUADOR
2	November 26, 2022 8:00 AM ET	QATAR	1 : 1	SENEGAL
	November 26, 2022 11:00 AM ET	NETHERLANDS	2 : 1	ECUADOR
3	November 30, 2022 10:00 AM ET	NETHERLANDS	2 : 1	QATAR
	November 30, 2022 10:00 AM ET	ECUADOR	1 : 1	SENEGAL

Group B

		Pts	MP	W	D	L	GF	GA	+/-
ADVANCE	England	9	3	3	0	0	7	2	5
ADVANCE	USA	4	3	1	1	1	4	4	0
	Wales	2	3	0	2	1	2	4	-2
	Iran	1	3	0	1	2	3	6	-3

Auto / Clear

1	November 22, 2022 8:00 AM ET	ENGLAND	3 : 1	IRAN
	November 22, 2022 2:00 PM ET	USA	1 : 1	WALES
2	November 26, 2022 5:00 AM ET	WALES	1 : 1	IRAN
	November 26, 2022 2:00 PM ET	ENGLAND	2 : 1	USA
3	November 30, 2022 2:00 PM ET	WALES	0 : 2	ENGLAND
	November 30, 2022 2:00 PM ET	IRAN	1 : 2	USA

Group C

		Pts	MP	W	D	L	GF	GA	+/-
ADVANCE	Argentina	7	3	2	1	0	5	3	2
ADVANCE	Poland	5	3	1	2	0	4	3	1
	Mexico	4	3	1	1	1	4	4	0
	Saudi Arabia	0	3	0	0	3	3	6	-3

Auto / Clear

1	November 23, 2022 5:00 AM ET	ARGENTINA	2 : 1	SAUDI ARABIA
	November 23, 2022 11:00 AM ET	MEXICO	1 : 1	POLAND
2	November 27, 2022 8:00 AM ET	POLAND	2 : 1	SAUDI ARABIA
	November 27, 2022 2:00 PM ET	ARGENTINA	2 : 1	MEXICO
3	December 1, 2022 2:00 PM ET	POLAND	1 : 1	ARGENTINA
	December 1, 2022 2:00 PM ET	SAUDI ARABIA	1 : 2	MEXICO

Group D

		Pts	MP	W	D	L	GF	GA	+/-
ADVANCE	France	9	3	3	0	0	6	3	3
ADVANCE	Tunisia	2	3	0	2	1	3	4	-1
	Denmark	2	3	0	2	1	3	4	-1
	Australia	2	3	0	2	1	3	4	-1

Auto / Clear

1	November 23, 2022 8:00 AM ET	DENMARK	1 : 1	TUNISIA
	November 23, 2022 2:00 PM ET	FRANCE	2 : 1	AUSTRALIA
2	November 27, 2022 5:00 AM ET	TUNISIA	1 : 1	AUSTRALIA
	November 27, 2022 11:00 AM ET	FRANCE	2 : 1	DENMARK
3	December 1, 2022 10:00 AM ET	TUNISIA	1 : 2	FRANCE
	December 1, 2022 10:00 AM ET	AUSTRALIA	1 : 1	DENMARK

Group E

		Pts	MP	W	D	L	GF	GA	+/-
ADVANCE	Spain	9	3	3	0	0	6	3	3
ADVANCE	Germany	4	3	1	1	1	4	4	0
	Japan	2	3	0	2	1	3	4	-1
	Costa Rica	1	3	0	1	2	3	5	-2

Auto / Clear

1	November 24, 2022 8:00 AM ET	GERMANY	1 : 1	JAPAN
	November 24, 2022 11:00 AM ET	SPAIN	2 : 1	COSTA RICA
2	November 28, 2022 5:00 AM ET	JAPAN	1 : 1	COSTA RICA
	November 28, 2022 2:00 PM ET	SPAIN	2 : 1	GERMANY
3	December 2, 2022 2:00 PM ET	JAPAN	1 : 2	SPAIN
	December 2, 2022 2:00 PM ET	COSTA RICA	1 : 2	GERMANY

Group F

		Pts	MP	W	D	L	GF	GA	+/-
ADVANCE	Belgium	9	3	3	0	0	6	3	3
ADVANCE	Morocco	4	3	1	1	1	4	4	0
	Croatia	2	3	0	2	1	3	4	-1
	Canada	1	3	0	1	2	3	5	-2

Auto / Clear

1	November 24, 2022 5:00 AM ET	MOROCCO	1 : 1	CROATIA
	November 24, 2022 2:00 PM ET	BELGIUM	2 : 1	CANADA
2	November 28, 2022 8:00 AM ET	BELGIUM	2 : 1	MOROCCO
	November 28, 2022 11:00 AM ET	CROATIA	1 : 1	CANADA
3	December 2, 2022 10:00 AM ET	CROATIA	1 : 2	BELGIUM
	December 2, 2022 10:00 AM ET	CANADA	1 : 2	MOROCCO

Group H

Group G

		Pts	MP	W	D	L	GF	GA	+/-
ADVANCE	Brazil	9	3	3	0	0	6	2	4
ADVANCE	Switzerland	6	3	2	0	1	5	4	1
	Serbia	3	3	1	0	2	4	5	-1
	Cameroon	0	3	0	0	3	2	6	-4

Auto / Clear

1	November 25 , 2022 5:00 AM ET	SWITZERLAND	2	1	CAMEROON
	November 25 , 2022 2:00 PM ET	BRAZIL	2	1	SERBIA
2	November 29 , 2022 5:00 AM ET	CAMEROON	1	2	SERBIA
	November 29 , 2022 11:00 AM ET	BRAZIL	2	1	SWITZERLAND
3	December 3 , 2022 2:00 PM ET	CAMEROON	0	2	BRAZIL
	December 3 , 2022 2:00 PM ET	SERBIA	1	2	SWITZERLAND

		Pts	MP	W	D	L	GF	GA	+/-
ADVANCE	Portugal	7	3	2	1	0	5	3	2
ADVANCE	Uruguay	4	3	1	1	1	4	4	0
	Korea Republic	2	3	0	2	1	3	4	-1
	Ghana	2	3	0	2	1	3	4	-1

Auto / Clear

1	November 25 , 2022 8:00 AM ET	URUGUAY	2	1	KOREA REP.
	November 25 , 2022 11:00 AM ET	PORTUGAL	2	1	GHANA
2	November 29 , 2022 8:00 AM ET	KOREA REP.	1	1	GHANA
	November 29 , 2022 2:00 PM ET	PORTUGAL	2	1	URUGUAY
3	December 3 , 2022 10:00 AM ET	KOREA REP.	1	1	PORTUGAL
	December 3 , 2022 10:00 AM ET	GHANA	1	1	URUGUAY

În ceea ce privește rezultatele din knock-out stage, acestea sunt urmatoarele:

Optimi de finală:

3 Dec: Netherlands – USA 2-1
3 Dec: Argentina – Tunisia 2-1
5 Dec: Spain – Morocco 2-1
5 Dec: Brazil – Uruguay 2-1
4 Dec: England – Qatar 3-0
4 Dec: France – Poland 2-1
6 Dec: Belgium – Germany 1-2
6 Dec: Portugal – Switzerland 1E-1

Sferturi:

9 Dec: Netherlans – Argentina 2-1
9 Dec: Spain – Brazil 1-2
10 Dec: England – France 2-1
10 Dec: Germany – Portugal 2-1

Semifinale:

13 Dec: Argentina – Brazil 1-1E
14 Dec: England-Germany 2E-2

Locul 3:	**Finală**:
17 Dec: Argentina – Germany 2E-2	18 Dec: Brazil – England 1E-1

Problema care a apărut în cazul fazelor finale ale competiției, după cum se poate observa, este aceea că în semifinale, finala mică și finala mare, toate meciurile s-au terminat la egalitate. Astfel, am decis câștigătoare prin echipa care avea coeficientul ExpectedGoals mai mare. În cazul finalei, coeficientul era de 1.429 la 1.324, fiind foarte aproape. De altfel, același lucru s-a întâmplat și în prima semifinală, fiind 1.392 la 1.481 în favoarea Braziliei. Este de recunoscut faptul că dacă baza de date ne permitea implementarea unui sistem de prelungiri și lovituri de departajare, am fi putut avea altă câștigătoare, chiar pe cea din realitate (Argentina).

6.2. Intrebări de cercetare

6.2.1. Cât de mult se aseamănă predicția cu realitatea?

Să o luăm cu începutul: în grupe s-au jucat un total de 48 de meciuri, dintre care a fost prezis outcome-ul exact ca și în realitate la 23 dintre ele (cine câștigă sau dacă se termină egal), ceea ce ne duce la 47,92% șansa de a prezice cine câștigă. Dacă vorbim de scor exact, aici avem doar 3 meciuri din cele 23, deci dacă ne raportăm la totalul de meciuri(de precizat faptul ca s-a mai ghicit scorul partial adica in loc de 2-1 programul a prezis 1-2, dar acelea nu le luăm in considerare), avem o șansă de doar 6,25% de a obține scorul exact.

Pe măsură ce vorbim de echipele care au ieșit din grupe, avem 12 din 16 care sunt aceleași, surprizele în cadrul predicției fiind Germania, Qatar, Tunisia și Uruguay. De asemenea, în cazul Qatar și Tunisia, toate echipele de pe locurile 2-4 aveau același număr de puncte și același golaveraj, iar în lipsa altor informații relevante precum cartonașele, programul de bracket a decis cine va merge mai departe. Totuși, am realizat predicția cu fiecare dintre cele 3 echipe din fiecare grupă pentru a vedea dacă reușește vreuna să treacă mai departe, răspunsul fiind negativ, deci privind rezultatul de la departajare nu a avut vreo importanță deosebită.

Faptul că avem doar 12 din 16 ca în realitate se observă în sferturi, unde vor fi nimerite 6 din 8, ca pe urmă în finală să fie doar 1 din 4, iar finala să fie una neașteptată între Brazilia și Anglia, soldată cu victoria brazilienilor.

6.2.2. Ce șanse are un outsider să caștige?

În cadrul acestui model, șansele unui outsider să câștige nu sunt foarte măritoare, lucru care nu a putut fi îmbunătățit din cauza volumului mic de date relevante, neavând acces la mai multe statistici despre echipe, la lot etc.

Totuși, în cadrul predicției au existat echipe care au reușit să facă o figură frumoasă în grupe, precum Qatar care în realitate nu a obținut niciun punct, iar aici a obținut două.

Capitolul 7. Concluzii

Fotbalul, ca și unul dintre cele mai iubite sporturi din lume, a atras o mulțime de cercetători din diferite domenii, astfel că predicția rezultatelor meciurilor a devenit un subiect de mare interes care are un impact semnificativ asupra industriei pariurilor sportive, analizei performanței echipelor, a strategiilor de antrenament etc.

Potrivit acestui proiect, s-a testat capacitatea de a prezice rezultatele meciurilor de fotbal international folosind inteligența artificială, lucru evidențiat în mai multe capitole precum:

- introducere: a oferit o introducere generala ăi s-au discutat dificultățile și avantajele acestui domeniu.
- starea actuala a domeniului: s-a oferit o analiză a literaturii existente privind predicția, dar si a inceputului cercetării in fotbal.
- rezultate: prezentarea rezultatelor obtinuțe prin implementarea modelului precum și consecințele rezultatelor.
- etc.

Acest proiect a demonstrat că inteligența artificială are potențialul ca în viitor să prezică meciuri ținând cont de o multitudine mult mai mare de factori pentru o îmbunătățire a acurateții și deschiderea de noi oportunități în alte domenii, datorită dezvoltării continue a acesteia și accesului la integrarea unor noi surse de date.

Capitolul 8. Bibliografie

[1] M. A. Hall, Correlation-based Feature Selection for, The University of Waikato, 1999.

[2] I. R. K. M. N. W. Stefano Nembrini, „The revival of the Gini importance?," *Bioinformatics*, vol. 34, nr. 21, p. 3711–3718, 2018.

[3] D. G. Thomas Reilly, „Science and football: a review of applied research," *Journal of Sports Sciences*, vol. 21, nr. 9, pp. 693-705, 2003.

[4] T. a. B. J. Reilly, „Anaerobic and aerobic," *Training in Sport: Applying Sport Science*, pp. 351-409, 1988.

[5] C. N. F. a. W. C. Nicholas, „The Loughborough Intermittent Shuttle Test: a field test that simulates the activity pattern of soccer," *Journal of Sports Science*, vol. 18, pp. 97-104, 2000.

[6] N. R. T. M. T. R. S. a. R. D. Gleeson, „Influence of acute endurance activity on leg neuromuscular and musculoskeletal performance," *Medicine and Science in Sports and Exercise*, vol. 30, pp. 596-608, 1998.

[7] C. W. A. W. T. a. S. M. Weigelt, „Transfer and motor skill learning in association football," *Ergonomics*, vol. 43, pp. 1698-1 707, 2000.

[8] G. A. a. W. M., „An analysis of penalty kick before and after the recent rule changes," *Insight: The FA Coaches Association Journal*, vol. 1 (2), nr. 21, 1998.

[9] M. Rahman, „A deep learning framework for football match prediction," *SN Appl. Sci*, vol. 2, p. 165, 2020.

[10] M. J. Maher, „Modelling association football scores," *Stat Neerl*, vol. 3, nr. 36, p. 109–118, 1982.

[11] C. S. Dixon MJ, „Modelling association football scores and inefficiencies in the football betting market," *J Royal Stat Soc Ser C (Appl Stat)*, vol. 2, nr. 46, p. 265–280, 2002.