MINISTERUL EDUCAȚIEI
UNIVERSITATEA TEHNICĂ
DIN CLUJ-NAPOCA

Rezultatele Fotbalului Internațional Feminin

Student: **Dan Lorena Elena**

Cuprins

➢ Capitolul 1. Introducere, Motivația alegerii bazei de date

1.1 Introducere

Fotbalul feminin a devenit tot mai important și popular în ultimele decenii, devenind o sursă de inspirație pentru femei din întreaga lume. Cu toate că istoria sa poate fi urmărită până în secolul al XIX-lea, fotbalul feminin a cunoscut o creștere semnificativă în ultimul secol. A devenit un sport în care femeile și-au câștigat locul și respectul în fața publicului, făcându-l la fel de important ca și fotbalul masculin.

Beneficiile implicate în practicarea fotbalului feminin sunt multiple. Pe lângă beneficiile fizice evidente, cum ar fi îmbunătățirea condiției fizice și a sănătății cardiovasculare, fotbalul are și un impact pozitiv asupra sănătății mentale. Este un sport care promovează încrederea în sine si spiritul de echipă . De asemenea, fotbalul feminin oferă oportunități pentru femei să își dezvolte abilități de lider și să își îndeplinească potențialul într-un mediu competitiv.

Un alt aspect interesant al fotbalului feminin este că acesta deschide noi oportunități de explorare a lumii. În calitate de sportiv de performanță, fotbalul feminin aduce jucătoarele în contact cu diverse culturi și locuri, oferindu-le oportunitatea de a călători și a cunoaște noi locuri în timp ce practică pasiunea lor pentru fotbal.

1.2 Motivația alegerii bazei de date

Motivația principală pentru alegerea bazei de date "Women's International Football Results" este legată de pasiunea personală pentru fotbal. Fiind o jucătoare activă de fotbal și făcând parte din echipa CF CFR 1907 Cluj, am avut ocazia să experimentez beneficiile și frumusețea acestui sport. Fotbalul feminin nu este doar o pasiune pentru mine, ci și un mod de viață și o sursă de inspirație constantă.

Prin această documentație, doresc să împărtășesc colegilor mei această pasiune și să le arăt că fotbalul nu este doar pentru bărbați. Prin intermediul acestui proiect, doresc să demonstrez că fetele pot excela în acest sport la fel de mult ca și băieții și că pot concura la același nivel. Este o oportunitate de a promova egalitatea de gen și de a încuraja mai multe femei să se implice în fotbal.

➢ Capitolul 2. Contextul bazei de date și al proiectului, cerințe, ce dorim să obținem

2.1 Contextul de bazei de date și al proiectului

Pentru acest proiect, ne propunem să explorăm și să analizăm datele din mai multe surse referitoare la fotbalul feminin. Avem trei baze de date principale: "goalscorers", "result1" și "shootouts", care conțin informații despre jucătoare, meciuri și rezultatele acestora.

- Baza de date “goalscorers”: această bază de date conține informații despre marcatorii golurilor în meciurile de fotbal feminin. Fiecare înregistrare include: Data meciului, Echipa gazdă (home_team), Echipa oaspete (away_team), Echipa care a marcat (team), Numele marcatorului (scorer), Minutul în care a fost marcat golul (minute), Dacă golul a fost un autogol (own_goal), Dacă golul a fost marcat din penalty (penalty).
- Baza de date ”result1”: această bază de date conține rezultatele meciurilor de fotbal feminin între echipele de acasă și echipele oaspete. Informațiile include: Data meciului (date), Echipa gazdă (home_team), Echipa oaspete (away_team), Scorul echipei gazdă (home_score), Scorul echipei oaspete (away_score), Turneul în cadrul căruia s-a desfășurat meciul (tour0ment) (de exemplu, Euro sau Cupa Mondială).
- Baza de date “shootouts”: această bază de date conține informații despre meciurile care s-au decis prin loviturile de departajare. Fiecare înregistrare include: Data meciului(date), Echipa gazdă (home_team), Echipa oaspete (away_team), Echipa câștigătoare (winner).

2.2 Cerințe ale proiectului

Dezvoltarea unei aplicații pentru căutarea ușoară a informațiilor: ne propunem să dezvoltăm o aplicație care să permită căutarea și filtrarea rapidă a informațiilor despre jucătoare, meciuri și rezultatele acestora, folosind datele din toate cele trei baze de date.

- *Găsirea anumitor jucătoare*: vom utiliza baza de date "goalscorers" pentru a identifica jucătoarele care au marcat cele mai multe goluri și pentru a analiza evoluția lor în timp.
- *Evoluția jucătoarelor între anumite meciuri:* folosind datele din "goalscorers", vom examina modul în care performanța jucătoarelor a evoluat între anumite meciuri sau în cadrul unor perioade de timp.
- *Analiza meciurilor din EURO, Cupa Mondiala și alte campionate importante:* vom utiliza datele din "result1" pentru a identifica și analiza meciurile din turneele importante precum Euro și Cupa Mondială, precum și altele.

- *Informații despre echipe si scorurile obținute:* vom extrage informații despre echipele care au jucat în meciurile din baza de date "result1" și scorurile pe care le-au obținut.
- *Analiza performanței echipelor:* folosind datele din "result1", să se analizeze performanța echipelor în funcție de diferite criterii, cum ar fi numărul de victorii, înfrângeri, goluri marcate și primite etc.
- *Identificarea tendințelor temporale:* să se analizeze tendințele temporale în ceea ce privește rezultatele meciurilor sau performanța jucătoarelor în timp, folosind datele din "result1" și "goalscorers".
- *Compararea performanței echipelor și jucatoarelor:* să se efectueze comparații între performanța diferitelor echipe sau jucătoare, folosind metrici precum media golurilor marcate, procentajul de victorii etc.
- *Predicția rezultatelor meciurilor viitoare:* să se dezvolte modele de predicție pentru a anticipa rezultatele meciurilor viitoare pe baza datelor istorice din "result1" și "goalscorers".
- *Segmentarea datelor:* să se segmenteze datele în funcție de criterii precum țara, turneul, perioada de timp etc., pentru a permite analize comparative și detaliate.
- *Normalizarea și curățarea datelor:* realizarea unui proces de normalizare și curățare a datelor pentru a asigura coerența și calitatea acestora, eliminând erorile și incoerențele.
- *Integrare cu alte surse de date:* explorarea posibilității de a integra datele din bazele de date existente cu alte surse de date relevante, cum ar fi statistici ale jucătorilor sau evenimente media, pentru a obține o perspectivă mai cuprinzătoare.

2.3 Obiectivele Proiectului

Obiectivul principal al acestui proiect este de a dezvolta o aplicație folosind RapidMiner care să faciliteze accesul și analiza datelor despre fotbalul feminin internațional. Această aplicație va permite utilizatorilor să:

- Găsească informații detaliate despre jucătoare și meciuri.
- Analizeze evoluția performanțelor jucătoarelor și echipelor.
- Compară performanțele diferitelor echipe și jucătoare.
- Anticipeze rezultatele meciurilor viitoare.
- Vizualizeze datele într-un mod interactiv și accesibil.

Prin realizarea acestor obiective, proiectul va contribui la o mai bună înțelegere a fotbalului feminin și va oferi un instrument valoros pentru fani, antrenori și cercetători.

➢ Capitolul 3. Aspecte teoretice relevante, inclusiv starea actuală a domeniului

3.1 Analiza Performanței in Fotbal

Analiza performanței sportive a devenit un domeniu de cercetare extrem de activ. Studiile se concentrează pe diverse aspecte ale jocului, cum ar fi eficiența jucătorilor, tactici și strategii de joc, și predicția rezultatelor meciurilor.

- **M. Lames & G. Hansen (2001)** - Acest studiu se concentrează pe analiza jocului de fotbal prin metode de observare și modelare, subliniind importanța datelor colectate în timp real pentru îmbunătățirea performanței echipelor.
- **C. Carling, AM Williams, & T. Reilly (2005)** - Autorii discută despre importanța analizei datelor în fotbal, oferind o privire de ansamblu asupra metodelor utilizate pentru a evalua performanțele fizice și tehnice ale jucătorilor.

3.2 Fotbalul Feminin: Evoluție și Provocări

Fotbalul feminin a cunoscut o creștere considerabilă în popularitate și calitate, dar continuă să se confrunte cu provocări unice.

- **J. Williams(2003)** - Acest studiu explorează istoria și dezvoltarea fotbalului feminin, evidențiind momentele cheie și provocările care au marcat acest sport.
- **G. Pfister(2015)** - Cercetarea se concentrează pe barierele și oportunitățile pentru femei în sport, cu accent pe fotbalul feminin și diferențele de gen în accesul la resurse și suport.

3.3 Tehnologii și Metodologii în Analiza Datelor Stortive

Avansurile tehnologice și metodologiile moderne au revoluționat modul în care sunt analizate datele sportive.

- **J.Gudmundsson & M.Horton(2017)** - O analiză comprehensivă a tehnicilor de vizualizare a datelor sportive și a modului în care acestea pot fi utilizate pentru a îmbunătăți strategia și performanța echipelor.
- **T. D'Orazio, M.Leo & P.Spangnolo(2019)** - Studiul explorează utilizarea inteligenței artificiale și a învățării automate în analiza performanței sportive, cu aplicații specifice în fotbal.

3.4 Predicția Rezultatelor în Fotbal

Predicția rezultatelor meciurilor de fotbal este un subiect de mare interes, implicând diverse metode statistice și de învățare automată.

- **L.M. Hvattum & H. Arntzen(2010)** - Autorii propun un model de predicție bazat pe performanțele trecute ale echipelor și jucătorilor, demonstrând acuratețea predicțiilor realizate. Sistemul de rating ELO este folosit pentru a deriva covariate care sunt apoi utilizate în modele de regresie logit ordonată.
- **A. Rathke(2017)** - Studiul investighează utilizarea modelelor de regresie și a rețelelor neuronale pentru a prezice rezultatele meciurilor de fotbal, cu aplicații specifice în turneele internaționale.

3.5 Analiza Golurilor și a Marcatorilor

Analiza golurilor și a performanței marcatorilor poate oferi perspective asupra eficienței ofensive a echipelor și jucătorilor.

- **H. Lepschy, H. Wasche & A. Woll(2020)** - Acest studiu se concentrează pe modelele de marcare a golurilor și factorii care influențează succesul ofensiv în fotbalul profesionist.
- **A.M. Gomez, C.Lago-Penas & R. Pollard(2013)** - Autorii explorează variabilele care influențează marcarea golurilor în fotbal, inclusiv poziția pe teren și strategiile de joc.

3.6 Importanța Analizei Tactice

Analiza tactică este esențială pentru înțelegerea și îmbunătățirea strategiilor de joc ale echipelor.

- **F.M. Clemente, F.M.L. Martins & R.S. Mendes(2016)** - Studiul explorează utilizarea analizei rețelelor pentru a înțelege structura și dinamica jocului de fotbal, oferind perspective asupra colaborării și interacțiunilor dintre jucători.
- **D. Memmert & R. Rein(2018)** - Autorii discută despre importanța analizei spațiale și a comportamentului colectiv al echipelor în fotbal, subliniind utilizarea tehnologiilor de tracking pentru a optimiza performanța.

3.7 Concluzie

Analiza datelor în fotbalul feminin este un domeniu complex și multidimensional, care implică diverse tehnici și metodologii pentru a înțelege și îmbunătăți performanța jucătoarelor și echipelor. Cercetările de până acum au demonstrat potențialul imens al analizei datelor pentru a transforma modul în care este jucat și gestionat fotbalul feminin. Prin integrarea acestor aspecte teoretice și practice, putem dezvolta aplicații și modele predictive eficiente care să contribuie la progresul acestui sport.

➢ Capitolul 4. Implementarea aspectelor teoretice în cadrul proiectulu

Implementarea aspectelor teoretice este esențială pentru a transforma datele brute în informații utile și valoroase. În cadrul acestui capitol, vom descrie procesul de implementare a tehnicilor de analiză a datelor folosind RapidMiner, un instrument puternic pentru data mining și machine learning. Ne vom concentra pe metodele și pașii utilizați pentru a atinge obiectivele proiectului, precum și pe modul în care am integrat cerințele specifice ale bazei de date.

4.1 Pregătirea Datelor

Curățarea datelor

Primul pas în pregătirea datelor este curățarea acestora. Am identificat și eliminat valorile lipsă, dublurile și erorile din înregistrările bazelor de date "goalscorers", "result1" și "shootouts" făcând aceste modificări in bazele de date propriu-zise.

Trasformarea datelor

Am transformat datele pentru a le face compatibile cu analizele ulterioare. Aceasta a inclus conversia formatelor de date, normalizarea valorilor numerice și codificarea variabilelor categorice. RapidMiner facilitează aceste transformări prin operatorii "Nominal to Numerical" precum si modificările specifice pe care le-am făcut in bazele de date.

Identificarea Jucatoarelor Cheie

Pentru a găsi jucătoarele cu performanțe notabile, am realizat agregări și filtre. Am folosit operatorul "Aggregate" pentru a calcula numărul total de goluri marcate de fiecare jucătoare și operatorul "Filter Examples" pentru a extrage jucătoarele cu cele mai multe goluri.

Evaluarea Performanței Meciurilor

Am dezvoltat modele predictive pentru a anticipa rezultatele meciurilor viitoare. Am utilizat algoritmi de clasificare, precum Decision Trees, Random Forests și Gradient Boosted Trees, pentru a crea aceste modele. Operatorii "Decision Tree", "Random Forest" și “Gradient Boosted Trees” din RapidMiner au fost folosiți pentru antrenarea și validarea modelelor.

4.2 Aanaliza Corelației și Explicarea Rezultatelor

1. Din baza de date "goalscorers":

Din matricea de corelație, observăm că **cea mai mare corelație**, în afara valorii 1 (care reprezintă corelația unei variabile cu ea însăși), este între variabilele **"scorer" și "date"**, cu un coeficient de corelație de **0.9316**. Această corelație ridicată indică o relație puternică între jucătoarea care a marcat (scorer) si data meciului(date). Cu alte cuvinte, performanțele marcatoarelor sunt strâns legate de perioada în care au avut loc meciurile. Aceasta poate reflecta tendințe temporale în fotbalul feminin. De exemplu, anumite jucătoare au performanțe remarcabile în anumite perioade ale carierei lor sau în anumite sezoane. O corelație ridicată poate indica și faptul că, pe măsură ce fotbalul feminin a evoluat și a devenit mai popular, anumite jucătoare au început să iasă în evidență prin performanțe repetate și marcante.

Alte corelații:

- „scorer" si „home_team" (0.4558): Aceasta indică o corelație moderată pozitivă, sugerând că jucătoarele care marchează sunt, de asemenea, influențate de echipa gazdă. Este posibil ca jucătoarele să aibă performanțe mai bune atunci când joacă pe teren propriu.

- "home_team" si "team" (0.5979): Aceasta sugerează o corelație moderată spre puternică între echipa gazdă și echipa care marchează. Este logic, deoarece echipa gazdă ar fi adesea responsabilă pentru majoritatea golurilor într-un meci jucat acasă.

Concluzii: Cea mai mare corelatie din setul de date este între "scorer" si "date" (0.9316), ceea ce indică o relație puternică între cine marchează și data meciului.

Aceasta sugerează că există factori temporali semnificativi care influențează performanța jucătoarelor, cum ar fi evoluția sportului și carierele individuale ale jucătoarelor.

Alte corelații moderat puternice sugerează legături între echipe și performanțele individuale ale jucătoarelor, precum și între echipele gazdă și performanțele marcatoarelor.

Din matricea de corelație, observăm că **cea mai mică corelație** este între **variabilele "home_team" si "minute"**, cu un coefficient de corelatie de **-0.0021**. Acest coeficient de corelație extrem de mic, practic 0, indică faptul că nu există relație lineară între echipa gazdă (home_team) si minutul in care se marcheaza golurile(minute). Aceasta sugerează că minutul în care se marchează golurile nu este influențat de echipa gazdă. Golurile pot fi marcate în orice minut al meciului, indiferent dacă echipa este

gazdă sau nu. În mod similar, faptul că o echipă joacă acasă nu afectează în mod semnificativ minutele specifice în care se marchează golurile.

Alte coleratii:

- „minute” si „scorer”(0.01416): Aceasta indică o corelație foarte mică între minutul în care se marchează și cine marchează golul. Aceasta arată că nu există o legătură puternică între jucătorii specifici și minutele în care se marchează golurile.
- „minute” si „away_team” (-0.0304): Aceasta indică o corelație negativă foarte mică, sugerând că echipa vizitatoare nu are un impact semnificativ asupra minutei în care se marchează golurile.

Concluzii: Cea mai mica corelatie din setul de date este intre „home_team” si „minute” (-0.0021), indicând lipsa unei relații între echipa gazdă și minutul în care se marchează golurile.

Aceasta sugerează că factorul temporal al marcării golurilor este independent de echipa care joacă acasa.

Alte corelații extrem de mici indică, de asemenea, că variabilele legate de echipe și minutele specifice nu sunt legate într-un mod semnificativ.

Attributes	team	scorer	home_t...	away_te...	date	minute
team	1	0.419	0.598	0.440	0.365	0.022
scorer	0.419	1	0.456	0.430	0.932	0.015
home_tea...	0.598	0.456	1	0.183	0.441	-0.002
away_tea...	0.440	0.430	0.183	1	0.431	-0.030
date	0.365	0.932	0.441	0.431	1	0.003
minute	0.022	0.015	-0.002	-0.030	0.003	1

2. Din baza de date „result1”:

Din matricea de corelație, observăm că **cea mai mare corelație**, în afara valorii 1 (care reprezintă corelația unei variabile cu ea însăși), este între variabilele **“date” și “tour0ment”**, cu un coeficient de corelație de **0.5501**. Pe măsură ce fotbalul feminin a devenit mai popular și mai organizat, au fost introduse noi turnee și competiții internaționale, regionale și naționale. Astfel, anumite perioade (ani) sunt mai reprezentative pentru apariția și creșterea numărului de turnee. Este posibil ca anumite turnee să fie organizate în anumite perioade ale anului sau în anumite cicluri, ceea ce creează un pattern temporal specific. De exemplu, Campionatul Mondial de

Fotbal Feminin și Campionatul European de Fotbal Feminin au loc la intervale regulate, iar meciurile din aceste turnee sunt distribuite temporal în mod previzibil. Anumite turnee se desfășoară în mod regulat la anumite date, iar corelația reflectă aceste tipare fixe. De exemplu, Olimpiada și turneele continentale (cum ar fi Campionatul Asiei) au date prestabilite pentru desfășurare.

Alte corelații:

- „home_team" și "away_team"(0.4025): Aceasta indică o corelație moderată între echipele gazdă și cele oaspete, sugerând că există o anumită relație între ele, posibil datorită structurilor de ligă sau campionate unde echipele gazdă și cele oaspete sunt adesea predeterminate.
- "home_team" și „home_score"(0.2584): Aceasta sugerează că echipele gazdă tind să aibă un impact asupra scorului de acasă, indicând un avantaj minor pentru echipa gazdă.

În concluzie, corelația puternică dintre "date" și "tour0ment" subliniază importanța temporalității în organizarea și desfășurarea competițiilor de fotbal feminin, oferind informații valoroase pentru analiza și înțelegerea evoluției acestui sport.

Din matricea de corelație, observăm că **cea mai mică corelație** este între variabilele **"date" și "neutral**", cu un coeficient de corelație de **-0.0023**. Acest coeficient de corelație extrem de mic, practic zero, indică faptul că nu există aproape nicio relație lineară între data meciului și faptul că meciul a fost jucat pe teren neutru. Faptul că meciul este jucat pe un teren neutru nu este legat de data la care a avut loc meciul. Meciurile pe teren neutru pot avea loc în orice moment al anului, fără a urma un tipar temporal specific. Meciurile pe teren neutru sunt determinate de factori organizaționali și logistici care nu sunt legate de datele calendaristice. De exemplu, finale de turnee sau meciuri internaționale pot fi jucate pe teren neutru, dar datele acestor meciuri sunt alese independent de alte meciuri.

Alte corelații:

- „neutral" și "home_score"(-0.0140): Aceasta indică o corelație negativă foarte mică, sugerând că nu există o relație semnificativă între terenul neutru și scorul echipei gazdă.
- "neutral" și "away_score"(-0.0317): Aceasta sugerează că terenul neutru are o influență nesemnificativă asupra scorului echipei oaspete.

Concluzii: Cea mai mica corelație din setul de date este între "date" și "neutral" (-0.0023), indicând lipsa unei relații între data meciului și faptul că meciul a fost jucat pe teren neutru. Aceasta reflectă că meciurile pe teren neutru nu sunt distribuite temporal într-un mod care să fie corelat cu datele specifice ale meciurilor. Alte corelații extrem de mici indică, de asemenea, că variabilele legate de scor și teren neutru nu sunt legate într-un mod semnificativ.

Attributes	date	tour0m...	home_t...	away_te...	home_s...	away_sc...	neutral
date	1	0.550	0.297	0.283	-0.076	0.037	-0.002
tour0ment	0.550	1	0.237	0.242	-0.069	0.009	0.016
home_tea...	0.297	0.237	1	0.403	-0.137	0.258	-0.113
away_tea...	0.283	0.242	0.403	1	0.243	-0.115	-0.105
home_sc...	-0.076	-0.069	-0.137	0.243	1	-0.339	-0.014
away_sc...	0.037	0.009	0.258	-0.115	-0.339	1	-0.032
neutral	-0.002	0.016	-0.113	-0.105	-0.014	-0.032	1

3. Din baza de date „shootouts”:

Cea mai mare corelație este între **"date" și "away_team"**, iar coeficientul de corelație este **0.811**. Acest coeficient de corelație indică cât de strâns sunt datele despre data meciului și echipa oaspete corelate. Cu cât valoarea coeficientului este mai mare, cu atât există o corelație mai puternică între cele două variabile. În acest caz, valoarea mare indică că data meciului și echipa oaspete sunt puternic corelate, adică datele despre data meciului pot influența performanța echipei oaspete sau invers.

Alte corelații:

- "winner" și "date" (0.786): Această corelație arată cât de strâns sunt rezultatele meciurilor corelate cu data acestora. Cu alte cuvinte, cât de mult rezultatele meciurilor depind de data la care acestea au avut loc. O corelație de 0.786 indică că există o legătură semnificativă între rezultatul meciului și data acestuia, sugestivă pentru influența factorilor precum condițiile meteorologice, starea fizică și mentală a jucătorilor, strategiile tactice etc.
- "away_team" și "home_team" (0.646): Această corelație indică cât de strâns sunt echipa oaspete și echipa gazdă corelate. Un coeficient de corelație de 0.646 sugerează că există o legătură semnificativă între performanța echipelor în rolurile lor de oaspeți și gazde. Acest lucru poate fi influențat de factori precum avantajul terenului propriu, strategiile diferite adoptate de echipele acasă și în deplasare, precum și obiectivele și formele echipelor.

În concluzie, Analiza corelațiilor dintre variabilele "date", "winner", "home_team" și "away_team" evidențiază că există legături semnificative între acestea. În primul rând, corelația

ridicată (0.811) între "date" și "away_team" sugerează că performanța echipei oaspete poate fi influențată de data meciului, posibil datorită factorilor precum oboseala sau călătoriile lungi. În al doilea rând, corelația între "winner" și "date" (0.786) indică că rezultatele meciurilor sunt strâns legate de data acestora, posibil fiind influențate de evenimente sau condiții specifice ale zilei. În sfârșit, corelația (0.646) între "away_team" și "home_team" sugerează că performanța echipelor este influențată de statutul lor de gazdă sau oaspete, acest lucru fiind explicat prin avantajul terenului propriu și strategiile diferite adoptate de echipe în funcție de rolul lor în meci.

Cea mai mică corelație din matricea datelor este între variabilele **"home_team" și "winner"**, având o valoare de **0.555**. Această corelație sugerează că există o legătură mai slabă între faptul că echipa gazdă câștigă și rezultatul final al meciului. În alte cuvinte, există o mai mică influență a faptului că echipa gazdă este în avantajul terenului propriu asupra rezultatului final al meciului. Această corelație mai mică poate fi interpretată ca fiind cauzată de o varietate de factori, inclusiv performanța individuală a echipelor în diverse condiții de joc, strategiile tactice adoptate și alți factori specifici fiecărui meci în parte.

Alte corelațiiȘ

- "home_team" și "away_team", având o valoare de 0.646. Acest coeficient sugerează că există o legătură moderată între performanța echipelor gazdă și a celor oaspete. Cu alte cuvinte, performanța unei echipe la domiciliu nu este foarte strâns corelată cu performanța echipei adversare atunci când aceasta joacă în deplasare. Această situație poate fi influențată de factori precum calitatea echipelor, starea terenului, strategiile adoptate și alți factori specifici fiecărui meci.
- "winner" și "home_team", având o valoare de 0.555. Aceasta indică o legătură relativ slabă între faptul că echipa gazdă câștigă și rezultatul final al meciului. Asta înseamnă că avantajul terenului propriu nu are o influență semnificativă asupra șanselor de victorie ale echipei gazdă. Această corelație mai mică poate fi rezultatul diversității de factori care influențează rezultatele meciurilor, cum ar fi forma echipelor, calitatea adversarilor și circumstanțele specifice ale fiecărui meci.

În concluzie, analiza corelațiilor dintre variabilele "home_team", "away_team" și "winner" indică că performanța echipelor, atât în calitate de gazdă, cât și în deplasare, nu este puternic corelată cu rezultatul final al meciului. Această constatare sugerează că rezultatele sportive sunt influențate de o varietate de factori, inclusiv performanța individuală a echipelor, strategiile tactice și circumstanțele specifice ale fiecărui meci în parte.

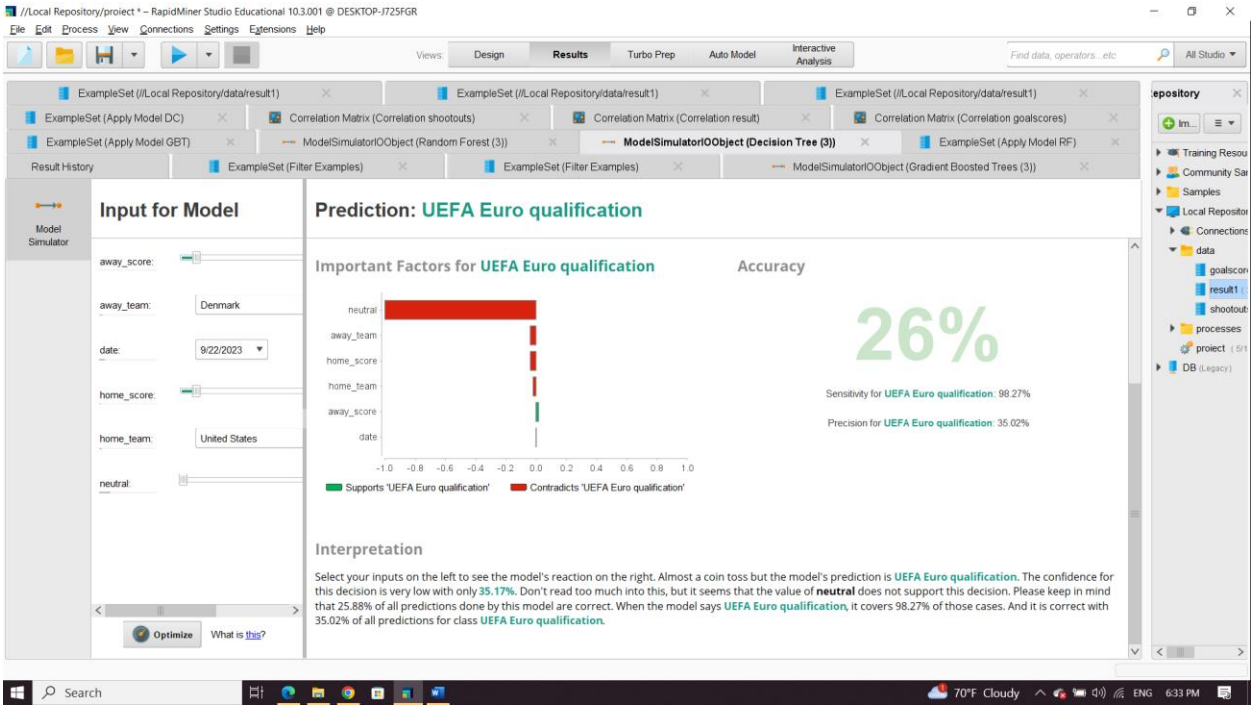
Attributes	date	winner	home_t...	away_te...
date	1	0.786	0.615	0.811
winner	0.786	1	0.556	0.654
home_tea...	0.615	0.556	1	0.646
away_tea...	0.811	0.654	0.646	1

4.3 Modelul Decision tree

Modelul de arbore de decizie este un algoritm de învățare automată care utilizează o serie de reguli de decizie pentru a clasifica datele în funcție de caracteristicile lor. În cazul bazei mele de date, modelul de arbore de decizie a fost antrenat să prezică clasificarea echipelor de fotbal în funcție de turneul la care participă.

Algoritmul de arbore de decizie analizează caracteristicile fiecărui meci, cum ar fi țara, data meciului și scorul, pentru a identifica modele și reguli care să prezică calificarea unei echipe la UEFA Euro. Modelul generează un arbore cu noduri și ramuri, în care fiecare nod reprezintă o caracteristică și fiecare ramură reprezintă o decizie sau o diviziune în funcție de acea caracteristică.

În ciuda unui nivel scăzut de încredere, modelul prezice că evenimentul este asociat cu calificarea la UEFA Euro. Totuși, nivelul scăzut de încredere sugerează că modelul nu este foarte sigur în această predicție și că valoarea 'neutră' nu susține această decizie. Cu toate acestea, atunci când modelul indică calificarea la UEFA Euro, acesta acoperă aproape toate cazurile relevante și este corect în aproximativ o treime din predicțiile sale pentru această clasă.



➢ Capitolul 5. Testare și Validare

Pentru a asigura fiabilitatea și performanța modelelor noastre de predicție, am efectuat teste riguroase de testare și validare. Un aspect esențial al acestui proces a fost împărțirea datelor noastre în două seturi distincte: un set de date de antrenare, utilizat pentru antrenarea modelelor, și un set de date de testare, utilizat pentru a evalua performanța modelelor.

Am utilizat funcția Split Data din RapidMiner pentru a împărți datele noastre în proporția de 80% pentru antrenare și 20% pentru testare. Această abordare ne-a permis să antrenăm modelele pe o cantitate semnificativă de date, asigurând în același timp disponibilitatea unui set independent de date pentru evaluarea obiectivă a performanței modelelor.

➢ Capitolul 6. Rezultate

În acest capitol, sunt prezentate rezultatele obținute în cadrul aplicației, inclusiv rezolvarea anomaliilor în bazele de date, identificarea corelațiilor, dezvoltarea și evaluarea modelelor de predicție, precum și clasificarea golurilor date de către jucători.

6.1 Rezolvarea anomaliilor în baza de date

Pentru a asigura calitatea și integritatea datelor, s-au identificat și rezolvat anomaliile prezente în cele trei baze de date. Aceasta a implicat identificarea și corectarea erorilor, precum date lipsă, date aberante sau valori nevalide.

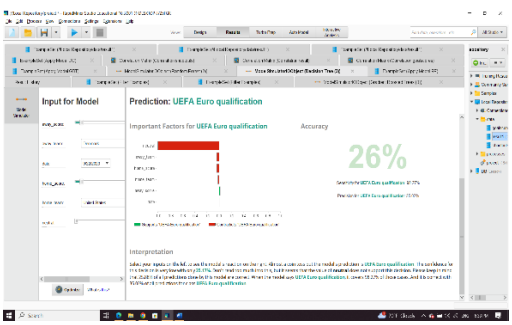
6.2 Identificarea corelațiilor

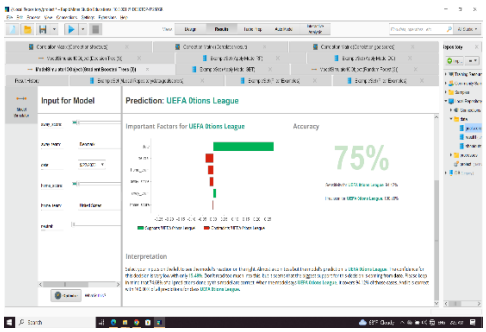
Prin analiza datelor, au fost identificate și evaluate corelațiile dintre diferite variabile din bazele de date. Acest lucru a furnizat o înțelegere mai profundă a relațiilor dintre diferitele aspecte ale datelor și a contribuit la dezvoltarea modelelor de predicție.

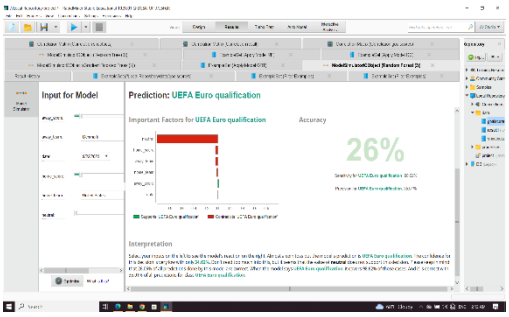
6.3 Dezvoltarea și evaluarea modelelor de predicție

S-au dezvoltat trei modele de predicție: arbore de decizie (Decision Tree), Random Forest și Gradient Boosted Trees (GBT). Deși aceste modele nu au o optimizare foarte bună în ceea ce privește precizia predicțiilor, acestea au fost implementate cu succes și au fost evaluate în cadrul aplicației:

- Decision tree(Arbore de decizie): Acest model utilizează o serie de reguli de decizie pentru a clasifica datele și a prezice rezultatele.
- Random forest(Pădure aleatorie): Acest model combină mai mulți arbori de decizie pentru a îmbunătăți precizia și generalizarea predicțiilor.
- Gradient boosted treets(GBT): Acest model construiește un ansamblu de arbori de decizie într-o manieră secvențială, punând accent pe exemplele dificile de clasificat.



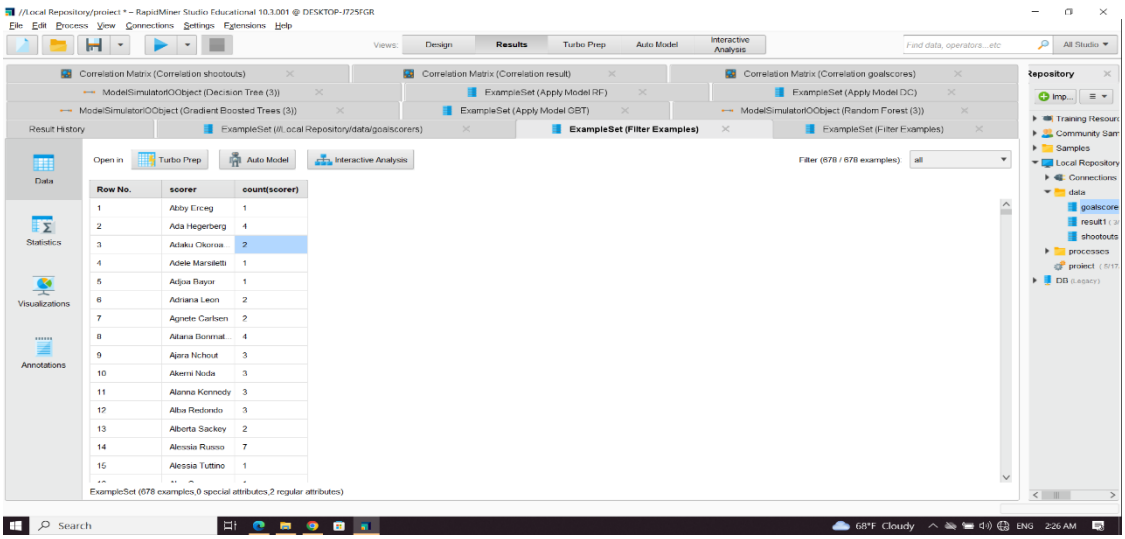




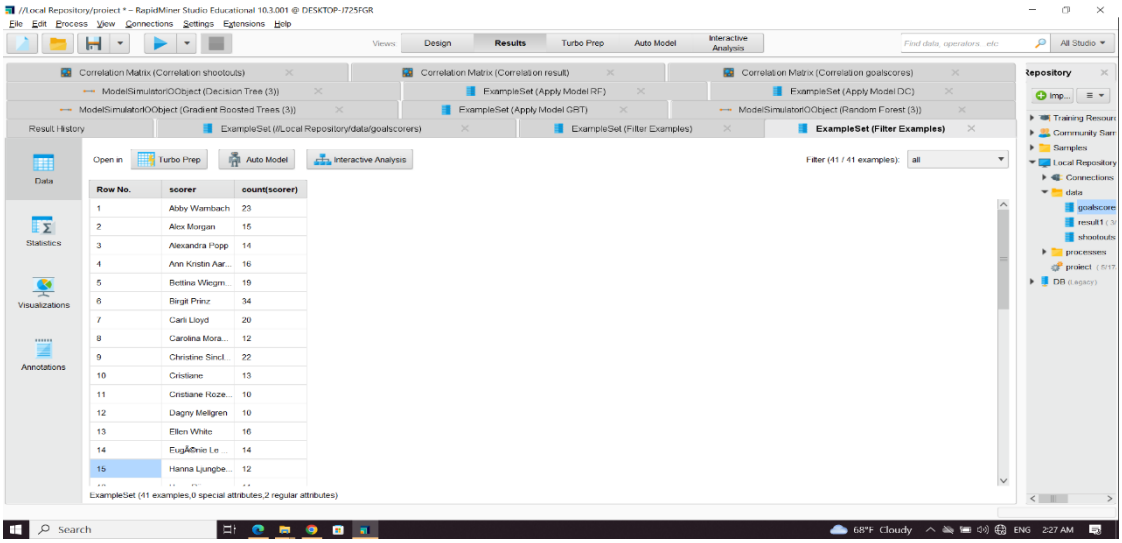
6.4 Clasificarea golurilor date de către jucători

S-a efectuat o analiză a golurilor marcate de către jucătoare pe o anumită perioadă de timp, iar jucătoarele au fost clasificați în două grupuri: cele care au marcat peste si egal 10 goluri și cele care au marcat sub 10 goluri. Această analiză a oferit o perspectivă asupra performanței individuale a jucătoarelor și a contribuției lor la echipă.

Jucatoarele care au marcat sub 10 goluri



Jucatoarele care au marcat peste si egal 10 goluri



➢ Capitolul 7. Concluzii

Proiectul meu a reprezentat o explorare captivantă a analizei datelor în contextul fotbalului feminin internațional. Am reușit să aduc în prim-plan importanța și valoarea analizei datelor în înțelegerea performanței echipelor și jucătoarelor. Am pornit de la rezolvarea anomaliilor și identificarea corelațiilor în bazele de date, continuând cu dezvoltarea și evaluarea modelelor de predicție, precum și clasificarea golurilor marcate de jucători. Aceste eforturi mi-au furnizat o înțelegere mai profundă a performanței echipei și a contribuției individuale a jucătoarelor.

Cu toate acestea, proiectul meu este doar începutul unei călătorii continue. Printre aspectele viitoare de dezvoltare se numără: optimizarea modelelor existente, explorare datelor suplimentare, extinderea funcționalităților aplicației, îmbunatățirea interfeței utilizatorului.