UNIVERSITATEA BABEŞ-BOLYAI CLUJ-NAPOCA
FACULTATEA DE MATEMATICĂ ŞI INFORMATICĂ
SPECIALIZAREA Matematica-Informatica

LUCRARE DE LICENŢĂ

Predicția rezultatelor unor meciuri din fotbal

Conducător științific
Lect. Dr. Onet-Marian Zsuzsanna

Absolvent
Zărnescu Bogdan-Gabriel

2024

ABSTRACT

Abstract

This thesis explores the application of Bayesian networks for predicting the outcomes of football matches, highlighting the versatility and efficiency of this tool in modeling probabilistic relationships and managing uncertainty. The study covers both the fundamental theoretical aspects of Bayesian networks and their practical applicability in the sports domain.

Key contributions of this work include:

Providing a solid theoretical framework for understanding and using Bayesian networks, including their structure and functioning, key concepts, and inference methods.

Detailed analysis of relevant variables for predicting football match outcomes, as well as evaluating different approaches and techniques used in this field.

Developing and evaluating a practical application based on Bayesian networks, demonstrating its efficiency and accuracy in predicting football match results.

The advantages of Bayesian Networks, such as flexibility, interpretability, and computational efficiency, have been highlighted throughout the thesis. However, challenges associated with their use, such as computational complexity and the need for large amounts of data, were also discussed.

Future research directions include exploring the combination of Bayesian Networks with other machine learning techniques to improve prediction accuracy and robustness, extending the applicability of Bayesian Networks in various fields, and developing efficient methods for constructing and calibrating complex networks.

In conclusion, Bayesian Networks represent a valuable tool for modeling probabilistic relationships and managing uncertainty, providing a robust and flexible framework for a variety of applications. The theoretical and practical contributions discussed in this thesis offer a solid foundation for future research and the application of Bayesian Networks in various domains.

Cuprins

Capitolul 1

Introducere

În era digitalizării și a exploziei informaționale, gestionarea și interpretarea datelor complexe reprezintă provocări majore în diverse domenii. Unul dintre cele mai eficiente instrumente pentru tratarea incertitudinii și realizarea raționamentului probabilistic este utilizarea rețelelor bayesiene. Aceste structuri grafice, care modelează relațiile dintre variabilele aleatoare prin intermediul unui graf aciclic direcționat (DAG), au aplicabilitate în numeroase domenii, de la medicină și inginerie, până la finanțe și sport.

Această lucrare își propune să exploreze utilizarea rețelelor bayesiene pentru predicția rezultatelor meciurilor de fotbal, un domeniu complex și captivant care a atras atenția cercetătorilor, analiștilor sportivi și pasionaților de fotbal deopotrivă. În esență, obiectivul este de a anticipa rezultatul unui meci înainte ca acesta să aibă loc, utilizând diverse metode și tehnici de analiză.

Structura lucrării este următoarea:

- **Capitolul 2** oferă o introducere detaliată în rețelele bayesiene, explicând structura și funcționarea acestora, conceptele cheie și metodele de inferență, precum și avantajele și provocările utilizării lor.
- **Capitolul 3** se concentrează pe predicția meciurilor de fotbal, analizând diverse abordări și tehnici utilizate în acest domeniu. Sunt discutate avantajele și limitările fiecărei metode, cu exemple concrete de aplicații.
- **Capitolul 4** prezintă dezvoltarea și evaluarea unei aplicații practice bazate pe rețele bayesiene pentru predicția rezultatelor meciurilor de fotbal. Este realizată o analiză detaliată a setului de date utilizat și sunt evaluate performanțele modelului dezvoltat.
- **Capitolul 5** oferă concluziile generale ale lucrării, evidențiind contribuțiile teoretice și practice, avantajele și limitările abordării propuse și direcțiile viitoare de cercetare.

Prin această structură, lucrarea își propune să ofere o înțelegere cuprinzătoare a rețelelor bayesiene și a modului în care acestea pot fi utilizate pentru a aborda probleme complexe în diverse domenii, cu un accent special pe predicția rezultatelor meciurilor de fotbal.

În timpul elaborării acestei lucrări, autorul a folosit ChatGPT pentru elemente de ortografie și editare a textului pentru a avea o exprimare științifică. După utilizarea acestui instrument, autorul a revizuit și editat conținutul generat și își asumă întreagă responsabilitate pentru conținutul lucrării.

Capitolul 2

Bayesian Networks

2.1 Introducere în Rețelele Bayesiene

În era informației, gestionarea și interpretarea datelor complexe reprezintă provocări majore în diverse domenii. Una dintre abordările eficiente pentru tratarea incertitudinii și realizarea raționamentului probabilistic este utilizarea rețelelor bayesiene. Acestea sunt structuri grafice care modelează relațiile dintre variabilele aleatoare prin intermediul unui graf aciclic direcționat (DAG - *directed acyclic graph*).

Rețelele bayesiene sunt denumite după celebrul matematician britanic Thomas Bayes, fiind bazate pe principiile teoremei lui Bayes. Această teoremă oferă un cadru matematic pentru actualizarea probabilităților în lumina unor noi dovezi, permițând astfel realizarea de inferențe robuste și informate.

Un aspect esențial al rețelelor bayesiene este capacitatea lor de a reprezenta și gestiona incertitudinea într-un mod natural și intuitiv. Fiecare nod din rețea corespunde unei variabile aleatoare, iar legăturile direcționate dintre noduri indică relații de dependență condiționată. Aceste relații permit descompunerea distribuției de probabilitate comună a tuturor variabilelor într-un produs de probabilități condiționate, facilitând astfel calculele probabilistice complexe.

Rețelele bayesiene sunt extrem de versatile și au fost aplicate cu succes într-o gamă largă de domenii, inclusiv în medicină pentru diagnostic și planificarea tratamentului, în inginerie pentru analiza fiabilității sistemelor, și în finanțe pentru evaluarea riscului. Capacitatea lor de a combina datele empirice cu cunoștințele anterioare într-un mod coerent le face un instrument puternic pentru raționament și luarea deciziilor.

În acest capitol, vom explora în detaliu structura și funcționarea rețelelor bayesiene, discutând conceptele cheie și metodele de inferență. De asemenea, vom prezenta avantajele și provocările utilizării acestor rețele, oferind exemple concrete pentru a ilustra aplicabilitatea lor în diverse scenarii. Prin această abordare, ne pro-

punem să oferim o înțelegere cuprinzătoare a rețelelor bayesiene și a modului în care acestea pot fi utilizate pentru a aborda probleme complexe în diverse domenii.

Informațiile prezentate în acest capitol sunt inspirate din articolul lui Ben-Gal (2007) privind rețelele bayesiene [BG07].

2.2 Structura Rețelelor Bayesiene

Rețelele bayesiene sunt instrumente puternice pentru modelarea relațiilor probabilistice dintre variabilele aleatoare. Structura lor constă în două componente esențiale: nodurile și muchiile. În această secțiune, vom explora în detaliu fiecare dintre aceste componente și modul în care acestea contribuie la formarea unui graf aciclic direcționat (DAG).

2.2.1 Nodurile

Fiecare nod dintr-o rețea bayesiană reprezintă o variabilă aleatoare. Aceste variabile pot fi discrete sau continue și pot reprezenta orice tip de informație relevantă pentru domeniul de aplicare al rețelei. De exemplu, într-o aplicație medicală, nodurile pot reprezenta simptome, diagnostice și rezultate ale testelor medicale.

2.2.2 Muchiile

Muchiile dintr-o rețea bayesiană sunt direcționate și indică relațiile de dependență condiționată dintre variabile. O muchie direcționată de la nodul A la nodul B sugerează că A are o influență directă asupra lui B. În acest context, A este numit părinte, iar B este numit copil.

Direcția muchiilor este crucială, deoarece determină modul în care se vor calcula probabilitățile condiționate. Prin intermediul acestor relații, rețeaua bayesiană poate captura complexitatea dependențelor dintre variabilele aleatoare.

2.2.3 Graful Aciclic Direcționat (DAG)

Structura rețelelor bayesiene este reprezentată printr-un graf aciclic direcționat (DAG). Un DAG este un graf care nu conține cicluri, ceea ce înseamnă că nu există nicio cale direcționată care să înceapă și să se termine la același nod.

Graful aciclic direcționat (DAG) este esențial pentru rețelele bayesiene, deoarece asigură că toate relațiile de dependență condiționată sunt reprezentate corect. Fiecare nod din DAG poate fi descompus într-o distribuție de probabilitate condiționată de părinții săi. Ben-Gal [BG07] subliniază importanța graficelor aciclice direcționate

(DAG) în reprezentarea relațiilor condiționate dintre variabile, evidențiind că aceste grafuri permit descompunerea distribuției de probabilitate într-un produs de probabilități condiționate.

2.2.4 Exemplu de Structură a unei Rețele Bayesiene

Pentru a ilustra structura unei rețele bayesiene, se consideră un exemplu simplu din domeniul medical. S-a presupus că se dorește modelarea relației dintre simptomele unui pacient (febră și tuse), prezența unei boli (gripă) și rezultatul unui test medical (test pozitiv).

- Nodurile reprezintă variabilele: febră (F), tuse (T), gripă (G) și test pozitiv (P).
- Muchiile direcționate sunt: $G \rightarrow F$, $G \rightarrow T$, $G \rightarrow P$.

Reprezentarea acestei structuri se poate vizualiza în Figura 2.1

În această rețea, probabilitatea ca un pacient să aibă febră, tuse și un test pozitiv este condiționată de prezența gripei, iar relațiile de dependență sunt capturate prin muchiile direcționate ale grafului.

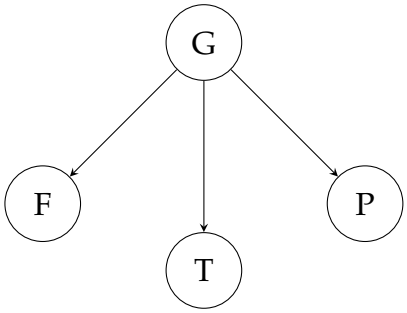

Figura 2.1: Exemplu simplu al unei rețelei bayesiene

2.3 Concepte Cheie

Rețelele bayesiene sunt apreciate pentru capacitatea lor de a gestiona complexitatea relațiilor probabilistice dintre variabile. În această secțiune, vom explora conceptele cheie care stau la baza funcționării rețelelor bayesiene: independența condiționată, distribuția de probabilitate comună, inferența și învățarea.

2.3.1 Independența Condiționată

Unul dintre principalele avantaje ale utilizării rețelelor bayesiene este capacitatea de a reprezenta independența condiționată dintre variabile. Independența

condiționată permite simplificarea calculelor probabilistice prin reducerea numărului de relații directe care trebuie luate în considerare.

Dacă două variabile X și Y sunt condiționat independente dată fiind o a treia variabilă Z, relația lor poate fi exprimată astfel:

$$P(X, Y|Z) = P(X|Z) \cdot P(Y|Z)$$

Această proprietate este reprezentată în rețeaua bayesiană prin absența unei muchii directe între X și Y, condiționată de Z.

2.3.2 Distribuția de Probabilitate Comună

Un aspect esențial al rețelelor bayesiene este abilitatea de a descompune distribuția de probabilitate comună a unui set de variabile în produsul probabilităților condiționate ale fiecărei variabile, dați fiind părinții săi din graf. Aceasta este una dintre principalele caracteristici care fac rețelele bayesiene atât de puternice și eficiente.

Formula pentru descompunerea distribuției de probabilitate comună este:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\text{Părinți}(X_i))$$

Această formulă permite calcularea probabilităților într-un mod modular, utilizând relațiile de dependență condiționată specificate în rețea.

2.3.3 Inferența în Rețele Bayesiene

Inferența este procesul de calculare a distribuțiilor de probabilitate pentru anumite variabile, dată fiind observația altor variabile. În contextul rețelelor bayesiene, inferența poate fi utilizată pentru a răspunde la întrebări despre distribuțiile probabile ale unor variabile necunoscute, având în vedere informațiile disponibile.

Metodele comune de inferență includ:

- **Inferența exactă:** Metode precum eliminarea variabilelor, care calculează exact distribuțiile de probabilitate necesare.
- **Inferența aproximativă:** Metode precum simularea Monte Carlo, care aproximează distribuțiile de probabilitate prin generarea de eșantioane multiple.

Inferența exactă este de preferat atunci când este posibil, dar inferența aproximativă poate fi necesară în cazurile în care rețeaua este prea complexă pentru a permite calcule exacte eficiente.

2.3.4 Învățarea în Rețele Bayesiene

Învățarea într-o rețea bayesiană implică două aspecte principale: învățarea parametrilor și învățarea structurii.

Învățarea parametrilor: Procesul de estimare a valorilor pentru probabilitățile condiționate din date. Aceasta poate fi realizată utilizând metode precum estimarea maximului de verosimilitate sau inferența bayesiană.

Învățarea structurii: Determinarea configurației rețelei (adică, care variabile sunt conectate prin muchii direcționate). Acest proces este mai complex și poate implica explorarea unui spațiu mare de structuri posibile. Algoritmii de învățare a structurii pot utiliza scoruri de adecvare a datelor și diverse metode euristice pentru a identifica structura optimă.

Învățarea rețelelor bayesiene permite adaptarea modelului la datele specifice ale problemei și îmbunătățirea performanței inferenței prin actualizarea parametrilor și structurii rețelei.

2.4 Aplicații ale Rețelelor Bayesiene

Rețelele bayesiene au o aplicabilitate vastă datorită capacității lor de a gestiona incertitudinea și de a modela relațiile probabilistice complexe. În această secțiune, vom explora câteva dintre domeniile cheie în care rețelele bayesiene au demonstrat o utilitate semnificativă: medicina, ingineria și finanțele.

2.4.1 Aplicații în Medicină

În domeniul medical, rețelele bayesiene sunt utilizate pentru diagnosticarea bolilor, planificarea tratamentului și prognoza evoluției pacienților. Capacitatea rețelelor bayesiene de a integra date clinice și cunoștințe anterioare într-un cadru probabilistic le face deosebit de valoroase în medicină.

- **Diagnosticare:** Rețelele bayesiene pot combina simptomele observate și rezultatele testelor pentru a estima probabilitatea prezenței diferitelor boli. De exemplu, un sistem bazat pe rețele bayesiene poate folosi simptome precum febra, tusea și dificultățile respiratorii pentru a diagnostica pneumonia.
- **Planificarea tratamentului:** Rețelele bayesiene pot ajuta medicii să aleagă cel mai potrivit tratament pentru un pacient, luând în considerare istoricul medical și răspunsurile anterioare la tratamente.

- **Prognoza evoluției:** Rețelele bayesiene pot fi utilizate pentru a prognoza evoluția stării de sănătate a unui pacient pe baza datelor clinice și a tratamentului administrat.

2.4.2 Aplicații în Inginerie

În inginerie, rețelele bayesiene sunt folosite pentru analiza fiabilității sistemelor, gestionarea riscurilor și monitorizarea stării echipamentelor.

- **Analiza fiabilității:** Rețelele bayesiene pot modela dependențele dintre diferitele componente ale unui sistem și pot estima probabilitatea de defectare a întregului sistem. Aceste modele pot ajuta la identificarea componentelor critice și la planificarea întreținerii preventive.
- **Gestionarea riscurilor:** În proiectele de inginerie complexă, rețelele bayesiene pot evalua riscurile asociate diferitelor faze ale proiectului și pot ajuta la luarea deciziilor pentru a minimiza aceste riscuri.
- **Monitorizarea stării:** Rețelele bayesiene pot fi utilizate pentru a monitoriza starea echipamentelor industriale în timp real, detectând anomalii și prezicând momentul optim pentru întreținere.

2.4.3 Aplicații în Finanțe

În sectorul financiar, rețelele bayesiene sunt folosite pentru evaluarea riscurilor, detectarea fraudelor și modelarea piețelor financiare.

- **Evaluarea riscurilor:** Rețelele bayesiene pot evalua riscul de credit al clienților bancari, combinând datele istorice cu informațiile actuale despre comportamentul financiar al clienților.
- **Detectarea fraudelor:** În tranzacțiile financiare, rețelele bayesiene pot identifica modele de comportament fraudulos, ajutând la prevenirea pierderilor financiare.
- **Modelarea piețelor financiare:** Rețelele bayesiene pot modela relațiile dintre variabilele economice și financiare, ajutând la prognoza evoluției piețelor și la luarea deciziilor de investiție.

2.4.4 Alte Aplicații

Rețelele bayesiene au aplicații și în alte domenii, cum ar fi:

- **Inteligența artificială:** Pentru raționament și învățare în sisteme autonome.
- **Bioinformatică:** Pentru analiza datelor genetice și modelarea proceselor biologice.
- **Psihologie:** Pentru modelarea proceselor cognitive și inferențelor în cercetările psihologice.

Aceste aplicații demonstrează versatilitatea și puterea rețelelor bayesiene în abordarea problemelor complexe în diverse domenii. Un exemplu notabil îl reprezintă domeniul sportiv, unde conform unei analize comparative realizate de Raz [Raz17], rețelele bayesiene s-au dovedit a fi o metodă eficientă pentru predicția rezultatelor, în special în fotbal, obținând performanțe comparabile cu alte metode consacrate de predicție.

Mai mult, Herbinet [HER18] subliniază utilizarea rețelelor bayesiene în combinație cu alte tehnici de învățare automată pentru predicția evenimentelor sportive, demonstrând aplicabilitatea și eficiența acestor modele în prognoza rezultatelor meciurilor de fotbal.

Studiile au demonstrat că rețelele bayesiene oferă avantaje semnificative în predicția rezultatelor sportive, comparativ cu alte tehnici de învățare automată, cum ar fi arborii de decizie sau regresia logistică. Acestea permit o mai bună gestionare a incertitudinii și pot modela date incomplete, oferind o performanță robustă în predicțiile sporturilor de echipă, cum ar fi fotbalul și cricketul [PKJ21].

Rețelele bayesiene au fost aplicate și în sporturi precum baschetul, pentru a analiza și prezice performanțele jucătorilor. Un studiu recent a arătat că aceste rețele pot modela relațiile multivariate între indicatorii de performanță, ajutând la dezvoltarea strategiilor de joc și optimizarea deciziilor tactice [PD23].

2.5 Avantajele Rețelelor Bayesiene

Rețelele bayesiene sunt apreciate pentru numeroasele lor avantaje în modelarea relațiilor probabilistice și gestionarea incertitudinii. În această secțiune, vom explora câteva dintre avantajele cheie ale utilizării rețelelor bayesiene în diverse aplicații.

2.5.1 Modularitate

Unul dintre cele mai mari avantaje ale rețelelor bayesiene este modularitatea lor. Fiecare nod din rețea reprezintă o variabilă aleatoare, iar relațiile dintre noduri sunt definite prin probabilități condiționate. Aceasta permite ca rețeaua să fie actualizată modular, fără a necesita recalcularea întregii structuri.

Modularitatea facilitează actualizarea și extinderea rețelei atunci când sunt disponibile noi date sau cunoștințe. De asemenea, permite reutilizarea părților componente ale rețelei în alte aplicații, economisind timp și resurse.

2.5.2 Interpretabilitate

Rețelele bayesiene oferă o vizualizare clară a relațiilor de dependență dintre variabile, făcându-le ușor de înțeles și interpretat. Structura grafică a rețelei, reprezentată printr-un graf aciclic direcționat (DAG), permite utilizatorilor să vadă direct modul în care variabilele sunt interconectate.

Această interpretabilitate este esențială în domenii precum medicina și ingineria, unde experții trebuie să înțeleagă și să explice raționamentul din spatele deciziilor luate de model.

2.5.3 Eficiență Computațională

Rețelele bayesiene permit descompunerea distribuției de probabilitate comună într-un produs de probabilități condiționate, ceea ce facilitează calculul eficient al probabilităților. Aceasta reduce complexitatea calculului, permițând realizarea de inferențe rapide chiar și în rețele mari și complexe.

Metodele de inferență exactă și aproximativă, cum ar fi eliminarea variabilelor și simularea Monte Carlo, contribuie la eficiența computațională a rețelelor bayesiene. Aceste metode permit calculul probabilităților necesare pentru luarea deciziilor într-un mod rapid și precis.

2.5.4 Flexibilitate în Modelare

Rețelele bayesiene sunt extrem de flexibile și pot fi utilizate pentru a modela o gamă largă de relații probabilistice. Ele pot reprezenta atât variabile discrete, cât și continue, și pot gestiona dependențe complexe între variabile.

Această flexibilitate permite adaptarea rețelelor bayesiene la diverse aplicații și domenii, de la diagnostic medical la analiza fiabilității sistemelor și evaluarea riscurilor financiare.

2.5.5 Gestionarea Informațiilor Incomplete

Un alt avantaj important al rețelelor bayesiene este capacitatea lor de a gestiona informațiile incomplete sau lipsă. Rețelele bayesiene pot realiza inferențe și actualizări probabilistice chiar și atunci când unele date nu sunt disponibile.

Aceasta este o caracteristică esențială în multe aplicații practice, unde datele complete pot fi dificil de obținut. Capacitatea de a lucra cu informații incomplete face rețelele bayesiene deosebit de utile în situații de incertitudine.

2.5.6 Integrarea Cunoștințelor Anterioare

Rețelele bayesiene permit integrarea cunoștințelor anterioare (prior knowledge) cu datele empirice pentru a construi modele probabilistice robuste. Aceasta se realizează prin specificarea distribuțiilor anterioare pentru variabilele din rețea, care sunt apoi actualizate pe baza datelor noi.

Această caracteristică este deosebit de valoroasă în domenii unde există expertiză și cunoștințe anterioare semnificative, care pot fi combinate cu datele actuale pentru a îmbunătăți precizia și relevanța modelului.

2.6 Provocări în Utilizarea Rețelelor Bayesiene

Deși rețelele bayesiene oferă numeroase avantaje în modelarea relațiilor probabilistice și gestionarea incertitudinii, există și provocări semnificative asociate cu utilizarea lor. În această secțiune, vom explora câteva dintre aceste provocări, inclusiv complexitatea calculului, învățarea structurii, necesitatea unor cantități mari de date și gestionarea incertitudinii și a datelor incomplete.

2.6.1 Complexitatea Calculului

Una dintre principalele provocări în utilizarea rețelelor bayesiene este complexitatea calculului, în special în cazul rețelelor mari și complexe. Inferența exactă în astfel de rețele poate deveni rapid ineficientă din punct de vedere computațional, datorită numărului mare de combinații posibile de variabile.

Pentru a depăși această provocare, sunt adesea utilizate metode de inferență aproximativă, cum ar fi simularea Monte Carlo. Aceste metode permit realizarea de inferențe în rețele complexe, dar la prețul unei precizii reduse și a unei creșteri a timpului de calcul.

2.6.2 Învățarea Structurii

Învățarea structurii unei rețele bayesiene din date reprezintă o provocare majoră, deoarece spațiul posibil de structuri este foarte mare. Identificarea structurii optime necesită explorarea unui număr mare de configurații posibile, ceea ce poate fi extrem de consumator de resurse.

Algoritmii de învățare a structurii, cum ar fi căutarea bazată pe scoruri și metodele de restricționare, sunt folosiți pentru a aborda această problemă. Cu toate acestea, acești algoritmi pot fi limitați de capacitatea lor de a gestiona seturi de date mari și complexe.

2.6.3 Necesitatea Cantităților Mari de Date

Pentru a estima precis parametrii unei rețele bayesiene, este adesea necesară o cantitate mare de date. Datele insuficiente pot duce la estimări nesigure și la performanțe slabe ale modelului.

În multe aplicații practice, obținerea unor cantități mari de date poate fi dificilă sau costisitoare. Aceasta reprezintă o limitare semnificativă în utilizarea rețelelor bayesiene în anumite domenii.

2.6.4 Gestionarea Informațiilor Incomplete

Deși rețelele bayesiene sunt capabile să gestioneze informațiile incomplete, acest proces poate fi complex și consumator de resurse. În unele cazuri, datele lipsă pot introduce incertitudine suplimentară în model și pot afecta acuratețea inferențelor.

Pentru a gestiona informațiile incomplete, sunt utilizate tehnici precum imputarea datelor și inferența probabilistică. Totuși, aceste tehnici pot introduce și ele incertitudine și pot necesita resurse computaționale semnificative.

2.6.5 Evaluarea Performanței

Evaluarea performanței rețelelor bayesiene poate fi dificilă, în special în cazul problemelor complexe cu multe variabile. Stabilirea criteriilor de evaluare adecvate și interpretarea rezultatelor pot necesita expertiză semnificativă.

Metodele de validare încrucișată și evaluarea pe seturi de date independente sunt adesea folosite pentru a evalua performanța rețelelor bayesiene. Cu toate acestea, metodele menționate pot fi consumatoare de timp și resurse, în special pentru seturi de date mari.

2.7 Exemplu de Rețea Bayesiană

Pentru a ilustra modul în care funcționează rețelele bayesiene, vom construi și analiza un exemplu simplu, menționat și în Secțiunea 2.2.4. Acest exemplu va demonstra cum variabilele pot fi reprezentate și cum relațiile de dependență condiționată sunt folosite pentru a face inferențe probabilistice.

2.7.1 Descrierea Problemei

Să considerăm un exemplu din domeniul medical. Dorim să modelăm relația dintre simptomele unui pacient, prezența unei boli și rezultatul unui test medical. Variabilele implicate în această rețea bayesiană sunt:

- **Febră (F):** Variabilă binară care indică dacă pacientul are febră (F=true) sau nu (F=false).
- **Tuse (T):** Variabilă binară care indică dacă pacientul are tuse (T=true) sau nu (T=false).
- **Gripă (G):** Variabilă binară care indică dacă pacientul are gripă (G=true) sau nu (G=false).
- **Test (P):** Variabilă binară care indică rezultatul testului pentru gripă (P=pozitiv) sau (P=negativ).

2.7.2 Structura Rețelei

Rețeaua bayesiană pentru această problemă poate fi reprezentată grafic printr-un graf aciclic direcționat (DAG). Structura rețelei este definită prin relațiile de dependență condiționată dintre variabile. Vom considera următoarele relații:

- Gripă (G) influențează apariția febrei (F).
- Gripă (G) influențează apariția tusei (T).
- Gripă (G) influențează rezultatul testului medical (P).

Aceste relații pot fi reprezentate prin săgeți în rețea:

$$G \to F, \quad G \to T, \quad G \to P$$

Structura rețelei este prezentată în Figura 2.2

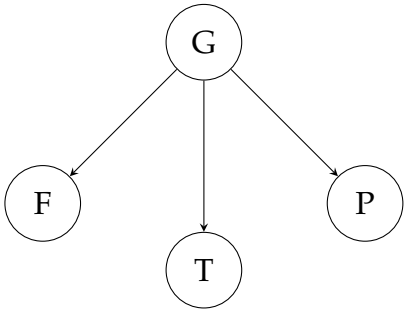

Figura 2.2: Structura rețelei bayesiene pentru a determina dacă cineva are gripă

2.7.3 Distribuțiile de Probabilitate

Fiecare nod din rețea are asociată o distribuție de probabilitate condiționată. Aceste distribuții sunt esențiale pentru realizarea inferențelor în rețea.

- $\boldsymbol{P(G)}$**:** Probabilitatea ca un pacient să aibă gripă.
- $\boldsymbol{P(F|G)}$**:** Probabilitatea ca un pacient să aibă febră, dat fiind că are gripă.
- $\boldsymbol{P(T|G)}$**:** Probabilitatea ca un pacient să aibă tuse, dat fiind că are gripă.
- $\boldsymbol{P(P|G)}$**:** Probabilitatea ca rezultatul testului să fie pozitiv, dat fiind că pacientul are gripă.

Să presupunem următoarele valori pentru aceste probabilități:
$P(G = \text{true}) = 0.1$
$P(F = \text{true}|G = \text{true}) = 0.8$
$P(F = \text{true}|G = \text{false}) = 0.2$
$P(T = \text{true}|G = \text{true}) = 0.7$
$P(T = \text{true}|G = \text{false}) = 0.3$
$P(P = \text{pozitiv}|G = \text{true}) = 0.9$
$P(P = \text{pozitiv}|G = \text{false}) = 0.1$

2.7.4 Inferența în Rețea

Să presupunem că un pacient prezintă simptome de febră și tuse. Dorim să calculăm probabilitatea ca acest pacient să aibă gripă, având în vedere aceste simptome.

Utilizăm teorema lui Bayes pentru a calcula această probabilitate:

$$P(G = true|F = true, T = true) = \frac{P(F = true, T = true|G = true) \cdot P(G = true)}{P(F = true, T = true)}$$

Pentru a calcula $P(F = true, T = true)$, folosim regula sumei:
$P(F = \text{true}, T = \text{true}) = P(F = \text{true}, T = \text{true} \mid G = \text{true}) \cdot P(G = \text{true}) +$
$+ P(F = \text{true}, T = \text{true} \mid G = \text{false}) \cdot P(G = \text{false})$
Calculăm valorile necesare:
$P(F = true, T = true|G = true) = P(F = true|G = true) \cdot$
$\cdot P(T = true|G = true) =$
$= 0.8 \cdot 0.7 = 0.56$
$P(F = true, T = true|G = false) = P(F = true|G = false) \cdot$
$\cdot P(T = true|G = false) =$
$= 0.2 \cdot 0.3 = 0.06$

$P(F = true, T = true) = 0.56 \cdot 0.1 + 0.06 \cdot 0.9 = 0.056 + 0.054 = 0.11$

În final, calculăm probabilitatea dorită:

$$P(G = true|F = true, T = true) = \frac{0.56 \cdot 0.1}{0.11} = \frac{0.056}{0.11} \approx 0.509$$

Aceasta arată că, având în vedere simptomele de febră și tuse, probabilitatea ca pacientul să aibă gripă este de aproximativ 50.9%.

Capitolul 3

Predictia Meciurilor de Fotbal

3.1 Introducere în Tema Predicției Meciurilor de Fotbal

Predicția rezultatelor meciurilor de fotbal reprezintă un domeniu complex și captivant care a atras atenția cercetătorilor, analiștilor sportivi și pasionaților de fotbal deopotrivă. În esență, obiectivul acestei teme este de a anticipa rezultatul unui meci înainte ca acesta să aibă loc, utilizând diverse metode și tehnici de analiză. Predicția poate viza aspecte precum determinarea echipei câștigătoare, estimarea scorului exact sau chiar identificarea unor evenimente specifice din timpul meciului, cum ar fi numărul de goluri marcate, posesia mingii sau numărul de cartonașe.

Complexitatea acestei probleme derivă din natura aleatorie și imprevizibilă a sportului, unde factori precum forma de moment a echipelor, accidentările, condițiile meteorologice și deciziile tactice ale antrenorilor pot influența semnificativ rezultatul final. Totodată, fotbalul este un joc de echipă, iar performanța colectivă poate varia considerabil de la un meci la altul. Din acest motiv, predicția rezultatelor meciurilor de fotbal necesită o abordare multidisciplinară, combinând statistica, învățarea automată, analiza datelor și cunoștințele specifice sportului.

Problema principală în predicția meciurilor de fotbal se poate segmenta în două mari categorii: predicția rezultatelor exacte (scorul final) și predicția rezultatului binar (câștigătorul meciului sau dacă se va termina la egalitate). Fiecare dintre aceste categorii prezintă provocări distincte și necesită modele și metode de analiză adaptate specificului lor. În continuare, vom explora diverse abordări și tehnici utilizate în acest domeniu, evidențiind avantajele și limitările fiecăreia.

3.2 Abordari diverse în Predictia meciurilor de fotbal

3.2.1 Analiza detaliată a variabilelor pentru predicția rezultatelor meciurilor de fotbal în sezonul 2013/14 [R22]

Fátima Rodrigues și Ângelo Pinto prezintă în "Prediction of football match results with Machine Learning" [R22] o analiză a anumitor modele de predicție a scorului pentru sezonul 2013/2014. În cadrul testărilor pentru modelele de predicție în sezonul 2013/14, au fost considerate 7 runde de jocuri, totalizând 70 de partide din prima liga a Angliei. Un aspect crucial a fost identificarea și utilizarea variabilelor preexistente la începerea unui meci, cum ar fi media golurilor unei echipe.

Studiul a inclus analiza detaliată a variabilelor colectate, dintre care 31 de variabile au fost considerate inițial. Aceste variabile includ factori precum evaluarea generală a echipei, evaluările atacului, mijlocului și apărării, care au fost considerate constante pe durata unui sezon pentru fiecare echipă. Rodrigues și Pinto [R22] au demonstrat, într-un studiu realizat asupra sezonului 2013/2014, că variabile precum evaluările generale ale echipelor, performanțele anterioare, precum și factori legați de jocurile anterioare, au o influență semnificativă asupra predicției rezultatelor. Utilizarea corelației între aceste variabile a permis o optimizare eficientă a procesului de predicție. În plus, au fost luate în considerare variabile legate de performanțele anterioare, cum ar fi victoriile echipei gazdă și cele ale echipei vizitatoare, precum și media golurilor încasate acasă sau în deplasare.

O matrice de corelație a fost utilizată pentru a evalua relațiile dintre aceste variabile, demonstrând coerența dintre variabilele care evaluează calitatea echipei de fotbal. Matricea de corelație a fost utilă în identificarea variabilelor redundante, iar eliminarea celor cu corelații ridicate a optimizat procesul de predicție.

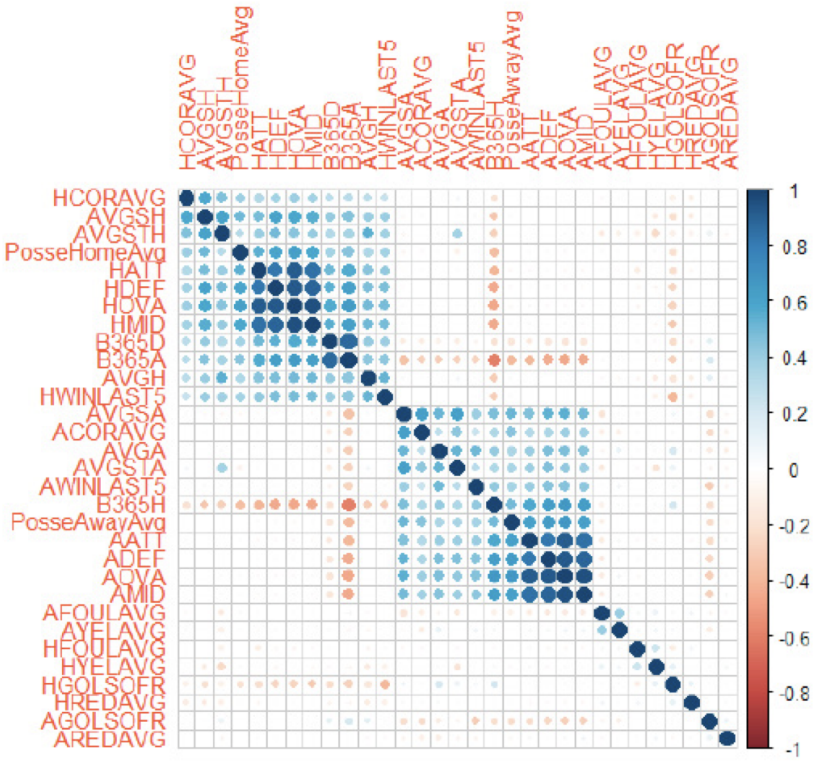

Figura 3.1: Matricea de Corelatie intre Variabile [R22]

Tehnici precum algoritmul Boruta au contribuit la identificarea celor mai semnificative variabile. Acest algoritm a eliminat 7 variabile nesemnificative, permițând utilizarea a 24 de variabile pentru construcția modelelor de clasificare.

Rezultatele obținute din predicțiile efectuate în prima fază au fost esențiale pentru evaluarea succesului modelului de prognosticare. Algoritmul SVM s-a evidențiat ca fiind cel mai performant, înregistrând un procent de succes de peste 61,32%. Chiar și profitul obținut de 95,06 euro, deși nu considerabil, a reprezentat o marjă de profit rezonabilă de 12,51%. De asemenea, notabil este faptul că toate modelele de predicție au generat profit.

În cadrul acestei evaluări, au fost utilizati mai multi algoritmi de învățare automată, printre care:

- Naive Bayes (NB) – pachetul e1071;
- K-nearest neighbors (KNN) – pachetul kknn;
- Random Forest (RF) – pachetul randomForest;
- Support Vector Machines (SVM) – metoda svm din pachetul e1071;
- C5.0 (decision trees) – pachetul C50;
- Xgboost – pachetul xgboost;

- Multinomial Logistic Regression (MLR) – metoda multinom din pachetul nnet;
- Artificial Neural Networks (ANN) – metoda nnet a pachetului nnet.

Tehnici avansate de învățare automată și deep learning, precum rețelele neuronale convoluționale (CNN) și mașinile cu vectori de suport (SVM), au fost aplicate cu succes pentru predicția rezultatelor meciurilor din Major League Baseball (MLB), obținând rate de acuratețe ridicate. Aceste modele demonstrează eficiența tehnicilor moderne de predicție în sporturi diverse [Li21].

În cadrul celei de-a doua faze, s-a decis să se testeze toate combinațiile posibile cu cele 24 de variabile preselectate, având ca obiectiv obținerea celei mai bune rate de succes posibile. Dat fiind numărul mare de variabile, peste 260.000 de combinații ar fi fost necesare pentru testare. Întrucât această abordare nu ar fi fost fezabilă, s-a optat pentru identificarea celor mai semnificative variabile utilizând metoda de selecție a caracteristicilor înapoi "rfe" (the Backwards Feature Selection "rfe" method) din pachetul "caret" al software-ului R. Acest algoritm a evaluat inițial importanța tuturor variabilelor, urmând apoi iterații succesive în care elimina variabilele mai puțin semnificative, păstrând doar cele esențiale în fiecare iterație.

Variabilele identificate drept cele mai semnificative au fost: B365H, B365D, B365A, AVGH, AGVA, HOVA și AGOLSOFR. Din cele 24 de atribute inițiale, s-au păstrat 11, generând un total de 2048 de combinații, această abordare fiind viabilă.

Analiza rezultatelor a evidențiat faptul că testarea diferitelor combinații de variabile a generat rezultate promițătoare. Toate modelele au obținut rate de succes superioare față de modelul inițial, care înregistrase o rată de succes de 61,32%. Algoritmii care au contribuit la obținerea celor mai bune rezultate au fost SVM, RF, Xgboost și RNA. Modelul cel mai performant a rezultat din testarea combinațiilor de 8 variabile, incluzând cele 7 identificate ca fiind cele mai importante, rezultând într-un total de 15 variabile. Analiza și prelucrarea atentă a datelor au permis trageri de concluzii semnificative cu privire la variabilele relevante pentru modelele utilizate.

3.2.2 Modelarea și Evaluarea Prognozelor Fotbalistice [HER18]

În continuare, se vor explora ipotezele diferite pentru proiectarea modelelor și se va evalua performanța acestora în comparație cu tehnici de referință.

Performanța modelelor din articolul "Predicting Football Results Using Machine Learning Techniques" [HER18] se situează la un nivel comparabil cu cele tradiționale și atinge o precizie similară cu modelele utilizate de agențiile de pariuri. Scopul este de a testa diverse modele de învățare automată și de a investiga diferite designuri și ipoteze de model pentru a maximiza capacitatea predictivă a acestora.

Un succes al proiectului ar implica crearea unui model de clasificare, capabil să prezică rezultatul viitor al meciurilor, și a unui model de regresie, capabil să estimeze scorul unui meci viitor, ale căror performanțe predictive să fie comparabile cu diferite metode de referință.

Modelele liniare generalizate reprezintă un set de metode de regresie în care valoarea de ieșire este presupusă a fi o combinație liniară a valorilor de intrare.

Componentele modelului includ:

- Generarea valorilor xG (expected goals) pentru fiecare șut: pentru a estima probabilitatea de a marca un gol pentru fiecare șut.
- Generarea valorilor xG pentru meciuri: pentru a estima numărul de goluri așteptate pe baza șuturilor în timpul meciului și a altor informații despre meci.
- Calculul ELO: pentru a genera cote ELO pentru echipele ofensive și defensive după fiecare meci, actualizându-le constant pentru a fi folosite în modelele de clasificare și regresie.
- Antrenarea modelului de clasificare: pentru a prezice rezultatul unui meci folosind ratingurile ELO ale celor două echipe.
- Antrenarea modelului de regresie: pentru a estima numărul așteptat de goluri pentru fiecare echipă și a prezice rezultatul meciului.

Pentru fiecare componentă a modelului, se vor prezenta alegerile făcute și modul în care valorile sunt obținute.

Au fost implementate diverse modele pentru fiecare componentă, alegând modele precum Gaussian Naive Bayes pentru clasificarea valorilor xG ale șuturilor, Random Forest Regressor pentru estimarea xG pentru meciuri și SVM pentru clasificarea rezultatului final al meciurilor. Modelele au fost optimizate pentru a obține cea mai bună performanță.

În cele din urmă, se va evalua performanța modelului, comparându-l cu modele de referință și observând că utilizarea golurilor așteptate în locul celor reale a generat o performanță predictivă mai bună.

Aceste experimente au arătat că modelele, bazate pe așteptările golurilor și ratingurile ELO, au obținut o performanță comparabilă cu modele complexe utilizate de agențiile de pariuri și au depășit modelele tradiționale, precum modelul Dixon & Coles. Acest lucru sugerează că utilizarea golurilor așteptate ca metrică este o parte esențială a modelării fotbalului pentru o predictivitate mai mare.

S-au explorat o varietate de modele și parametri pentru a optimiza performanța și s-a construit un proces de antrenament și testare pentru a ajusta rapid și eficient

modelul. Totuși, pentru o generalizare mai bună și consolidarea modelului, ar fi necesară utilizarea unui set mai amplu de date.

Prin urmare, modelele au obținut rezultate semnificative, indicând faptul că utilizarea așteptărilor golurilor și ratingurilor ELO poate îmbunătăți semnificativ performanța modelului de prognoză a meciurilor de fotbal.

3.2.3 Framework Hibrid pentru Prognozarea Rezultatelor Sportive [BM08]

Framework-ul propus de articolul "A compound framework for sports results prediction: A football case study" [BM08] constă din două componente esențiale: un raționament bazat pe reguli și o rețea Bayesiană, care colaborează pentru a prezice rezultatele meciurilor sportive.

Framework-uri hibride bazate pe rețele bayesiene și algoritmi genetici au demonstrat o eficiență crescută în optimizarea structurii rețelelor bayesiene pentru predicția rezultatelor sportive. Aceste metode permit modelarea dependențelor complexe dintre variabilele sportive și gestionarea incertitudinii, îmbunătățind astfel precizia predicțiilor [CC19].

Acestă abordare derivă din observația că meciurile sportive sunt extrem de stocastice, dar strategiile unei echipe pot fi aproximate și captate prin reguli logice precise. Borshchev și Mitrovic [BM08] au propus acest cadru hibrid, demonstrând că integrarea acestor două componente permite o predicție mai robustă și mai fiabilă, chiar și în condițiile unor date statistice limitate. Prin integrarea acestor două componente, framework-ul poate furniza predicții fiabile, chiar și în cazul datelor statistice limitate, având capacitatea de a prezice rezultatele între echipe care au avut puține întâlniri anterioare.

Când vine vorba de prognozarea rezultatelor meciurilor sportive, majoritatea cercetărilor anterioare s-au concentrat în general pe un singur factor, de obicei, scorul. În contrast, framework-ul propus poate lua în considerare o gamă largă de factori, inclusiv scorurile curente, moralul, oboseala, abilitățile etc. Prin abordarea bazată pe cunoștințe în serie temporală în joc, framework-ul poate să reflecte fluxurile și refluxurile din cadrul unui meci sportiv, contribuind la sporirea realității și preciziei predicțiilor.

A fost dezvoltat un predictor de rezultate de fotbal numit FRES pe baza acestui framework și s-a demonstrat că oferă predicții raționale și stabile.

Prezicerea rezultatelor meciurilor sportive reprezintă un subiect captivant pentru o varietate de persoane, de la fani până la pariori. Este și o problemă de cercetare interesantă, datorită complexității sale: rezultatul unui meci sportiv depinde de mulți factori, cum ar fi moralul echipei, abilitățile individuale, strategia de an-

trenament etc. Acest articol prezintă un cadru nou pentru prezicerea rezultatelor sportive.

Framework-ul formalizează strategiile sportive prin intermediul unui raționament bazat pe reguli și gestionează incertitudinea cu ajutorul rețelelor Bayesiane. Prin această abordare compusă, putem să ținem cont de faptul că majoritatea rezultatelor sportive sunt extrem de stocastice, dar în același timp, strategiile unei echipe pot fi reprezentate de reguli logice precise. În plus, atunci când prezice rezultatele meciurilor sportive, framework-ul acesta consideră o gamă largă de factori, cum ar fi scorurile actuale, moralul, oboseala, abilitățile etc. Acest demers este motivat de ideea că acuratețea predicției în domeniile de predicție non-triviale poate fi îmbunătățită prin luarea în considerare corectă a mai multor factori care influențează rezultatele.

Din moment ce majoritatea sporturilor sunt caracterizate de fluxuri și refluxuri, cu schimbări constante în cadrul meciurilor, framework-ul lor poate fi văzut ca un simulator pentru un meci sportiv. De asemenea, sistemul creat este stocastic, astfel încât rezultatele din diferite rulări pot varia. Au adoptat o abordare Monte-Carlo pentru evaluarea rezultatelor generale ale sistemului, rulând sistemul de mai multe ori și agregând rezultatele.

Prezicerea meciurilor de fotbal este o provocare fascinantă și dificilă pentru aplicațiile bazate pe cunoștințe, ridicând noi probleme de reprezentare și achiziție. Majoritatea cercetărilor anterioare referitoare la această problemă au tratat meciul individual ca fiind atomic, aplicând tehnici statistice sau de învățare automată pentru a genera prognoze și, prin urmare, nu au abordat aceste probleme. În fiecare cadru al unui meci, se incearca modelarea raționamentului pe care l-ar putea face un antrenor bun.

3.2.4 Prezicerea rezultatelor meciurilor de fotbal cu Rețele Bayesiene: O analiză comparativă [Raz17]

Predicția rezultatelor în fotbal a devenit din ce în ce mai populară în ultimii ani, generând diverse abordări de modele de predicție pentru a evalua factorii determinanți ai victoriei, remizei sau înfrângerii unei echipe. Trei tipuri principale de abordări sunt utilizate în acest context: abordări statistice, abordări de învățare automată și abordări bayesiene. Recent, multe studii s-au concentrat pe utilizarea modelelor bayesiene în predicția rezultatelor meciurilor de fotbal.

Lucrarea [Raz17] propune utilizarea Rețelelor Bayesiene pentru a prezice rezultatele meciurilor de fotbal, luând în considerare victoria echipei gazdă, victoria echipei oaspete și remiza. Scopul este ca rezultatele obținute să servească drept referință pentru cercetările ulterioare în acest domeniu.

Fotbalul este unul dintre cele mai recunoscute și îndrăgite sporturi la nivel mondial. Predicția rezultatelor meciurilor de fotbal a atras atenția atât a managerilor de echipe,

cât și a suporterilor. Complexitatea acestei predicții este determinată de o serie de factori, inclusiv colaborarea în echipă, abilitățile individuale, condițiile meteo și avantajul de a juca acasă. Predicția exactă a rezultatelor meciurilor rămâne o provocare majoră, chiar și pentru experți. 90 de minute de joc pot aduce surprize majore, cum ar fi accidentările sau eliminările prin cartonașe roșii. Norocul poate influența de asemenea rezultatele meciurilor, făcând ca o echipă puternică să nu câștige întotdeauna împotriva unei echipe considerate mai slabă.

Datorită diversității factorilor care influențează meciurile de fotbal, această cercetare utilizează Rețelele Bayesiene, recunoscute pentru aplicabilitatea lor în predicția vremii, a sporturilor și în alte domenii. Abordarea bayesiană propusă folosește atribute furnizate de site-ul http://www.football-data.co.uk, care oferă date istorice despre rezultatele fotbalului englez.

Setul de date utilizat pentru modelare include diferiți factori relevanți pentru predicția meciurilor de fotbal. Rezultatele obținute pentru meciurile din Premier League în sezoanele 2010-2011, 2011-2012 și 2012-2013 arată o acuratețe medie generală de 75,09%, depășind semnificativ procentul de acuratețe anterior, care era de 59,21%. Acuratețea a fost evaluată prin împărțirea setului de date într-un set de antrenament și unul de testare, folosind Clasificatorul Bayesian aplicat pe diferite segmente ale setului de date.

Predicțiile meciurilor de fotbal au captat interesul cercetătorilor, care au încercat să găsească metode mai eficiente. Utilizarea Rețelelor Bayesiene pentru a prezice rezultatele meciurilor de fotbal în Premier League a demonstrat o acuratețe medie de 75,09% în cele trei sezoane analizate. Aceste rezultate pot servi drept referință și sursă de inspirație pentru cercetările viitoare în domeniul prezicerii rezultatelor meciurilor de fotbal.

3.2.5 Compararea Tehnicilor de *Data Mining* în Prezicerea Rezultatelor Fotbalistice: Arborii de Decizie [Ros18]

Acest articol [Ros18] explorează tehnici diferite de *data mining* pentru a prezice rezultatele meciurilor de fotbal, unde obiectivele sunt victoria, remiza sau înfrângerea. Scopul principal al cercetării este identificarea celei mai precise tehnici de minerit de date adaptate naturii datelor din fotbal.

Rezultatele experimentelor comparative indică faptul că arborii de decizie au oferit cea mai mare acuratețe medie de predicție în domeniul prezicerii meciurilor de fotbal, atingând 99,56%.

Există deja numeroase cercetări referitoare la predicția rezultatelor în sport, în special în baschet și fotbal. Majoritatea analizelor se concentrează și prezic folosind o singură tehnică, însă există și cercetări care combină diferite tehnici, unele utili-

zate pentru a îmbunătăți acuratețea predicției. Acest articol își propune să acopere această lacună printr-o analiză comparativă a mai multor tehnici de minerit de date folosind exclusiv date brute, fără a implica elemente care ar putea influența rezultatul, cum ar fi derivatele datelor sau tehnici specifice.

Pentru atingerea acestor obiective, articolul se concentrează pe trei aspecte: prezicerea rezultatelor meciurilor de fotbal folosind date brute despre rezultatele meciurilor anterioare, generarea unor modele pentru aceste rezultate utilizând diverse tehnici de minerit de date și, în cele din urmă, o analiză comparativă pentru a evalua rezultatele acurateței.

S-au folosit diverse tehnici, precum arborii de decizie (MC4), rețelele neurale și K-Nearest Neighbors (KNN), fiecare având specificitățile și metodele sale pentru predicția rezultatelor meciurilor de fotbal. Rezultatele comparative indică faptul că arborii de decizie au cea mai mare acuratețe medie de 99,56%, urmați de rețelele neurale și tehnica KNN cu 96,83% și, respectiv, 77,54%, în timp ce rețelele bayesiene au prezentat cea mai scăzută acuratețe medie de 76,41%.

Concluzia experimentelor a demonstrat că arborele de decizie este cea mai precisă tehnică pentru a prezice rezultatele meciurilor de fotbal, cu o acuratețe de 99,56%, în comparație cu alte tehnici, cum ar fi rețelele neurale, rețelele bayesiene și K-Nearest Neighbors, în ordine. Această cercetare poate fi îmbunătățită prin explorarea și implementarea acelorași tehnici pentru fiecare metodă pentru a crește acuratețea rezultatelor prin includerea altor atribute corelate cu rezultatul meciurilor, precum distanța parcursă de echipa oaspete sau performanța actuală a echipelor, care pot influența rezultatul meciului de fotbal. Validarea predicțiilor ar trebui să implice și date din meciurile în timp real pentru a consolida acuratețea acestora.

Predicția rezultatelor meciurilor ar trebui să devină o practică obișnuită în sporturile noastre naționale, contribuind la o mai bună organizare și performanță a sportului respectiv, oferindu-le managerilor și sportivilor o perspectivă mai clară asupra jocului înainte de începerea acestuia.

3.2.6 Sistemul Dolores: Un Model Hibrid de Evaluare a Echipelor și Predicție a Rezultatelor Meciurilor de Fotbal la Nivel Global [Con19]

Sistemul ”Dolores”: Un model ce anticipează rezultatele meciurilor de fotbal la nivel global. Această abordare integrează un sistem dinamic de evaluare, furnizând măsuri relative ale superiorității între adversari pentru fiecare ligă, extinzând sistemul pi-rating.

Odată ce o echipă se alătură unei noi ligi, primește o evaluare implicită de 0 pentru acea ligă. Evaluarea anterioară este salvată pentru vechea ligă și devine eva-

luarea implicită a echipei în cazul unui eventual retur.

Un model hibrid folosește evaluările rezultate ca intrare pentru a prezice distribuția 1X2 (HDA) (Home Win - Draw - Away Win). Acest sistem reevaluează echipele bazându-se pe discrepanțele observate în goluri pentru fiecare meci.

Rata de învățare λ determină cât de mult noile rezultate influențează evaluările echipei. Cu cât λ este mai mare, cu atât rezultatele recente devin mai influente.

Fiecare echipă are evaluări distincte pentru meciurile de acasă și pentru cele în deplasare.

Evaluările provizorii sunt stabilite în funcție de trei parametri suplimentari:

1. Pragul de formă ϕ: Reprezintă numărul de performanțe consecutive care nu declanșează factorul de formă.

2. Impactul evaluării μ: Diferența de rating utilizată pentru evaluările provizorii în comparație cu cele de fundal.

3. Factorul de diminuare δ: Impactul formei se diminuează cu fiecare perioadă suplimentară de sub/supraperformanță.

Dacă sunt observate performanțe neobișnuite, evaluările provizorii se modifică, altfel rămân egale cu cele de fond.

Parametrii sistemului sunt optimizați pentru acuratețea predictivă printr-o căutare a spațiului hiperparametrilor.

Această metodă se bazează pe rate de învățare $\lambda = 0,054$ și $\gamma = 0,79$ pentru a minimiza eroarea de predicție la 0,211208. În plus, parametrii optimi pentru evaluările provizorii sunt descoperiți la $\delta = 2,5$, $\mu = 0,01$ și $\phi = 1$, reducând eroarea de predicție la 0,211198.

Modelele sunt adaptate pentru ligile de fotbal la nivel global, utilizând rate de învățare puțin mai mari decât cele raportate în versiunea inițială pi-rating, pentru a acoperi lipsa datelor și a obține evaluări mai precise.

Discrepanțele de evaluare sunt convertite în predicții ale meciurilor printr-un model bazat pe diferențele de abilitate, generând probabilități ale distribuției 1X2 în funcție de discrepanțele de rating.

Această abordare oferă o perspectivă valoroasă asupra evaluărilor echipelor și a influenței acestora asupra predicțiilor meciurilor.

Capitolul 4

Aplicație pentru predicția meciurilor de fotbal

4.1 Analiza și Caracterizarea Setului de Date

4.1.1 Introducere în Setul de Date

Setul de date utilizat în această analiză a fost inițial descărcat de pe GitHub, de la adresa:
"https://github.com/dataquestio/project-walkthroughs/blob/master/football_matches/matches.csv"

Acest set de date conținea informații despre meciurile de fotbal din competiția de top a Angliei, si anume Premier League. Pentru a-l adapta nevoilor lucrării de față, am aplicat diverse tehnici de *feature engineering* utilizând Jupyter Notebook. În final, setul de date conține 1390 de rânduri (instanțe) și 35 de coloane (atribute).

Informațiile din acest set de date provin din două sezoane ale Premier League, respectiv sezoanele 2020-2021 și 2021-2022. Fiecare sezon include 20 de echipe participante. În sezonul 2020-2021, fiecare echipă a jucat câte 38 de meciuri, conform formatului obișnuit al competiției, în timp ce în sezonul 2021-2022, fiecare echipă a disputat 34 de meciuri. Setul de date conține variabile relevante pentru analiza performanțelor echipelor, precum numărul de puncte obținute, diferența de puncte față de adversari, locul de desfășurare a meciurilor (acasă sau în deplasare), și alte statistici legate de joc.

Cele 35 de atribute din setul de date sunt:

- Unnamed: 0: Index implicit.
- Index: Index personalizat.
- team: Numele echipei.

- opponent: Numele echipei adversare.
- round: Runda în care s-a desfășurat meciul.
- same city: Indicator dacă echipele joacă în același oraș (1 = Da, 0 = Nu).
- date: Data meciului.
- time: Ora meciului.
- day: Ziua săptămânii în care s-a desfășurat meciul.
- venue: Locul desfășurării meciului (Home/Away).
- attendance: Daca meciul s-a disputat cu portile inchise.(in perioada covid-ului nu a fost permisa prezenta fanilor pe stadion)
- result 1: Al treilea cel mai recent rezultat obtinut de echipa care se afla in dreptul coloanei ”team”.
- result 2: Al doilea cel mai recent rezultat obtinut de echipa care se afla in dreptul coloanei ”team”.
- result 3: Cel mai recent rezultat obtinut de echipa care se afla in dreptul coloanei ”team”.
- points in last 3: Numarul de puncte acumulate in ultimele 3 meciuri de echipa din dreptul coloanei ”team”.
- trend(Negative/Neutral/Positive/Insufficient Data): forma echipei ”team” luand in considerare numarul de puncte acumulat ultimele 3 meciuri.
- win percentage in last 3: procentajul victoriilor obtinute in ultimele 3 meciuri de echipa ”team”.
- draw percentage in last 3: procentajul egalurilor obtinute in ultimele 3 meciuri de echipa ”team”.
- loss percentage in last 3: procentajul infrangerilor obtinute in ultimele 3 meciuri de echipa ”team”.
- points: Numărul de puncte obținute de echipa din dreptul coloanei ”team” pana la momentul meciului respectiv.
- category: Categoria din care face echipa din dreptul coloanei ”team” in functie de numarul de puncte pe care le-a obtinut (din 10 in 10 puncte)

- result 1 opponent: Al treilea cel mai recent rezultat obtinut de echipa care se afla in dreptul coloanei ”opponent”.
- result 2 opponent: Al doilea cel mai recent rezultat obtinut de echipa care se afla in dreptul coloanei ”opponent”.
- result 3 opponent: Cel mai recent rezultat obtinut de echipa care se afla in dreptul coloanei ”opponent”.
- points in last 3 opponent: Numarul de puncte acumulate in ultimele 3 meciuri de echipa din dreptul coloanei ”opponent”.
- trend opponent: forma echipei ”opponent” luand in considerare numarul de puncte acumulat ultimele 3 meciuri.
- win percentage in last 3 opponent: procentajul victoriilor obtinute in ultimele 3 meciuri de echipa ”opponent”.
- draw percentage in last 3 opponent: procentajul egalurilor obtinute in ultimele 3 meciuri de echipa ”opponent”.
- loss percentage in last 3 opponent: procentajul infrangerilor obtinute in ultimele 3 meciuri de echipa ”opponent”.
- points opponent: Numărul de puncte obținute de echipa adversă pana la momentul meciului respectiv.
- category opponent: Categoria din care face echipa adversă in functie de numarul de puncte pe care le-a obtinut.(din 10 in 10 puncte)
- points difference: Diferența de puncte dintre echipa din dreptul coloanei ”team” și adversar.
- category difference: Diferenta de categorie dintre cele doua echipe.
- result: Rezultatul final al meciului (Win/Loss/Draw).
- season: Sezonul în care s-a desfășurat meciul.

Cele 35 de atribute din setul de date pot fi împărțite în două categorii: atributele originale, preluate direct din sursa inițială, și atributele create ulterior, menite să îmbunătățească analiza și predicția rezultatelor meciurilor. Atributele create manual au fost dezvoltate pentru a surprinde mai bine tendințele și dinamica echipelor, oferind informații suplimentare esențiale pentru predicția performanțelor echipelor.

Atributele originale din setul de date sunt următoarele:

- Unnamed: 0: Index implicit.
- Index: Index personalizat.
- team: Numele echipei.
- opponent: Numele echipei adversare.
- round: Runda în care s-a desfășurat meciul.
- date: Data meciului.
- time: Ora meciului.
- day: Ziua săptămânii în care s-a desfășurat meciul.
- venue: Locul desfășurării meciului (Home/Away).
- attendance: Dacă meciul s-a disputat cu porțile închise (în perioada pandemiei COVID-19 nu a fost permisă prezența fanilor pe stadion).
- result: Rezultatul final al meciului (Win/Loss/Draw).
- season: Sezonul în care s-a desfășurat meciul.

Atributele create ulterior au fost adăugate pentru a completa setul de date cu variabile relevante pentru analiza performanțelor echipelor. Aceste atribute au fost gândite pentru a oferi o imagine de ansamblu mai precisă asupra stării echipelor înaintea meciurilor și pentru a captura informații suplimentare utile în predicția rezultatelor.

Atributele create manual includ:

- same city: Indicator dacă echipele joacă în același oraș (1 = Da, 0 = Nu).
- result 1: Al treilea cel mai recent rezultat obținut de echipa care se află în dreptul coloanei ”team”.
- result 2: Al doilea cel mai recent rezultat obținut de echipa care se află în dreptul coloanei ”team”.
- result 3: Cel mai recent rezultat obținut de echipa care se află în dreptul coloanei ”team”.
- points in last 3: Numărul de puncte acumulate în ultimele 3 meciuri de echipa din dreptul coloanei ”team”.

- points: Numărul de puncte obținute de echipa din dreptul coloanei ”team” până la momentul meciului respectiv.
- result 1 opponent: Al treilea cel mai recent rezultat obținut de echipa care se află în dreptul coloanei ”opponent”.
- result 2 opponent: Al doilea cel mai recent rezultat obținut de echipa care se află în dreptul coloanei ”opponent”.
- result 3 opponent: Cel mai recent rezultat obținut de echipa care se află în dreptul coloanei ”opponent”.
- points in last 3 opponent: Numărul de puncte acumulate în ultimele 3 meciuri de echipa din dreptul coloanei ”opponent”.
- points opponent: Numărul de puncte obținute de echipa adversă până la momentul meciului respectiv.
- trend (Negative/Neutral/Positive/Insufficient Data): Forma echipei ”team”, luând în considerare numărul de puncte acumulate în ultimele 3 meciuri.
- win percentage in last 3: Procentajul victoriilor obținute în ultimele 3 meciuri de echipa ”team”.
- draw percentage in last 3: Procentajul egalurilor obținute în ultimele 3 meciuri de echipa ”team”.
- loss percentage in last 3: Procentajul înfrângerilor obținute în ultimele 3 meciuri de echipa ”team”.
- category: Categoria din care face parte echipa ”team”, în funcție de numărul de puncte obținute (din 10 în 10 puncte).
- trend opponent: Forma echipei ”opponent”, similar cu trend-ul echipei ”team”.
- win percentage in last 3 opponent: Procentajul victoriilor obținute în ultimele 3 meciuri de echipa ”opponent”.
- draw percentage in last 3 opponent: Procentajul egalurilor obținute în ultimele 3 meciuri de echipa ”opponent”.
- loss percentage in last 3 opponent: Procentajul înfrângerilor obținute în ultimele 3 meciuri de echipa ”opponent”.
- category opponent: Categoria echipei adverse, în funcție de numărul de puncte.

- points difference: Diferența de puncte dintre echipa din dreptul coloanei "team" și adversar.
- category difference: Diferența de categorie dintre cele două echipe.

Aceste atribute suplimentare au fost introduse deoarece ele oferă o perspectivă mai detaliată asupra formei echipelor și asupra diferențelor semnificative dintre acestea. Atributele referitoare la tendințe și procentaje, de exemplu, sunt extrem de utile pentru a analiza forma recentă a echipelor și impactul acesteia asupra performanței în meciurile viitoare. În plus, variabilele precum "points difference" și "category difference" permit o evaluare mai exactă a echilibrului de forțe dintre echipe, un factor crucial în determinarea rezultatului unui meci.

Setul de date final a fost ajustat pentru a include toate informațiile necesare pentru a dezvolta un model de predicție a rezultatelor meciurilor de fotbal.

4.1.2 Analiza Performanței Echipelor

În această secțiune, vom explora mai profund distribuția punctelor obținute și rezultatele meciurilor din setul nostru de date, analizând diverse aspecte statistice și vizuale.

Caracteristici ale Punctelor Obținute

Vom examina statisticile cheie asociate cu numărul de puncte obținute de către echipe:

Media punctelor: Media numărului de puncte obținute de echipe este de aproximativ 1.39 puncte pe meci.

Deviația standard a punctelor: Deviația standard este de aproximativ 0.79, indicând variabilitatea punctelor obținute între echipe.

Mediana punctelor: Mediana este de 1 punct pe meci.

Valoarea minimă și maximă a punctelor: Numărul minim de puncte obținute de o echipă într-un meci este 0, iar maximul este 3.

Vizualizarea Distribuției Punctelor

Pentru a înțelege mai bine modul în care sunt distribuite punctele între echipe, vom analiza două vizualizări importante:

Diagrama Relației dintre Puncte și Diferența de Puncte: Această diagramă ne va ajuta să evaluăm corelația între punctele obținute și diferența de puncte între echipe.

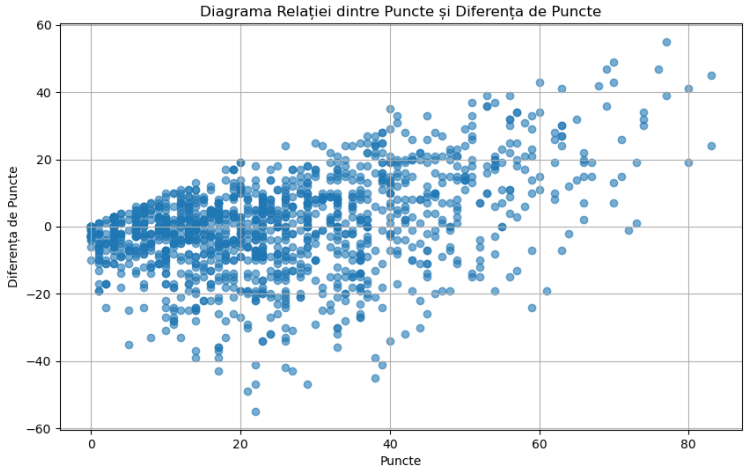

Figura 4.1: Diagrama Relației dintre Puncte și Diferența de Puncte

Figura 4.1 prezintă corelația dintre punctele obținute de fiecare echipă pe parcursul campionatului și diferența de puncte între echipă și adversarul său la momentul respectiv. Pe axa orizontală este reprezentat numărul de puncte acumulate, iar pe axa verticală se află diferența de puncte dintre echipe. O diferență pozitivă indică faptul că echipa respectivă are mai multe puncte decât adversarul său, în timp ce o diferență negativă arată că echipa a acumulat un număr mai mic de puncte comparativ cu adversarul.

După o analiză mai detaliată, se observă că densitatea punctelor este foarte mare în jurul originii graficului, ceea ce se explică prin faptul că toate echipele încep campionatul cu 0 puncte. Astfel, în prima etapă, diferența de puncte este de asemenea 0 pentru toate echipele, ceea ce generează o densitate mare în această zonă. Pe măsură ce echipele acumulează puncte, diferențele între ele variază, iar acest lucru este vizibil în zona de 0-40 de puncte, unde apar atât diferențe pozitive, cât și negative. Acest fapt reflectă competiția dintre echipe, unele fiind mai performante, iar altele rămânând în urmă. Începând de la aproximativ 50 de puncte, majoritatea diferențelor de puncte devin pozitive, indicând faptul că doar echipele de top reușesc să acumuleze un număr mare de puncte, ceea ce le plasează constant deasupra adversarilor din punct de vedere al performanței.

Histograma Distribuției Punctelor: Prin intermediul acestei histograme, vom examina modul în care sunt distribuite punctele în întregul set de date, identificând eventuale tendințe sau modele în performanța echipelor.

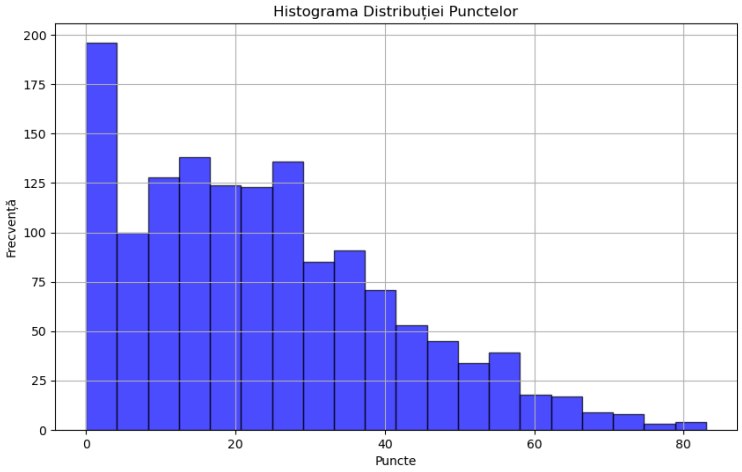

Figura 4.2: Histograma Distribuției Punctelor

Figura 4.2 prezintă distribuția numărului de puncte acumulate de echipe pe parcursul campionatului. Pe axa orizontală sunt reprezentate punctele, iar pe axa verticală, frecvența echipelor care au obținut acel număr de puncte. Cea mai mare frecvență se regăsește în intervalul 0-40 de puncte, deoarece majoritatea echipelor au acumulat acest număr de puncte de-a lungul campionatului, în timp ce doar un număr redus de echipe a reușit să atingă între 50 și 100 de puncte. Distribuția este asimetrică, sugerând că performanțele echipelor au variat semnificativ, iar doar câteva echipe au dominat campionatul.

Prin combinarea informațiilor statistice și vizuale, vom obține o înțelegere cuprinzătoare a performanței echipelor din setul nostru de date, facilitând identificarea aspectelor cheie și a tendințelor relevante pentru analiza noastră.

4.1.3 Analiza Evoluției Performanței Echipelor de Fotbal

În această secțiune, ne propunem să investigăm tendințele generale în evoluția performanței echipelor de fotbal pe parcursul întregului sezon. Deși nu avem statistici pentru fiecare meci, putem utiliza datele disponibile pentru a trasa o imagine a modului în care performanța echipelor a fluctuat pe tot parcursul sezonului.

Am examinat evoluția punctajului și a diferenței de puncte pentru fiecare echipă în funcție de etapele sezonului. Am identificat perioade de creștere sau scădere a performanței și am încercat să determinăm factorii care ar putea influența aceste tendințe.

De asemenea, am investigat relațiile între variabilele de performanță, cum ar fi numărul total de puncte obținute și diferența de puncte dintre echipe, pentru a identifica posibile modele sau corelații între acestea. Rezultatele analizei noastre arată că

unele echipe au avut o evoluție constantă sau ascendentă în timpul sezonului, în timp ce altele au experimentat fluctuații semnificative în performanță.

4.1.4 Concluzii și Direcții Viitoare

În concluzie, analiza noastră a datelor despre performanța echipelor de fotbal din sezoanele 2020-2022 a furnizat o înțelegere detaliată a evoluției și a factorilor care influențează performanța acestora. Rezultatele obținute pot fi utile pentru cluburile de fotbal, antrenorii și analiștii sportivi în luarea deciziilor strategice și îmbunătățirea performanței echipelor.

Pentru direcții viitoare, sugestii ar putea include extinderea analizei către alte variabile relevante sau aplicarea unor tehnici de analiză mai avansate pentru a obține înțelegeri mai profunde despre performanța echipelor de fotbal.

4.2 Dezvoltarea Modelului Bayesian

În această secțiune, vom discuta dezvoltarea unui model Bayesian pentru predicția rezultatelor meciurilor de fotbal, utilizând un set de date care acoperă două sezoane ale Premier League (2020-2021 și parțial 2021-2022). Modelul Bayesian oferă un cadru robust pentru tratarea incertitudinii și gestionarea relațiilor probabilistice dintre variabile.

În cadrul procesului de dezvoltare a modelului Bayesian, a fost testată inițial o variantă de rețea Bayesiană cu un număr mai mare de legături între noduri. Deși acest model părea promițător din punct de vedere al complexității și al detaliilor incluse, performanța sa s-a dovedit a fi mult sub așteptări. Modelul a înregistrat o acuratețe de aproximativ 30% în predicțiile corecte și a necesitat un timp de procesare mult mai mare, ceea ce a afectat semnificativ eficiența sa. (Vezi Figura 4.3)

```
# Definirea modelului ajustat
model = BayesianNetwork([
    ('team', 'result'),
    ('opponent', 'result'),
    ('venue', 'result'),
    ('same city', 'result'),
    ('result 1', 'result'),
    ('result 2', 'result'),
    ('result 3', 'result'),
    ('result 1 opponent', 'result'),
    ('result 2 opponent', 'result'),
    ('result 3 opponent', 'result'),
    ('points', 'category'),
    ('points opponent', 'category opponent'),
    ('category', 'result'),
    ('category opponent', 'result'),
    ('points difference', 'category difference'),
    ('category difference', 'result')
])
```

Figura 4.3: Prima varianta de model

Unul dintre principalele motive pentru această performanță scăzută este legat de lipsa unui volum suficient de mare de date în setul de antrenament. În absența unor date extinse și diversificate, rețeaua nu a putut să învețe corect tiparele necesare pentru anumite predicții. Acest aspect este o provocare majoră în predicția sportivă, mai ales în fotbalul modern, unde variabilitatea și imprevizibilitatea sunt foarte mari. Pentru a atinge o precizie ridicată în predicții, este necesar un volum mult mai mare de date statistice.

Ulterior, am decis să dezvolt modelul treptat, analizând impactul fiecărei legături asupra performanței. Am început prin utilizarea a doar doua legături, pe care le-am considerat ca fiind cele mai importante, între 'category difference' și 'result', și între 'venue' și 'result'. (Vezi Figura 4.4) Acest model inițial a obținut un procent de 53.15% predicții corecte, însă a reușit să prezică doar victoriile și înfrângerile. Acest lucru se datorează faptului că modelul se bazează pe inferență, iar probabilitatea unui rezultat de egalitate este foarte mică în comparație cu victoriile sau înfrângerile.

```
model = BayesianNetwork([('category difference', 'result'),('venue', 'result')])
```

Figura 4.4: Varianta modelului care prezice doar victoria sau infrangerea

Cu toate acestea, am decis să extind modelul pentru a include și predicțiile de egalitate, după ce am identificat tipare specifice care să maximizeze precizia predicțiilor

de egalitate. Deși modelul final a obținut un scor puțin mai mic, consider că este mai important să pot face predicții pentru toate stările posibile din coloana 'result', oferind astfel o imagine mai completă asupra rezultatului meciurilor.

4.2.1 Structura Modelului Bayesian

Modelul Bayesian creat este structurat pentru a include variabile esențiale care pot influența rezultatul unui meci de fotbal. Aceste variabile includ:

- **venue (V)**: Locul unde se desfășoară meciul (acasă sau în deplasare).
- **category difference (CD)**: Diferenta de categorie dintre cele doua echipe.
- **draw percentage in last 3 (DT)**: Procentajul de egaluri din ultimele 3 meciuri ale echipei "team".
- **draw percentage in last 3 opponent (DO)**: Procentajul de egaluri din ultimele 3 meciuri ale echipei "opponent".

Modelul utilizează aceste variabile pentru a prezice variabila țintă result, care reprezintă rezultatul meciului (victorie, înfrângere sau egalitate). (Vezi Figura 4.5)

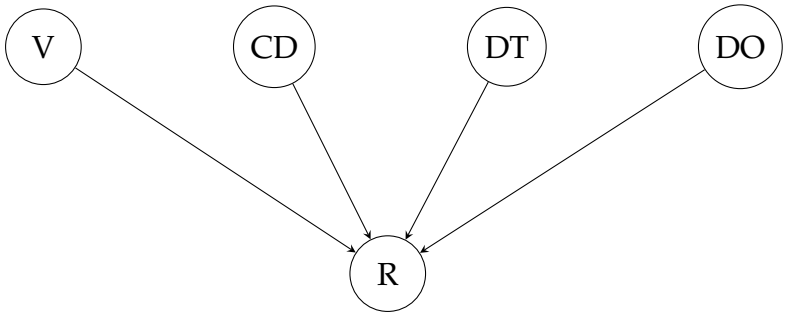

Figura 4.5: Rețeaua bayesiană utilizată

4.2.2 Procesul de Antrenament

Setul de date a fost împărțit în două subseturi: 90% pentru antrenament și 10% pentru testare. Această împărțire asigură că modelul poate învăța dintr-o cantitate semnificativă de date și poate fi apoi evaluat pe un set separat pentru a verifica performanța.

Modelul Bayesian a fost antrenat utilizând estimatorul MaximumLikelihoodEstimator, care permite estimarea parametrilor rețelei pe baza datelor de antrenament. Codul pentru antrenarea modelului este prezentat în Figura 4.6:

```
model = BayesianNetwork([('category difference', 'result'), ('venue', 'result'),
                         ('draw percentage in last 3', 'result'), ('draw percentage in last 3 opponent', 'result')])

# Estimarea parametrilor modelului
model.fit(df_new, estimator=MaximumLikelihoodEstimator)

# Verificarea modelului
assert model.check_model(), "Modelul nu este valid!"
```

Figura 4.6: Codul pentru antrenarea modelului

4.2.3 Evaluarea Modelului

Evaluarea modelului a fost realizată pe setul de date de testare, utilizând inferența variabilelor pentru a prezice rezultatul meciurilor. Codul pentru evaluarea modelului este prezentat în Figura 4.7::

```
# Inferența variabilelor
inference = VariableElimination(model)

test_data = pd.read_csv("test_data_9_AUGUST_CORECTAT.csv")

# Definirea funcției de predicție prin eșantionare
1 usage
def predict_with_inference(data_row):
    evidence = data_row.to_dict()
    evidence_inference = {key: value if key in model.nodes() else value for key, value in evidence.items() if
                          key in model.nodes() and key != 'result'}
    result = inference.map_query(variables=['result'], evidence=evidence_inference)
    return result['result']


# Aplicarea funcției de predicție pentru fiecare rând din setul de date
test_data['Predicted_result'] = test_data.apply(predict_with_inference, axis=1)

# Compararea predicțiilor cu valorile reale
correct_predictions = sum(test_data['result'] == test_data['Predicted_result'])
total_predictions = len(test_data)
accuracy = correct_predictions / total_predictions
print(f"Accuracy: {accuracy * 100:.2f}%")

# Salvarea modelului antrenat
import pickle

with open('trained_model_bayesian_network_9_AUGUST.pkl', 'wb') as f:
    pickle.dump(model, f)
```

Figura 4.7: Codul pentru evaluarea si salvarea modelului

4.2.4 Rezultate

Modelul Bayesian a obținut un scor de acuratețe de aproximativ 50.35% pe setul de date de testare. Aceasta sugerează că, deși modelul are capacitatea de a face inferențe bazate pe variabilele disponibile, există loc pentru îmbunătățiri suplimentare.

4.3 Dezvoltarea Aplicației Web pentru Predicții Fotbalistice

4.3.1 Introducerea aplicației

Aplicația web dezvoltată permite utilizatorilor să introducă variabile referitoare la echipele implicate în meciuri de fotbal și să obțină o predicție a rezultatului meciului (victorie, egalitate sau înfrângere) utilizând un model Bayesian antrenat anterior. Aceasta rulează pe un server local și este construită cu ajutorul framework-ului Flask în Python.

4.3.2 Funcționalități

Aplicația are următoarele funcționalități principale:

- **Introducerea datelor:** Utilizatorii pot introduce date referitoare la echipele implicate în meci, incluzând numele echipelor, locul de desfășurare a meciului (acasă/deplasare), diferența de categorie, și procentajele câștigurilor, egalităților și înfrângerilor din ultimele 3 meciuri.
- **Validarea datelor:** Aplicația verifică corectitudinea datelor introduse. De exemplu, „category difference" trebuie să fie un număr între -5 și 5, iar câmpurile referitoare la procentele ultimelor meciuri trebuie să conțină doar valori prestabilite (0.0, 33.33, 50.0, 66.66, 100.0).
- **Calculul rezultatului:** După validarea datelor, aplicația apelează modelul Bayesian pentru a prezice rezultatul meciului (victorie, egalitate sau înfrângere). Acest rezultat este afișat în josul paginii cu un mesaj de tip „Predicted Result: W/D/L", unde W este pentru victorie, D pentru egalitate, și L pentru înfrângere.

4.3.3 Tehnologii utilizate

Aplicația a fost construită utilizând următoarele tehnologii:

- **Python & Flask:** Folosite pentru a gestiona partea de backend și pentru a încărca și utiliza modelul Bayesian.
- **HTML:** Utilizat pentru structura paginii web, oferind o interfață simplă și ușor de utilizat.
- **CSS:** Folosit pentru a stiliza pagina și pentru a asigura o dispunere adecvată a elementelor.

- **JavaScript:** Utilizat pentru a prelucra datele introduse de utilizatori și pentru a gestiona afișarea rezultatelor și validarea inputurilor.

4.3.4 Integrarea modelului Bayesian

Modelul Bayesian este încărcat dintr-un fișier pickle salvat anterior, care conține modelul antrenat. Codul care gestionează încărcarea [Figura 4.8] și utilizarea modelului pentru predicție este scris în Python și este integrat în fișierul *"app.py"* al aplicației.

```
8   # Încărcarea modelului antrenat
9   with open('trained_model_bayesian_network_9_AUGUST.pkl', 'rb') as file:
10      model = pickle.load(file)
```

Figura 4.8: Codul pentru încărcarea modelului

```
16  # Definirea funcției de predicție
    1 usage
17  def predict_result(data_row):
18      evidence = {key: value for key, value in data_row.items() if key in model.nodes() and key != 'result'}
19      result = inference.map_query(variables=['result'], evidence=evidence)
20      return result['result']
```

Figura 4.9: Codul pentru predicție

Funcția *"predict_result"* [Figura 4.9] preia datele de la utilizator și le folosește ca „evidence"(variabilele de care tine cont pentru a realiza predictiile) în cadrul modelului Bayesian, utilizând *"inference.map_query"* pentru a calcula predicția finală. Rezultatul este apoi returnat și afișat în interfața utilizatorului.

4.3.5 Interfața utilizatorului

Pagina web este simplă și oferă un formular prin care utilizatorul poate introduce datele necesare. După ce toate datele sunt introduse corect și validate, utilizatorul poate apăsa butonul *"Predict Result"* pentru a obține rezultatul. O previzualizare a interfeței utilizatorului se poate observa în Figura 4.10.

Validările input-urilor:

- **Category Difference:** Valoarea trebuie să fie între -5 și 5.
- **Procentaje:** Doar valorile prestabilite (0.0, 33.33, 50.0, 66.66, 100.0) sunt acceptate pentru câmpurile referitoare la procentajele de câștig, egalitate și înfrângere. (Vezi Figura 4.11)

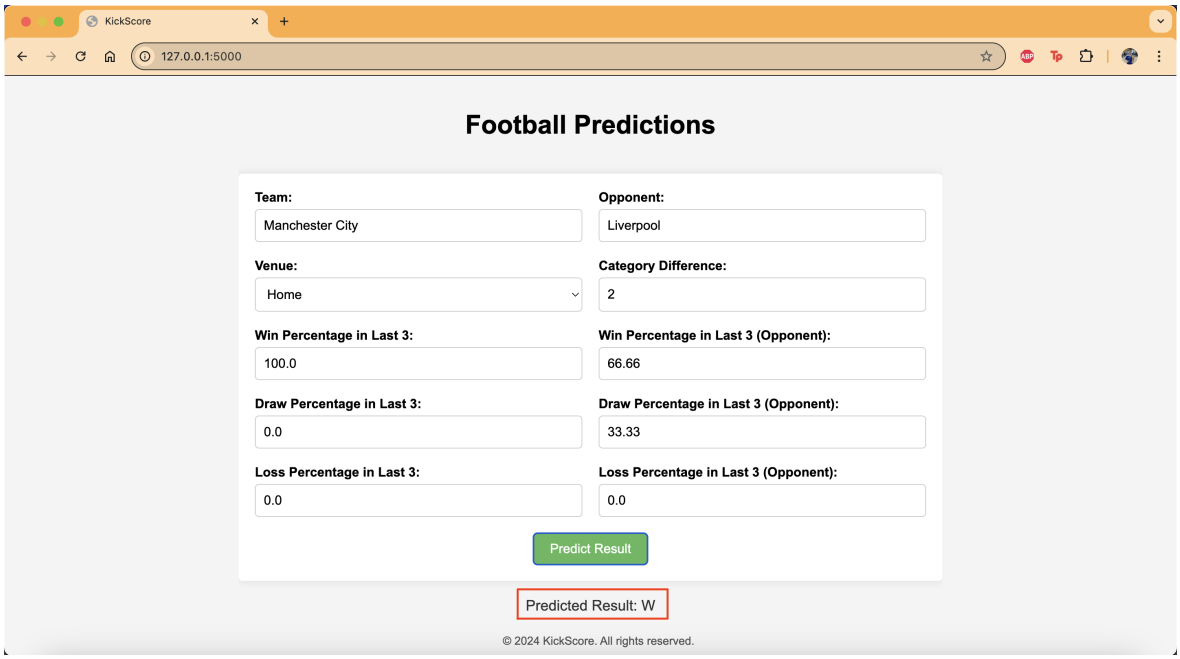

Figura 4.10: Interfața utilizatorului

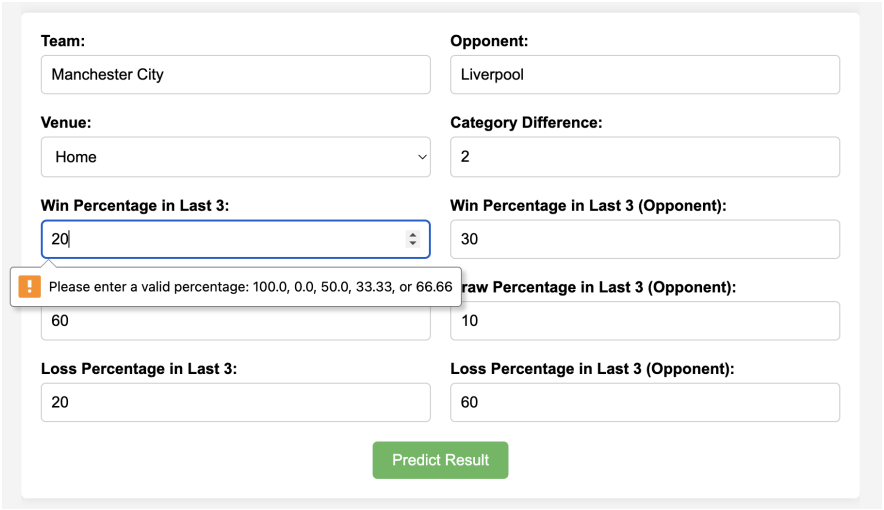

Figura 4.11: Procentaje invalide

4.3.6 Afișarea rezultatului

După ce modelul returnează o predicție, aceasta este afișată sub formular, în formatul *"Predicted Result: W/D/L"*. De exemplu, dacă modelul prezice o victorie, mesajul afișat va fi *"Predicted Result: W"*.

4.3.7 Testare și îmbunătățiri

Aplicația a fost testată pe un set de date diferite, iar validările implementate asi-

gură că utilizatorul introduce date corecte și complete. În continuare, pot fi adăugate îmbunătățiri pentru a permite predicții mai complexe sau pentru a include alte variabile care ar putea influența rezultatul unui meci de fotbal.

4.4 Concluzii și Direcții Viitoare

Modelul Bayesian dezvoltat pentru predicția rezultatelor meciurilor de fotbal demonstrează aplicabilitatea acestor rețele în gestionarea incertitudinii și modelarea relațiilor probabilistice. Cu toate acestea, performanța modelului indică faptul că există factori suplimentari care ar putea fi incluși pentru a îmbunătăți acuratețea predicțiilor.

Direcțiile viitoare de cercetare și dezvoltare includ:

- explorarea altor variabile care pot influența rezultatul meciurilor;

- îmbunătățirea metodei de antrenament prin utilizarea unui set mai mare de date;

- combinarea rețelelor Bayesiene cu alte tehnici de învățare automată pentru a îmbunătăți performanța.

Un aspect important al acestui proiect a fost încercarea de a construi un model mai complex, cu un număr mai mare de legături între noduri în rețeaua Bayesiană. Cu toate acestea, rezultatele obținute au indicat o performanță scăzută, cu doar 30% din predicții fiind corecte, în mare parte din cauza insuficienței datelor din setul de antrenament. În concluzie, se observă că pentru a îmbunătăți semnificativ acuratețea predicțiilor sportive, este necesară o cantitate mult mai mare de date statistice.

Aplicația web dezvoltată pentru predicția meciurilor de fotbal s-a dovedit a fi funcțională și ușor de utilizat, însă există mai multe direcții de modernizare și extindere. O direcție viitoare ar putea fi îmbunătățirea interfeței utilizatorului (UI) pentru a oferi o experiență mai intuitivă și mai plăcută. De asemenea, pot fi adăugate funcționalități suplimentare, cum ar fi includerea unor statistici detaliate despre echipe și jucători, și validări avansate pentru a preveni introducerea de date incorecte. Scalabilitatea aplicației poate fi, de asemenea, extinsă prin implementarea acesteia pe un server public și prin utilizarea unei baze de date care să stocheze datele relevante pentru a face predicții mai precise. În concluzie, aplicația oferă o bază solidă pentru predicția rezultatelor meciurilor de fotbal, iar îmbunătățirile propuse vor contribui la creșterea performanței și a utilizabilității acesteia.

Capitolul 5

Concluzii

Această lucrare a explorat utilizarea rețelelor bayesiene pentru predicția rezultatelor meciurilor de fotbal, evidențiind versatilitatea și eficiența acestui instrument în modelarea relațiilor probabilistice și gestionarea incertitudinii. Studiul a acoperit atât aspectele teoretice fundamentale ale rețelelor bayesiene, cât și aplicabilitatea lor practică în domeniul sportiv.

Printre principalele contribuții ale lucrării se numără:

- **Oferirea unui cadru teoretic solid** pentru înțelegerea și utilizarea rețelelor bayesiene, incluzând structura și funcționarea acestora, conceptele cheie și metodele de inferență.
- **Analiza detaliată a variabilelor relevante** pentru predicția rezultatelor meciurilor de fotbal, precum și evaluarea diferitelor abordări și tehnici utilizate în acest domeniu.
- **Dezvoltarea și evaluarea unei aplicații practice** bazate pe rețele bayesiene, demonstrând eficiența și acuratețea acesteia în predicția rezultatelor meciurilor de fotbal.

Avantajele rețelelor bayesiene, cum ar fi flexibilitatea, interpretabilitatea și eficiența computațională, au fost evidențiate pe parcursul lucrării. Totuși, au fost discutate și provocările asociate cu utilizarea acestora, cum ar fi complexitatea calculului și necesitatea unor cantități mari de date.

Direcțiile viitoare de cercetare includ explorarea combinării rețelelor bayesiene cu alte tehnici de învățare automată pentru a îmbunătăți acuratețea și robustetea predicțiilor, extinderea aplicabilității rețelelor bayesiene în diverse domenii și dezvoltarea de metode eficiente pentru construcția și calibrarea rețelelor complexe.

În concluzie, rețelele bayesiene reprezintă un instrument valoros pentru modelarea relațiilor probabilistice și gestionarea incertitudinii, oferind un cadru robust și

flexibil pentru o varietate de aplicații. Contribuțiile teoretice și practice discutate în această lucrare oferă o bază solidă pentru cercetările viitoare și aplicarea rețelelor bayesiene în diverse domenii.

Bibliografie

[BG07] Irad Ben-Gal. Bayesian networks. In Fabrizio Ruggeri, Ron Kenett, and Frederick Faltin, editors, *Encyclopedia of Statistics in Quality and Reliability*. Wiley, Chichester, UK, 2007.

[BK02] Christian Borgelt and Rudolf Kruse. *Graphical Models: Methods for Data Analysis and Mining*. John Wiley & Sons, 2002.

[BM08] Chongyoun Choe Hyeonsang Eom R.I. (Bob) McKay Byungho Min, Jinhyuck Kim. A compound framework for sports results prediction: A football case study. *Knowledge-Based Systems*, 21:551–562, 2008. Accessed: 2024-01-07.

[Bor08] ME Borsuk. Ecological informatics: Bayesian networks. In Sven Erik Jørgensen and Brian Fath, editors, *Encyclopedia of Ecology*. Elsevier, 2008.

[CC19] Fatemeh Vafaee Peter C. Nelson Carlo Contaldi. Bayesian network hybrid learning using an elite-guided genetic algorithm. *Artificial Intelligence Review*, 52:245–272, 2019. Accessed: 2024-08-16.

[CD03] J. W. Comley and D. L. Dowe. General bayesian networks and asymmetric languages. In *Proceedings of the 2nd Hawaii International Conference on Statistics and Related Fields*, June 2003.

[CD05] J. W. Comley and D. L. Dowe. Minimum message length and generalized bayesian nets with asymmetric languages. In P. D. Grünwald, I. J. Myung, and M. A. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*, Neural information processing series, pages 265–294. Bradford Books (MIT Press), Cambridge, Massachusetts, 2005. This paper puts decision trees in internal nodes of Bayes networks using Minimum Message Length (MML).

[CGH97] E. Castillo, J. M. Gutiérrez, and A. S. Hadi. Learning bayesian networks. In *Expert Systems and Probabilistic Network Models*, Monographs in computer science, pages 481–528. Springer-Verlag, New York, 1997.

[Con19] Anthony C. Constantinou. Dolores: a model that predicts football match outcomes from all over the world. *Knowledge-Based Systems*, 108:49–75, 2019.

[Dar09] A. Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, 2009.

[Dow11] David L. Dowe. Hybrid bayesian network graphical models, statistical consistency, invariance and uniqueness. In *Philosophy of Statistics*, pages 901–982. Elsevier, 2011. Published on 2011-05-31.

[FN04] N. Fenton and M. E. Neil. Combining evidence in risk analysis using bayesian networks. *Safety Critical Systems Club Newsletter*, 13(4):8–13, July 23 2004.

[FN07] N. Fenton and M. E. Neil. Managing risk in the modern world: Applications of bayesian networks. Technical report, London Mathematical Society and the Knowledge Transfer Network for Industrial Mathematics, London, England, November 2007. A Knowledge Transfer Report.

[GCSR03] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, 2003. Part II: Fundamentals of Bayesian Data Analysis: Ch.5 Hierarchical models.

[HER18] Corentin HERBINET. Predicting football results using machine learning techniques. *INDIVIDUAL PROJECT REPORT*, 2018. Accessed: 2024-01-07.

[Li21] Mei-Ling Huang Yun-Zhi Li. Use of machine learning and deep learning to predict the outcomes of major league baseball matches. *Applied Sciences*, 11, 2021. Accessed: 2024-08-16.

[PD23] Livia De Giovanni Vincenzina Vitale Pierpalo D'Urso. A bayesian network to analyse basketball players' performances: a multivariate copula-based approach. *Annals of Operations Research*, 325:419–440, 2023. Accessed: 2024-08-16.

[PKJ21] Waris Quamer Rajendra Pamula Praphula Kumar Jain. Sports result prediction using data mining techniques in comparison with base line model. *OPSEARCH*, 58:54–70, 2021. Accessed: 2024-08-16.

[Raz17] Nazim Razali. Predicting football matches results using bayesian networks for english premier league (epl). *IOP Conference Series: Materials Science and Engineering*, 2017. Accessed: 2024-01-07.

[Ros18] Che Mohamad Firdaus Che Mohd Rosli. A comparative study of data mining techniques on football match prediction. *Journal of Physics: Conference Series*, 2018. Accessed: 2024-01-07.

[R22] Fátima Rodrigues and Ângelo Pinto. Prediction of football match results with machine learning. *Procedia Computer Science*, 204:463–470, 2022. Accessed: 2024-01-07.