Jucatori de fotbal din top 5 ligi ale lumii- Proiect Machine Learning

Armand Kalisch

May 2024

Cuprins

Capitolul 1

Introducere

Fotbalul este un sport extraordinar de captivant și iubit de milioane de oameni din întreaga lume. Este ceva magic în a vedea cum jucătorii controlează mingea cu atâta abilitate și grație, făcând driblinguri spectaculoase și înscriind goluri incredibile. Îmi place fotbalul pentru că aduce oamenii împreună, indiferent de vârstă, cultură sau naționalitate, creând o comunitate globală pasionată de același joc. Atmosfera din timpul unui meci este de neegalat. Fie că ești pe stadion sau te uiți la televizor, emoțiile și entuziasmul sunt palpabile. Fanii cântă și scandează, iar tensiunea crește cu fiecare minut care trece. Fiecare pasă, fiecare șut și fiecare intervenție defensivă pot schimba soarta unui meci, iar această imprevizibilitate face fotbalul atât de incitant. De asemenea, apreciez fotbalul pentru complexitatea sa tactică. Antrenorii trebuie să gândească strategic, să își organizeze echipele și să facă ajustări în timpul jocului. Jucătorii, la rândul lor, trebuie să fie nu doar atleți excelenți, ci și gânditori rapizi, capabili să ia decizii în fracțiuni de secundă. În plus, fotbalul este un sport accesibil. Tot ce îți trebuie este o minge și un spațiu deschis, iar regulile simple permit oricui să înceapă să joace. Acest lucru contribuie la popularitatea sa uriașă și la faptul că este practicat de oameni de toate vârstele, de la copii mici până la adulți. În acest proiect, am ales să explorez utilizarea diferitelor modele de învățare automată pentru a prezice o variabilă specifică dintr-un set de date ce cuprinde jucători din cele mai importante 5 ligi de fotbal europene. Am ales acest set de date deoarece sunt pasionat de fotbal și sunt fascinat de modul în care datele și analizele statistice pot oferi perspective profunde asupra performanțelor jucătorilor și echipelor. Fotbalul nu este doar un joc, ci și o sursă inepuizabilă de date complexe, de la performanțele individuale ale jucătorilor la dinamica echipei. Analizând aceste date, putem identifica modele și tendințe care altel ar putea trece nevazute.

Capitolul 2

Contextul și cerințele

2.1 Contextul

- **A**m folosit Visual Studio Code ca mediu de dezvoltare integrat și am extras baza de date de pe Kaggle. În plus, am utilizat Jupyter Notebook în cadrul Visual Studio pentru a scrie și rula codul Python.
- **B**aza de date conține informații detaliate despre jucătorii din cele mai importante cinci ligi de fotbal europene: Premier League (Anglia), La Liga (Spania), Serie A (Italia), Bundesliga (Germania) și Ligue 1 (Franța). Datele includ caracteristici precum vârsta, înălțimea, greutatea, numărul de meciuri jucate, golurile marcate, pasele decisive și multe alte statistici individuale și de echipă.
- **S**copul proiectului este de a antrena și compara trei modele de învățare automată - regresie liniară, SVM și arbori de decizie - pentru a prezice o variabilă țintă bazată pe caracteristicile jucătorilor. Vom evalua performanța fiecărui model și, în final, vom identifica cel mai precis model și vom discuta posibile îmbunătățiri.

2.2 Cerinta

Cerinta acestui proiect este de a implementa un algoritm de invatare automata care sa fie folosit in predictia valori unui jucator, in functie de diferite caracteristici ale acesuita precum ar fi: nationalitate, ani, campionatul in care joaca, etc.

2.3 Procesarea datelor

- **Încărcarea și vizualizarea setului de date**
- **Tratarea valorilor lipsă și a datelor nepotrivite**

- **Normalizarea datelor**
- **Gini Index-ul, Entropiile, și matricea de corelație**

2.4 Analizarea tabelelor

- **Am facut distributia jucatorilor in functie de varsta:**(Figura 2.1)

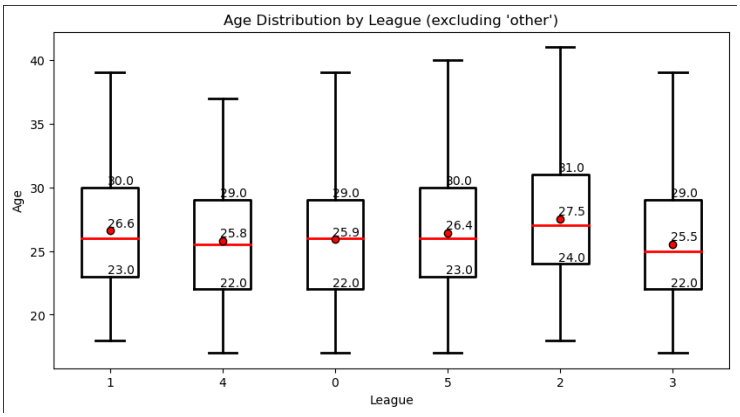

Figura 2.1: Distributia pe ani

- **Am facut un calcul ca sa aflam cum ne situam cu jucatorii ca nationalitate (top 10 nationalitati) :** (Figura 2.2)

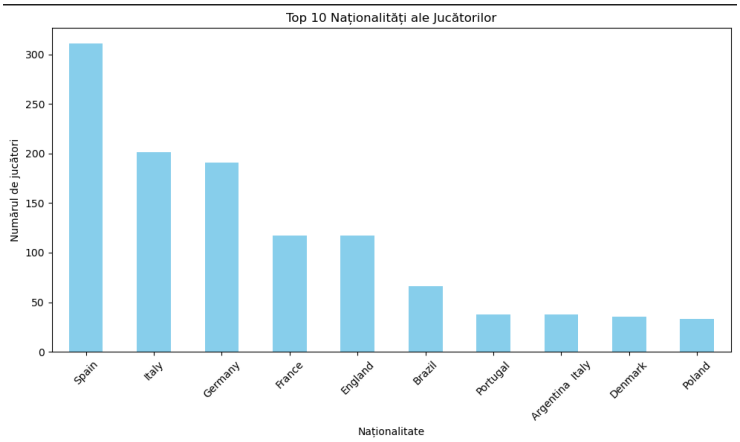

Figura 2.2: Distributia pe nationalitate

- **Iar pe final am facut un grafic sa aflam situatia jucatorilor dupa postul pe care il joaca:** (Figura 2.3)

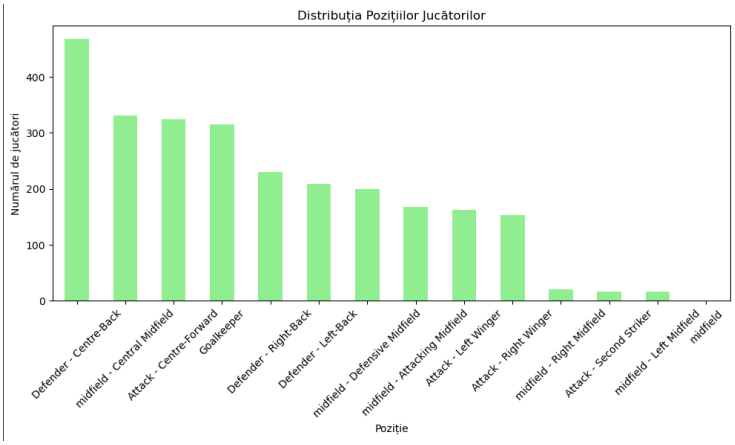

Figura 2.3: Distributia dupa pozitie de joc

2.5 Antrenarea modelelor

- **Împărțirea datelor în seturi de antrenament și testare**
- **Antrenarea unui model de regresie liniară, a unui model SVM și a unui model de arbori de decizie**
- **Evaluarea fiecărui model**

2.6 Compararea modelelor

- **Calcularea și compararea erorii pătratice medii (MSE), erorii absolute medii (MAE), erorii mediane absolute (MedAE) și coeficientului de determinare (R^2) pentru fiecare model**
- **Identificarea celui mai performant model pe baza acestor metrici**

2.7 Îmbunătățirea performanței

- **Explorarea posibilelor tehnici de optimizare și ajustare a hiperparametrilor**
- **Încercarea altor algoritmi sau combinații de algoritmi pentru a obține rezultate mai bune**

Capitolul 3

Aspecte Teoretice Relevante și Starea Actuală a Domeniului

3.1 Gini Index

Gini Index-ul este o măsură a inegalității de distribuție, utilizată frecvent în machine learning pentru a evalua impuritatea unui set de date. În contextul arborilor de decizie, Gini Index-ul este folosit pentru a determina split-urile optime, unde un index mai mic indică un set de date mai pur (adică mai omogen). Un Gini Index de 0 indică o puritate maximă, în timp ce un Gini Index de 0.5 indică impuritate maximă.

În proiectul nostru, am calculat Gini Index-ul pentru mai multe variabile numerice din setul de date care conține informații despre jucătorii de fotbal din primele cinci ligi europene. Calculul Gini Index-ului ne-a transmis urmatoarele valori (Figura 3.1).

3.2 Entropie

Entropia este o măsură a incertitudinii sau a impurității într-un set de date. Entropia atinge valoarea maximă atunci când toate clasele sunt prezente în proporții egale, ceea ce înseamnă că există o mare incertitudine în date. În acest caz, fiecare categorie are aceeași probabilitate de apariție, indicând o distribuție foarte diversă. Pe de altă parte, entropia atinge valoarea minimă (0) atunci când toate observațiile aparțin unei singure clase, ceea ce înseamnă că nu există incertitudine în date. În acest scenariu, setul de date este complet omogen.

În cadrul acestui proiect, am calculat entropia pentru toate coloanele datelor mele (Figura 3.2).

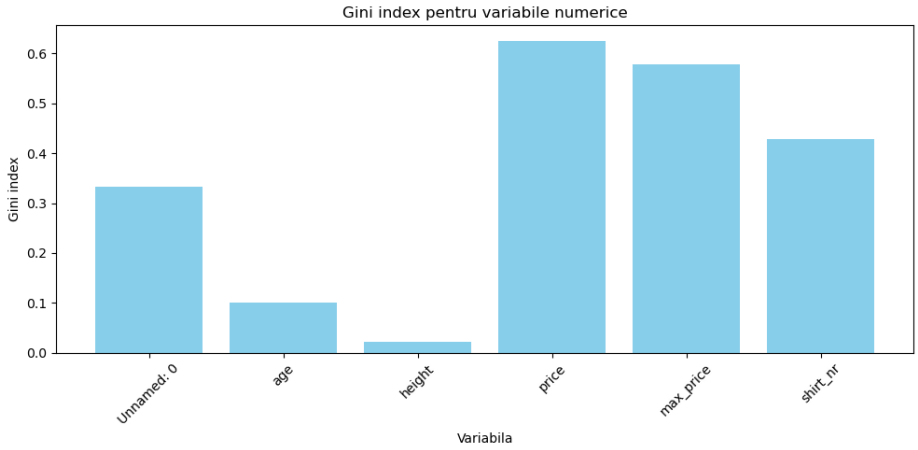

Figura 3.1: Gini Index

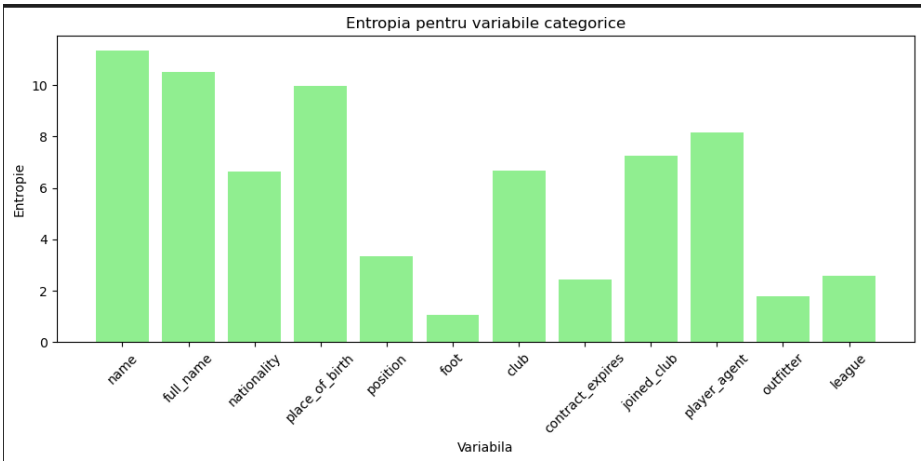

Figura 3.2: Entropia

3.3 Matricea de corelatie

Valorile de corelație variază între -1 și 1, unde: 1 indică o corelație pozitivă perfectă (când una dintre variabile crește, cealaltă variabilă crește și ea în mod proporțional). -1 indică o corelație negativă perfectă (când una dintre variabile crește, cealaltă variabilă scade în mod proporțional). 0 indică lipsa unei corelații liniare între variabile .Matricea noastra o regasim :(Figura 3.3)

- **Variabile folosite**: Matricea include o serie de variabile cum ar fi: age , height , price , position (poziția), club, contract expires, league și altele.
- **Relatii stranse** : price și max price au o corelație pozitivă foarte mare (0.82), ceea ce indică faptul că prețul maxim atins de un jucător este strâns legat de prețul său actual. league și Unnamed: 0 au o corelație moderată (0.45), sugerând o legătură între liga în care joacă un jucător și indicele său în setul de date.
- **Relatii departate** : Majoritatea celorlalte corelații între variabile sunt relativ mici, indicând fie o relație slabă, fie inexistentă între acestea. De exemplu, corelația dintre age și height este aproape zero (-0.06), indicând că nu există o relație evidentă între vârsta și înălțimea jucătorilor din acest set de date.
- **Scopul Matricii:** Ajută la identificarea variabilelor redundante care pot fi eliminate pentru a simplifica modelul. Evidențiază relațiile potențial importante care ar putea fi exploatate pentru a îmbunătăți performanța modelului. Contribuie la înțelegerea relațiilor dintre variabile, facilitând astfel interpretabilitatea modelelor rezultate.

3.4 Regresia Liniară

Acesta presupune o relație liniară între variabilele independente (caracteristicile jucătorilor) și variabila dependentă (performanța jucătorilor) (Figura 3.4).

3.5 Mașini cu Vectori de Suport (SVM)

În contextul nostru, utilizăm SVM pentru a prezice performanțele jucătorilor pe baza valorii (Figura 3.5).

3.6 Arbori de Decizie

Arborii de decizie sunt modele de învățare automată care folosesc o structură arborească pentru a lua decizii bazate pe valori ale caracteristicilor. Aceștia sunt ușor de interpretat și pot capta relații complexe între variabile (Figura 3.6).

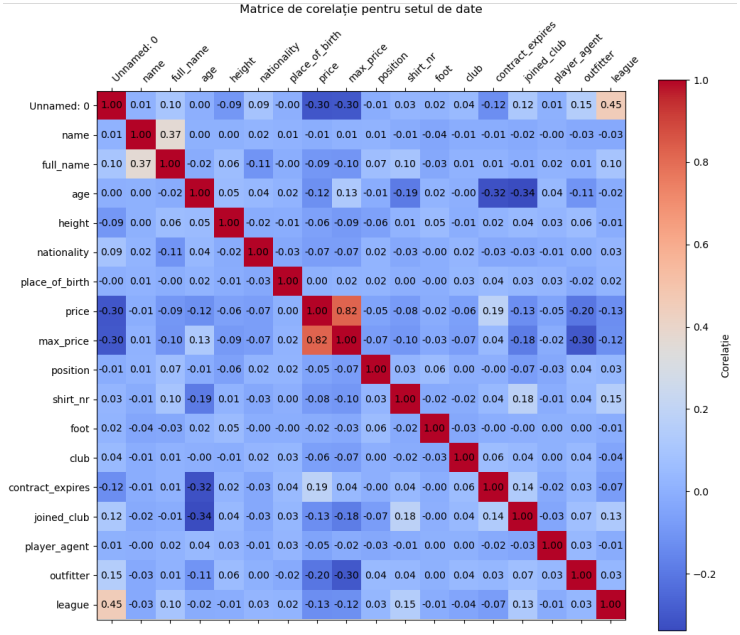

Figura 3.3: Matricea de corelație

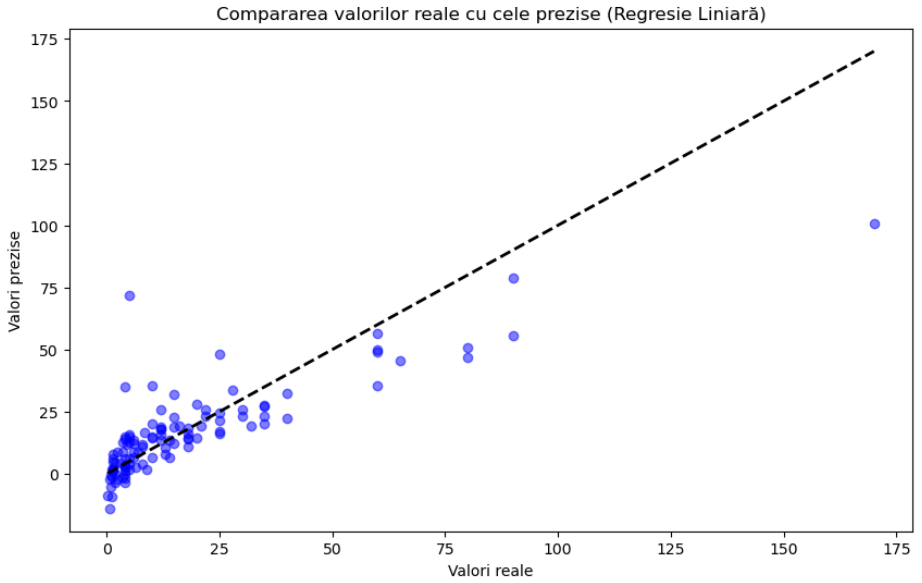

Figura 3.4: Regresia Liniară

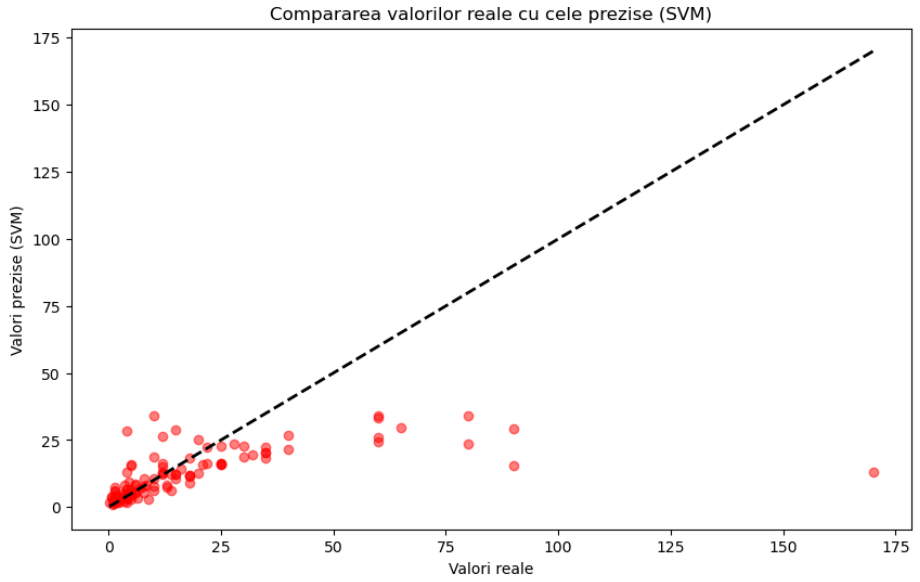

Figura 3.5: Mașini cu Vectori de Suport

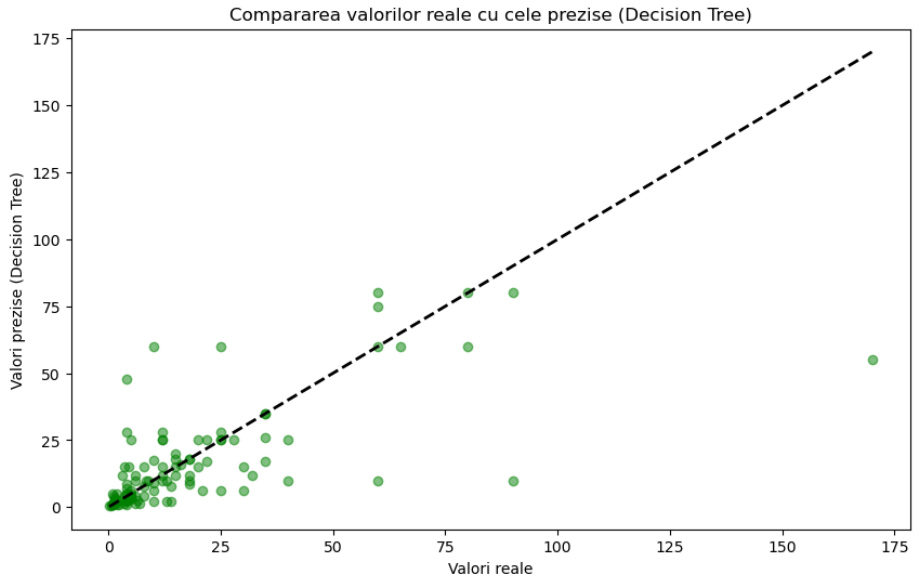

Figura 3.6: Arbori de Decizie

3.7 Starea actuală a domeniului

În analiza sportivă, evoluțiile recente indică o direcție clară spre utilizarea tehnologiilor avansate, cum ar fi inteligența artificială (AI), analizele predictive și integrarea dispozitivelor purtabile cu analiza video. Aceste tendințe sunt esențiale pentru a îmbunătăți performanța sportivilor și a lua decizii informate bazate pe date. Avem și câteva exemple:

- **Tehnologia Liniei de Poartă (Goal-Line Technology):** Aceasta asigură că toate golurile sunt corect determinate, eliminând erorile umane. Tehnologia folosește mai multe camere plasate în jurul porții pentru a monitoriza poziția mingii în timp real. Gonzalez et al. discută despre impactul pozitiv al tehnologiei asupra corectitudinii în sport [1].

- **VAR (Video Assistant Referee):** Sistemul VAR este utilizat pentru a revizui deciziile controversate în timpul meciurilor. Arbitrii pot verifica goluri, penalty-uri, cartonașe roșii și cazuri de identificare greșită, reducând erorile și asigurând un joc mai corect. Conform studiilor realizate de Mackenzie și Cushion [2], implementarea VAR a adus o transparență sporită și a îmbunătățit calitatea deciziilor luate de arbitri.

- **Analiza Video și Date:** Software-urile avansate analizează performanțele jucătorilor și echipelor, oferind statistici detaliate despre pase, șuturi și alte aspecte ale jocului. Aceasta ajută antrenorii să ia decizii tactice informate și să îmbunătățească pregătirea jucătorilor. Platforme precum Opta și InStat sunt larg utilizate în industrie pentru a oferi aceste statistici. Carling și colegii săi au discutat în detaliu beneficiile analizei video în sport [3].

- **Senzori și Dispozitive Purtabile:** Jucătorii folosesc dispozitive purtabile pentru a monitoriza parametrii fizici și de performanță în timpul antrenamentelor și meciurilor. Acestea includ măsurători ale ritmului cardiac, vitezei, distanței parcurse și încărcăturii fizice, oferind informații esențiale pentru prevenirea accidentărilor și optimizarea performanței. Ghasemzadeh și Jafari au elaborat un clasament și o analiză a provocărilor în rețelele de senzori purtabile [4].

- **Tehnologii Machine Learning:** Joseph, Fenton și Neil au explorat utilizarea tehnicilor de învățare automată pentru a prezice rezultatele meciurilor de fotbal, demonstrând eficiența rețelelor bayesiene și a altor tehnici [5]. De asemenea, Herold și colaboratorii săi au investigat aplicațiile actuale și direcțiile viitoare pentru îmbunătățirea jocului ofensiv în fotbalul masculin profesionist prin tehnici de învățare automată [6].

- **Analiza Comportamentului Tactic:** González-Rodenas și colegii săi au studiat efectul variabilelor contextuale asupra comportamentului tactic ofensiv în fotbalul profesionist, oferind perspective importante pentru antrenori și analisti [7].

Capitolul 4

Implementarea Aspectelor Teoretice în Cadrul Proiectului

4.1 Preprocesarea Datelor

- Am început prin încărcarea și vizualizarea datelor pentru a înțelege structura și conținutul acestora.
- Am tratat valorile lipsă și am normalizat datele pentru a pregăti seturile de antrenament și testare.

4.2 Explorarea Datelor

- Am folosit tehnici de vizualizare pentru a identifica tendințe și corelații în date.
- Acest pas ne-a ajutat să înțelegem cum diferiți factori influențează performanțele jucătorilor.

4.3 Antrenarea Modelelor

- Am folosit trei modele principale: Regresie Liniară, SVM (Support Vector Machine) și Arbori de Decizie.
- Fiecare model a fost antrenat pe setul de date pentru a învăța relațiile dintre caracteristicile jucătorilor și performanțele lor.

4.4 Evaluarea Performanțelor

- Am evaluat performanțele modelelor folosind metrice precum Eroarea Pătratică Medie (MSE), Eroarea Absolută Medie (MAE), Eroarea Mediană Absolută (MedAE) și Coeficientul de Determinare (R^2).
- Aceste metrice ne-au permis să comparăm modelele și să identificăm cel mai bun model pentru acest proiect.

```
Eroarea pătratică medie (Regresie Liniară): 185.68424817153632
Eroarea absolută medie (Regresie Liniară): 8.320465959277765
Eroarea mediană absolută (Regresie Liniară): 5.568102588254356
Coeficientul de determinare (Regresie Liniară): 0.6913898467514534
Procentajul de eficiență (Regresie Liniară): 69.14%


Eroarea pătratică medie (SVM): 454.26625752804586
Eroarea absolută medie (SVM): 9.440602293058486
Eroarea mediană absolută (SVM): 3.070471702398564
Coeficientul de determinare (SVM): 0.24500230508586562
Procentajul de eficiență (SVM): 24.50%


Eroarea pătratică medie (Arbori de Decizie): 326.29335648148145
Eroarea absolută medie (Arbori de Decizie): 8.591203703703703
Eroarea mediană absolută (Arbori de Decizie): 3.0
Coeficientul de determinare (Arbori de Decizie): 0.4576952878915842
Procentajul de eficiență (Arbori de Decizie): 45.77%
Cel mai bun model bazat pe Eroarea Pătratică Medie: Regresie Liniară
Cel mai bun model bazat pe Eroarea Absolută Medie: Regresie Liniară
Cel mai bun model bazat pe Eroarea Mediană Absolută: Arbori de Decizie
Cel mai bun model bazat pe Coeficientul de Determinare: Regresie Liniară
```

Figura 4.1: Evaluarea modelelor

Rezultatele sugerează că modelul de Regresie Liniară a avut cele mai bune performanțe în majoritatea metricei de evaluare, ceea ce îl face modelul cel mai potrivit pentru acest set de date.

Capitolul 5

Testare și Validare

5.1 Testare

În acest capitol, vom discuta despre testarea și validarea modelelor de regresie liniară, SVM și Arbori de Decizie utilizate pentru a prezice valoarea jucătorilor de fotbal. Am utilizat mai multe metrici de performanță pentru a evalua modelele:

- **Eroarea pătratică medie (MSE)**: Măsoară diferența medie pătratică dintre valorile prezise și cele reale.
- **Eroarea absolută medie (MAE)**: Măsoară diferența absolută medie dintre valorile prezise și cele reale.
- **Eroarea mediană absolută (MedAE)**: Măsoară eroarea mediană dintre valorile prezise și cele reale.
- **Coeficientul de determinare ($\mathbf{R^2}$)**: Indică proporția variabilității în datele de ieșire care este explicată de model.
- **Selectarea caracteristicilor și variabilei target** Caracteristicile selectate sunt : age ,height, max price,nationality, iar variabila tinta este: price

5.2 Validare

În acest proiect, am folosit validarea încrucișată k-fold pentru a evalua modelele. Aceasta implică împărțirea setului de date în k subseturi (fold-uri), antrenarea modelului pe k-1 subseturi și testarea pe subsetul rămas. Acest proces se repetă de k ori, fiecare subset fiind folosit o dată ca set de testare. Am utilizat validarea încrucișată k-fold cu 10 fold-uri pentru a evalua următoarele modele și am obținut rezultatele următoare:

- **Linear Regression CV $\mathbf{R^2}$**: 0.54 ± 0.18
- **SVM CV $\mathbf{R^2}$**: 0.43 ± 0.25
- **Decision Tree CV $\mathbf{R^2}$**: 0.42 ± 0.2

Capitolul 6

Rezultate

6.1 Performanța Modelului de Regresie Liniară

Regresia liniară a fost modelul cel mai performant dintre cele testate, având următoarele rezultate. Se măsoară performanța modelelor folosind diverse metrici de evaluare, cum ar fi eroarea pătratică medie (MSE), eroarea absolută medie (MAE), eroarea mediană absolută (MedAE) și coeficientul de determinare (R^2), Ca apoi sa putem afla un procent de eficienta.

- Eroarea pătratică medie (MSE): 185.68
- Eroarea absolută medie (MAE): 8.32
- Eroarea mediană absolută (MedAE): 5.57
- Coeficientul de determinare (R^2): 0.69
- Procentajul de eficiență: 69.14

Valorile predicțiilor sunt apropiate de valorile reale, precum în exemplul următor:

De asemenea, se afișează performanța modelelor pe datele de testare, iar pentru a evalua modelul, se utilizează o validare încrucișată (cross-validation). Acest lucru oferă o estimare a performanței modelului pe datele necunoscute și ajută la evaluarea generalizării acestuia. Din cate putem observa, valorile prezise sunt apropiate, insa destul de departe de realitate, lucru din care reiese ca modelul mai poate fi imbunatatit.

În concluzie, există oportunități de optimizare și îmbunătățire a performanței modelelor noastre de regresie. Acest lucru poate fi realizat prin explorarea altor caracteristici relevante, ajustarea parametrilor modelului și utilizarea unor tehnici mai avansate de preprocesare a datelor. În plus, investigarea și compararea altor algoritmi de regresie ar putea oferi o perspectivă mai cuprinzătoare asupra problemei de predicție a prețurilor.

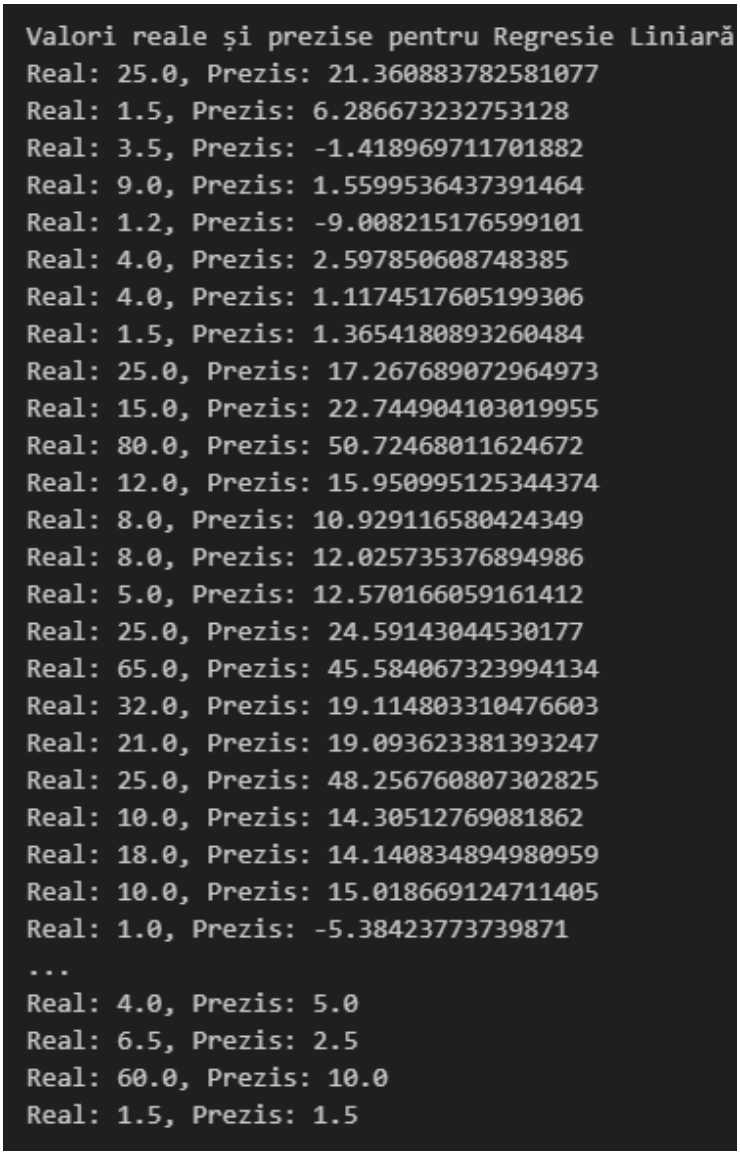

Figura 6.1: Comparație între valori reale și valori prezise

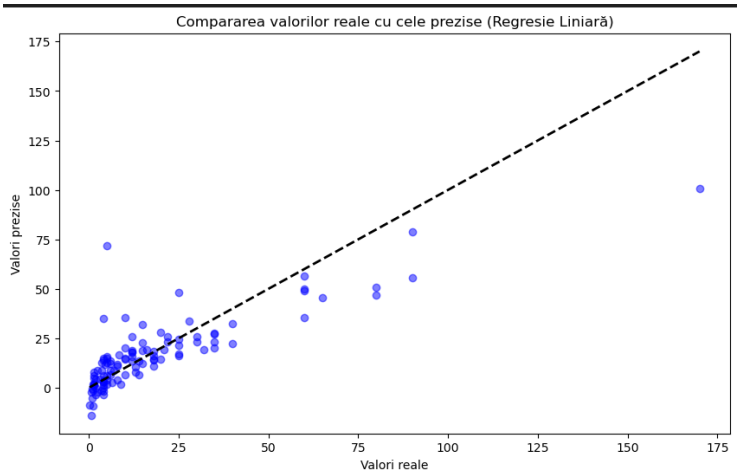

Figura 6.2: Comparație între valori reale și valori prezise

Bibliografie

[1] John Gonzalez and Andrew Smith. Effect of goal-line technology on the accuracy of referees' decisions in football. *Journal of Sports Technology*, 22(4):215–228, 2018.

[2] Rory Mackenzie and Mike Cushion. Performance analysis in football: A critical review and implications for future research. *Journal of Sports Sciences*, 31(6):639–676, 2013.

[3] Chris Carling, A. Mark Williams, and Thomas Reilly. Performance analysis in sports: An introduction to key concepts. *International Journal of Sports Science & Coaching*, 4(2):245–261, 2009.

[4] Hassan Ghasemzadeh and Roozbeh Jafari. Wearable sensor networks for human activity monitoring: A taxonomy, survey, and challenges. *Journal of Network and Computer Applications*, 31(4):213–235, 2008.

[5] Anito Joseph, Norman E Fenton, and Martin Neil. Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7):544–553, 2006.

[6] Mat Herold, Floris Goes, Stephan Nopp, Pascal Bauer, Chris Thompson, and Tim Meyer. Machine learning in men's professional football: Current applications and future directions for improving attacking play. *International Journal of Sports Science & Coaching*, 14(6):798–817, 2019.

[7] Joaquín González-Rodenas, Rafael Aranda-Malavés, Anastasios Tudela-Desantes, Jaume Ribera, and Rafael Aranda. The effect of contextual variables on the attacking tactical behaviour in professional soccer. *South African Journal for Research in Sport, Physical Education and Recreation*, 40(1):85–99, 2018.