# Predicția Aglomerării în Sala de Fitness folosind Modele de Învățare Automată

Crișan Antonel Gabriel

Mai 2024

# Cuprins

# Capitolul 1

# Introducere

## 1.1 Motivația alegerii bazei de date

În ultima perioadă, fiecare dintre noi a devenit mai preocupat de propria sănătate și stare de bine generală. Sportul a devenit un element important în viețile noastre și mai ales în rândul studenților, care doresc să trăiască mai sănătos și să-și reducă nivelul de cortizol acumulat pe tot parcursul zilei din cauza cursurilor, laboratoarelor și seminarilor. [5]Activitatea fizică are beneficii dovedite pentru bunăstarea fizică și psihologică și este asociată cu o reacție redusă la stres.

Fitness este un termen folosit pentru a acoperi o arie largă de activități fizice care cuprind antrenarea musculaturii, stretching-ul și exercițiile cardio, ce au ca scop menținerea organismului cât mai sănătos și tonifiat.

Datele sunt esențiale pentru luarea deciziilor informate în diverse domenii. Sala de fitness reprezintă un loc unde monitorizarea și optimizarea fluxului de persoane poate aduce beneficii atât pentru administratori, cât și pentru utilizatori. Baza de date aleasă conține informații relevante despre numărul de persoane prezente în sală, ziua săptămânii, dacă este weekend, dacă este vacanță, temperatura, dacă este început de semestru, dacă este în timpul semestrului, luna și ora. Aceste date permit aplicarea metodelor de învățare automată pentru a realiza predicții precise. Motivația principală pentru alegerea acestei baze de date este posibilitatea de a îmbunătăți gestionarea resurselor și a spațiilor în sălile de fitness din campus, oferind o experiență mai bună pentru studenți.

Avansurile tehnologice recente în domeniul inteligenței artificiale și al învățării automate ne permit să analizăm și să prezicem comportamentele pe baza unor seturi de date complexe. Acest proiect își propune să utilizeze aceste tehnologii pentru a optimiza gestionarea sălilor de fitness. Prin analiza datelor istorice, putem dezvolta modele predictive care să îmbunătățească programarea activităților și să maximizeze utilizarea eficientă a resurselor.

Obiectivele specifice ale acestui proiect includ optimizarea fluxului de persoane în sala de fitness, îmbunătățirea experienței utilizatorilor prin programarea mai eficientă a antrenamentelor și asigurarea unei gestionări mai bune a resurselor. Predicțiile precise ajută administratorii să ia decizii informate, cum ar fi ajustarea orarului pentru a evita aglomerările și pentru a asigura o distribuție echilibrată a utilizării sălii de fitness.

Pe termen lung, proiectul are potențialul de a servi ca bază pentru dezvoltarea unor soluții mai sofisticate care să includă analize predictive avansate și optimizări continue. Aceste îmbunătățiri pot contribui la creșterea satisfacției utilizatorilor și la menținerea unui mediu de antrenament eficient și plăcut.

# Capitolul 2

# Contextul bazei de date și al proiectului

## 2.1 Cerințe

Proiectul necesită o analiză detaliată a datelor disponibile pentru a dezvolta un model de predicție a aglomerării în sala de fitness din campus. Cerințele specifice includ:

- Colectarea datelor relevante despre numărul de studenți prezenți, condițiile meteorologice, timpul din zi și alți factori influenți.
- Preprocesarea datelor pentru a elimina anomaliile și a asigura integritatea acestora.
- Aplicarea unor modele de învățare automată pentru a realiza predicții precise.
- Evaluarea performanței modelelor prin metrici de evaluare adecvate.

## 2.2 Obiective

Scopul principal al proiectului este de a dezvolta un model de învățare automată capabil să prezică numărul de studenți prezenți în sala de fitness în diferite condiții. Obiectivele specifice includ:

- Explorarea și înțelegerea bazei de date.
- Implementarea și compararea mai multor modele de învățare automată, cum ar fi regresia liniară, rețelele neurale și random forest.
- Analiza performanței modelelor și selectarea celui mai bun model pentru predicții viitoare.

# Capitolul 3

# Aspecte teoretice relevante

## 3.1 Starea actuală a domeniului

Învățarea automată a devenit un instrument esențial în analiza datelor și predicția comportamentului uman în diverse contexte. În contextul sălilor de fitness, monitorizarea și predicția aglomerării pot ajuta la îmbunătățirea experienței utilizatorilor și la optimizarea resurselor.

### 3.1.1 Modele de Învățare Automată

În acest proiect, am explorat diferite modele de învățare automată, inclusiv regresia liniară, rețelele neurale artificiale (ANN) și pădurile aleatorii (Random Forest). Fiecare dintre aceste modele au avantaje și dezavantaje specifice.

### 3.1.2 Entropia și Indexul Gini

Entropia și indexul Gini sunt două măsuri de impuritate utilizate pentru a evalua calitatea unei împărțiri în arborii de decizie.

- **Entropia** este o măsură a incertitudinii sau a impurității într-un set de date. [6] Utilizarea conceptului de entropie în scopul cuantificării conținutului de informații într-o bază de date și dezvoltă o măsură de vulnerabilității bazată pe entropie. În figura 3.1 este reprezentată formula pentru calcularea entropiei.

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{3.1}$$

  unde $p_i$ este proporția elementelor din clasa $i$ în setul $S$.

- **Indexul Gini** [2] este o altă măsură a impurității, definită ca probabilitatea ca un element ales aleator să fie clasificat greșit dacă ar fi etichetat conform distribuției clasei în subset. În figura 3.2 este reprezentată formula pentru calcularea indexului gini.

$$G(S) = 1 - \sum_{i=1}^{c} p_i^2 \tag{3.2}$$

  unde $p_i$ este proporția elementelor din clasa $i$ în setul $S$.

Aceste măsuri sunt utilizate pentru a decide care caracteristică să fie aleasă la fiecare pas al construirii unui arbore de decizie, astfel încât impuritatea să fie minimizată.

### 3.1.3 Regresia liniară

Regresia liniară este o metodă statistică ce modelează relația dintre una sau mai multe variabile independente și o variabilă dependentă folosind o linie de regresie. Este utilizată pentru predicții simple și interpretări ușor de înțeles. [4] Poate unul dintre cei mai obișnuiți și mai cuprinzători algoritmi statistici și de învățare automată.

### 3.1.4 Rețelele Neurale Artificiale (ANN)

[7] Rețelele neurale artificiale sunt modele de învățare automată inspirate de structura și funcționarea creierului uman. Acestea sunt compuse din neuroni artificiali organizați în straturi și sunt capabile să învețe relații complexe dintre date.

### 3.1.5 Random Forest

[1]Random Forest este un algoritm de învățare automată robust și flexibil, care combină multipli arbori de decizie pentru a obține o predicție precisă. Este cunoscut pentru capacitatea sa de a gestiona seturi de date mari și complexe și de a evita supraînvățarea.

# Capitolul 4

# Implementarea aspectelor teoretice în cadrul proiectului

## 4.1 Preprocesarea Datelor

Datele au fost preprocesate pentru a elimina orice informație irelevantă sau redundantă. Au fost eliminate coloanele de timestamp și alte variabile care nu contribuie direct la predicția numărului de persoane.

## 4.2 Utilizarea Entropiei și Indexului Gini

Pentru a evalua dezordinea datelor, am calculat entropia și indexul Gini pentru variabilele noastre. Aceste măsuri ne ajută să înțelegem distribuția datelor și să luăm decizii informate în procesul de modelare.

### 4.2.1 Entropia datelor

Pentru a implementa calcularea entropiei datelor, am calculat frecvența valorii în setul de date, după care am calculat entropia cu ajutorul formulei enunțate în capitoul anterior.

### 4.2.2 Indexul Gini

Pentru a afla indexul gini am separat caracteristicile și variabila de răspuns, adică într-o variabilă $x$ excludem primele două coloane(datetime și numărul de persoane în sală), iar această variabilă stocarea caracteristicilor. În variabila $y$ stocăm prima coloana(numărul de persoane) ca și variabilă de răspuns. Inițiem un model de regresie cu arbore de decizie după care antrenăm modelul pentru a calcula importanța caracteristicilor.

## 4.3 Algoritmi de Învățare Automată

Pentru implementare, am utilizat Regresia Liniară, acest algoritm este utilizat pentru predicții simple și interpretări ușor de înțeles, Rețele Neurale Artificiale, acest algoritm este un model de învățare automată inspirată de structura și funcționarea creierului uman și Random Forest Regressor, cunoscut pentru performanța sa în predicțiile pe seturi de date mari și complexe.

### 4.3.1 Regresia liniară

Pentru implementare, am separat caracteristicile și variabila de răspuns, adică într-o variabilă $x$ excludem primele două coloane(datetime și numărul de persoane în sală), iar această variabilă stocarea caracteristicilor. În variabila $y$ stocăm prima coloana(numărul de persoane) ca și variabilă de răspuns. Am divizat datele în set de antrenament și set de testare(80%-20%). Am inițialiat și antrenat modelul de regresie liniară. Am evaluat modelul pe setul de testare, predicțiile făcute de modelul de regresie liniară, am calculat eroare absolută medie și eroarea medie patratică.

### 4.3.2 Retele Neurale Artificiale

Modul de implementare este asemănător cu a-l Regresiei Liniare, avem nevoie de două variabile, una pentru caracteristici și cealaltă pentru variabila de răspuns. Divizarea datelor s-a facut la fel, adică 80% set de antrenament și 20% se de testare. Am inițializat și antrenat modelul și bineînțeles am avaluat scorul pe setul de testare, iar apoi am calculat eroare medie absolută și eroare medie pătratică.

### 4.3.3 Random Forest Regressor

Pentru modelul Random Forest Regressor, povestea este aceeași la fel ca și in celelalte două modele prezentate mai sus, putem spune ca modul de implementare este asemănător pentru orice model de învățare automată, diferența o face performanța și timpul de lucru.

# Capitolul 5

# Testare și validare

## 5.1 Procedura de Testare și Validare

Pentru a evalua performanța modelelor, am folosit o procedură de validare încrucișată cu divizarea setului de date în seturi de antrenament și testare. Am împărțit datele în proporție de 80% pentru antrenament și 20% pentru testare, folosind funcția `train_test_split` din `scikit-learn`.

## 5.2 Seturile de Date

Setul de date utilizat pentru antrenare și testare a fost format din date despre numărul de persoane prezente într-o sală de fitness, împreună cu caracteristici precum ziua săptămânii, dacă este weekend, dacă este vacanță, temperatura în Fahrenheit, dacă este început de semestru, dacă este în timpul semestrului, luna și ora.

## 5.3 Metrici de Evaluare

Pentru a evalua performanța modelelor, am folosit următoarele metrici:

- **Scorul de determinare** ($R^2$): Măsoară proporția variabilității din variabila de răspuns explicată de model.
- **Eroarea Absolută Medie** (MAE): Măsoară media absolută a erorilor dintre valorile prezise și valorile reale.
- **Eroarea Medie Pătratică** (MSE): Măsoară media pătratelor erorilor dintre valorile prezise și valorile reale.

## 5.4 Metode de Validare

Am folosit validarea încrucișată pentru a asigura că modelele nu suferă de overfitting și pentru a evalua robustețea acestora. Validarea încrucișată presupune împărțirea setului de date în mai multe subseturi, antrenarea modelului pe un subset și testarea pe celălalt, repetând procesul pentru toate combinațiile posibile.

# Capitolul 6

# Rezultate

## 6.1 Performanța Modelelor

Am obținut următoarele rezultate pentru fiecare model:

- **Regresie Liniară**:
  - **Timpul de executie**: 0.7s
  - **Scorul de determinare** ($R^2$): 0.5146
  - **Eroarea Absolută Medie** (MAE): 12.0921
  - **Eroarea Medie Pătratică** (MSE): 250.8425

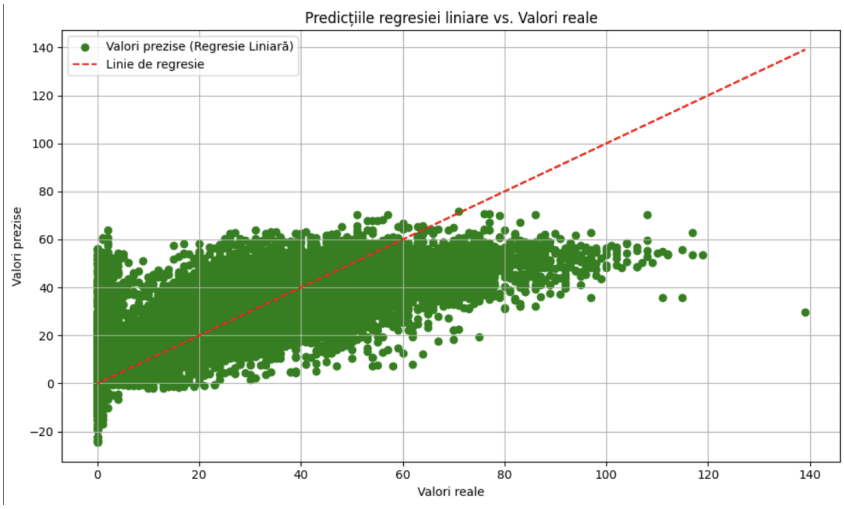


Figura 6.1: Predicțiile regresiei liniare vs. Valori reale

- **Rețele Neurale Artificiale**:
  - **Timpul de executie**: 2m 3.0s
  - **Scorul de determinare** ($R^2$): 0.6826
  - **Eroarea Absolută Medie** (MAE): 9.3169
  - **Eroarea Medie Pătratică** (MSE): 164.0282

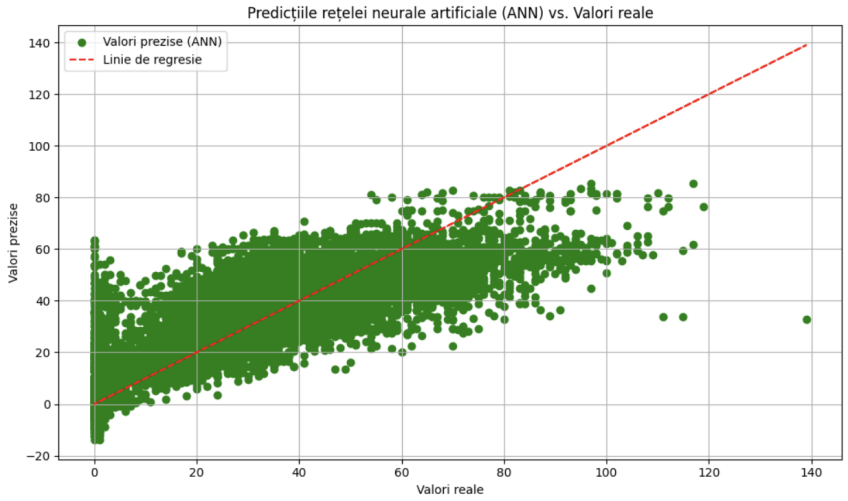


Figura 6.2: Predicțiile rețelelor neurale artificiale vs. Valori reale

- **Random Forest**:
  - **Timpul de executie**: 6.8s
  - **Scorul de determinare** ($R^2$): 0.9247
  - **Eroarea Absolută Medie** (MAE): 4.24
  - **Eroarea Medie Pătratică** (MSE): 38.92

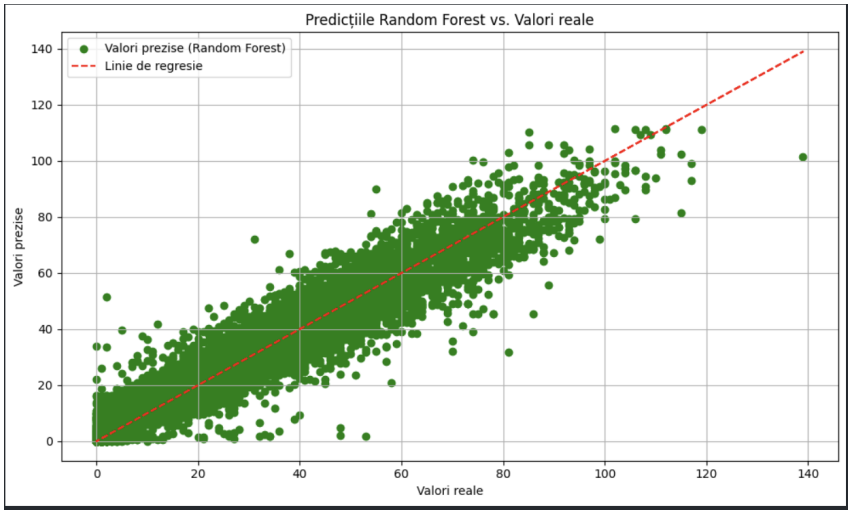


Figura 6.3: Predicțiile Random Forest vs. Valori reale

## 6.2 Entropie și Index Gini

În cadrul modelului Random Forest, am utilizat entropia și indexul Gini pentru a evalua impuritatea nodurilor din arborii de decizie. Aceste măsuri ne-au ajutat să alegem cele mai bune împărțiri la fiecare nod, minimizând astfel impuritatea și maximizând precizia modelului.

- **Entropie**:
  - Măsoară impuritatea sau dezordinea din date. O valoare mai mică a entropiei indică un nod mai pur. Figura 6.4 prezintă distribuția entropiei pentru datele utilizate în acest studiu.

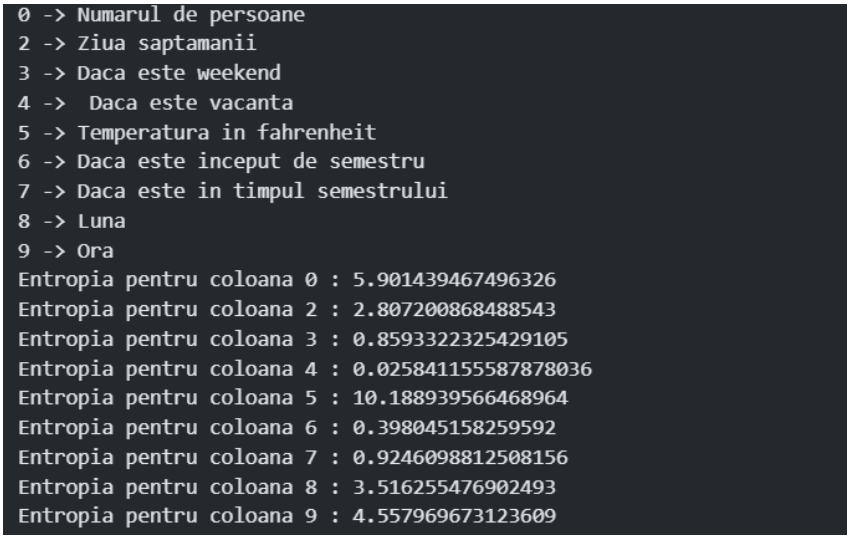


Figura 6.4: Entropia datelor

- **Index Gini**:
  - Măsoară probabilitatea ca un element ales aleator să fie clasificat incorect. Un index Gini mai mic indică un nod mai pur. Figura 6.5 prezintă distribuția indexului Gini pentru datele utilizate în acest studiu.

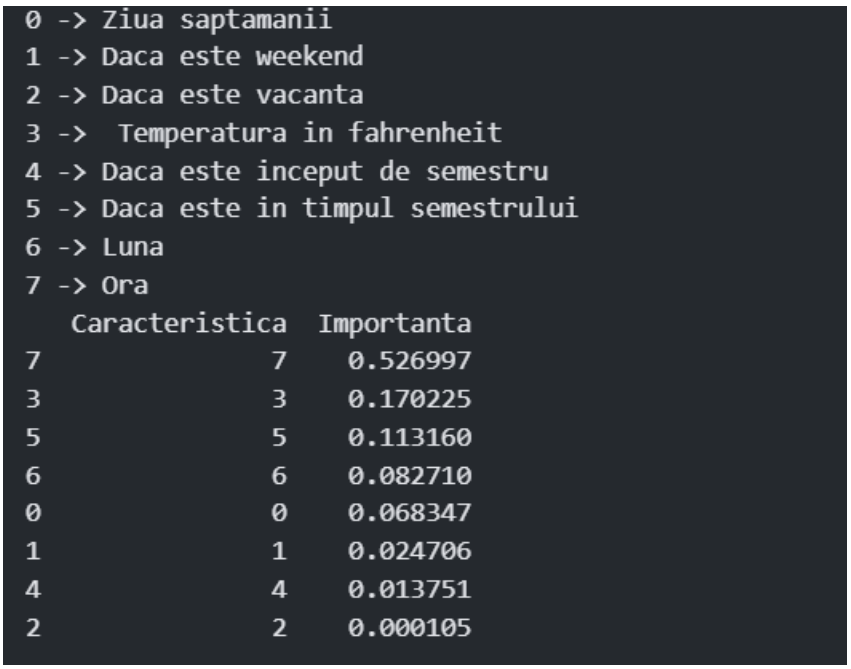


Figura 6.5: Indexul Gini

## 6.3 Rezultatele Modelului Random Forest pe Seturi Noi de Date

Pentru a evalua performanța modelului Random Forest pe seturi noi de date, am testat modelul cu un set de 301 instanțe de date diferite față de cele utilizate pentru antrenament. Scopul acestui test este de a verifica capacitatea modelului de a generaliza și de a face predicții precise pe date necunoscute.

### 6.3.1 Performanța Modelului

Rezultatele obținute în urma testării modelului pe noul set de date sunt următoarele:

- **Timpul de executie**: 1.1s
- **Scorul Random Forest Regressor**: 0.9262
- **Eroarea absolută medie (MAE)**: 3.143
- **Eroarea medie pătratică (MSE)**: 17.344

Aceste rezultate indică faptul că modelul Random Forest are o performanță ridicată și pe seturi noi de date, menținând un nivel scăzut de eroare și o capacitate bună de predicție.

### 6.3.2 Graficul Rezultatelor

Pentru a vizualiza mai clar performanța modelului Random Forest, am trasat un grafic care compară valorile reale cu valorile prezise de model pe setul de date de testare.

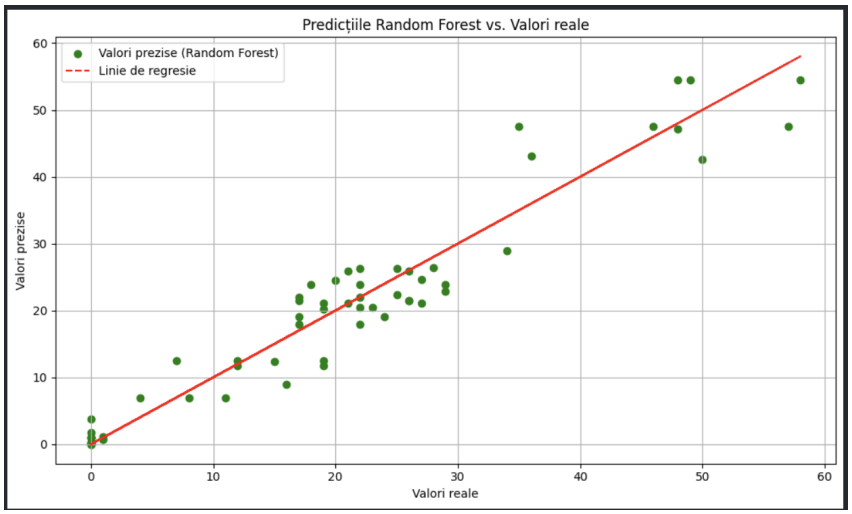


Figura 6.6: Predicțiile modelului Random Forest vs. Valori reale

În Figura 6.6, punctele verzi reprezintă valorile prezise de modelul Random Forest, în timp ce linia roșie punctată reprezintă linia de regresie ideală, unde valorile prezise sunt egale cu valorile reale. Apropierea punctelor verzi de linia roșie punctată indică precizia ridicată a modelului.

### 6.3.3 Interpretarea Rezultatelor

Performanța modelului Random Forest pe setul nou de date confirmă capacitatea acestuia de a generaliza și de a face predicții precise în condiții variate. Scorul ridicat și valorile scăzute ale MAE și MSE demonstrează că modelul poate gestiona eficient variabilitatea datelor și menține o precizie consistentă.

Rezultatele sugerează că modelul este robust și poate fi utilizat în scenarii reale pentru a prezice numărul de persoane din sala de fitness în funcție de diferite variabile. Acest lucru poate ajuta la optimizarea resurselor și la îmbunătățirea experienței utilizatorilor în sălile de fitness.

## 6.4 Compararea Performanței Modelelor

Comparând performanțele modelelor, Random Forest depășește cu mult atât regresia liniară, cât și rețelele neurale artificiale în ceea ce privește toate metricile de evaluare. Random Forest oferă un echilibru bun între performanță și complexitate.

# Capitolul 7

# Concluzii

## 7.1 Rezumatul Lucrării

În această lucrare, am abordat problema predicției numărului de studenți într-o sală de fitness din campus folosind tehnici de învățare automată. Am început prin a motiva alegerea bazei de date și a proiectului, evidențiind importanța optimizării fluxului de persoane pentru îmbunătățirea gestionării resurselor și a experienței utilizatorilor. Am descris contextul bazei de date și cerințele proiectului, urmând să prezentăm aspectele teoretice relevante, inclusiv entropia și indexul Gini, precum și starea actuală a domeniului.

## 7.2 Metodologie și Implementare

Am implementat trei modele de învățare automată: regresia liniară, rețelele neurale artificiale (ANN) și Random Forest. Fiecare model a fost testat și evaluat pe baza unor măsuri de performanță precum scorul $R^2$, eroarea absolută medie (MAE) și eroarea medie pătratică (MSE).Am utilizat entropia și indexul Gini pentru a evalua impuritatea nodurilor și a selecta cele mai bune împărțiri.

## 7.3 Testare și Validare

Modelul Random Forest a fost testat pe un set nou de date pentru a evalua capacitatea sa de generalizare. Rezultatele obținute au demonstrat o performanță ridicată, cu un scor $R^2$ de 0.9262, un MAE de 3.143 și un MSE de 17.344. Aceste rezultate confirmă că modelul poate face predicții precise și pe date necunoscute, menținând un nivel scăzut de eroare.

## 7.4 Rezultate și Interpretare

Rezultatele obținute cu modelul Random Forest au fost comparate cu cele ale altor modele, demonstrând o superioritate clară în ceea ce privește precizia predicțiilor. Analiza graficelor de predicție a arătat o apropiere considerabilă a valorilor prezise de valorile reale, evidențiind capacitatea modelului de a capta relațiile dintre variabilele de intrare și variabila de răspuns.

## 7.5 Implicații și Utilizări Viitoare

Acest studiu sugerează că utilizarea modelului Random Forest pentru predicția numărului de studenți într-o sală de fitness dintr-un campus poate aduce beneficii semnificative în optimizarea gestionării resurselor și îmbunătățirea experienței utilizatorilor. Pe viitor, se pot explora metode de optimizare a hiperparametrilor modelului și extinderea setului de date pentru a îmbunătăți și mai mult performanța predicțiilor.

## 7.6 Corelații

Corelația este o măsură statistică care indică gradul în care două variabile sunt lineare asociate. În contextul analizei datelor din sălile de fitness, identificarea corelațiilor dintre variabile poate oferi informații valoroase despre factorii care influențează fluxul de persoane.

Pentru această analiză, am calculat coeficienții de corelație Pearson [3] între diferitele variabile din setul de date. Rezultatele au evidențiat următoarele corelații semnificative:

- **Numărul de persoane prezente și ora din zi**: Există o corelație puternică între numărul de persoane prezente în sală și ora din zi, indicând perioadele de vârf și cele mai aglomerate intervale orare. De exemplu, orele de seară au înregistrat cel mai mare flux de persoane.
- **Numărul de persoane prezente și ziua săptămânii**: S-au observat diferențe semnificative între zilele săptămânii, cu un flux mai mare de persoane în zilele lucrătoare comparativ cu weekendurile. Această corelație sugerează că mulți studenți preferă să meargă la sală după cursuri.
- **Numărul de persoane prezente și vacanțele**: Vacanțele au avut un impact semnificativ asupra numărului de persoane prezente, cu o scădere notabilă în aceste perioade.
- **Numărul de persoane prezente și temperatura**: Temperatura ambientală a avut, de asemenea, o corelație moderată cu numărul de persoane prezente, indicând că extremele de temperatură (foarte cald sau foarte rece) pot influența frecvența vizitelor la sală.

Analiza corelațiilor ajută la identificarea tiparelor și a factorilor determinanți pentru fluxul de persoane în sălile de fitness. Aceste informații pot fi folosite pentru a planifica mai bine resursele și a optimiza orele de funcționare pentru a răspunde mai bine nevoilor studenților.

## 7.7 Concluzie Generală

În concluzie, aplicarea tehnicilor de învățare automată pentru predicția numărului de studenți într-o sală de fitness s-a dovedit a fi o soluție eficientă și precisă. Modelul Random Forest, în special, a demonstrat capacitatea de a face predicții exacte, ceea ce poate contribui semnificativ la optimizarea operațiunilor în sălile de fitness din orice campus. Acest studiu deschide calea pentru utilizarea pe scară largă a tehnologiilor de învățare automată în gestionarea resurselor și îmbunătățirea serviciilor în diverse domenii.

# Bibliografie


[1] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

[2] B Chandra and P Paul Varghese. Fuzzifying gini index based decision trees. *Expert Systems with Applications*, 36(4):8549–8559, 2009.

[3] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

[4] Dastan Maulud and Adnan M Abdulazeez. A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 1(2):140–147, 2020.

[5] Ulrike Rimmele, Bea Costa Zellweger, Bernard Marti, Roland Seiler, Changiz Mohiyeddini, Ulrike Ehlert, and Markus Heinrichs. Trained men show lower cortisol, heart rate and psychological responses to psychosocial stress compared with untrained men. *Psychoneuroendocrinology*, 32(6):627–635, 2007.

[6] Elizabeth A Unger, Lein Harn, and Vijay Kumar. Entropy as a measure of database information. pages 80–87, 1990.

[7] Jinming Zou, Yi Han, and Sung-Sau So. Overview of artificial neural networks. *Artificial neural networks: methods and applications*, pages 14–22, 2009.