Clasificarea

Sindromului Ovarelor Polichistice (PCOS)

Ruxandra-Teodora Zamfir

UNSTPB

Abstract

Sindromul ovarelor polichistice (PCOS) este o afecțiune hormonală frecventă în rândul femeilor, care duce adesea la multiple complicații de sănătate, precum dezechilibre hormonale, rezistență la insulină, diabet de tip 2 și boli cardiovasculare. În acest studiu, explorăm aplicarea tehnicilor de învățare automată pentru clasificarea PCOS, bazându-ne pe parametri clinici și fizici dintr-un set de date disponibil pe Kaggle. Comparăm performanța diferitelor algoritmi, inclusiv Naive Bayes, Support Vector Machine (SVM), Regresie Logistică și Random Forest, evaluăm performanța fiecărei metode în funcție de diferite metrici, cu scopul de a identifica abordarea optimă pentru sprijinirea diagnosticului rapid și precis al PCOS.

Cuvinte cheie - PCOS, Machine Learning, clasificare binară, SVM, Naive Bayes, Random Forest, Regresie Logistică

I. INTRODUCERE

Sindromul ovarelor polichistice (PCOS) este o tulburare endocrină complexă care afectează între 8-13% din femei la nivel mondial. [1] Se caracterizează prin prezența a numeroase mici chisturi pe ovare, dezechilibre hormonale și disfuncții metabolice. Femeile cu PCOS au adesea simptome precum cicluri menstruale neregulate, creștere excesivă a părului, acnee și infertilitate. De asemenea, PCOS este asociat cu alte afecțiuni: anxietate, depresie, diabet de tip 2 și boli cardiovasculare. [2]

Diagnosticul precoce al PCOS este esențial pentru un management eficient și prevenirea complicațiilor asociate. Totuși, metodele tradiționale de diagnostic se bazează adesea pe evaluări subiective, biomarkeri limitați și ecografii. [3]

Învățarea automată oferă o abordare bazată pe date pentru identificarea tiparelor și clasificarea cazurilor pe baza multiplelor parametri clinici și fizici, permițând diagnostice mai precise și mai eficiente.

II. CERCETĂRI ANTERIOARE

În acest studiu, s-au folosit două modele de învățare automată pentru a detecta PCOS: clasificatorul Bayesian și regresia logistică.

Rezultatele au arătat că pentru clasificatorul Bayesian s-a obținut o performanță mai bună decât regresia logistică, cu o acuratețe generală de 93,93% față de 91,04% pentru regresia logistică. [5]

Acest studiu a dezvoltat și evaluat modele de învățare automată XGBoost pentru predicția diagnosticului de PCOS utilizând înregistrări despre sănătate, obținând performanțe ridicate cu scoruri AUC de până la 85,2% utilizând doar 14-17 caracteristici cheie. Factorii cei mai importanți pentru predicția PCOS au fost: scorul MLP, numărul de sarcini, greutatea, vârsta și rezultatele

testelor de sarcină. Aceste date comune pot ajuta la detectarea precoce a bolii. [6]

Cercetări recente în predicția sindromului ovarian polichistic utilizează trei algoritmi de învățare automată Regresie Liniară, Random Forest și Arborele de Decizie, folosind un set de date cuprinzător care include parametri clinici, niveluri hormonale și informații demografice. Metodologia implică o colectare atentă a datelor din rapoartele pacienților, etape de preprocesare, inclusiv standardizarea și împărțirea datelor în seturi de antrenament și testare (80-20), urmate de implementarea modelului cu optimizarea hiperparametrilor. Dintre algoritmii testați, modelul Arborelui de Decizie a obținut cea mai mare acuratețe, 100%, depășind semnificativ atât Random Forest (88%) cât și Regresia Liniară (34%) în capacitățile de predicție a PCOS.

III. METODOLOGIE

Capitolul de metodologie descrie pașii principali utilizați pentru analiza și clasificarea PCOS prin tehnici de învățare automată. Inițial este prezentat setul de date utilizat, procesul începe cu preprocesarea acestor date. Ulterior, sunt prezentați mai mulți algoritmi de învățare automată.

Pentru evaluarea performanței modelelor, au fost utilizate mai multe metrici standard pentru a obține o imagine completă a capacității fiecărui model de a clasifica corect cazurile de PCOS. Această secțiune oferă o privire detaliată asupra procesului de analiză și pune bazele pentru interpretarea rezultatelor obținute.

1. SETUL DE DATE

Setul de date utilizat în acest studiu a fost preluat de pe Kaggle și include informații detaliate despre paciente diagnosticate cu sau fără PCOS. [7]

Acest set de date conține 43 de caracteristici împărțite în atribute fizice, precum vârsta, greutatea și IMC, dar și în parametri clinici, precum nivelurile hormonale și markerii rezistenței la insulină despre 541 de paciente. Variabila-țintă va indica dacă o pacientă suferă de PCOS sau nu, problema clasificării devenind o problemă de clasificare binară.

2. PREPROCESAREA DATELOR

Pentru a asigura calitatea datelor și a îmbunătăți performanța modelelor de clasificare, s-au aplicat mai mulți pași de preprocesare, descriși în detaliu mai jos.

a. Gestionarea valorilor lipsă

Datele lipsă pot afecta negativ performanța modelelor de învățare automată. Pentru a atenua acest impact s-au identificat valorile lipsă din setul de date. S-au înlocuit aceste valori cu mediana pentru valorile continue("Marraige Status (Yrs)", "II beta-HCG(mIU/mL)", "AMH(ng/mL)") pentru a minimiza influența valorilor extreme (outliers). [8] Pentru valorile categoriale("Fast food (Y/N)") au fost înlocuite cu modul.

b. Eliminarea variabilelor irelevante

Pentru a reduce zgomotul în date și a îmbunătăți performanța modelelor, s-au eliminat variabilele care nu contribuie direct la procesul de clasificare. De exemplu, variabile precum identificatorii unici (ex.: numărul dosarului pacientei) au fost eliminate, deoarece nu au relevanță pentru analiza PCOS. [9]

c. Standardizarea caracteristicilor

Standardizarea este un pas esențial în preprocesarea datelor, mai ales pentru algoritmi sensibili la magnitudinea valorilor caracteristicilor, precum SVM sau Regresie Logistică. Acest proces asigură că toate caracteristicile au o medie de 0 și o deviație standard de 1, permițând algoritmilor să funcționeze optim. [9]

d. Împărțirea setului de date

Setul de date, care conține 42 de caracteristici, a fost împărțit în două subseturi pentru antrenare și testare. Set de antrenare: Conține 378 de instanțe, reprezentând 80% din totalul datelor, utilizate pentru antrenarea modelelor. Set de testare: Include 163 de instanțe, reprezentând 20% din date, utilizate pentru evaluarea performanței modelelor pe date noi și neutilizate anterior. Această împărțire a fost realizată pentru a evalua corect generalizarea modelelor și a preveni overfitting.

```
RangeIndex: 541 entries, 0 to 540
Data columns (total 42 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Target                  541 non-null    int64
 1    Age (yrs)              541 non-null    int64
 2   Weight (Kg)             541 non-null    float64
 3   Height(Cm)              541 non-null    float64
 4   BMI                     541 non-null    float64
 5   Blood Group             541 non-null    int64
 6   Pulse rate(bpm)         541 non-null    int64
 7   RR (breaths/min)        541 non-null    int64
 8   Hb(g/dl)                541 non-null    float64
 9   Cycle(R/I)              541 non-null    int64
 10  Cycle length(days)      541 non-null    int64
 11  Marraige Status (Yrs)   540 non-null    float64
 12  Pregnant(Y/N)           541 non-null    int64
 13  No. of aborptions       541 non-null    int64
 14    I   beta-HCG(mIU/mL)  541 non-null    float64
 15  II    beta-HCG(mIU/mL)  541 non-null    object
 16  FSH(mIU/mL)             541 non-null    float64
 17  LH(mIU/mL)              541 non-null    float64
 18  FSH/LH                  541 non-null    float64
 19  Hip(inch)               541 non-null    int64
 20  Waist(inch)             541 non-null    int64
 21  Waist:Hip Ratio         541 non-null    float64
 22  TSH (mIU/L)             541 non-null    float64
 23  AMH(ng/mL)              541 non-null    object
 24  PRL(ng/mL)              541 non-null    float64
 25  Vit D3 (ng/mL)          541 non-null    float64
 26  PRG(ng/mL)              541 non-null    float64
 27  RBS(mg/dl)              541 non-null    float64
 28  Weight gain(Y/N)        541 non-null    int64
 29  hair growth(Y/N)        541 non-null    int64
 30  Skin darkening (Y/N)    541 non-null    int64
 31  Hair loss(Y/N)          541 non-null    int64
 32  Pimples(Y/N)            541 non-null    int64
 33  Fast food (Y/N)         540 non-null    float64
 34  Reg.Exercise(Y/N)       541 non-null    int64
 35  BP _Systolic (mmHg)     541 non-null    int64
 36  BP _Diastolic (mmHg)    541 non-null    int64
 37  Follicle No. (L)        541 non-null    int64
 38  Follicle No. (R)        541 non-null    int64
 39  Avg. F size (L) (mm)    541 non-null    float64
 40  Avg. F size (R) (mm)    541 non-null    float64
 41  Endometrium (mm)        541 non-null    float64
dtypes: float64(19), int64(21), object(2)
```

Figura 1. Baza de date pentru clasificarea pacientelor cu risc de PCOS

Figura 1 ilustrează cele 42 de caracteristici obținute în urma preprocesării pentru 541 de paciente.

3. ALGORITMI UTILIZATI

Algoritmi utilizați de învățare automată folosiți au fost următorii. Au fost implementați utilizând biblioteca sklearn.

❖ Naive Bayes

Clasificatorul Naive Bayes este o metoda simplă de clasificare bazată pe statistici. Presupune ca toate caracteristicile contribuie la clasificare și ca sunt independente între ele. [10]

❖ Regresie Logistică

Regresia Logistică este un model liniar utilizat pentru clasificare binară, estimând probabilitatea ca o observație să aparțină unei anumite clase, adică determinarea prezenței sau absenței PCOS. [11]

❖ SVM

Support Vector Machine (SVM) este un algoritm de învățare automată care utilizează hiperplane optime pentru a separa clasele, fiind capabile să gestioneze relații non-liniare. Această capacitate de a modela relații non-liniare este esențială, având în vedere că parametrii clinici și fizici, precum nivelurile hormonale sau rezistența la insulină, care nu sunt întotdeauna corelați liniar cu boala. [12]

❖ Random Forest

Modelul construiește mai mulți arbori de decizie independenți, fiecare antrenat pe un subset aleatoriu al datelor, și combină predicțiile lor prin vot majoritar pentru a obține o clasificare finală robustă. [13]

4. METRICI DE EVALUARE

Performanța fiecărui model a fost evaluată utilizând următoarele metrici:

- Matricea de Confuzie este o tabelă care prezintă performanța unui model de

clasificare prin compararea valorilor reale cu cele prezise. [14]

	Predicții	
Valori Reale	TN	FP
	FN	TP

Tabel 1. Reprezentarea unei matrice de confuzie pentru o clasificare binară

TP (True Positive): Numărul de paciente care au fost corect clasificate ca având PCOS.

TN (True Negative): Numărul de paciente care au fost corect clasificate ca nu având PCOS.

FN (False Negative): Numărul de paciente care au fost clasificate greșit ca nu având PCOS, în timp ce în realitate suferă de această afecțiune.

FP (False Positive): Numărul de paciente care au fost clasificate greșit ca având PCOS, în timp ce în realitate nu suferă de această afecțiune.

- Acuratețe: Proporția instanțelor clasificate corect.

$$Acuratețea = \frac{TP + TN}{TP + FP + TN + FN}$$

- Precizie: Proporția predicțiilor pozitive corecte între toate predicțiile pozitive.

$$Precizia = \frac{TP}{TP + FP}$$

- Recall: Proporția predicțiilor pozitive corecte între toate cazurile pozitive reale.

$$Recall = \frac{TP}{TP + FN}$$

- F1 Score: Media armonică a preciziei și recall.

$$F1Score = \frac{2\ x\ Precizie\ x\ Recall}{Precizie + Recall}$$

Alegerea metricilor adecvate este un pas esențial în contextul unui diagnostic medical, pentru a avea o perspectiva asupra clasificărilor corecte și eronate, cât și pentru vizualizarea acestora.

IV. REZULTATE ȘI DISCUȚIE

În acest capitol se vor interpreta rezultatele obținute pentru cei 4 algoritmi implementați utilizând metricile menționate în capitolul anterior.

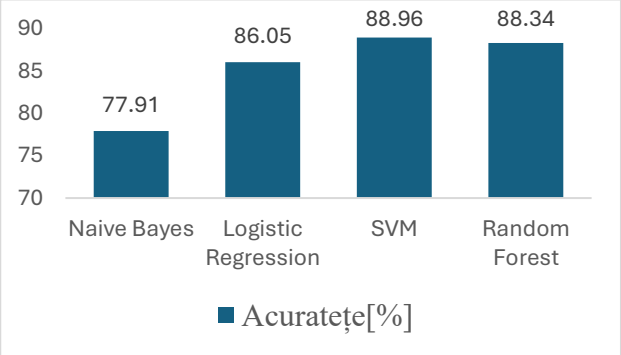


Figura 2. Acuratețea pentru cele 4 modele

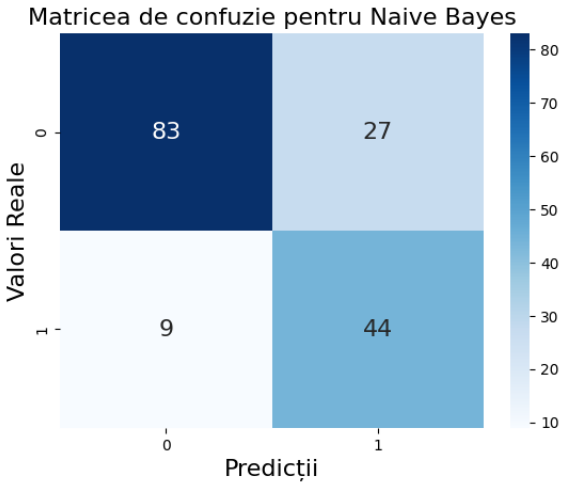
Model	Precizie [%]	Recall [%]	F1 Score [%]
Naive Bayes	61.97	83.02	70.97
Logistic Regression	86.05	69.81	77.08
SVM	92.68	71.7	80.85
Random Forest	88.64	73.58	80.41

Tabel 2. Analiza preciziei, recall-ului și F1-Score pentru cele 4 metode

❖ Naive Bayes

Pentru metoda Naive Bayes rezultatele au mai scăzute, acuratețea fiind cea mai mică de 77.91%, iar precizia 61.97% a fost mai

redusă comparativ cu ceilalți algoritmi. Deși recall-ul este relativ ridicat de 83.02%, precizia scăzută de 61.97% sugerează că modelul generează multe alarme false pozitive. Acest rezultat poate fi influențat de ipoteza de independență condiționată dintre variabile, ipoteză care nu reflectă realitatea datelor clinice, unde variabilele, precum markerii hormonali sau parametrii metabolici, sunt adesea corelate.

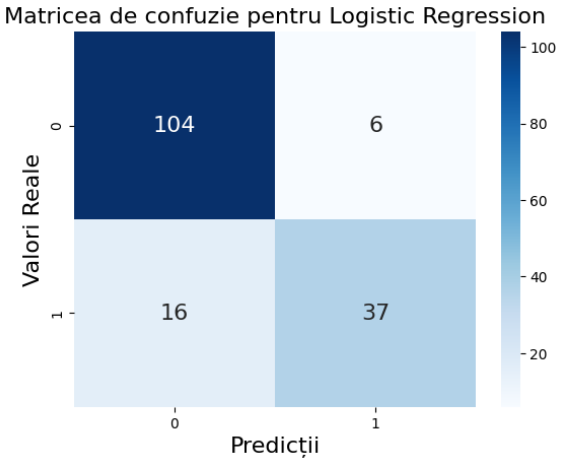


Numărul de False Positive este destul de mare (27), iar aceste cazuri sunt deosebit de importante în contextul medical, deoarece cazurile pot duce la investigații medicale suplimentare inutile și pot provoca anxietate pacientei.

❖ Logistic Regression

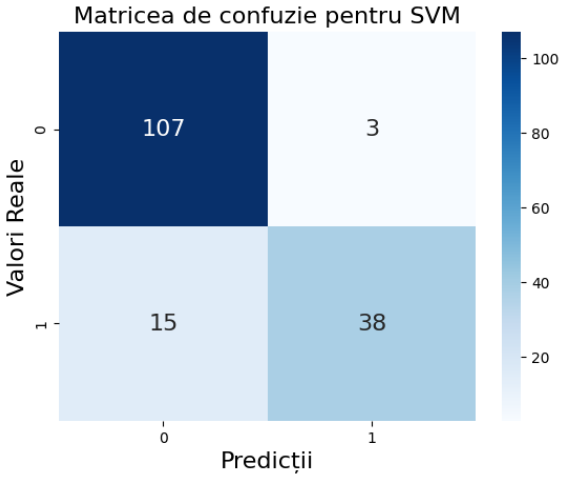
A obținut o performanță echilibrată, cu o acuratețe de 86.50% și valori bune ale preciziei de 86.05% și F1-Score-ului de 77.08%. Acest model a beneficiat de relațiile liniare dintre variabile, gestionând bine datele standardizate și corelate moderat. Totuși, un recall de 69.81% indică faptul că modelul a omis unele cazuri pozitive.

Numărul de False Negative este destul de mare (16), iar aceste cazuri sunt deosebit de importante în contextul medical, deoarece o pacientă cu PCOS care nu este diagnosticată poate avea complicații de sănătate.



❖ Support Vector Machine (SVM)

Prin metoda SVM am obținut cea mai mare acuratețe, 88.96%, datorită capacității sale de a separa clasele folosind hiperplane optime, chiar și în cazul relațiilor non-liniare dintre variabile. Precizia ridicată de 92.68% subliniază abilitatea sa de a face predicții corecte, în timp ce recall-ul de 71.70% sugerează o ușoară limitare în detectarea tuturor cazurilor pozitive. F1-Score-ul de 80.85% reflectă un echilibru general bun între precizie și recall.

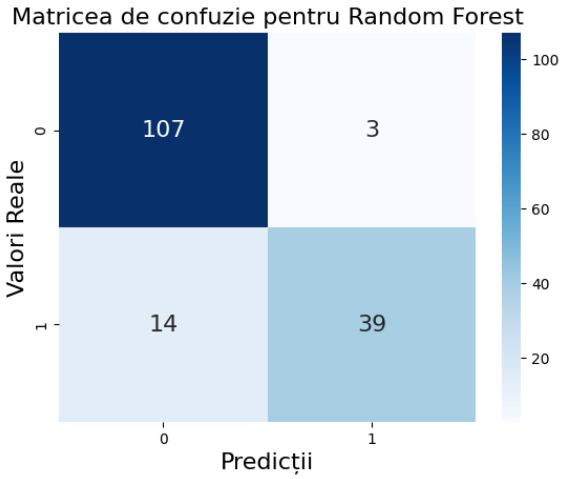


Numărul de False Negativ este moderat, asemănător metodei de Regresie Logistică.

❖ Random Forest

Cu o acuratețe de 88.34%, acest algoritm s-a apropiate de rezultatele metodei SVM. Abilitatea de a gestiona relații complexe și de a evita overfitting-ul prin utilizarea ansamblurilor de arbori de decizie l-a făcut

potrivit pentru analiza datelor de mare dimensionalitate. Precizia de 88.64% și recall-ul de 73.58% sunt echilibrate, iar F1-Score-ul de 80.41% indică o performanță generală bună.



Pentru că utilizarea algoritmilor de predicție a unui diagnostic este important să ne concentrăm pe puterea lor de a diferenția un diagnostic fals negativ.

V. CONCLUZII

Acest studiu demonstrează eficacitatea tehnicilor de învățare automată în clasificarea PCOS pe baza parametrilor clinici și fizici. Dintre modelele evaluate, SVM s-a dovedit a fi cel mai precis, urmat de Random Forest, însă pentru un diagnostic ne dorim ca recall-ul sa fie ridicat, întrucât un False Negativ care trece neobservat poate cauza complicații în timp, ceea ce ne arată faptul ca Naive Bayes are performanțe remarcabile din acest punct de vedere.

Aceste descoperiri sugerează că învățarea automată poate juca un rol crucial în sprijinul diagnosticării timpurii și tratamentului personalizat al PCOS.

Lucrările viitoare ar putea implica integrarea unor biomarkeri suplimentari, explorarea abordărilor bazate pe învățare profundă și validarea modelelor pe seturi de date mai mari și diverse. Analiza detaliată a feature-rilor ar fi o altă metodă de îmbunătățire a algorimilor, [14] precum și explorarea altor metode de învățare automată. [15]

REFERINȚE

[1] Hoeger, Kathleen M., Anuja Dokras, and Terhi Piltonen. "Update on PCOS: consequences, challenges, and guiding treatment." *The Journal of Clinical Endocrinology & Metabolism* 106.3 (2021): e1071-e1083.

[2] Teede, Helena, Amanda Deeks, and Lisa Moran. "Polycystic ovary syndrome: a complex condition with psychological, reproductive and metabolic manifestations that impacts on health across the lifespan." *BMC medicine* 8 (2010): 1-10.

[3] Battaglia, Cesare, et al. "Ultrasound evaluation of PCO, PCOS and OHSS." *Reproductive biomedicine online* 9.6 (2004): 614-619.

[4] Mehrotra, Palak, et al. "Automated screening of polycystic ovary syndrome using machine learning techniques." *2011 Annual IEEE India Conference*. IEEE, 2011.

[5] Zad, Zahra, et al. "Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records." *Frontiers in Endocrinology* 15 (2024): 1298628.

[6] Priyadharshini, M., et al. "PCOS Disease Prediction Using Machine Learning Algorithms." *International Research Journal on Advanced Engineering Hub (IRJAEH)* 2.03 (2024): 651-655.

[7] PCOS Classification - Input Data

https://www.kaggle.com/code/ilaydadu/pcos-classification/input

[8] Elmannai, Hela, et al. "Polycystic ovary syndrome detection machine learning model based on optimized feature selection and explainable artificial

intelligence." *Diagnostics* 13.8 (2023): 1506.

[9] Lim, Jiekee, et al. "Predicting TCM patterns in PCOS patients: An exploration of feature selection methods and multi-label machine learning models." *Heliyon* 10.15 (2024).

[10] Jadhav, Sayali D., and H. P. Channe. "Comparative study of K-NN, naive Bayes and decision tree classification techniques." *International Journal of Science and Research (IJSR)* 5.1 (2016): 1842-1845.

[11] Starbuck, Craig. "Logistic Regression." *The Fundamentals of People Analytics: With Applications in R*. Cham: Springer International Publishing, 2023. 223-238.

[12] Thakre, Vaidehi, et al. "PCOcare: PCOS detection and prediction using machine learning algorithms." *Biosci Biotechnol Res Commun* 13.14 (2020): 240-244.

[13] Amin, Fahmy, and M. Mahmoud. "Confusion matrix in binary classification problems: A step-by-step tutorial." *Journal of Engineering Research* 6.5 (2022): 0-0.

[14] Mehrotra, Palak, et al. "Automated screening of polycystic ovary syndrome using machine learning techniques." *2011 Annual IEEE India Conference*. IEEE, 2011.

[15] Bharati, Subrato, Prajoy Podder, and M. Rubaiyat Hossain Mondal. "Diagnosis of polycystic ovary syndrome using machine learning algorithms." *2020 IEEE region 10 symposium (TENSYMP)*. IEEE, 2020.