Cazan Ștefan-Rareș, 314CA

Model de clasificare – Predicția diabetului

Abordare:

În acest proiect am pornit de la un set de date public de pe platforma Kaggle, diabetes-prediction-dataset, care conține informații despre peste 100000 de pacienți.

Setul include caracteristici precum vârsta, genul, indicele de masă corporală (BMI), hipertensiunea, afecțiunile cardiace, istoricul de fumat, nivelul HbA1c și nivelul glicemiei.

La acestea am adăugat noi coloane, precum bin-uri pentru vârstă (<30, 30-50, 50-70 și peste 70 de ani), nivelul de activitate fizica (calculat in funcție de vârstă, la care se adaugă un noise de 10%) și feature-ul de obezitate, care este 1 daca indicele de masă corporal al pacientului este peste 30. Coloanele adăugate au rolul de a ajuta modelul și de a crește acuratețea.

Datele pentru fumat erau initial string-uri pe care le-am transformat prin one-hot encoding (cu funcția *get_dummies*), în funcție de alegere: nu a fumat niciodata, a fumat în trecut, nu fumează acum, fumează acum.

Analiza exploratorie a datelor:

Folosind seaborn și matplotlib, am generat grafice relevante despre datele mele, care sunt prezentate în continuare.

a) Analiza valorilor lipsă:

Deoarece peste 35% din datele despre fumat erau marcate cu „No Info” (mai exact 35816 de persoane, adică 35.82%), le-am înlocuit cu cea mai comună alegere a pacienților (am folosit de funcția *mode*). Celelalte date din dataset erau complete, după cum se poate vedea în tabelele generate de describe() în secțiunea următoare.

b) Statistici descriptive:

Statisticile descriptive pentru setul de antrenament:

	gender	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	smoking_history_ever	smoking_history_former	smoking_history_never	smoking_history_not current	is_obese	age_group_30-50	age_group_50-70	age_group_70+	physical_activity	diabetes
count	146400.000000	146400.000000	146400.000000	146400.000000	146400.000000	146400.000000	146400.000000	146400.000000	146400.000000	146400.000000	146400.000000	146400.000000	146400.000000	146400.000000	146400.000000	146400.000000	146400.000000
mean	0.336031	50.502297	0.071298	0.035273	29.374259	6.151779	163.394993	0.023245	0.071093	0.539713	0.039016	0.328074	0.216981	0.357350	0.201571	45.598320	0.500000
std	0.472583	21.436647	0.257323	0.184470	7.199426	1.209308	56.868204	0.150680	0.256981	0.498422	0.193635	0.469513	0.412191	0.479221	0.401175	33.722159	0.500002
min	0.000000	0.080000	0.000000	0.000000	10.010000	3.500000	80.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	36.000000	0.000000	0.000000	25.997865	5.700000	130.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	15.000000	0.000000
50%	0.000000	54.069595	0.000000	0.000000	27.320000	6.171367	155.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	40.000000	0.500000
75%	1.000000	67.000000	0.000000	0.000000	32.730000	6.649840	200.000000	0.000000	0.000000	1.000000	0.000000	1.000000	0.000000	1.000000	0.000000	76.000000	1.000000
max	2.000000	80.000000	1.000000	1.000000	95.690000	9.000000	300.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	100.000000	1.000000

Statisticile descriptive pentru setul de testare:

	gender	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	smoking_history_ever	smoking_history_former	smoking_history_never	smoking_history_not current	is_obese	age_group_30-50	age_group_50-70	age_group_70+	physical_activity	diabetes
count	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000
mean	0.413000	41.826878	0.074900	0.039100	27.336379	5.527400	137.974250	0.041000	0.091100	0.712500	0.063250	0.238050	0.279550	0.262350	0.133250	69.229100	0.085000
std	0.492588	22.464977	0.263236	0.193838	6.620125	1.070543	40.891357	0.198295	0.287758	0.452608	0.243418	0.425901	0.448789	0.439923	0.339853	25.767567	0.278889
min	0.000000	0.080000	0.000000	0.000000	10.210000	3.500000	80.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	24.000000	0.000000	0.000000	23.620000	4.800000	100.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	53.000000	0.000000
50%	0.000000	43.000000	0.000000	0.000000	27.320000	5.800000	140.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	73.000000	0.000000
75%	1.000000	59.000000	0.000000	0.000000	29.680000	6.200000	159.000000	0.000000	0.000000	1.000000	0.000000	0.000000	1.000000	1.000000	0.000000	91.000000	0.000000
max	2.000000	80.000000	1.000000	1.000000	87.700000	9.000000	300.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	100.000000	1.000000

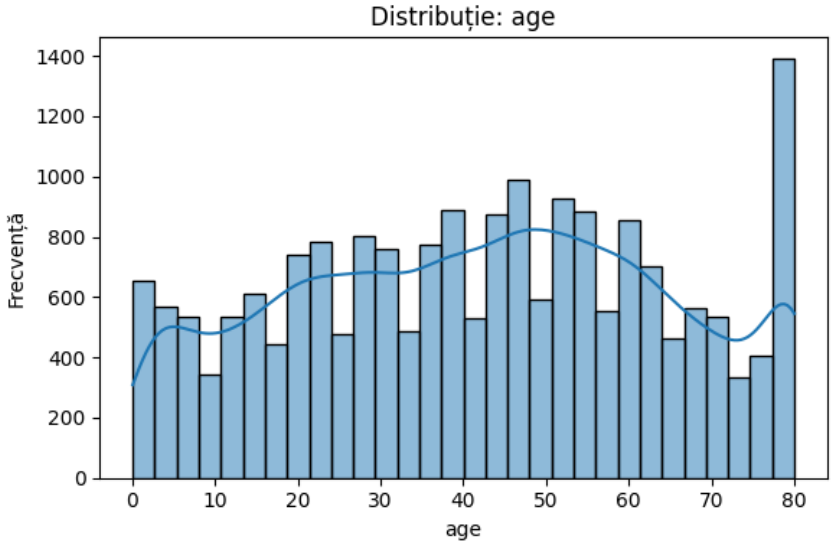
Analizând abaterea standard pentru vârstă, nivelul de glucoză din sânge și nivelul de activitate fizică, putem trage concluzia că aceste date variază mai mult decât celelalte.

Tabelul setului de testare și cel al setului de antrenament au valori similare, ceea ce era de așteptat de la un split 80-20 generat aleator. Diferența apare la coloana „diabetes”, unde media setului de antrenament este 0.5. Folosind SMOTE, datele din antrenament au fost echilibrate pentru a avea mai multe cazuri de diabet, deoarece numărul acestora era scăzut, ceea ce dăuna acurateței algoritmului K Nearest Neighbour.

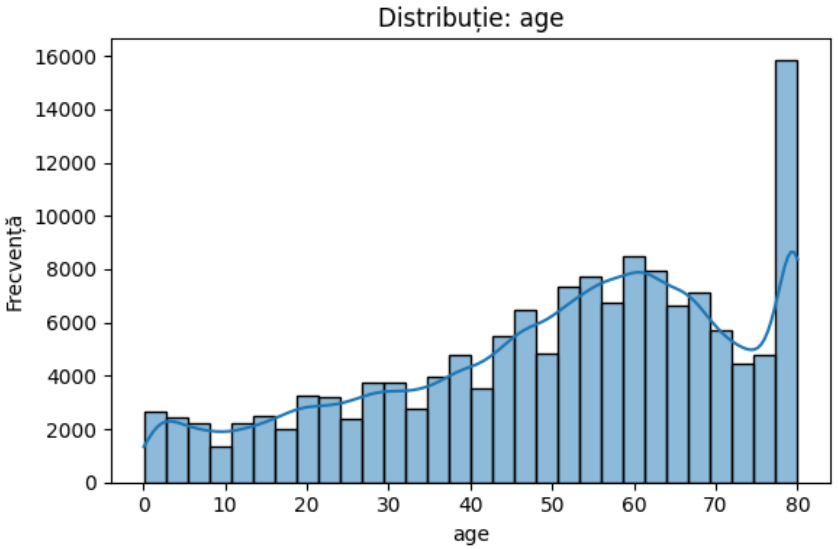
c) Analiza distribuției variabilelor:

Histograme pentru caracteristicile numerice:

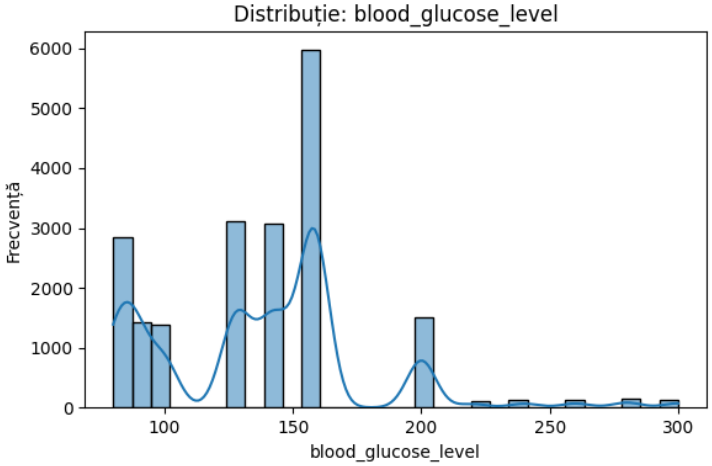
Setul test:



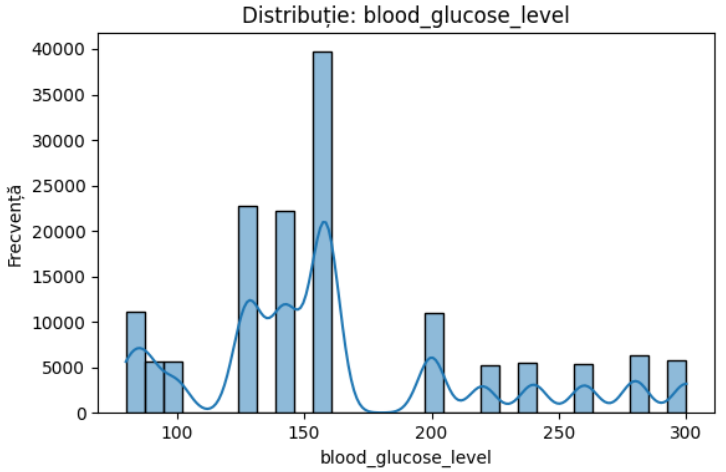
Setul train:



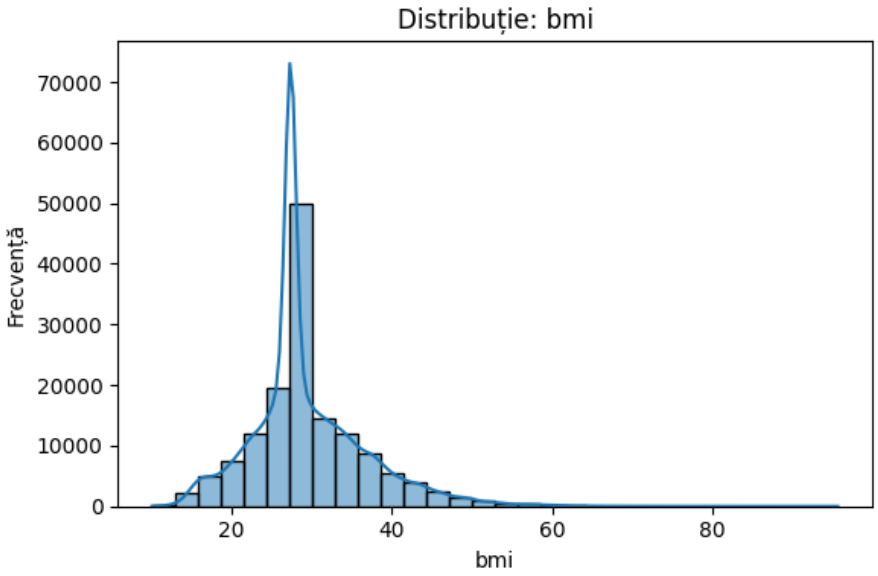
Setul test:



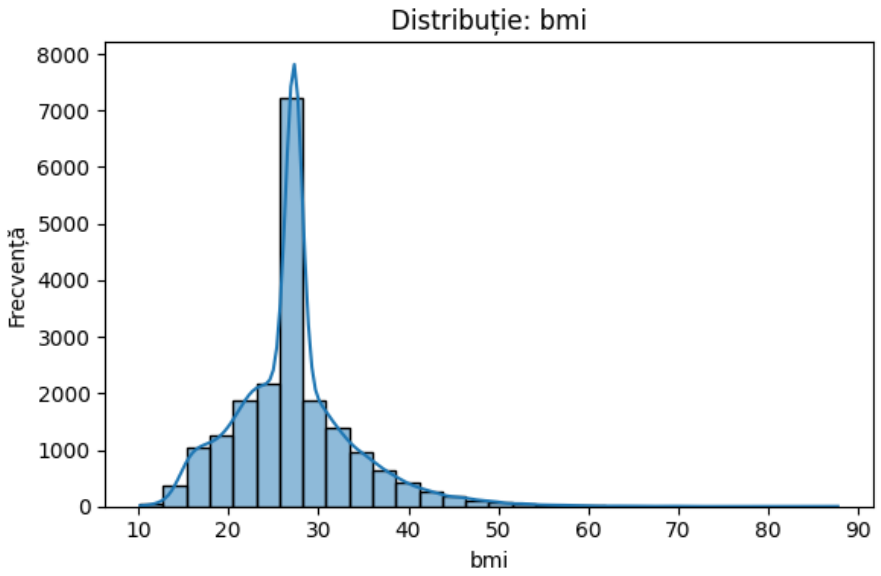
Setul train:



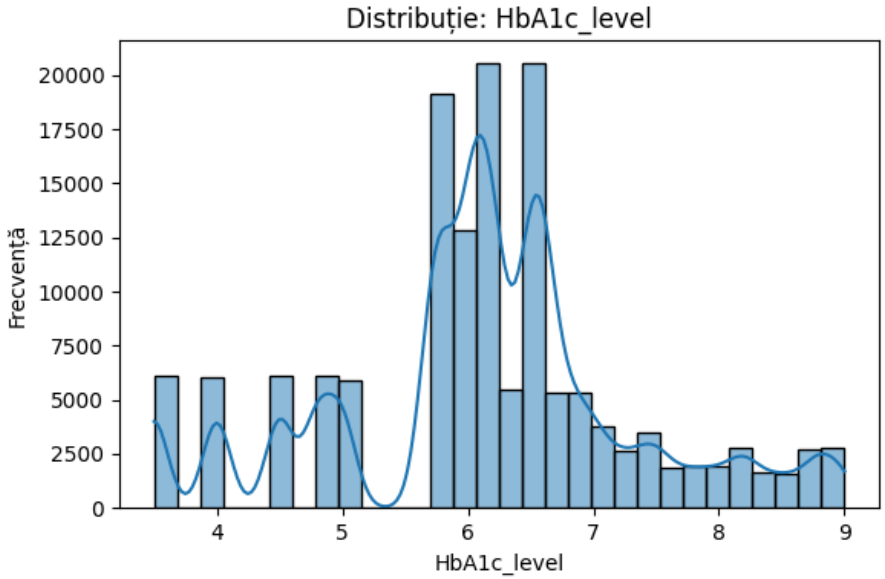
Setul test:



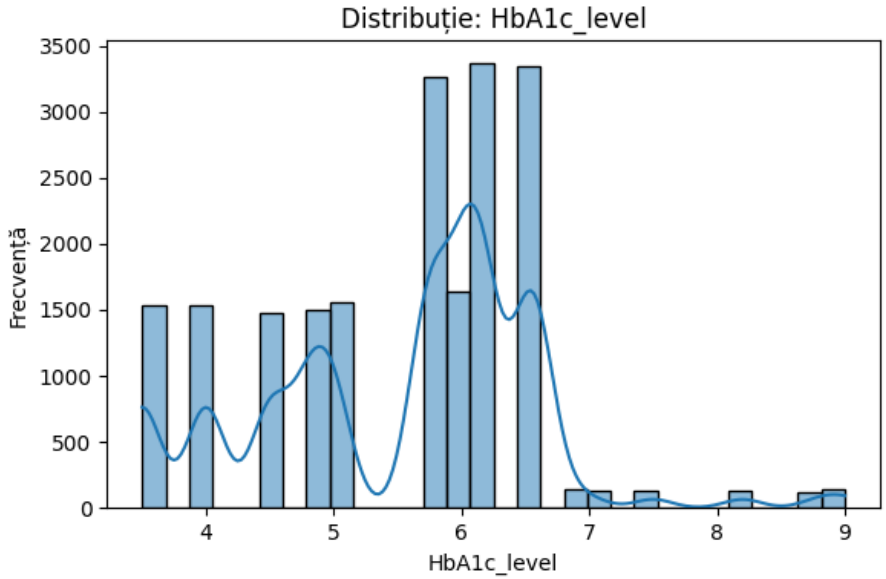
Setul train:



Setul test:

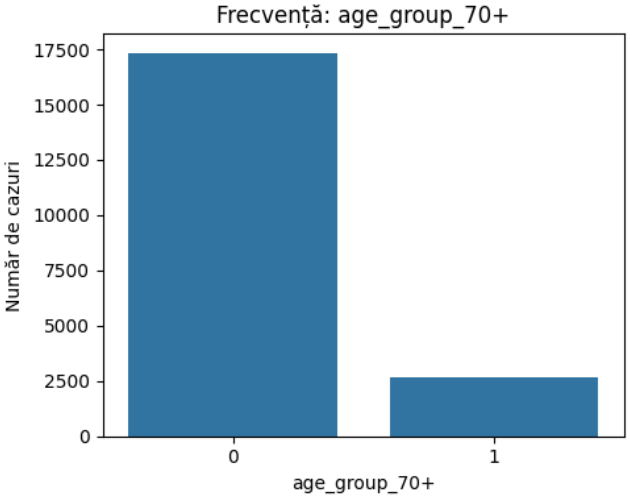


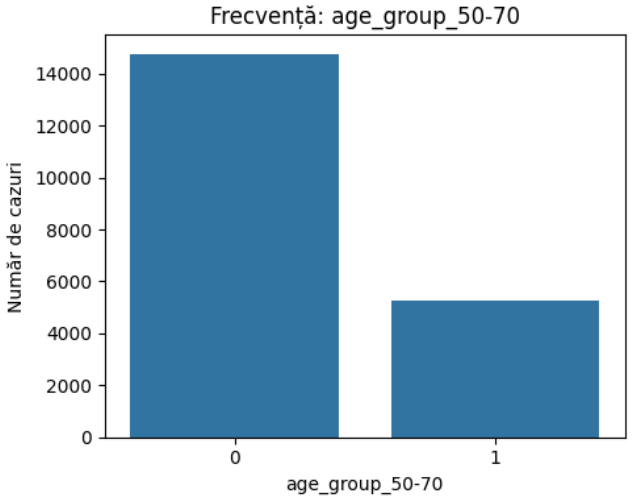
Setul train:

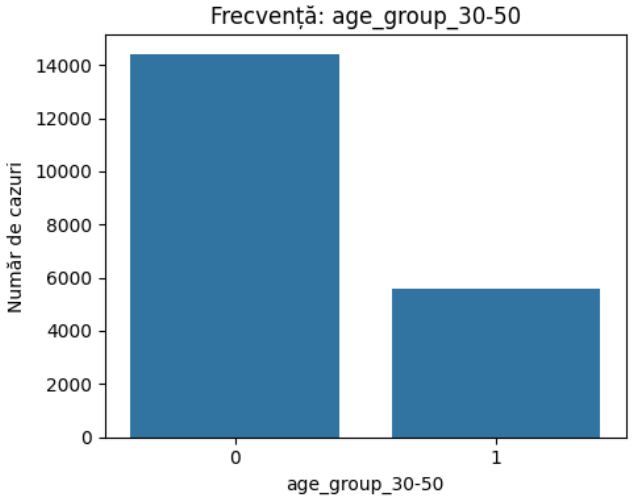


Grafice de tip countplot pentru variabilele categorice:

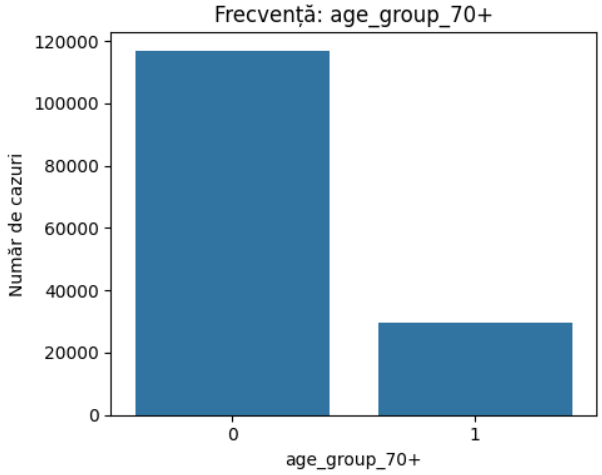
Setul de testare:

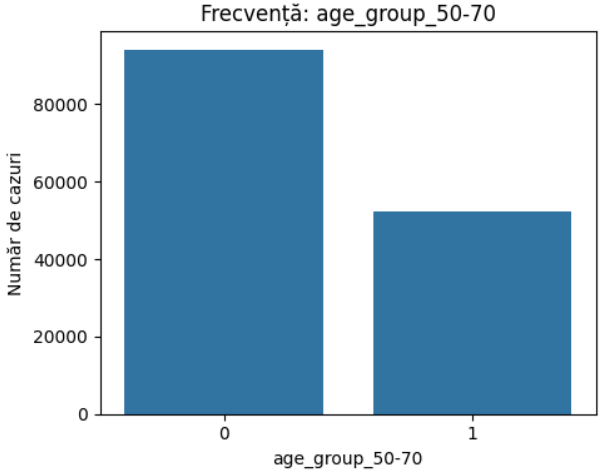


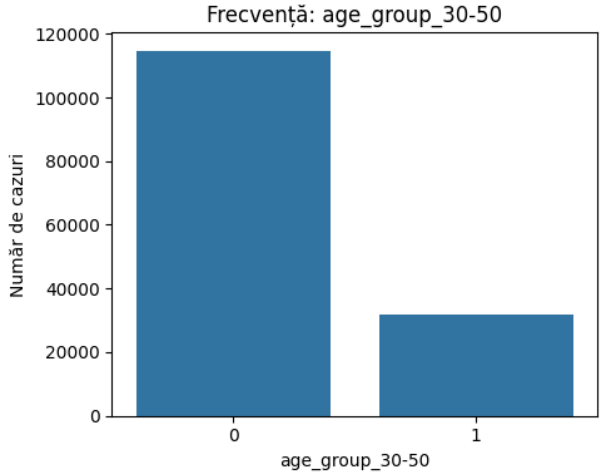




Setul de antrenament:

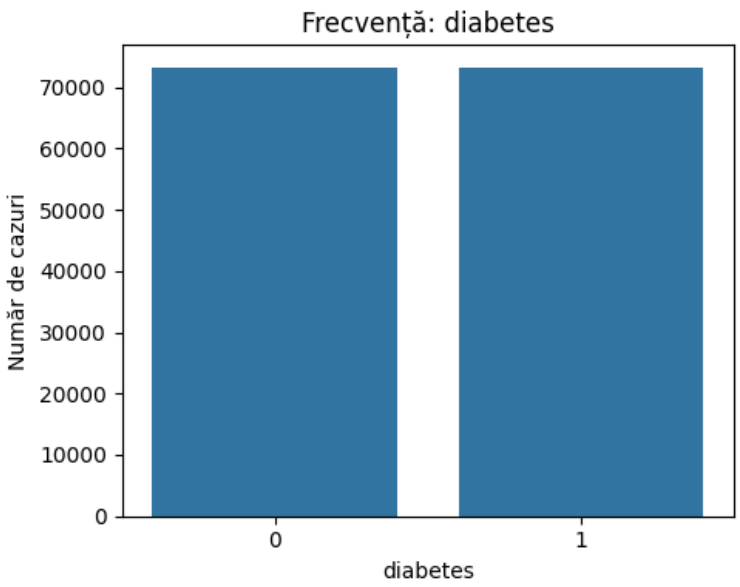




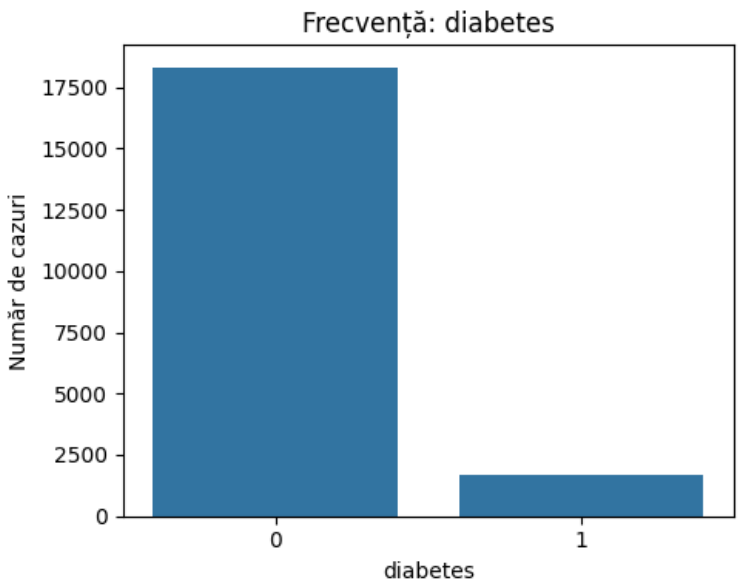


- Putem observa că avem mai multe date cu pacienți tineri și între 50-70 de ani decât cu cei de peste 70 de ani. Faptul că cei mai comuni sunt pacienții între 50 și 70 de ani este de folos pentru antrenare, deoarece majoritatea diagnosticelor de diabet se pun când pacienții au aceste vârste.

Setul de antrenament:

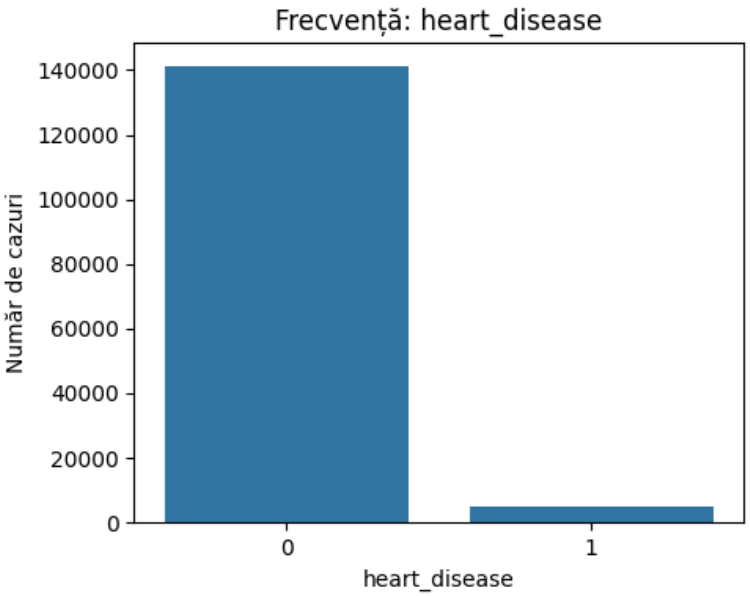


Setul de test:

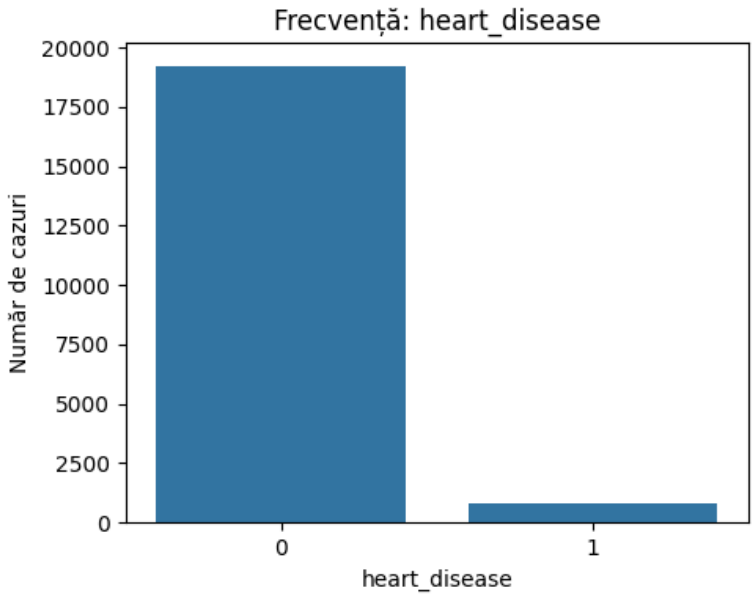


- Așa cum am specificat și mai devreme, deoarece am folosit SMOTE, datele din setul de antrenament diferă vizibil față de cele din setul de test în acest caz!

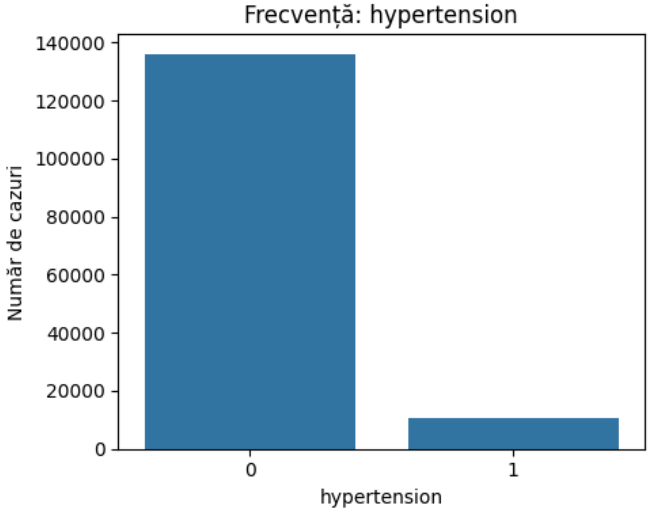
Setul de antrenament:



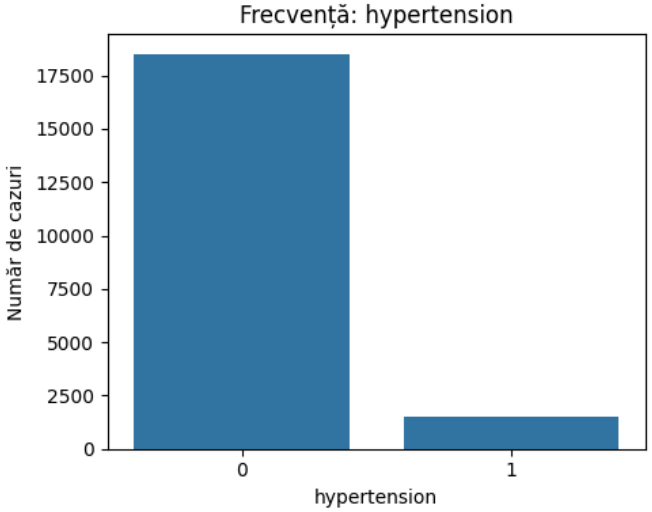
Setul de test:



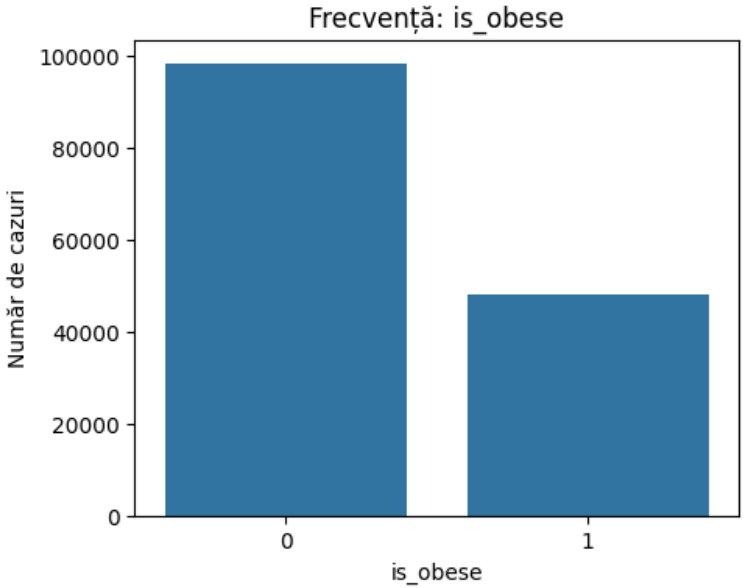
Setul de antrenament:



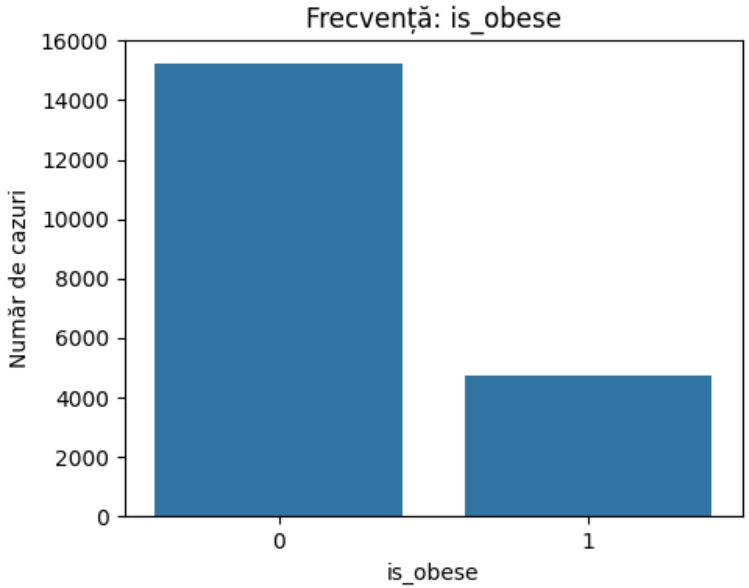
Setul de test:



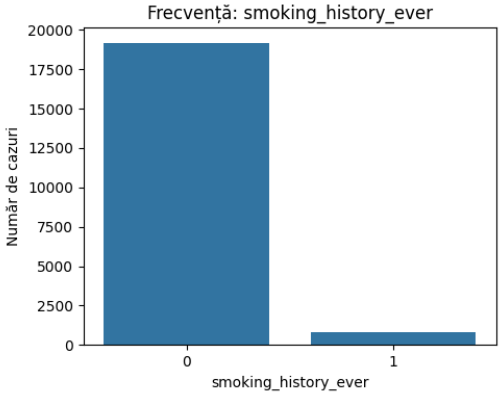
Setul de antrenament:

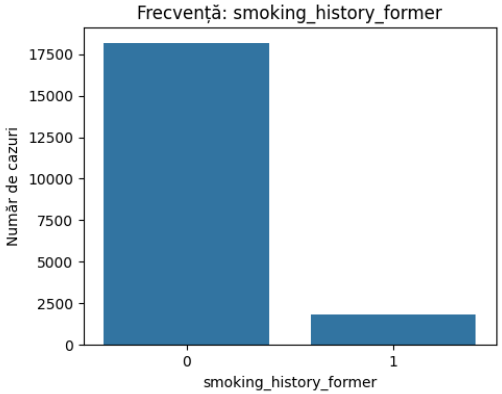


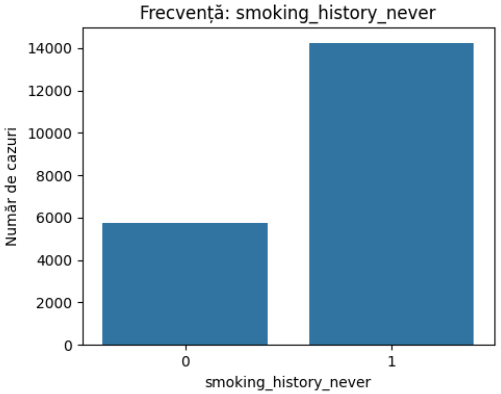
Setul de test:

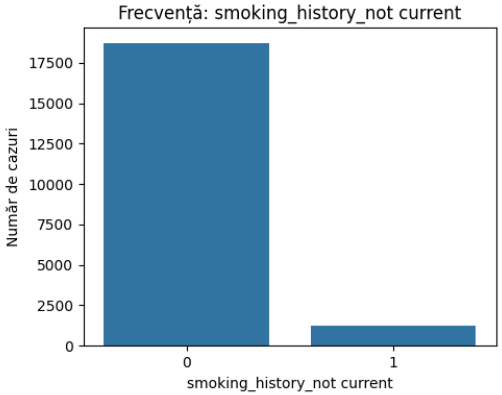


Setul de testare:

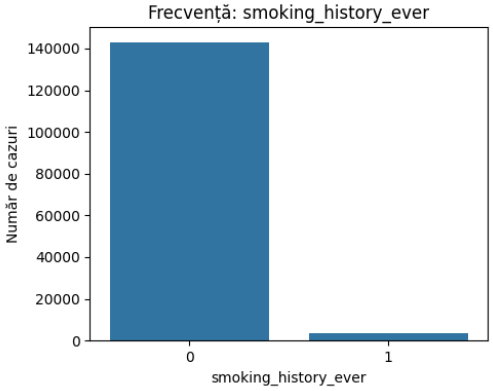


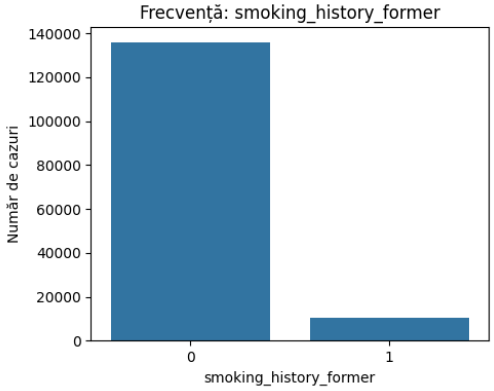


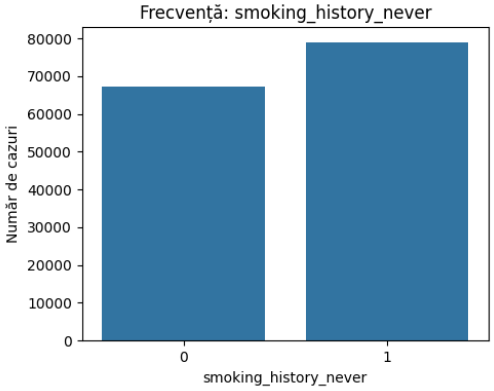


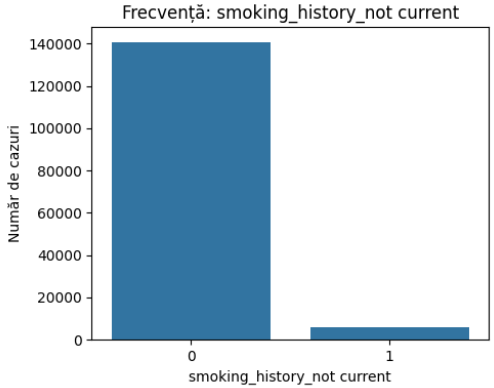


Setul de antrenament:





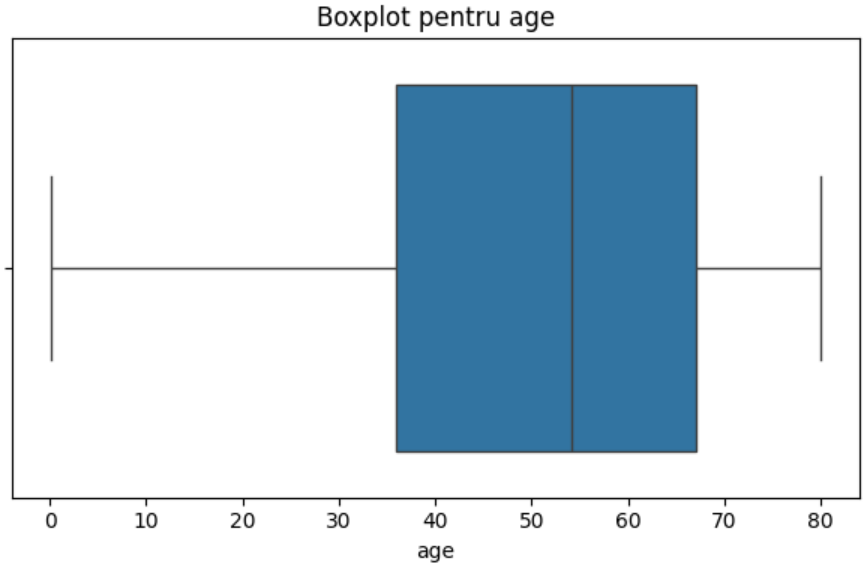




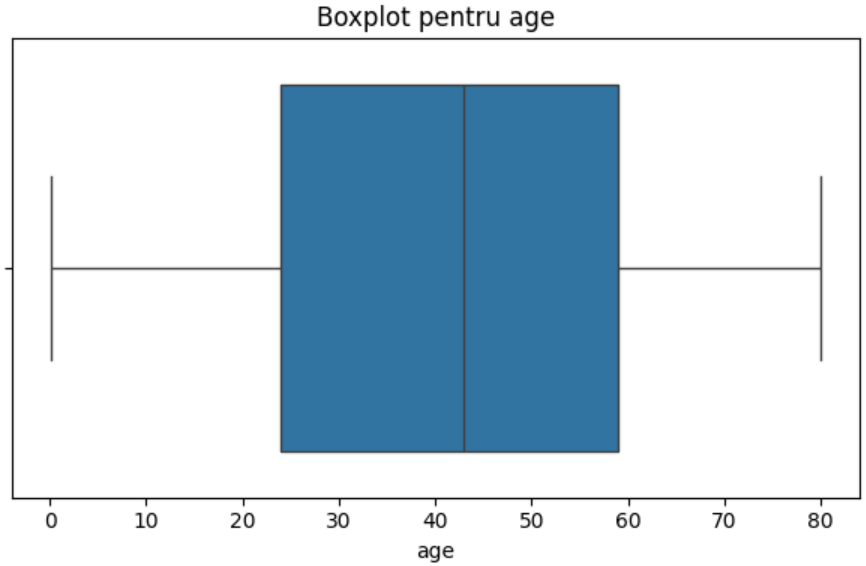
- Aici putem constata faptul că am folosit funcția mode pentru a umple valorile fără informații legate de fumat! Mode a determinat ca fiind cea mai comună alegere a smoking history „never”, ceea ce a exagerat și mai mult procentele. Așadar, în cazul în care fumatul ar avea o implicație directă în cauzarea diabetului, metoda de umplere a valorilor lipsă ar produce rezultate eronate! Totuși, în cazul acesta, după o testare, rezultatele nu se modifică în mod considerabil dacă pur și simplu am fi adăugat o coloana cu „No Info” pentru istoricul fumatului. Cum în laboratoare am folosit funcția mode pentru adăugarea datelor lipsă, am decis să aplic metoda și în proiect.

d) Detectarea outlierilor:

Setul de antrenament:

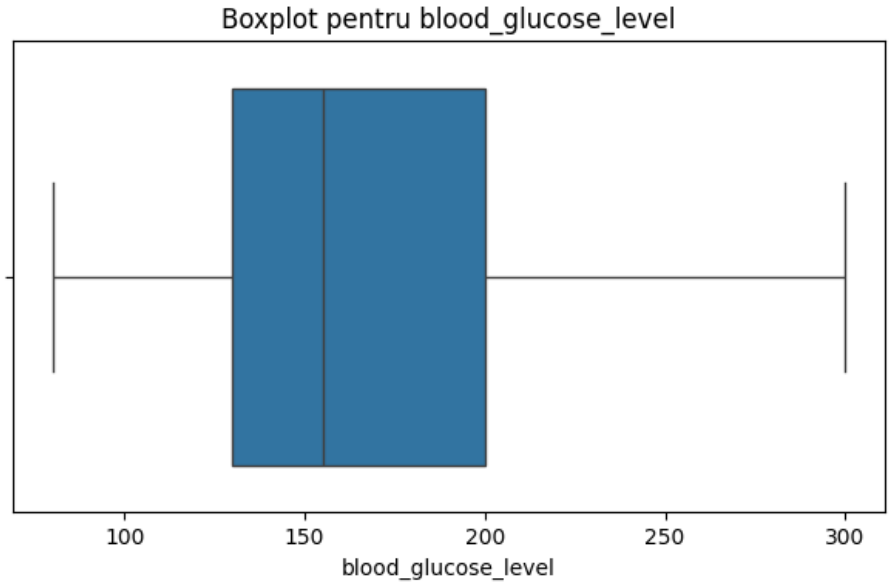


Setul de test:



- Se observă că nu avem outliers, criteriile pentru alegerea vârstei au fost, probabil, regulate atunci când s-a creat baza de date cu pacienți.

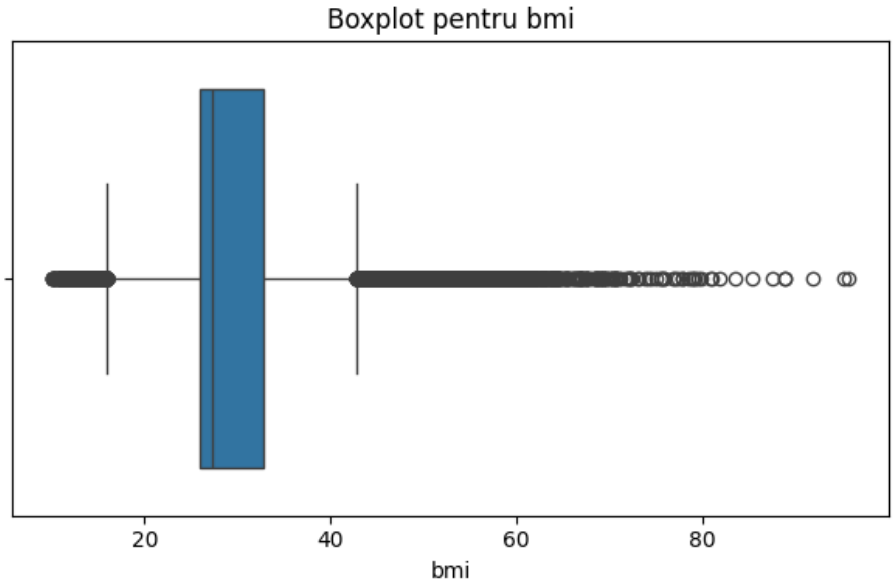
Setul de antrenament:



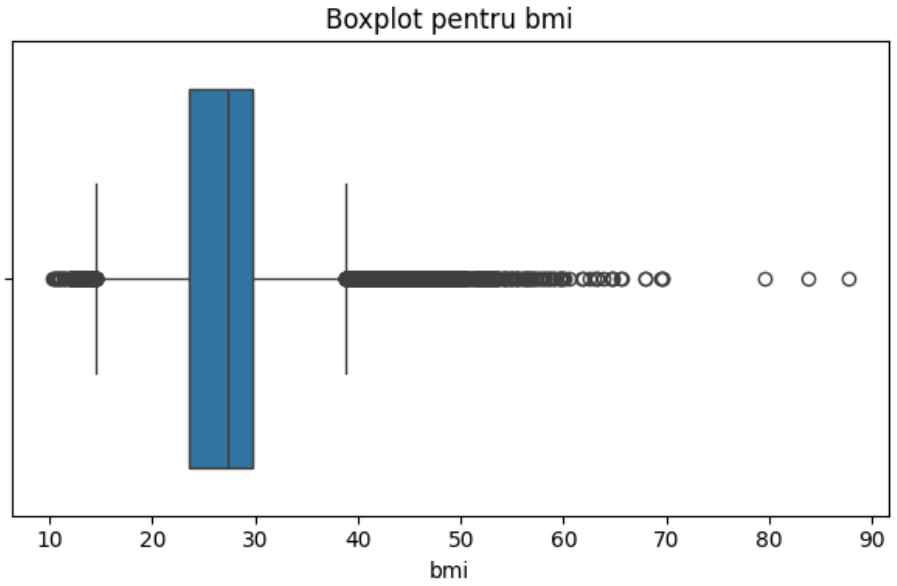
Setul de test:

- În setul de test avem câțiva outliers pentru blood_glucose_level, un feature important! O valoare crescută indică un risc ridicat de diabet, deci probabil modelul de clasificare a determinat corect diagnosticul.

Setul de antrenament:

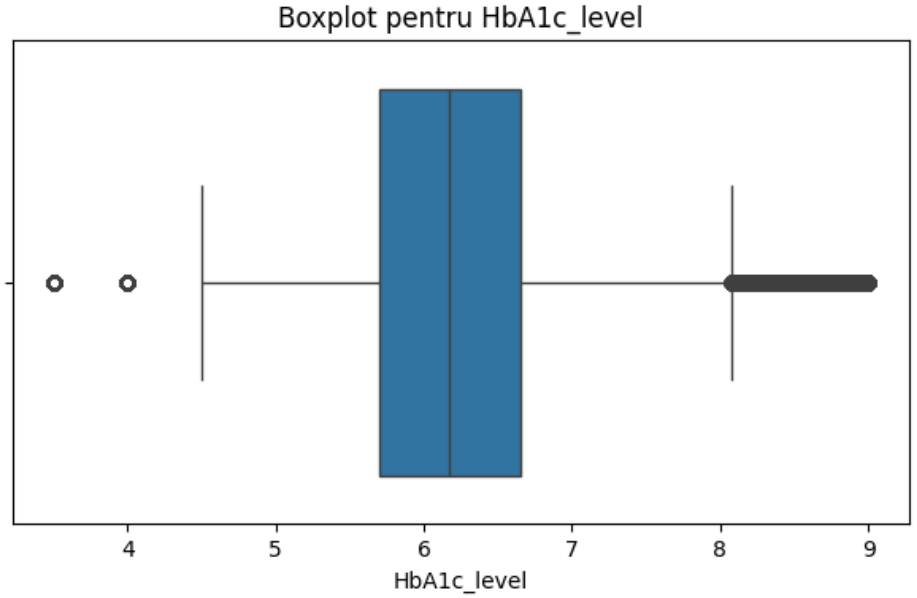


Setul de test:

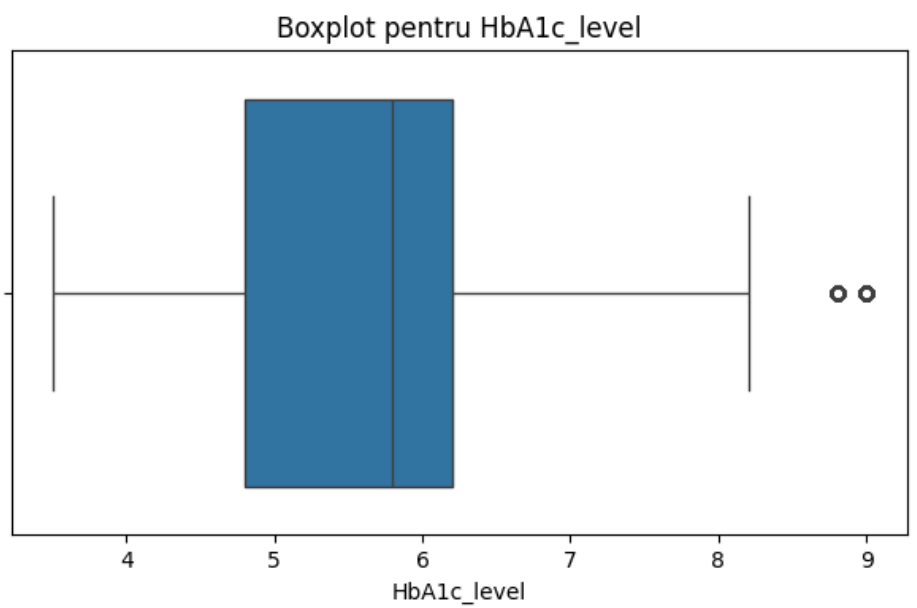


- Se poate observa că avem multi outliers în ambele cazuri! Pentru indicele de masă corporală, valorile mai mari de 60 sunt extrem de mari și indică , cel mai probabil, greșeli. Vom șterge toate datele pentru pacienții cu BMI mai mare de 60 pentru a crește acuratețea modelului.

Setul de antrenament

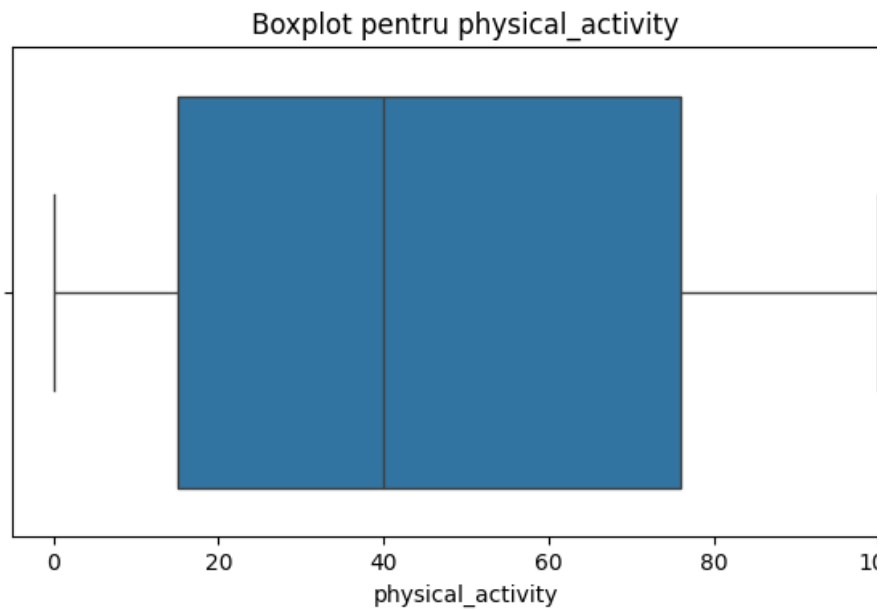


Setul de test

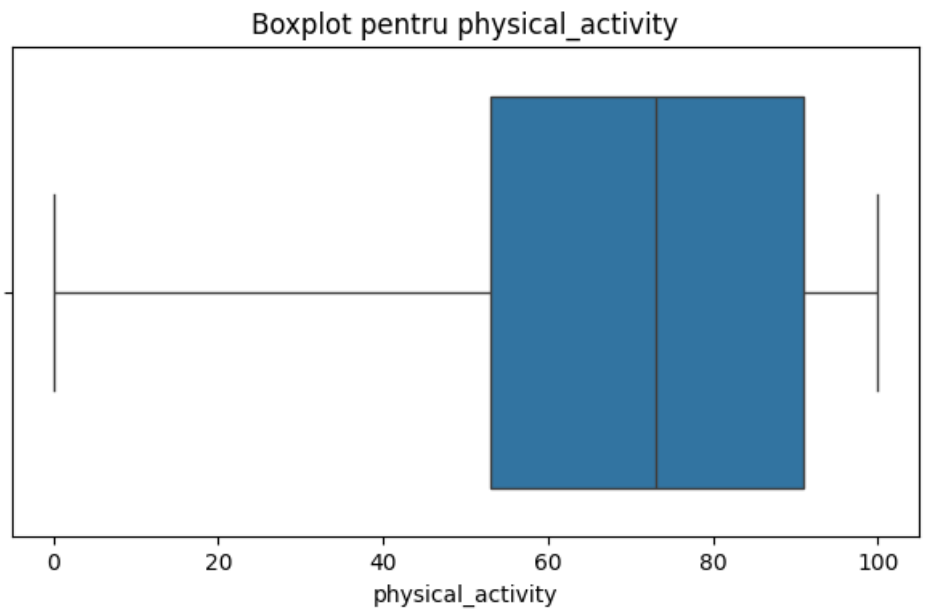


- Se observă 2 outliers în partea stângă pentru setul de antrenament și mai mulți la extremitatea din dreapta, iar pentru setul de test, 2 outliers în dreapta. Datele acestea nu sunt îngrijorătoare, întrucât un nivel mediu de 8-9 al glicemiei poate fi un caz extrem al diabetului. Matricea de corelație de la finalul documentului evidențiază că nivelul glicemiei este strâns legat de diagnosticul de diabet. (este unul din testele esențiale luate în considerare de doctori)

Setul de antrenament:



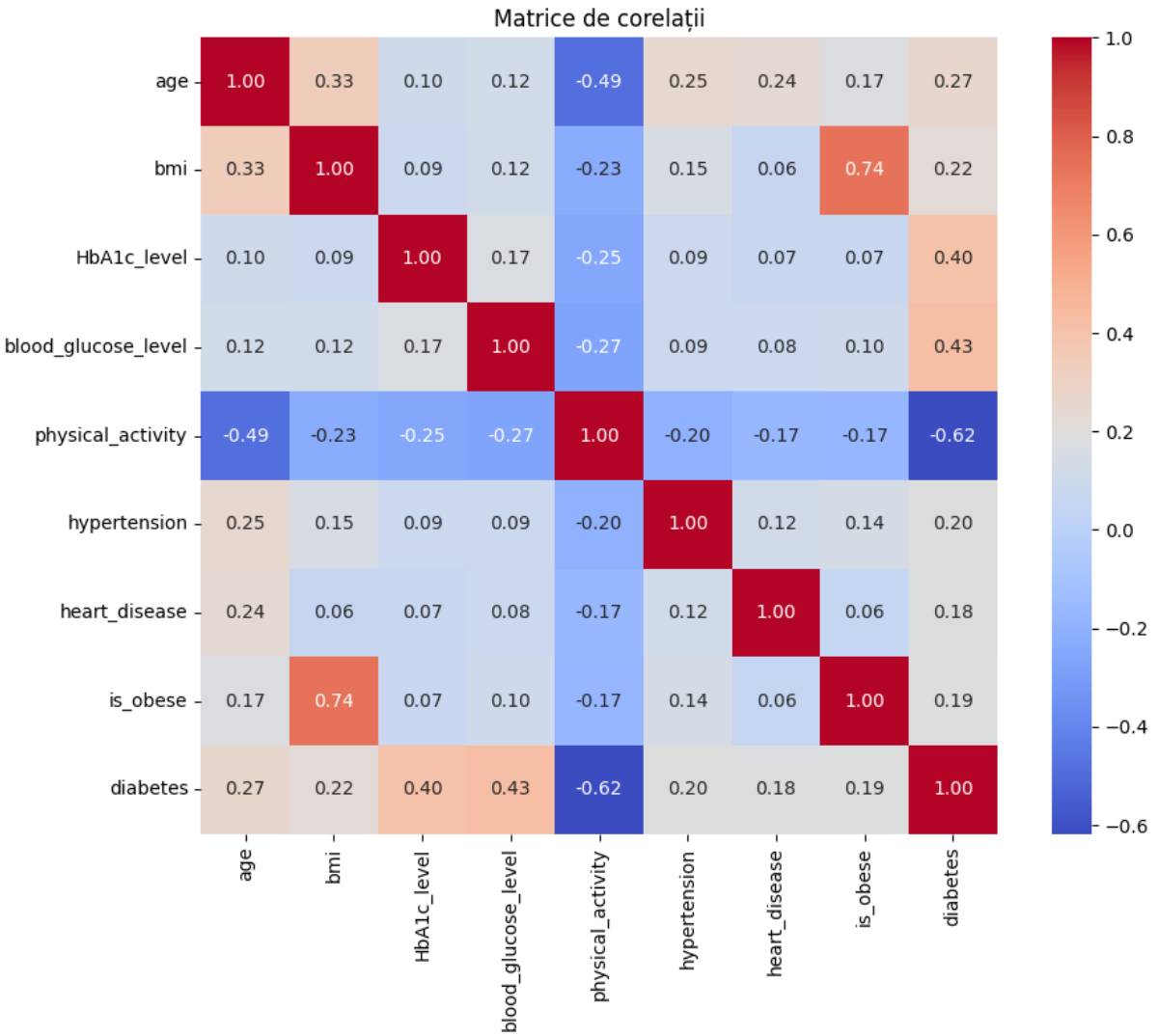
Setul de test:



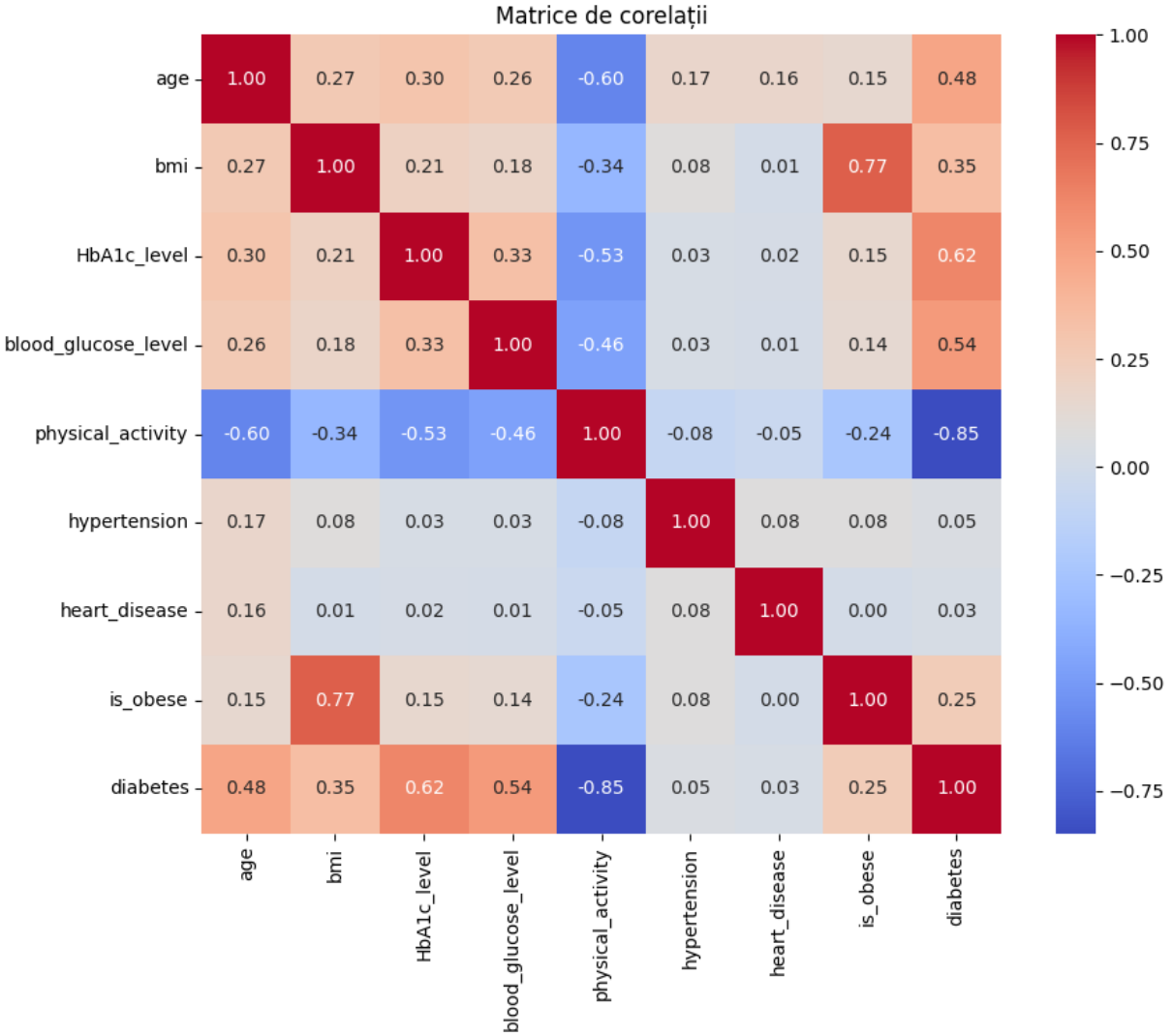
- Nu avem outliers dar observăm un range mai mare pe setul de antrenament și valori în general mai mari pe setul de test (intre procentele 25 si 75). Acest lucru poate fi cauzat de faptul că în setul de test sunt mai multe persoane mai tinere, iar vârsta este unul dintre criteriile pe baza căruia am generat acest feature.

e) Matricele de corelații (heatmap):

Setul de antrenament:



Setul de test:

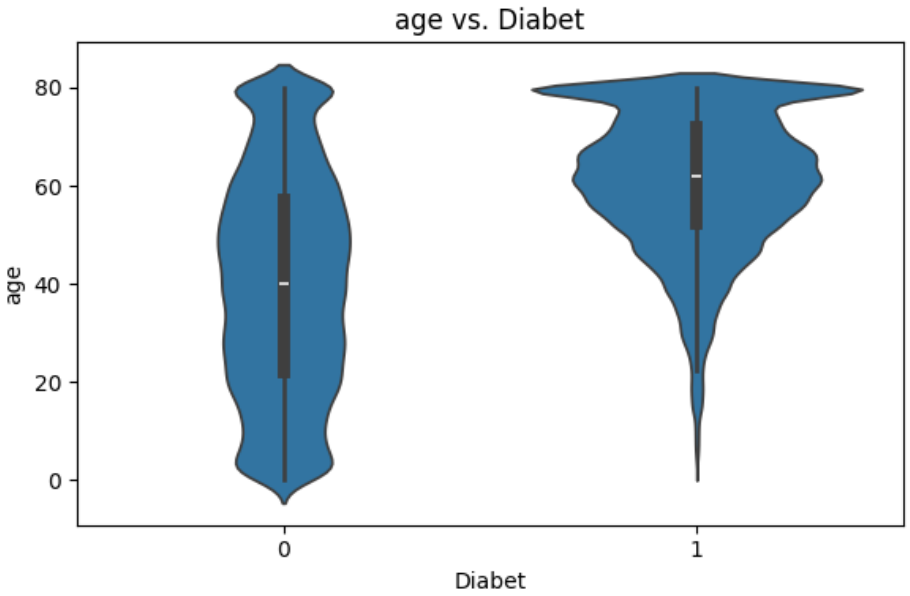


Analizând cele două matrice, putem trage următoarele concluzii:

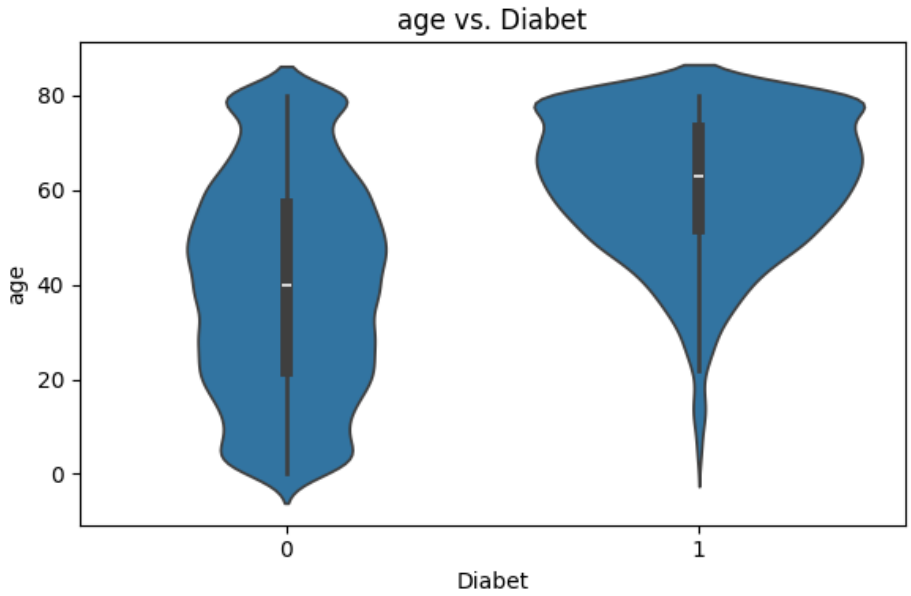
- Există o corelație clară între nivelul de activitate fizică și diabet, respectiv vârstă – acest lucru era de așteptat, deoarece când am generat acest feature, ne-am bazat pe vârsta persoanei și, pentru că am vrut să „ajut" modelul puțin, pe diagnosticul de diabet. Așadar, o relație clară între activitatea fizică și diabet/vârstă este normal.
- De asemenea, pentru că is_obese a fost construit în funcție de BMI, cu adăugarea a puțin noise, observăm că și acestea sunt în corelație.
- Așa cum am spus mai devreme, nivelul HbA1c (glicemia) este un factor foarte puternic în determinarea diabetului. Această corelație se observă pe matrice și confirma relevanța acestor factori în clasificare.
- Același lucru poate fi spus și despre nivelul de glucoză din sânge, pentru care observăm din nou o corelație mare cu prezența diabetului. Din nou, acest lucru confirmă faptul că modelul nostru funcționează corect.
- Alte corelații, precum cea dintre nivelul Hb1Ac și activitatea fizică, sau cea dintre vârstă și diabet sunt influențate, probabil, de feature-urile introduse de noi. Cum ele nu sunt îngrijorător de mari, putem trage concluzia că ele nu afectează modelul nostru, ci mai degrabă sunt conectate cu mai multe feature-uri, ceea ce generează rezultatul din matricea de corelație.

f) Analiza relațiilor cu variabila țintă (violin plots):

Setul de antrenament:

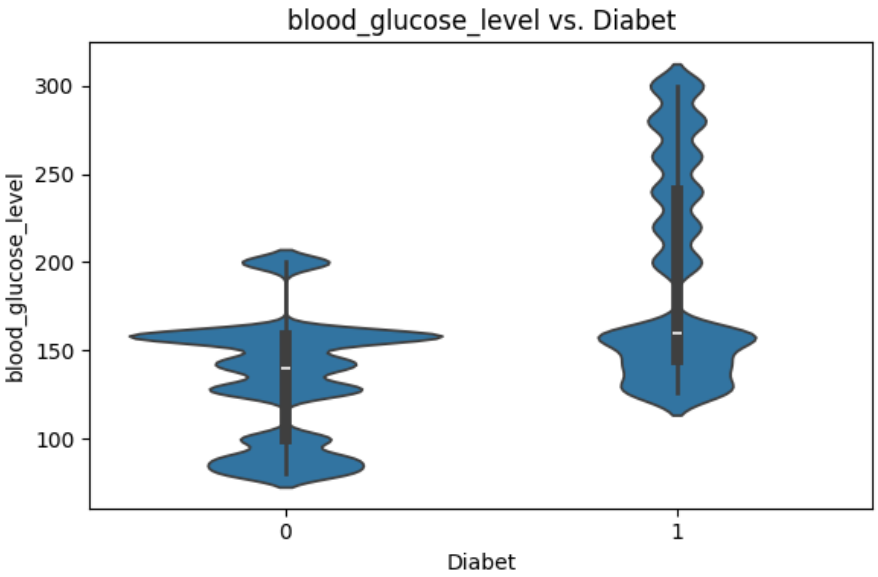


Setul de test

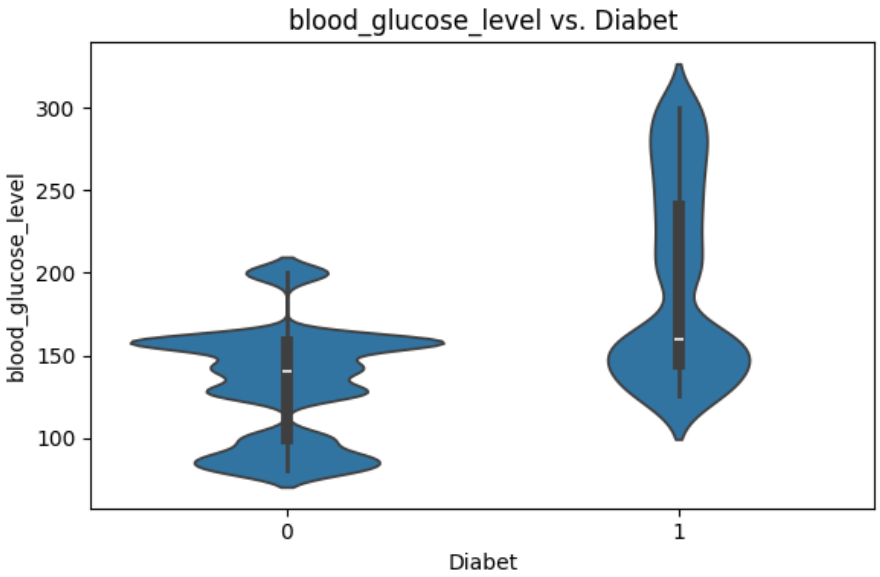


- Din ambele grafice observăm că riscul de diabet crește odata cu înaintarea în vârstă.

Setul de antrenament:

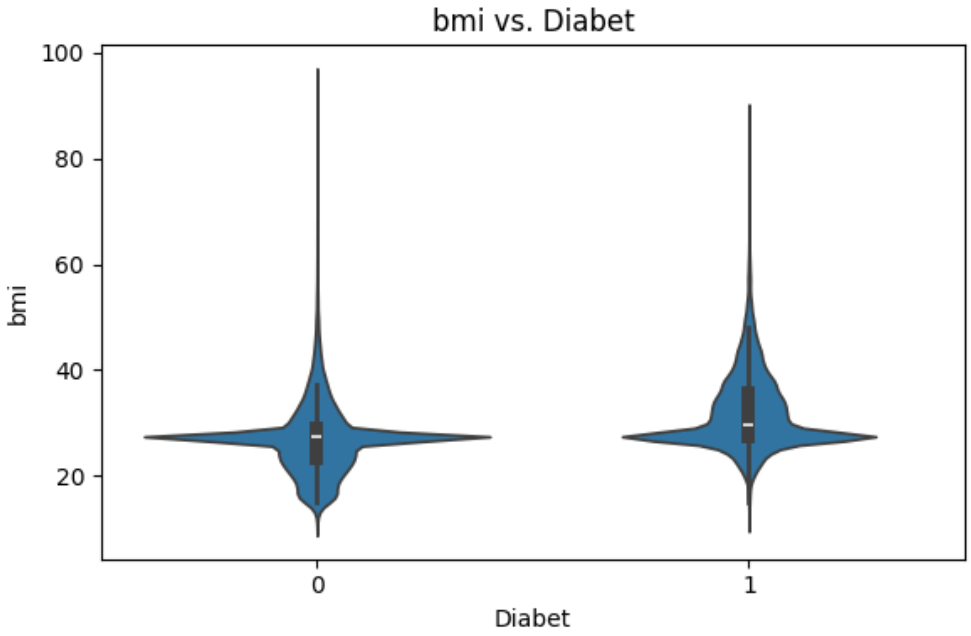


Setul de test:

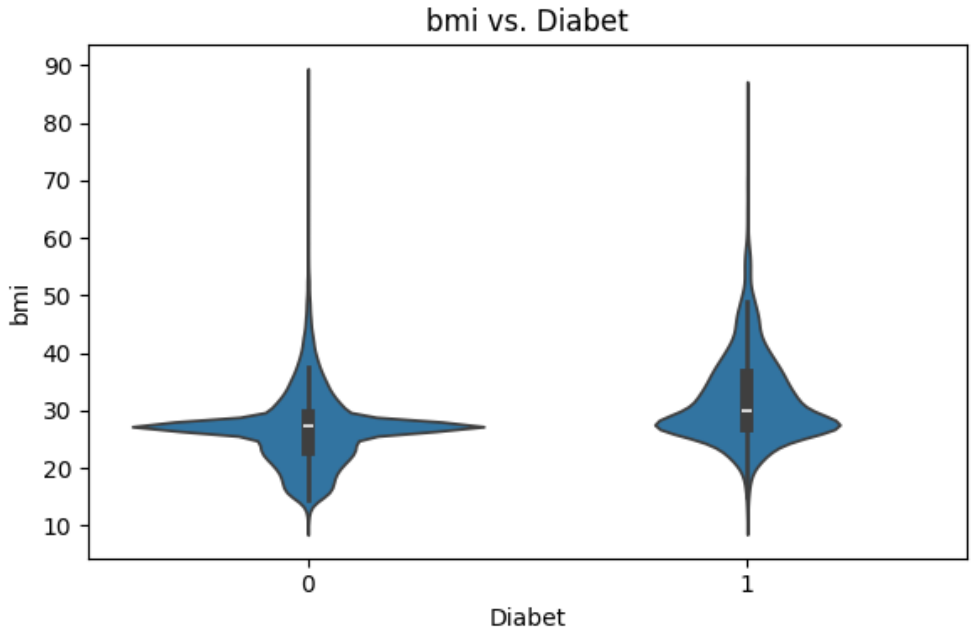


- Se observă o legătura clara între nivelul de glucoză din sânge și diabet! Acest lucru ne confirmă că datele sunt corecte și vor fi folositoare pentru antrenarea corectă a modelului.

Setul de antrenament:

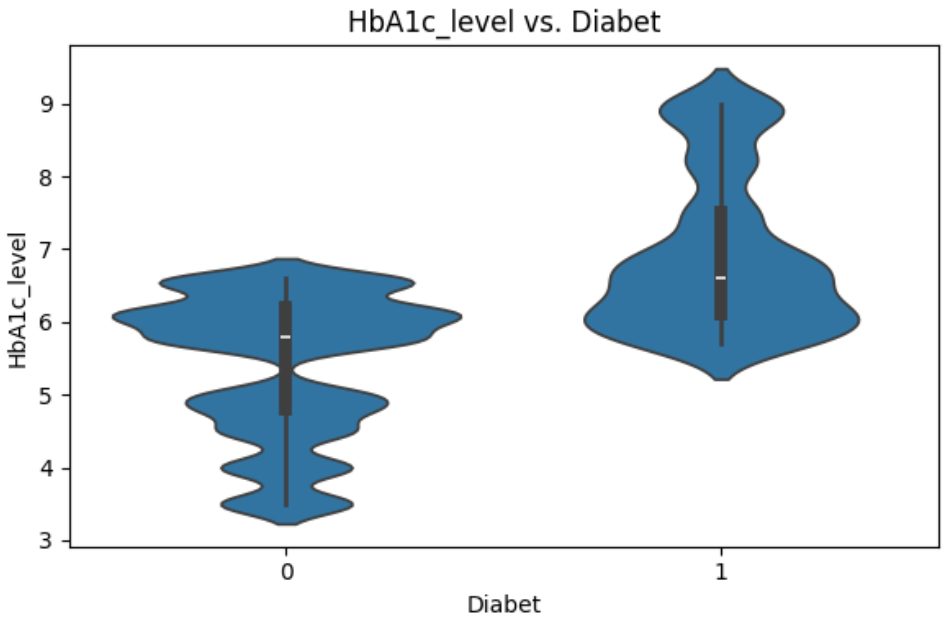


Setul de test:

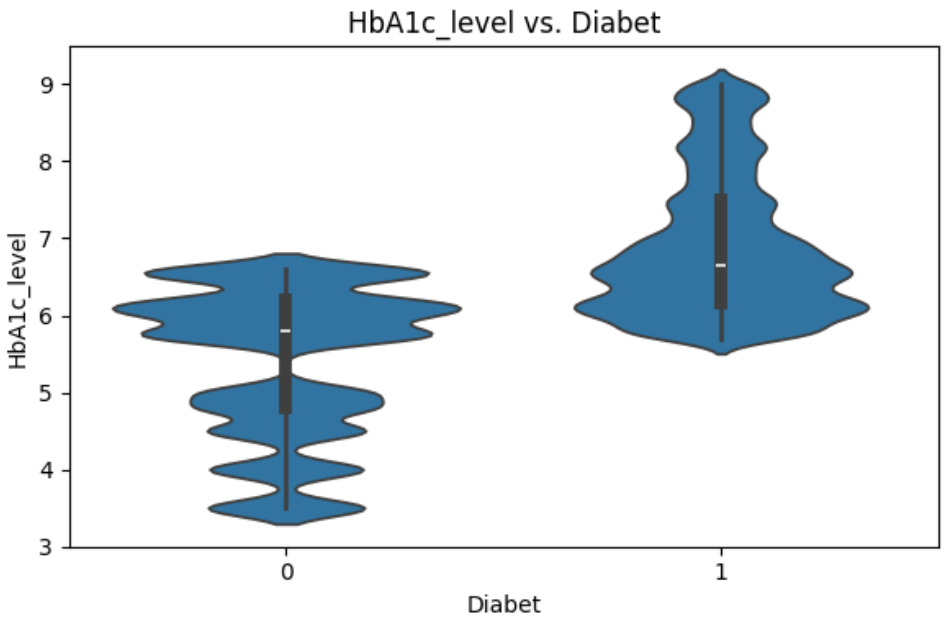


- Observăm o legătură subtilă între indicele de masă corporală și diabet. Această legătură va fi accentuată de feature-ul is_obese, cu ajutorul căruia modelul nostru va putea determina mai bine diagnosticul.

Setul de antrenament:

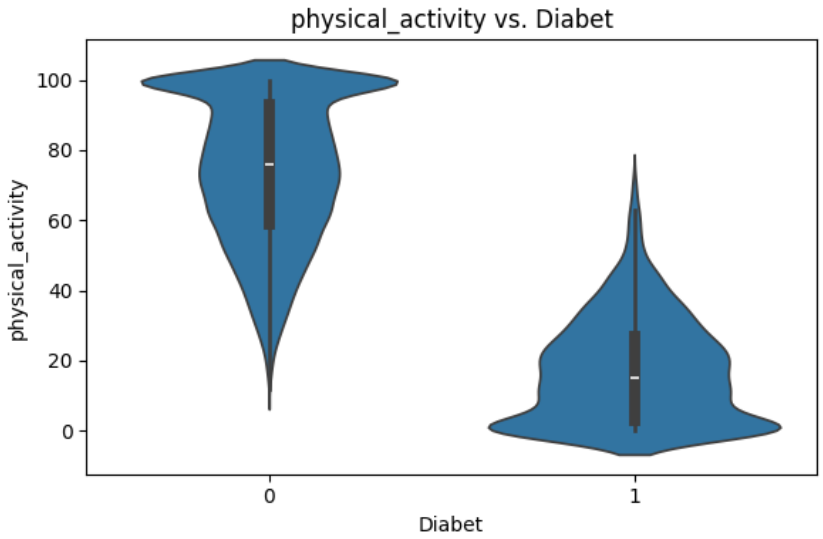


Setul de test:

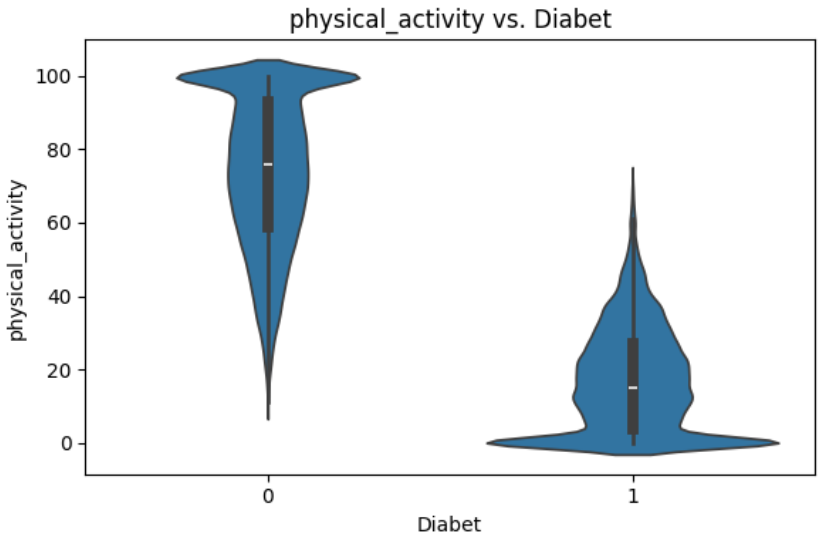


- Observăm din nou o legătură clară între nivelul de Hb1Ac și diabet, la fel cum am concluzionat și din matricea de corelații.

Setul de antrenament:



Setul de test:



După cum era de așteptat, din moment ce am avut control asupra generării acestui feature, cu cât nivelul de activitate fizică este mai mare, cu atât șansele de diabet sunt mai mici.

Antrenarea și evaluarea modelului:

După descărcarea datelor de pe kaggle, folosind get_dataset_and_path.py, încărcăm datele și le prelucrăm în dataset_manip.py.

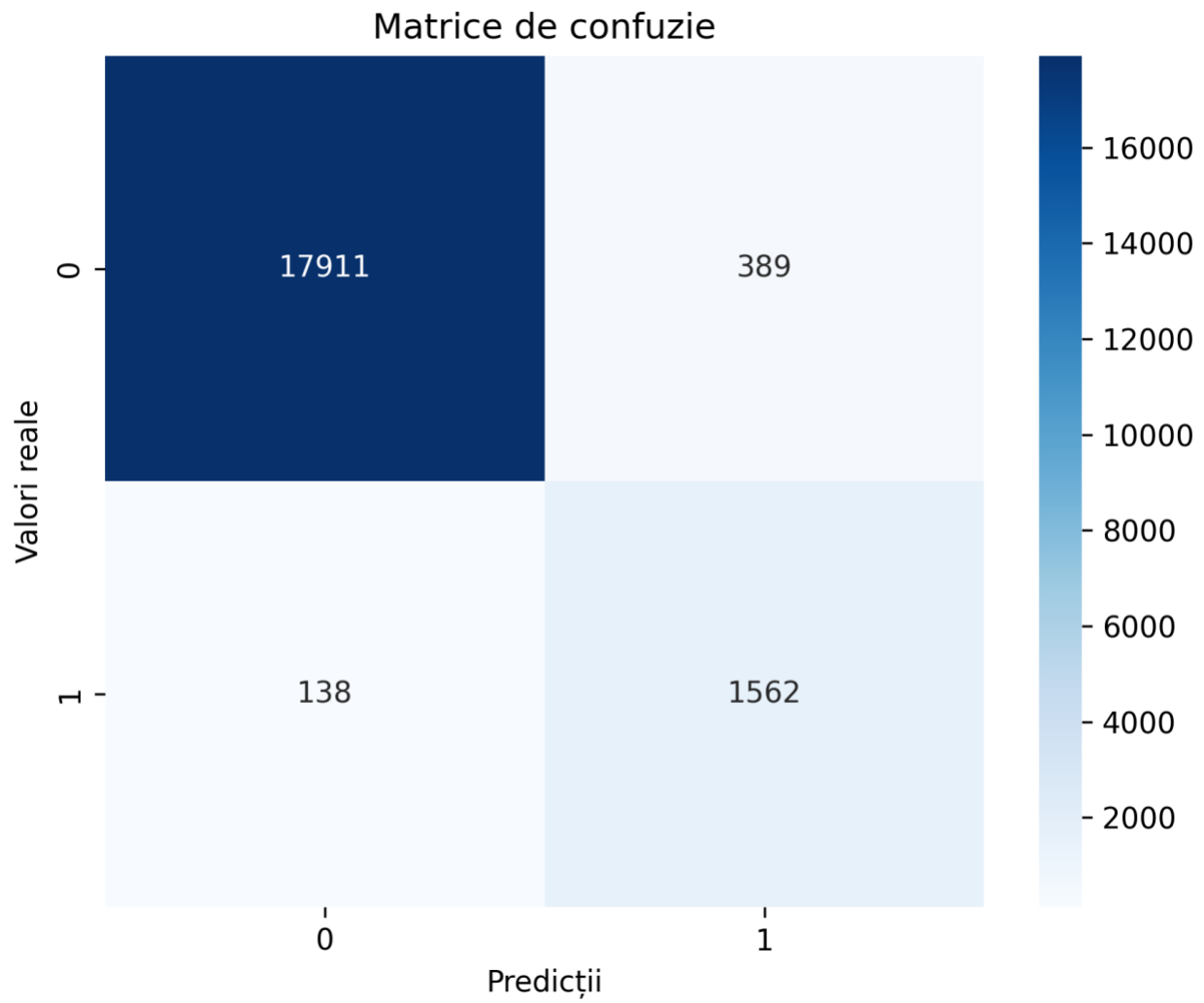
Transformăm toate datele în valori numerice, calculăm procentul de valori nule, adăugăm noi feature-uri, eliminăm outliers, facem split-ul, normalizăm și antrenăm datele. Salvăm modelul și scaler-ul pentru a le putea folosi la interfața grafică.

La final, daca user-ul dorește, creăm matricele de confuzie și grafice pentru analiza erorilor.

În analyze_and_get_graphs.py, generăm toate celelalte grafice pe care le vom folosi, iar în graphic_interface.py rulăm interfața grafică pentru modelul nostru.

***IMPORTANT: În arhivă sunt incluse doar data_frame-urile pentru train și test. Pentru a rula tot codul, este necesar să descărcați data_set-ul de pe kaggle și să îl încărcați în folderul curent!**

În continuare vom analiza matricea de confuzie și erorile.

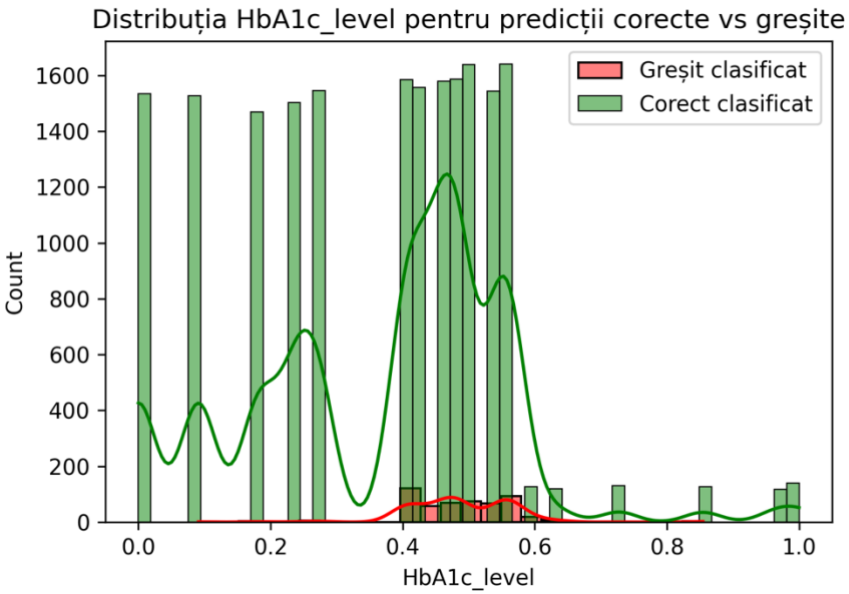


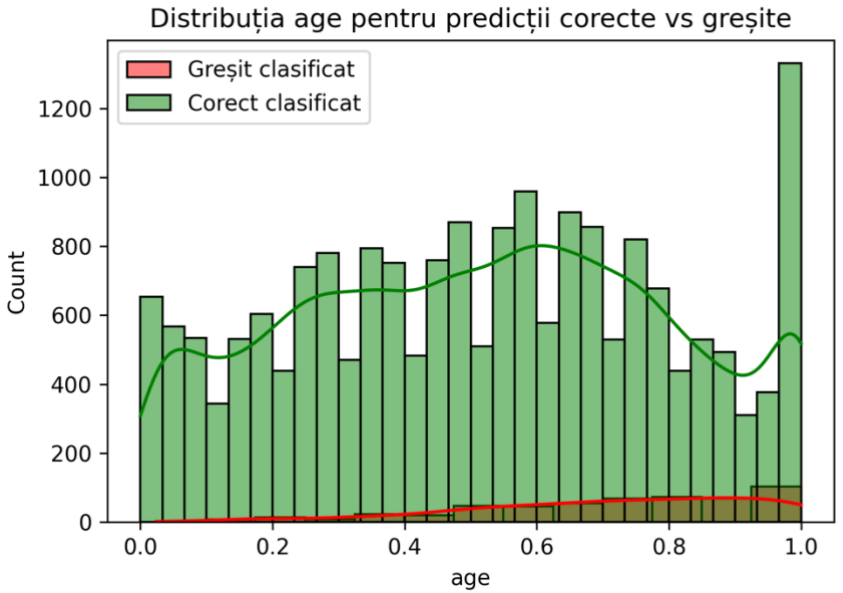
Observăm că modelul prezice mai des greșit diagnosticul de diabet. Acuratețea mare a modelului, 0.97, se datorează setului mare de date și puținelor diagnostice de diabet din acesta. Modelului ii este mai ușor să prezică faptul că un individ nu are diabet, având o precizie de 0.99, decât să prezică diagnosticul de diabet, cu o precizie de 0.81.

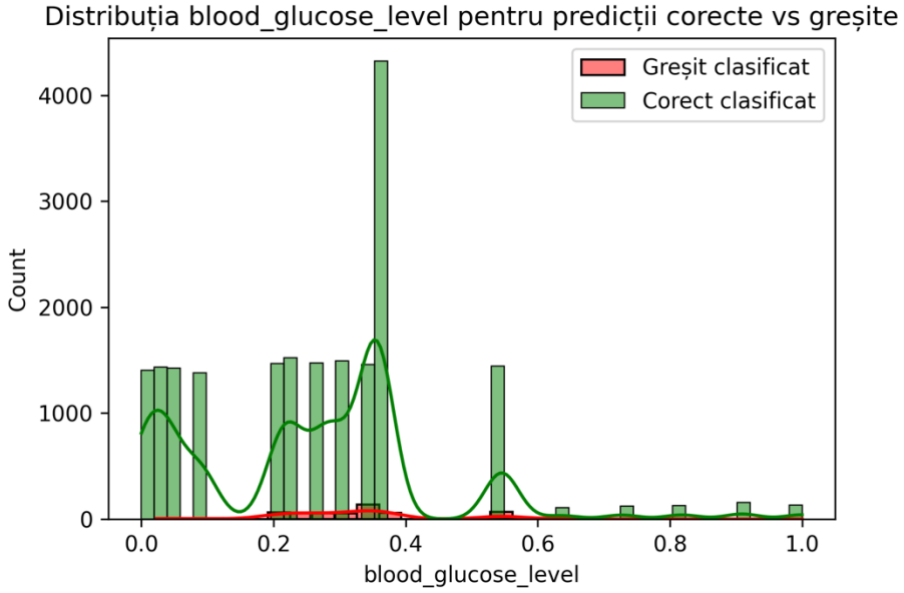
Raportul de clasificare al modelului:

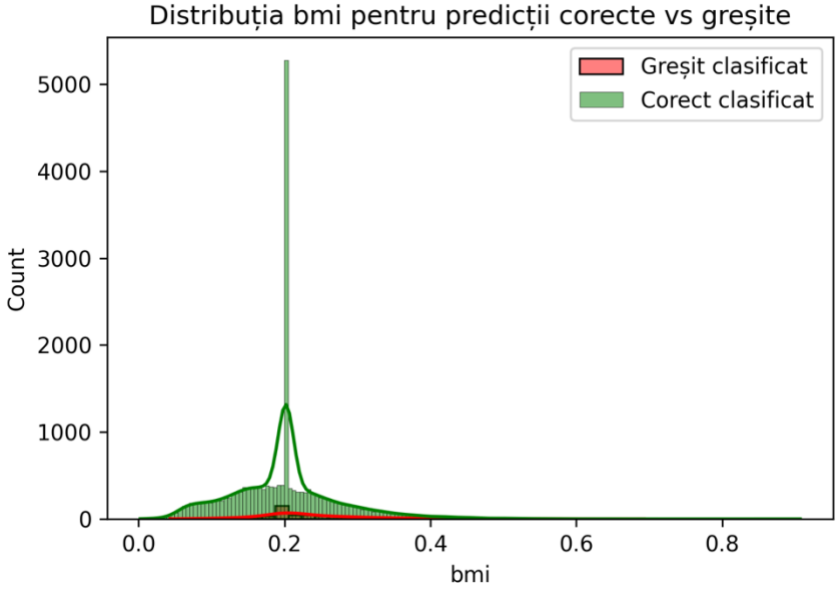
	Precision	Recall	F1-Score	Support
Not diabetes	0.99	0.98	0.99	18248
Diabetes	0.81	0.90	0.86	1693
Accuracy	-	-	0.97	19977
Macro Average	0.90	0.94	0.92	19977
Weighted Average	0.98	0.97	0.97	19977

Analiza erorilor:









Concluzii:

Pentru nivelul de glucoză, cel de HbA1c și pentru BMI, modelului i-a fost greu să clasifice pentru valorile cele mai comune, ceea ce este de înțeles, deoarece acestea sunt valorile normale, din care nu se poate trage o concluzie clară asupra diagnosticului.

În cazul vârstei, se observă din nou intervenția noastră, cu adăugarea feature-ului de physical activity, care probabil a determinat modelul să fie mai predispus în a diagnostica împotriva diabetului pentru persoanele mai tinere și vice versa, ceea ce nu are un fundament științific concret.

Concluzia finală:

Modelul prezice diabetul cu o acuratețe ridicată (97%), dar are dificultăți în identificarea cazurilor pozitive, afectând precizia pentru pacienții cu diabet. Rezultatele reflectă influența datelor și a caracteristicilor artificiale adăugate (is_obese, physical_activity).