Costache Ovidiu-Stefan 313CC
Partea I
Problema de Clasificare - Diabet

Setul de date pe care l-am utilizat in proiect se numeste “diabetes_prediction_dataset.csv”. Setul contine date medicale referitoare la pacienti cu scopul de a prezice daca un pacient are sau nu diabet pe baza mai multor caracteristici.
Dataset-ul este preluat de pe Kaggle (https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset).
Fiecare linie reprezinta un pacient, iar coloanele includ caracteristici medicale.
Pe parcursul proiectului am completat valorile lipsa din coloana “bmi” cu media acesteia. Variabilele categorice au fost codificate numeric pentru a putea realiza predictii folosindu-ma de LabelEncoder. Setul de date l-am impartit in train.csv si test.csv (80% - 20%). Spre final am standardizat datele numerice ca in laborator, pentru a antrena modelul.

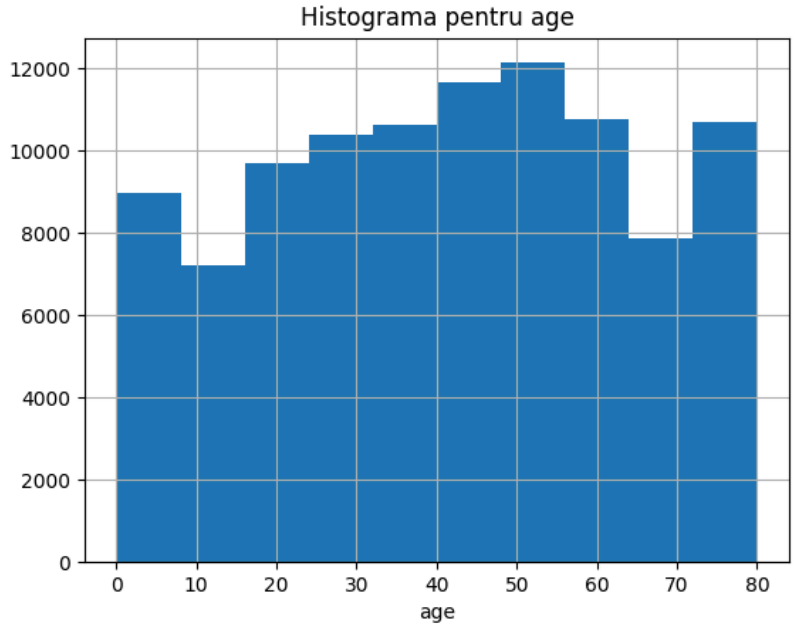
Citirea datelor:
Am folosit biblioteca Pandas pentru a incarca setul de date intr-un dataframe. Dupa incarcarea setului, am afisat primele randuri pentru a verifica structura datelor. Au fost extrase informatii pe baza setului de date, cum ar fi numarul de randuri si de coloane, tipurile de date pentru fiecare coloana si valorile unice din fiecare variabila.

Analiza valorilor lipsa:
Pentru a identifica valorile lipsa din fiecare coloana, am utilizat functia isnull() din Pandas impreuna cu sum(). Aceasta returneaza numarul total de valori lipsa de pe fiecare coloana.
Pentru coloana “bmi” am inlocuit valorile lipsa cu media de pe coloana (df['bmi'].mean()). Operatia se face cu functia fillna() ce are ca parametru inplace=True, pentru a modifica direct dataframe-ul original.

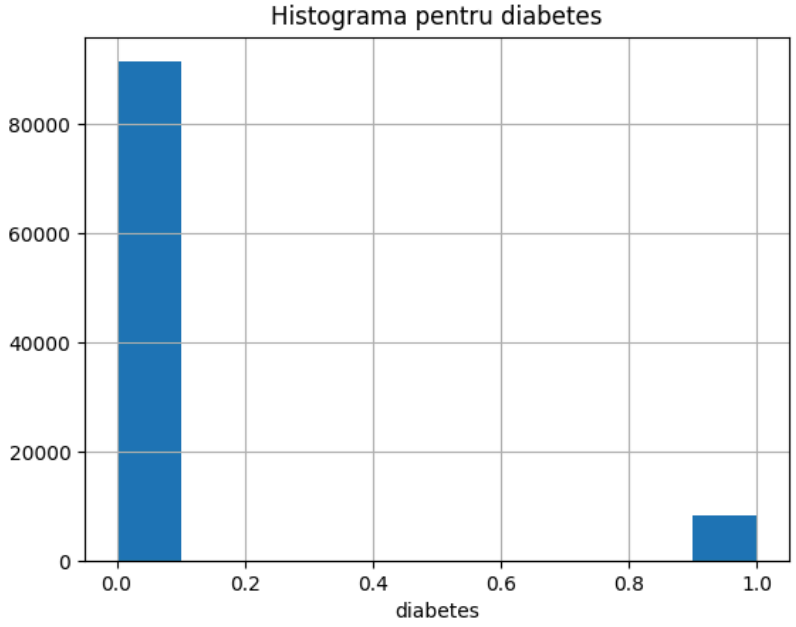
Statistici descriptive:
Statisticile descriptive le-am generat folosind describe() din biblioteca Pandas, cu optiunea include = “all”, pentru a obtine informatii pentru coloanele numerice, dar si pentru cele categorice.

Analiza distributiei variabilelor:

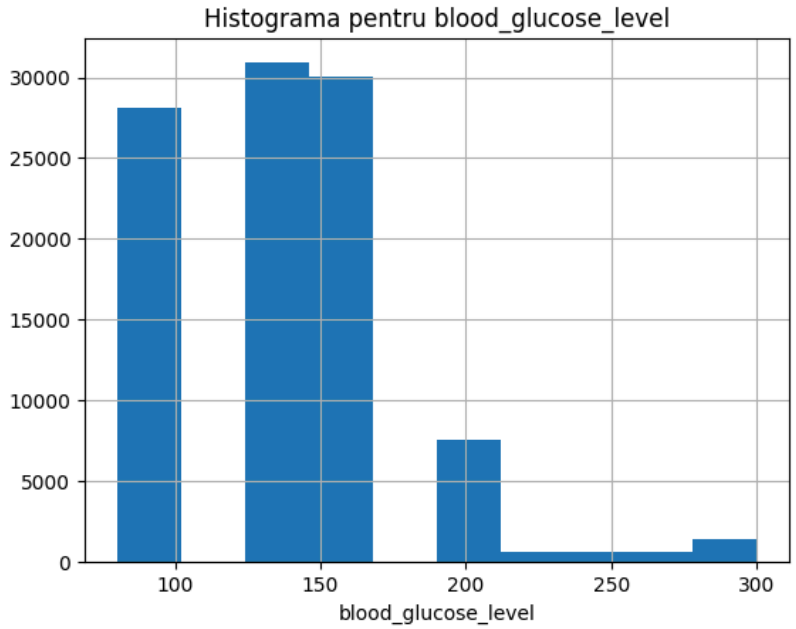
Histograme:
Ca sa identific coloanele numerice am folosit select_dtypes din Pandas cu parametrul include = [np.number]. Pentru fiecare astfel de coloana am generat o histigrama care arata frecventa valorilor variabilei din setul de date.
Voi raspunde la cele 3 intrebari pentru histogramele atasate:



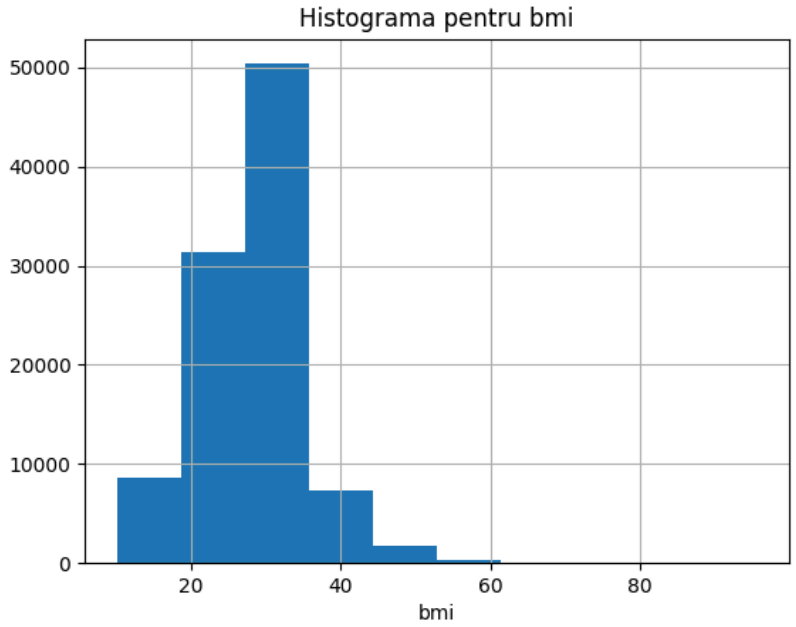
1) Distributia varstei este relativ uniforma, cele mai mari valori se regasec in intervalul de varsta [40, 60] ani;
2) Este posibil ca riscul de diabet sa creasca odata cu varsta;
3) Se pot standardiza valorile varstei pentru model



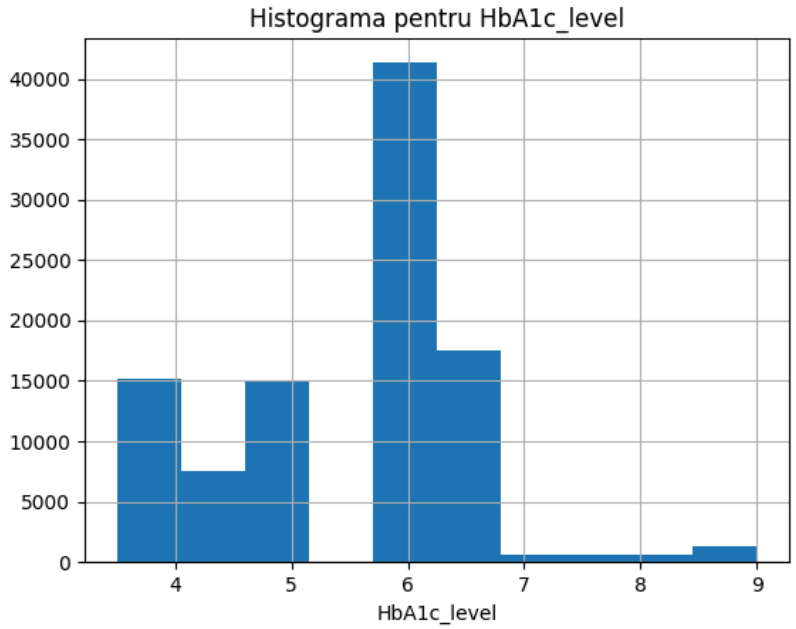
1) Majoritatea valorilor sunt zero, deci setul este dezechilibrat;
2) Modelul poate favoriza clasa majoritara;
3) Ar putea fi utilizate metode de echilibrare;



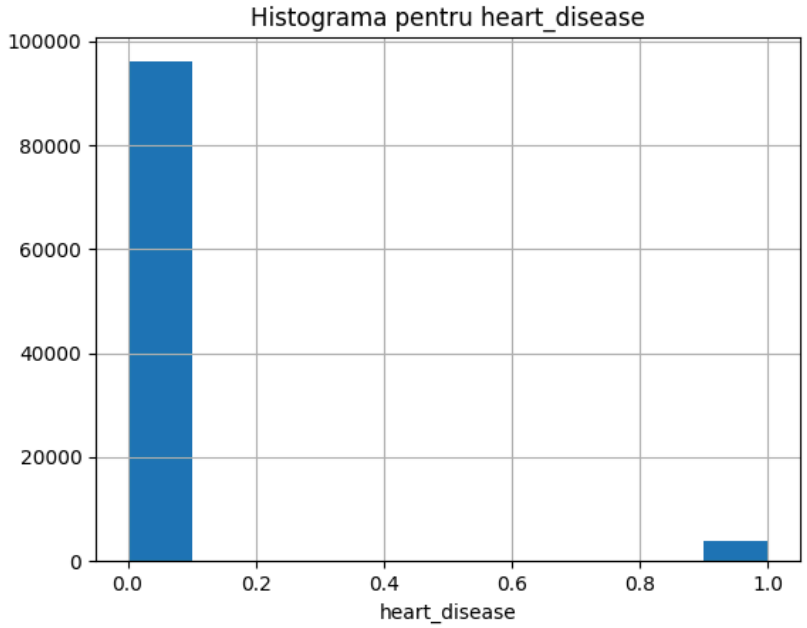
1) Majoritatea valorilor sunt concentrate sub 200;
2) Valorile foarte mari pot indica cazuri severe sau erori de introducere a datelor;
3) Analiza si tratarea outlier-ilor;



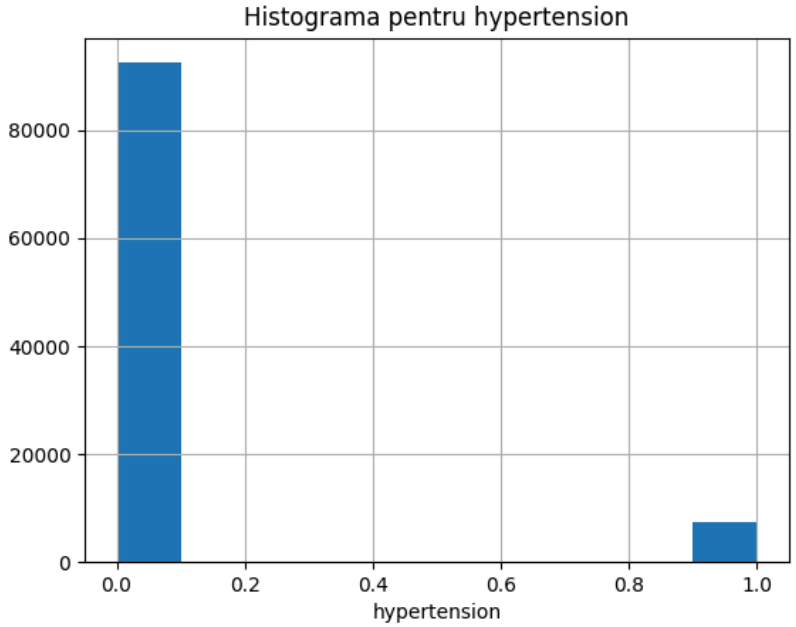
1) Valorile sunt concentrate intre 20 si 40, iar valorile foarte mari pot fi outlieri;
2) Valorile extreme (peste 60) pot fi erori sau cazuri rare;
3) Este utila identificarea valorilor extreme si tratarea acestora;



1) Exista doua grupe principale in jurul valorilor 4 si 6, dar exista si cateva cazuri extreme;
2) Valorile ridicate pot indica prezenta diabetului;
3) Standardizarea variabilei si verificarea valorilor extreme;

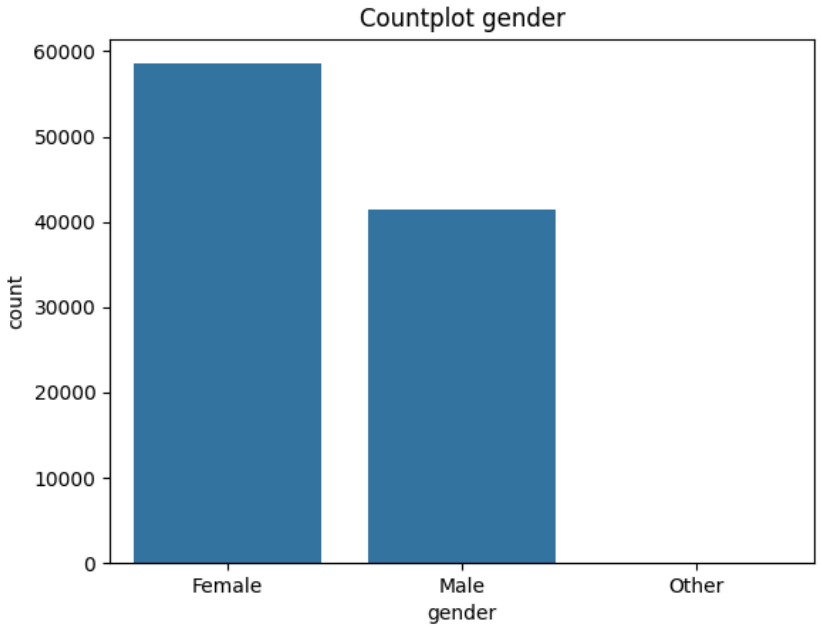


1) Majoritatea valorilor sunt zero, deci foarte putini pacienti au boala de inima;
2) Setul de date este dezechilibrat pe aceasta variabila;
3) Echilibrarea clasei daca aceasta variabila trebuie folosita ca tinta sau modelare;

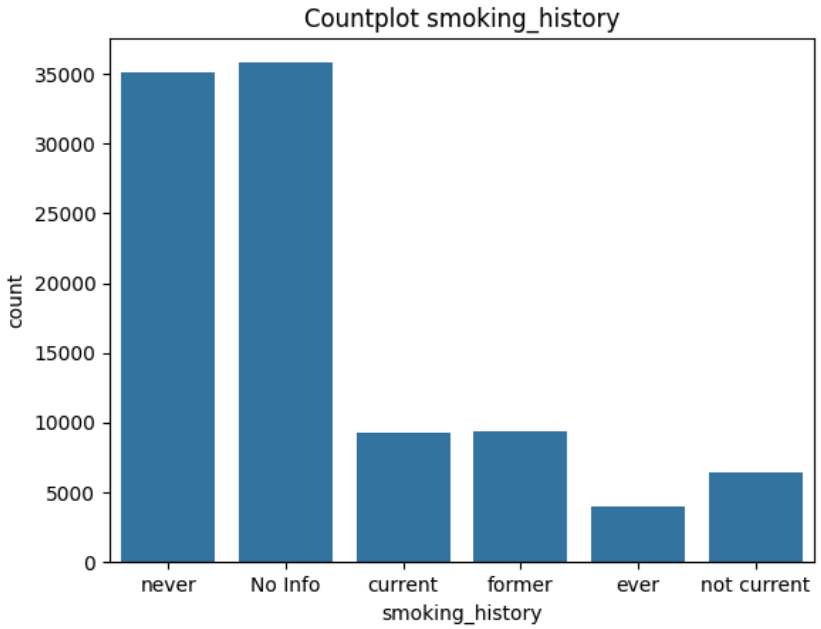


1) Majoritatea pacientilor nu au hipertensiune;
2) Dezechilibrul intre clase poate afecta modelele care folosesc aceasta variabila ca tinta;
3) Tehnici de echilibrare a claselor;

Grafice countplot:



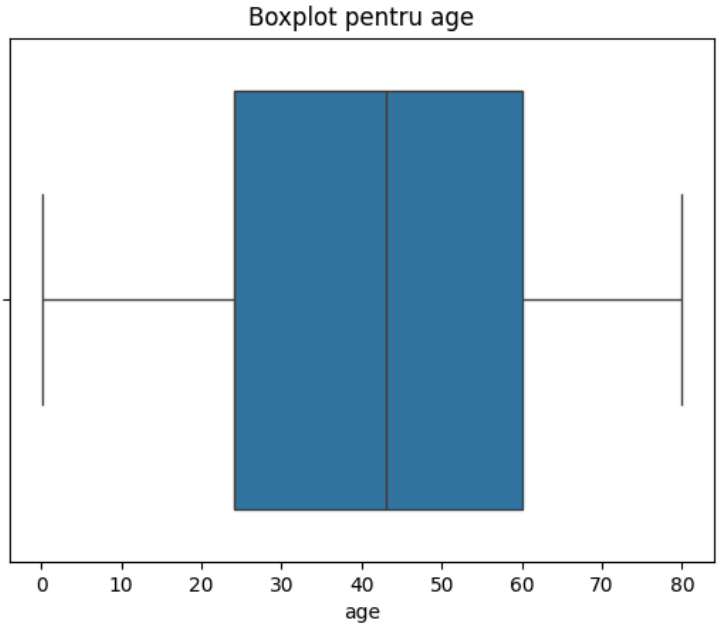
1) Sunt mai multe persoane de sex feminin decat masculin in acest set de date;
2) Distributia inegala poate afecta in diferite cazuri rezultatul analizei;
3) Excluderea valorilor rare (other)



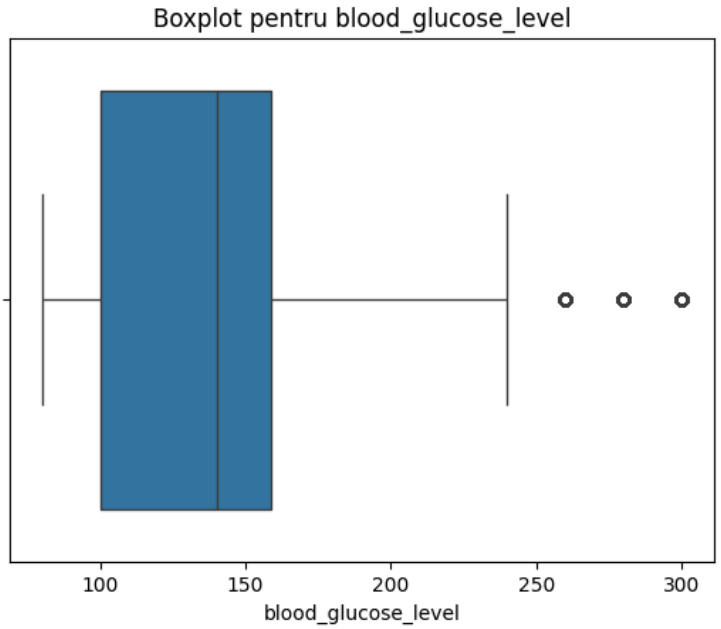
1) Cele mai multe inregistrari sunt in categoriilr “Never” si “No Info”;
2) Prezenta multor valori “No Info” poate afecta calitatea analizei si a modelelor;
3) Trebuie luata o decizie cu privire la categoria “No Info”;

Detectarea outlierilor:

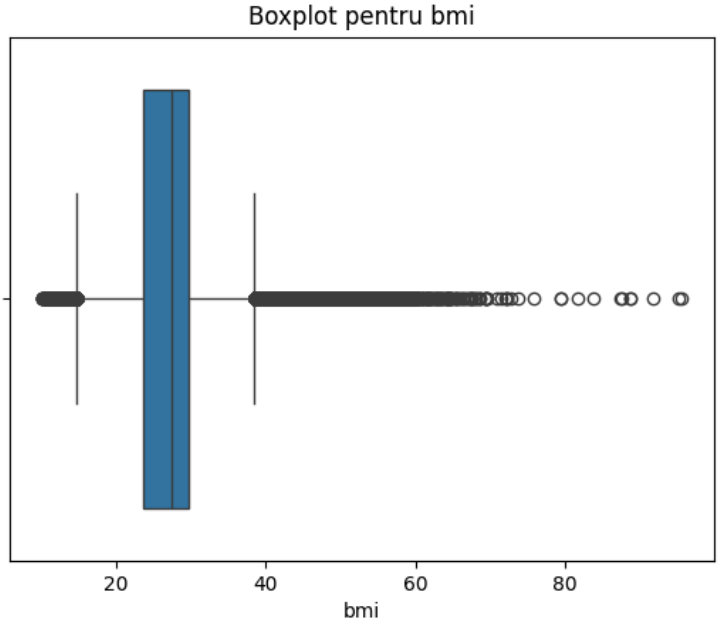
Creez boxplot-uri pentru fiecare coloana numerica din dataframe pentru a identifica vizual valorile atipice.



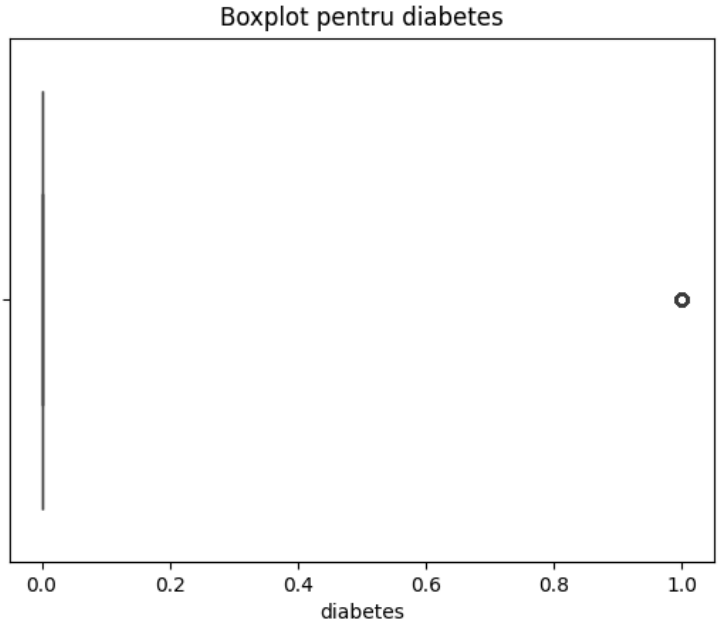
1) Valorile varstei sunt distribuite intre 0 si 80 de ani, fara outlieri evidenti;
2) Distributia varstelor pare relativ uniforma;
3) Nu sunt necesare interventii pentru outlieri;



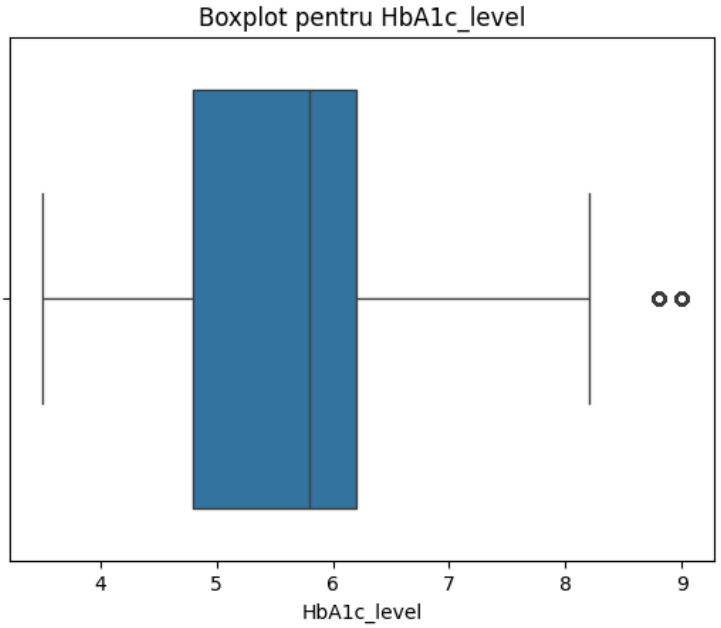
1) Majoritatea valorilor glicemiei sunt concentrare sub 250, dar exista cateva valori mult mai mari, vizibile ca puncte izolate;
2) Exista pacienti cu valori anormal de mari ale glicemiei, care pot indica posibile erori de masurare sau cazuri extreme;
3) Excluderea outlierilor sau transformarea acestora;



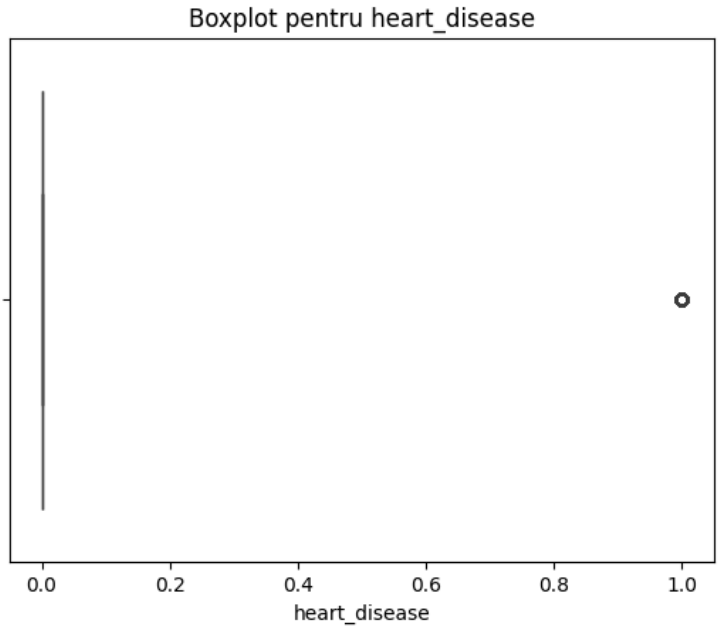
1) Exista foarte multi outlieri in parteadreapta a graficului, adica valori mari ale BMI;
2) Variabila contine numeroase valori extreme, unele posibil nereale sau calculate prost;
3) Eliminarea sau corectarea valorilor;



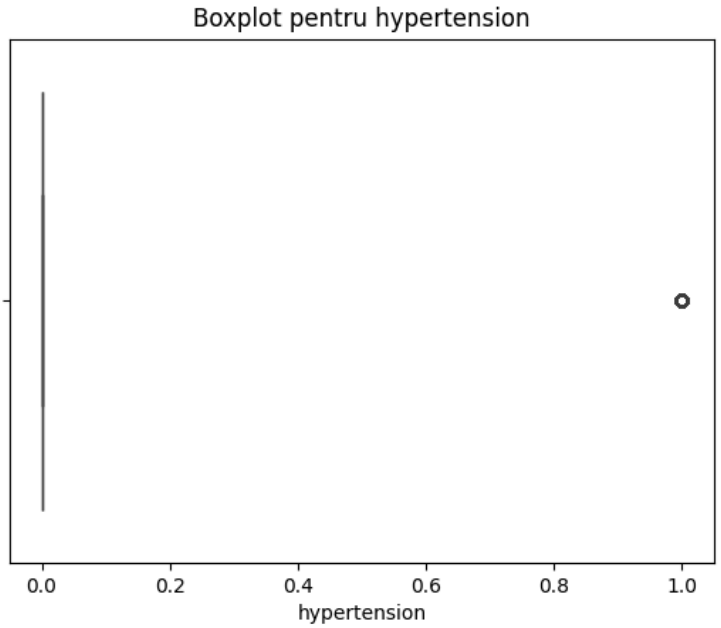
1) Variablia are doar doua valori posibile, 0 sau 1. Acest lucru este tipic pentru variabilele de acest fel, iar boxplot-ul arata cativa outlieri cu valoarea 1;
2) Nu apar outlieri propriu-zis, ci un dezechilibru intre clase;
3) Poate fi nevoie de metode de echilibrare a claselor;



1) Majoritatea valorilor este intre 3 si 8, insa exista si cateva valori izolatel;
2) Valorile izolate pot reprezenta cazuri severe sau erori de introducere a datelor;
3) Verificarea validitatii valorilor izolate sau eliminarea acestora;

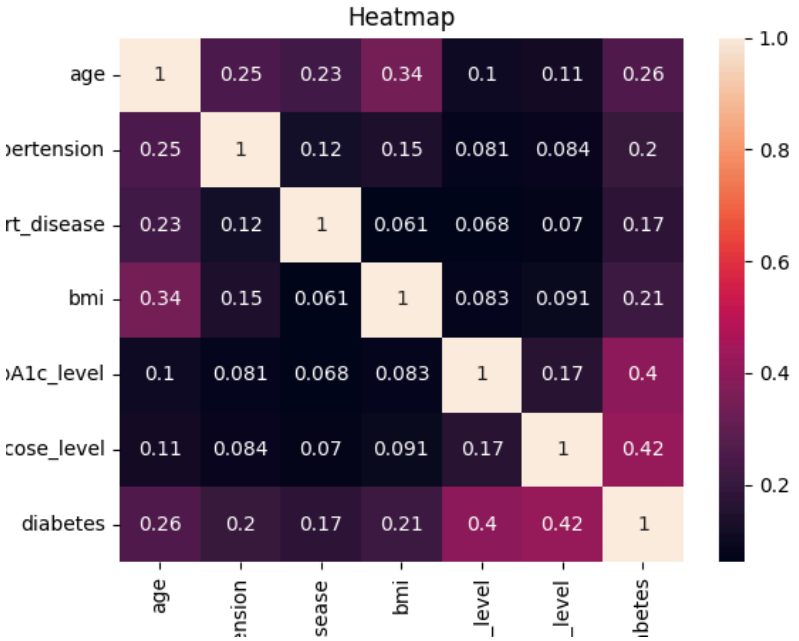


1) Variabilia binara, majoritatea este pe zero;
2) Dezechilibru intre clase;
3) Echilibrarea claselor;



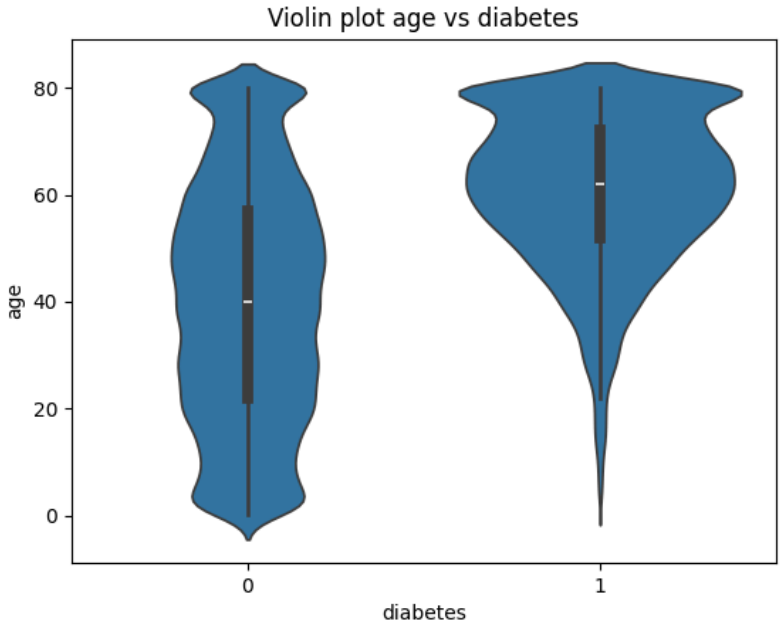
La fel ca la heart_disease;

Analiza corelatiilor:

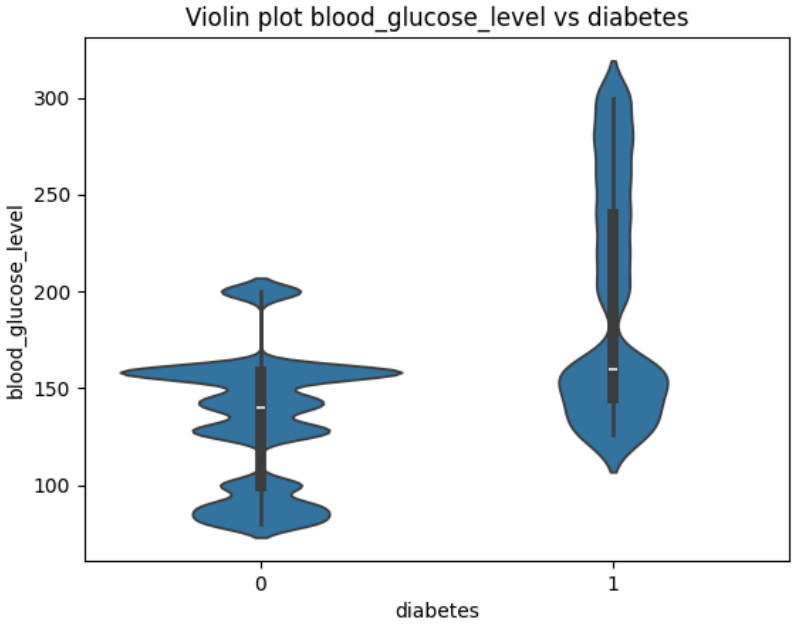


HbA1c_level si diabetes au o corelatie moderata (0.4), ceea ce este de asteptat, un nivel crescut al HbA1c indica diabet.
Blood_glucose_level si diabetes au de asemenea o corelatie moderata (0.42), ce confirma legatura dintre glicemie si diabet.
HbA1c_level si blood_glocose_level sunt corelate in proportie de 17%, nu foarte puternic, ceea ce arata ca desi ambele masoara controlul glicemiei, nu sunt complet redundante.

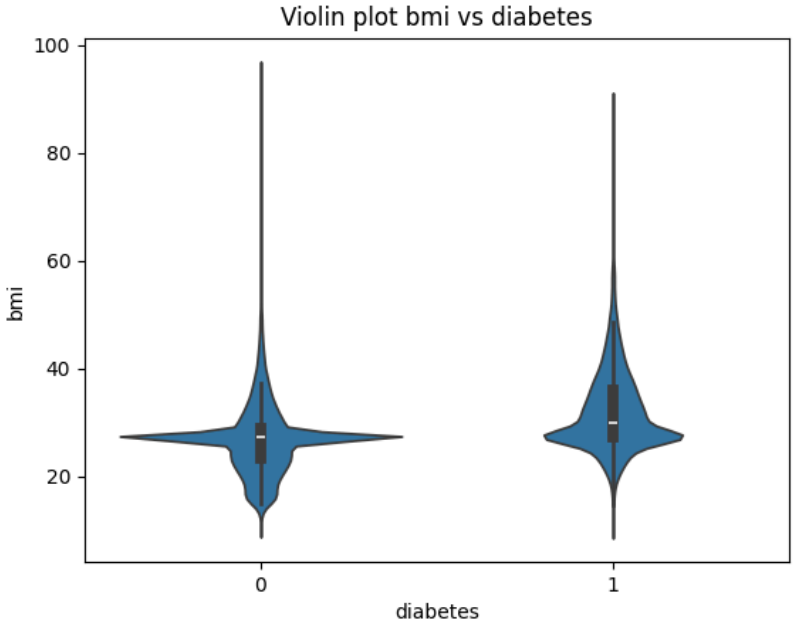
Analiza corelatiilor cu variablia tinta:
Pentru fiecare carcateristica numerica, mai putin diabetes, se genereaza un violin plot care arata distributia valorilor pentru pacientii cu si fara diabet. Folosesc sns.violinplot pentru compararea distributiilor si evidentiaza diferentele intre grupuri.



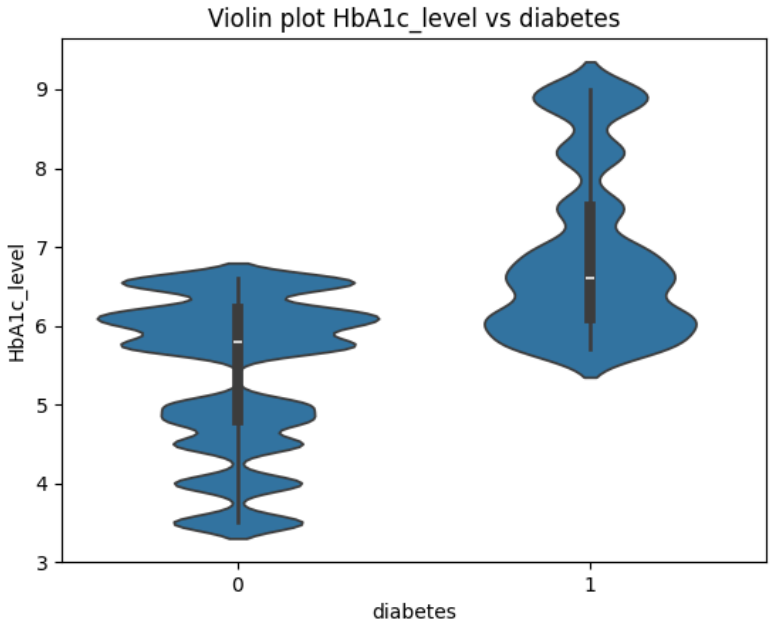
1) Personele cu diabet sunt in general mai in varsta;
2) Varsta crescuta este asociata cu un risc mai mare de diabet;
3) Variablia este relevanta pentru modelare;



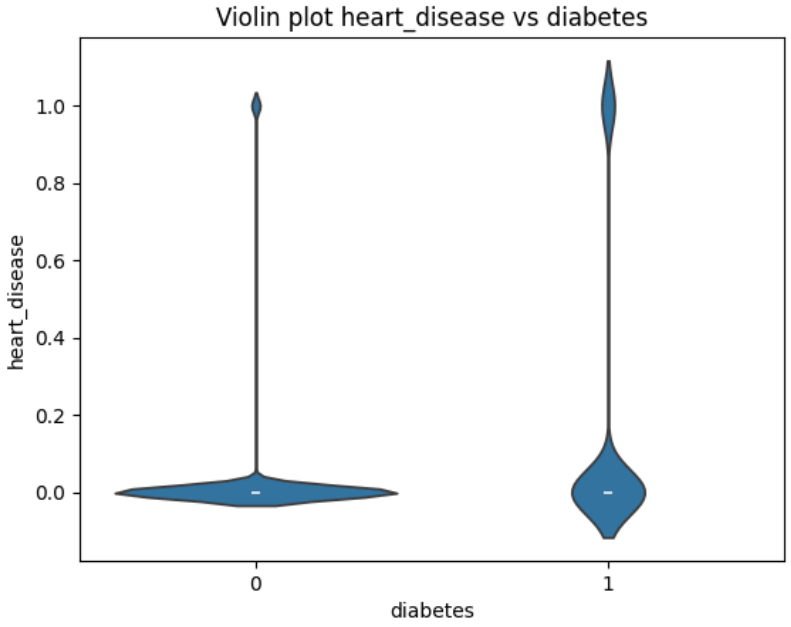
1) Persoanele cu diabet au valori mult mai mari ale glicemiei fata de cele fara; Blood_glucose_level diferentiaza clar intre clase si este un predictor important pentru diabetes;
3) Nu sunt necesare preprocesari;



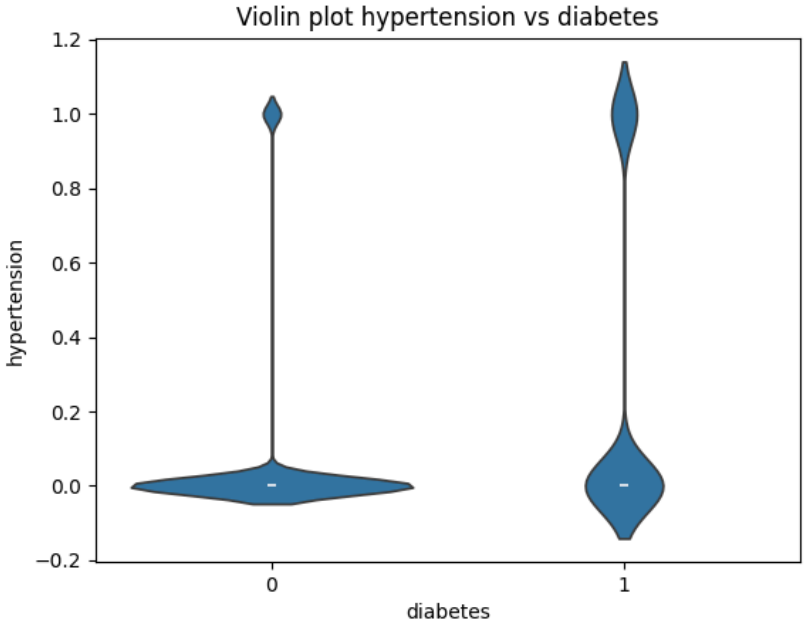
1) Persoanele cu diabet tind sa aiba un BMI mai mare;
2) BMI-ul crescut este asociat unui risc mai mare de diabet, dar diferenta nu este la fel de pronuntata ca la alti indicatori;
3) Nu sunt necesare preprocesari;



1) Persoanele cu diabet au Hb1Ac semnificativ mai mare;
2) Hb1Ac separa cele doua clase si este un bun indicator;
3) Nu sunt necesare preprocesari;

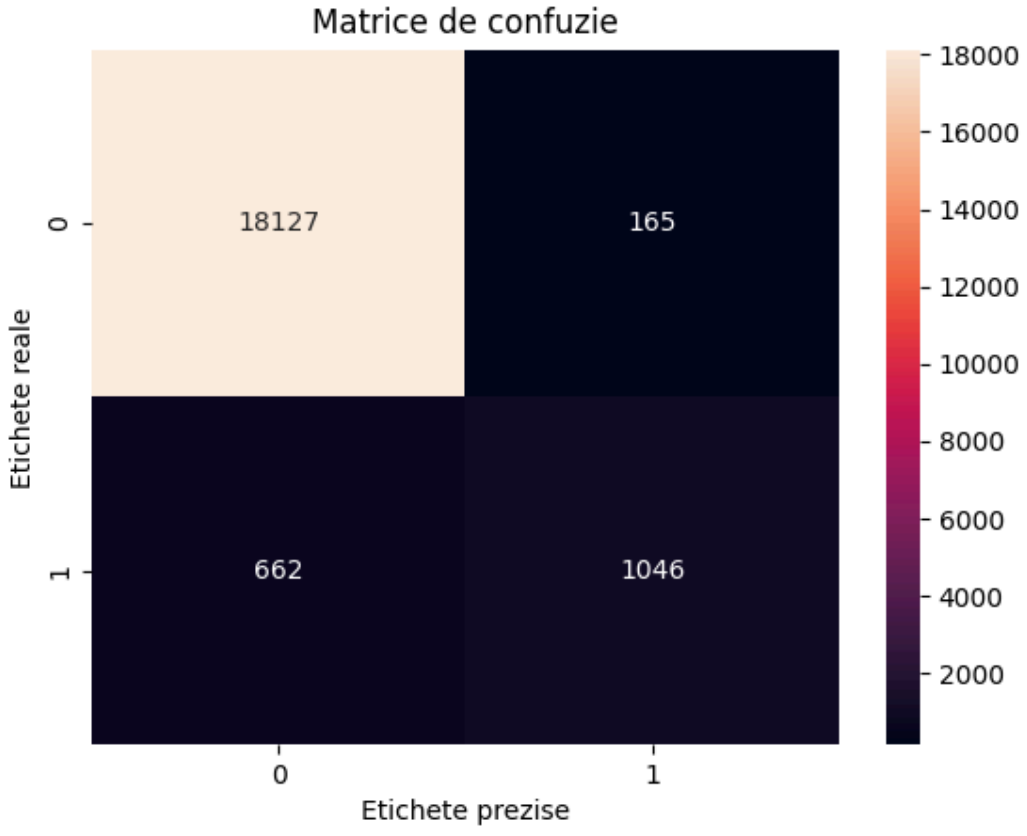


1) Diferenta nu este mare, insa diabetul este mai intalnit la persoanele cu boli de inima;
2) Diabetul este asociat cu un risc de boala de inima, dar relatia nu esste foarte puternica;
3) Nu sunt necesare preprocesari;



1) Hipertensiunea este cea mai intalnita la persoanele cu diabet;
2) Exista o asociere slaba intre hipertensiune si diabet;
3) Nu sunt necesare preprocesari;

Matricea de confuzie:



1) Modelul clasifica corect majoritatea cazurilor (18127 negative si 1046 pozitive). Sunt putine clasificari gresite;
2) Modelul are o acuratete buna, dar rateaza unele cazuri de diabet;
3) Nu sunt necesare preprocesari;

Output-ul obtinut:

Antetul este:

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0

Informatii despre dataframe:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 9 columns):

#	Column	Non-Null Count	Dtype
0	gender	100000 non-null	object
1	age	100000 non-null	float64
2	hypertension	100000 non-null	int64
3	heart_disease	100000 non-null	int64
4	smoking_history	100000 non-null	object
5	bmi	100000 non-null	float64
6	HbA1c_level	100000 non-null	float64
7	blood_glucose_level	100000 non-null	int64
8	diabetes	100000 non-null	int64

dtypes: float64(3), int64(4), object(2)
memory usage: 6.9+ MB
None

Dimensiunile sunt: (100000, 9)

Tipurile datelor folosite:

gender	object
age	float64
hypertension	int64
heart_disease	int64

smoking_history object
bmi float64
HbA1c_level float64
blood_glucose_level int64
diabetes int64
dtype: object

Ce valori se pot afisa pentru fiecare coloana in parte:
gender: ['Female' 'Male' 'Other']
age: [80. 54. 28. 36. 76.]
hypertension: [0 1]
heart_disease: [1 0]
smoking_history: ['never' 'No Info' 'current' 'former' 'ever']
bmi: [25.19 27.32 23.45 20.14 19.31]
HbA1c_level: [6.6 5.7 5. 4.8 6.5]
blood_glucose_level: [140 80 158 155 85]
diabetes: [0 1]

Analiza valorilor lipsa:

Numarul valorilor lipsa pe coloana:
gender 0
age 0
hypertension 0
heart_disease 0
smoking_history 0
bmi 0
HbA1c_level 0
blood_glucose_level 0
diabetes 0
dtype: int64

Procentul valorilor lipsa pe coloana:
gender 0.0
age 0.0
hypertension 0.0
heart_disease 0.0
smoking_history 0.0
bmi 0.0
HbA1c_level 0.0
blood_glucose_level 0.0
diabetes 0.0
dtype: float64

Tratarea valorilor lipsa pentru BMI:
Nu exista valori lipsa pentru BMI

Statistici descriptive:

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
count	100000	100000.000000	100000.00000	100000.000000	100000	100000.000000	100000.000000	100000.000000	100000.000000
unique	3	NaN	NaN	NaN	6	NaN	NaN	NaN	NaN
top	Female	NaN	NaN	NaN	No Info	NaN	NaN	NaN	NaN
freq	58552	NaN	NaN	NaN	35816	NaN	NaN	NaN	NaN
mean	NaN	41.885856	0.07485	0.039420	NaN	27.320767	5.527507	138.058060	0.085000
std	NaN	22.516840	0.26315	0.194593	NaN	6.636783	1.070672	40.708136	0.278883
min	NaN	0.080000	0.00000	0.000000	NaN	10.010000	3.500000	80.000000	0.000000
25%	NaN	24.000000	0.00000	0.000000	NaN	23.630000	4.800000	100.000000	0.000000
50%	NaN	43.000000	0.00000	0.000000	NaN	27.320000	5.800000	140.000000	0.000000
75%	NaN	60.000000	0.00000	0.000000	NaN	29.580000	6.200000	159.000000	0.000000
max	NaN	80.000000	1.00000	1.000000	NaN	95.690000	9.000000	300.000000	1.000000

Matricea de corelatii:

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
age	1.000000	0.251171	0.233354	0.337396	0.101354	0.110672	0.258008
hypertension	0.251171	1.000000	0.121262	0.147666	0.080939	0.084429	0.197823
heart_disease	0.233354	0.121262	1.000000	0.061198	0.067589	0.070066	0.171727
bmi	0.337396	0.147666	0.061198	1.000000	0.082997	0.091261	0.214357
HbA1c_level	0.101354	0.080939	0.067589	0.082997	1.000000	0.166733	0.400660
blood_glucose_level	0.110672	0.084429	0.070066	0.091261	0.166733	1.000000	0.419558

diabetes 0.258008 0.197823 0.171727 0.214357 0.400660
0.419558 1.000000

Coloanele categorice pentru codificare: ['gender', 'smoking_history']
Cum arata datele dupa codificare:

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	0	80.0	0	1	4	25.19	6.6	140	0
1	0	54.0	0	0	0	27.32	6.6	80	0
2	1	28.0	0	0	4	27.32	5.7	158	0
3	0	36.0	0	0	1	23.45	5.0	155	0
4	1	76.0	1	1	1	20.14	4.8	155	0

Dimensiunea setului train (80000, 8)
Dimensiunea setului test (20000, 8)

Acuratetea pe setul de test este 0.95865

Raportul de clasificare:

	precision	recall	f1-score	support
0	0.96	0.99	0.98	18292
1	0.86	0.61	0.72	1708
accuracy			0.96	20000
macro avg	0.91	0.80	0.85	20000
weighted avg		0.96	0.96	0.96	20000

Bonus:

Pentru bonus am ales varianta GitHub.
https://github.com/ovidiu-costache/Programarea-Calculatoarelor-si-Limbaje-de-Programare-3