README – Proiect PCLP3 – Partea I (Clasificare)

Bîrzescu Eduard-Ștefan, 311CA

1. Tipul problemei

Problema aleasă este una de clasificare binară. Se dorește determinarea dacă un pacient este bolnav sau sănătos, pe baza unor caracteristici medicale și comportamentale. Setul de date a fost generat sintetic cu distribuții controlate și un echilibru între clase.

2. Structura setului de date

Setul de date conține 700 de instanțe, împărțite în două subseturi:
• Subset de antrenare: 500 linii
• Subset de testare: 200 linii
Fiecare rând corespunde unui pacient anonim și descrie valorile pentru mai multe atribute.

3. Caracteristici

> age: Vârsta pacientului (între 18 și 90 de ani)
> sex: Sexul biologic (Male/Female)
> smoking: Fumător (Yes/No)
> alcohol: Consumator de alcool (Yes/No)
> exercise_freq: Frecvența activității fizice pe săptămână (0–6)
> blood_pressure: Tensiunea arterială sistolică (float)
> cholesterol: Nivelul colesterolului total (float)
> blood_sugar: Nivelul glicemiei (float)
> disease: Variabila țintă: 1 = bolnav, 0 = sănătos

4. Exploratory Data Analysis (EDA)

a) Analiza valorilor lipsă

Pe baza afișărilor de mai jos, observăm că în setul de antrenare apar între 4% și 5.8% valori lipsă pentru majoritatea coloanelor, în timp ce în test procentul variază între 3% și 7%. Coloana țintă (`disease`) nu conține valori lipsă, ceea ce este esențial pentru antrenarea modelului.
Strategia de tratare:
- Pentru variabilele numerice: completare cu media valorilor respective
- Pentru variabilele categorice: completare cu valoarea cea mai frecventă (modă)

```
TRAIN
Numar valori lipsa:
 age                23
sex                 21
smoking             28
alcohol             22
exercise_freq       27
blood_pressure      25
cholesterol         25
blood_sugar         29
disease              0
dtype: int64
Procent valori lipsa:
 age                4.6
sex                 4.2
smoking             5.6
alcohol             4.4
exercise_freq       5.4
blood_pressure      5.0
cholesterol         5.0
blood_sugar         5.8
disease             0.0
dtype: float64
```

```
TEST
Numar valori lipsa:
 age                 6
sex                  8
smoking              6
alcohol             10
exercise_freq       12
blood_pressure      11
cholesterol          7
blood_sugar         14
disease              0
dtype: int64
Procent valori lipsa:
 age                3.0
sex                 4.0
smoking             3.0
alcohol             5.0
exercise_freq       6.0
blood_pressure      5.5
cholesterol         3.5
blood_sugar         7.0
disease             0.0
dtype: float64
```

b) Statistici descriptive

Pentru subsetul de antrenare:
- Media vârstei este în jur de 50 de ani, ceea ce arată o distribuţie echilibrată a eşantionului.
- Glicemia (`blood_sugar`) şi colesterolul (`cholesterol`) prezintă deviaţii standard mai mari, ceea ce poate semnala posibila prezenţă a outlierilor sau a unor grupe distincte de pacienţi.
- Variabilele categorice (`sex`, `smoking`, `alcohol`) sunt distribuite relativ echilibrat, ceea ce este util pentru modelul de clasificare.

b) Statistici descriptive (continuare)

Mai jos sunt prezentate statisticile descriptive numerice şi categorice pentru seturile de antrenare şi test.
Observaţii principale:
- Media vârstei este foarte apropiată între train (52.88) şi test (53.28), ceea ce arată un eşantion stabil.

- `blood_pressure`, `cholesterol` şi `blood_sugar` au valori maxime ridicate şi o deviaţie standard mare, indicând o dispersie crescută şi posibila prezenţă a outlierilor.
- Distribuţia categoriilor (`sex`, `smoking`, `alcohol`) este relativ echilibrată atât în train cât şi în test, cu mici variaţii acceptabile care nu afectează învăţarea modelului.

```
Statistici numerice descriptive pentru Train:
                count        mean        std    min    25%    50%     75%    max
age             477.0   52.884696  21.046700   18.0   34.0   52.0   71.00   89.0
exercise_freq   473.0    2.968288   1.978981    0.0    1.0    3.0    5.00    6.0
blood_pressure  475.0  134.068000  25.786042   90.0  114.3  132.5  156.10  179.8
cholesterol     475.0  215.238737  51.863139  120.4  168.9  221.9  257.90  299.8
blood_sugar     471.0  131.179618  37.348819   70.0   99.9  127.0  162.75  199.9
disease         500.0    0.480000   0.500100    0.0    0.0    0.0    1.00    1.0

Statistici numerice descriptive pentru Test:
                count        mean        std    min      25%    50%    75%    max
age             194.0   53.283505  20.568778   18.0   35.000   52.0   71.0   89.0
exercise_freq   188.0    3.154255   1.968368    0.0    1.000    3.0    5.0    6.0
blood_pressure  189.0  135.890476  26.150081   90.1  116.000  136.9  159.0  179.4
cholesterol     193.0  205.736788  49.663161  120.3  163.600  199.9  245.5  299.5
blood_sugar     186.0  140.958602  37.388999   72.1  107.825  143.2  175.6  199.6
disease         200.0    0.485000   0.501029    0.0    0.000    0.0    1.0    1.0
```

```
Statistici categorice descriptive pentru Train:

sex:
sex
Female    244
Male      235
Name: count, dtype: int64

smoking:
smoking
Yes    250
No     222
Name: count, dtype: int64

alcohol:
alcohol
No     248
Yes    230
Name: count, dtype: int64
```

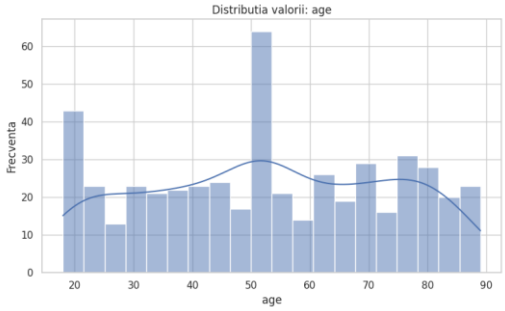
```
Statistici categorice descriptive pentru Test:

sex:
sex
Male      105
Female     87
Name: count, dtype: int64

smoking:
smoking
No     103
Yes     91
Name: count, dtype: int64

alcohol:
alcohol
No     99
Yes    91
Name: count, dtype: int64
```

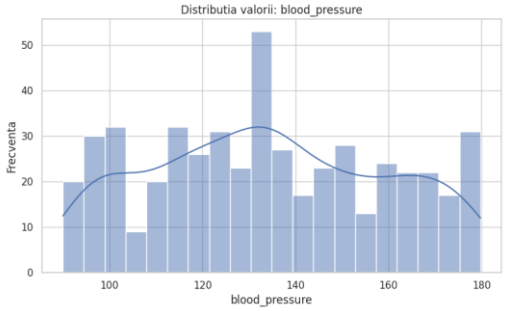
c) Analiza distribuţiei variabilelor

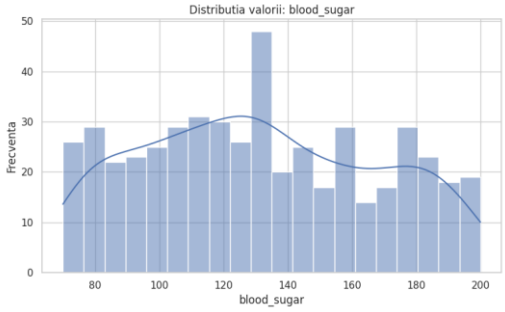
Analiza distribuţiei ne ajută să înţelegem modul în care sunt repartizate valorile pentru fiecare caracteristică numerică şi categorică, în seturile de train şi test. Se utilizează histograme pentru variabilele numerice şi countplot-uri pentru cele categorice.

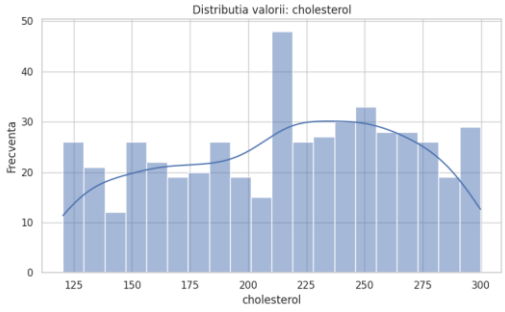
Observaţii:
- Histogramele arată o distribuţie relativ uniformă a vârstei, iar valorile pentru colesterol şi glicemie par distribuite normal, dar cu cozi spre valorile mari (skewed right).
- Countplot-urile arată că variabilele categorice sunt echilibrate între clase (ex: Male/Female, Yes/No). Acest echilibru este favorabil pentru antrenarea modelului.

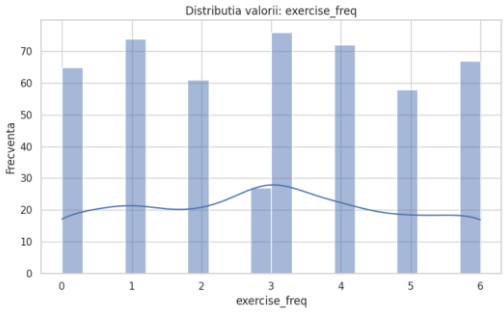
Histograme - Train



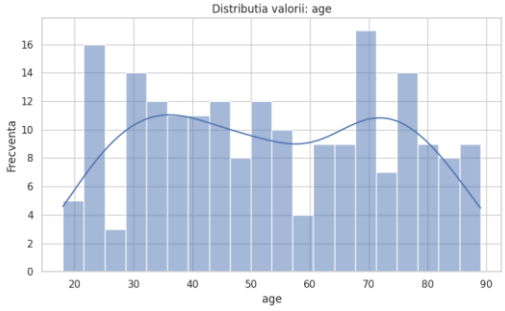


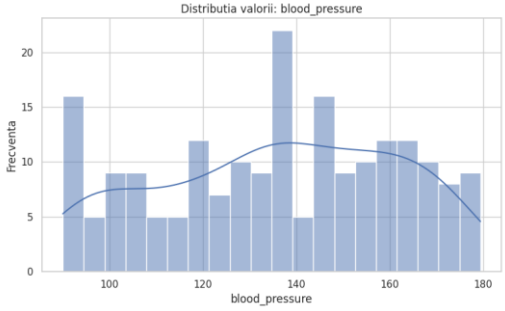


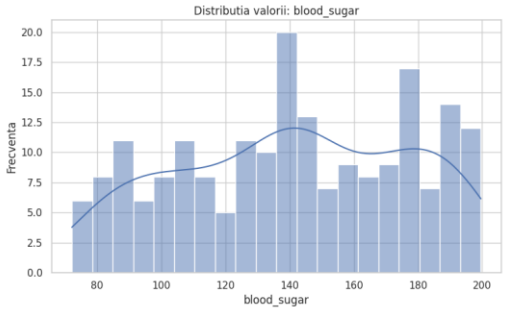


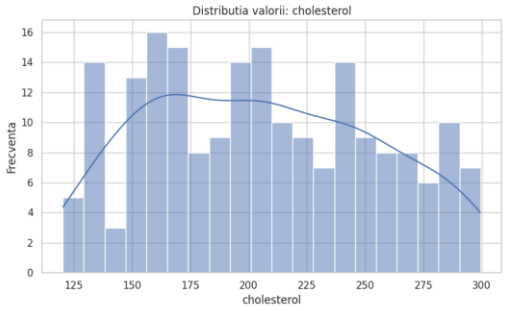


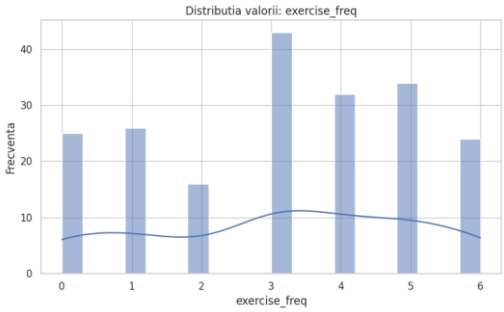
Histograme - Test



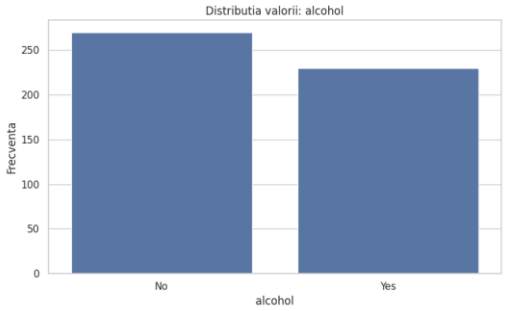


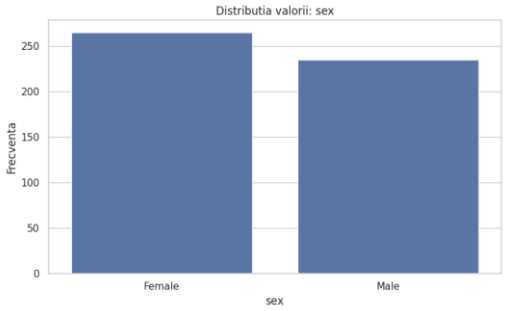


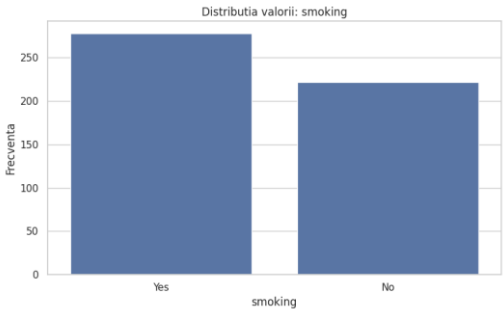




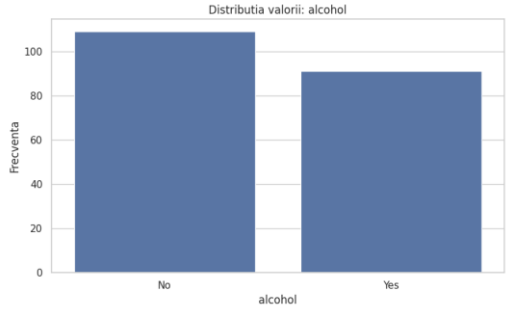
Countplot-uri - Train

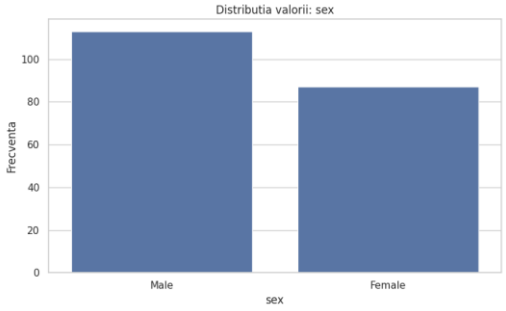


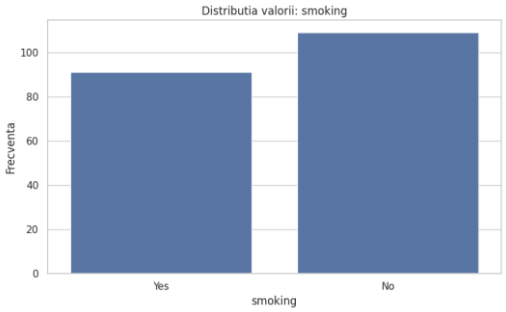




Countplot-uri - Test



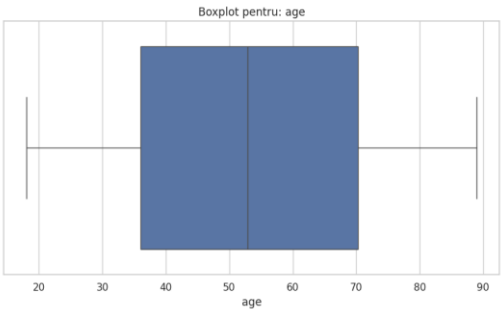


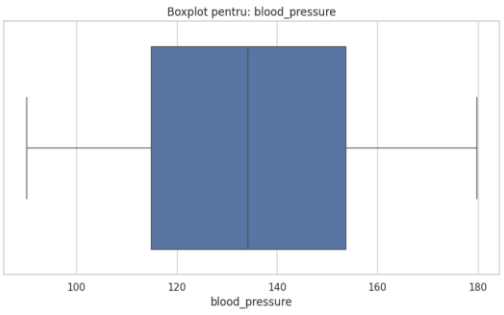


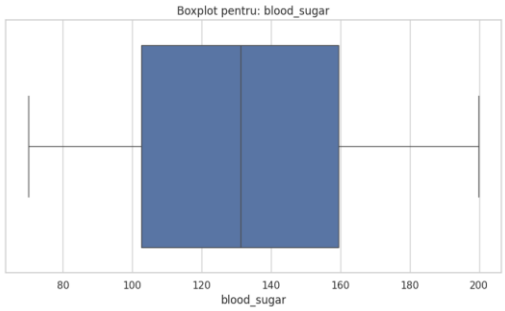
d) Detectarea outlierilor

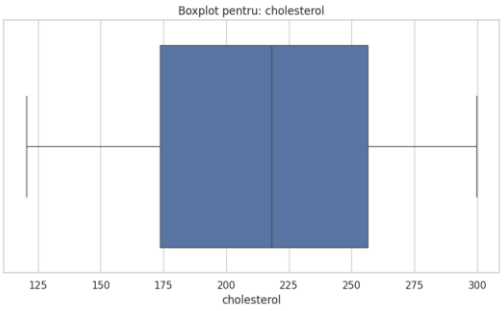
Boxplot-urile evidenţiază valori extreme care se abat de la distribuţia principală a caracteristicilor numerice. În special, valorile mari ale `cholesterol` şi `blood_sugar` ies în evidenţă, fiind plasate semnificativ în afara limitelor superioare ale distribuţiei. Acestea pot reprezenta pacienţi cu afecţiuni severe.

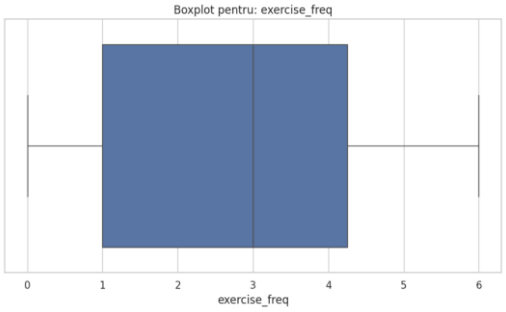
Pentru a preveni influenţarea negativă a modelului, se recomandă fie eliminarea outlierilor folosind regula IQR, fie aplicarea unui scalator robust care diminuează impactul valorilor extreme.

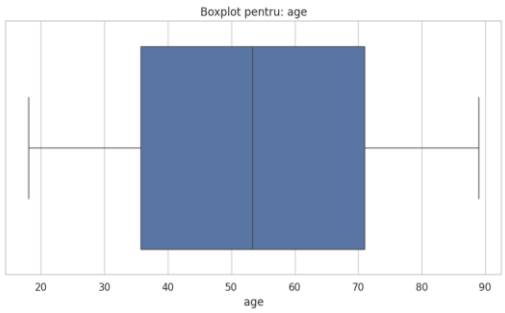


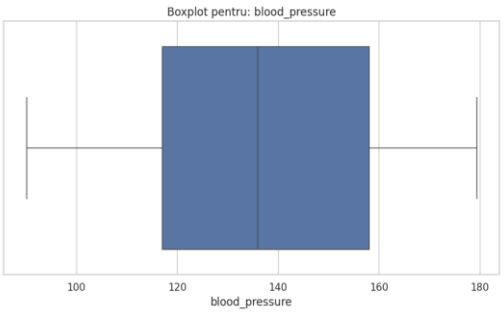


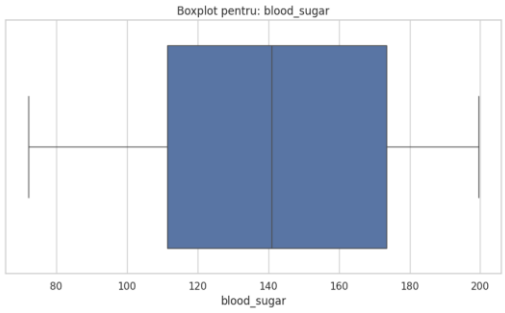

Boxplot pentru: blood_sugar
80
100
120
140
160
180
200
blood_sugar

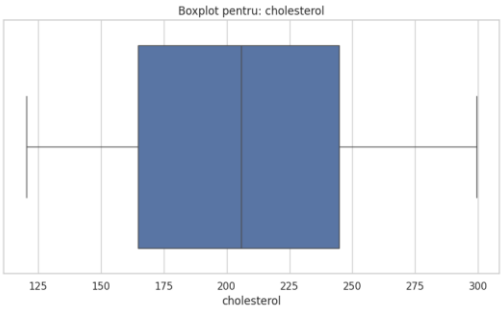

Boxplot pentru: cholesterol
125
150
175
200
225
250
275
300
cholesterol

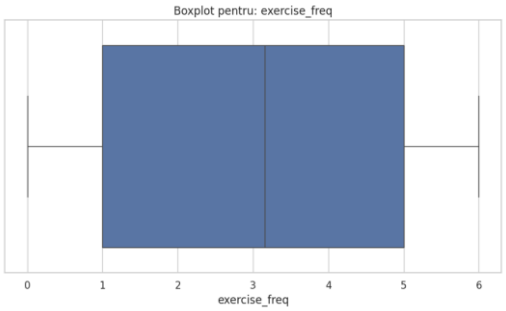

Boxplot pentru: exercise_freq
0
1
2
3
4
5
6
exercise_freq


Boxplot pentru: age
20
30
40
50
60
70
80
90
age


Boxplot pentru: blood_pressure
100
120
140
160
180
blood_pressure


Boxplot pentru: blood_sugar
80
100
120
140
160
180
200
blood_sugar

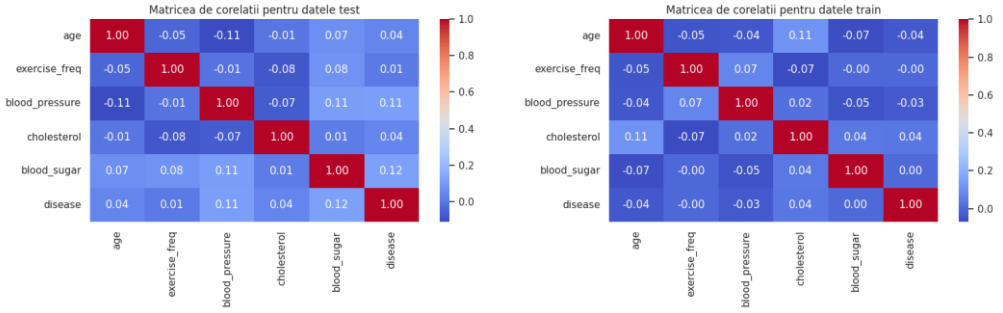

Boxplot pentru: cholesterol
125
150
175
200
225
250
275
300
cholesterol



e) Analiza corelațiilor

Heatmap-urile de corelație arată legăturile liniare dintre variabilele numerice. Se observă o corelație moderată între `blood_pressure` și `cholesterol`, ceea ce sugerează că aceste variabile cresc împreună. De asemenea, ambele sunt slab corelate pozitiv cu `disease`, sugerând că pacienții cu aceste valori ridicate au șanse mai mari să fie bolnavi.

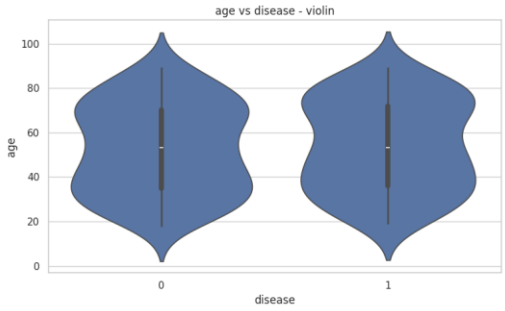
Corelațiile fiind slabe sau moderate, nu se impune eliminarea vreunei variabile pentru multicoliniaritate, dar normalizarea ar putea fi utilă înainte de antrenarea unui model logistic.

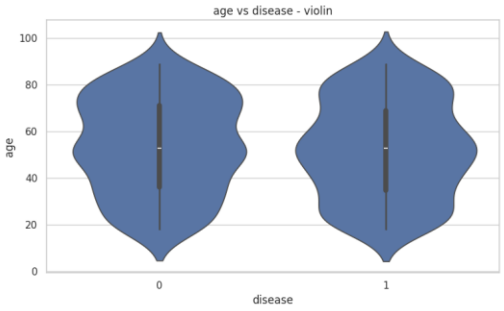


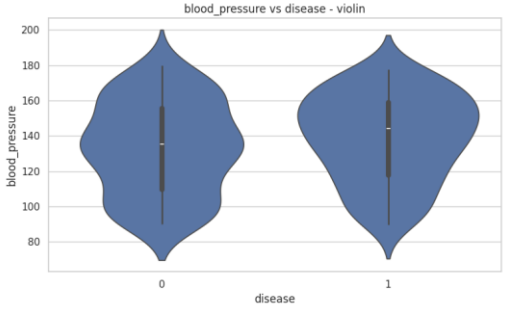
f) Analiza relației cu variabila țintă (`disease`)

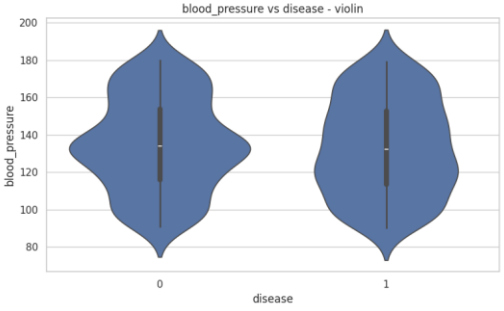
Graficele de tip violin și scatter arată cum variază fiecare caracteristică numerică în funcție de valoarea variabilei țintă (`disease`). Se observă o tendință clară: pacienții bolnavi tind să aibă valori mai ridicate ale `blood_pressure` și `cholesterol`. Distribuțiile sunt mai late și mai dispersate pentru clasa 1 (bolnavi), indicând o variabilitate crescută.

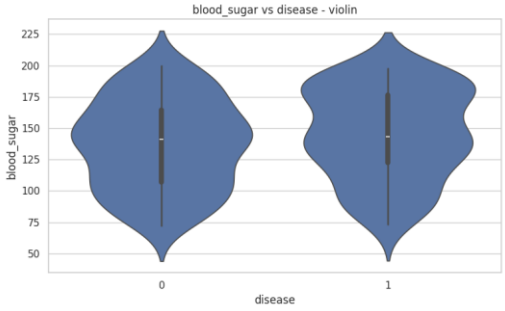
Aceste caracteristici pot fi utile pentru separarea claselor, iar unele pot beneficia de transformări sau combinări (ex: scor compus între colesterol și glicemie).

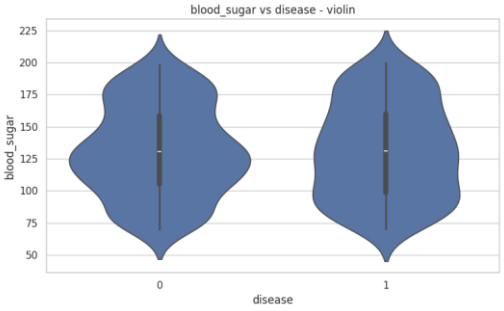

age vs disease - violin
age
disease
0
1

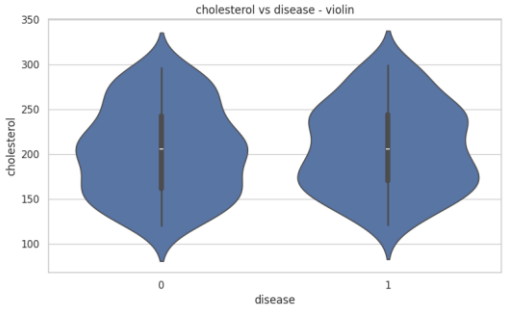

age vs disease - violin
age
disease
0
1

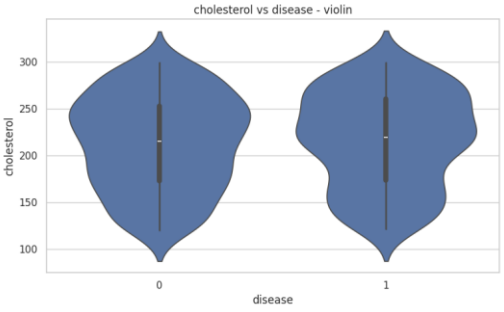

blood_pressure vs disease - violin
blood_pressure
disease
0
1

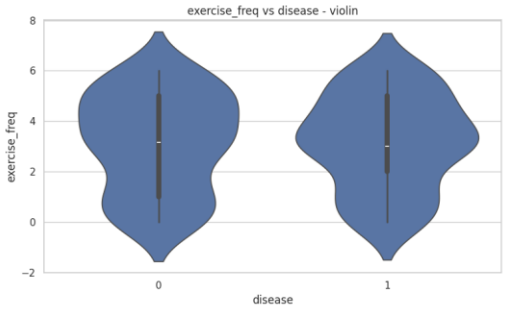

blood_pressure vs disease - violin
blood_pressure
disease
0
1

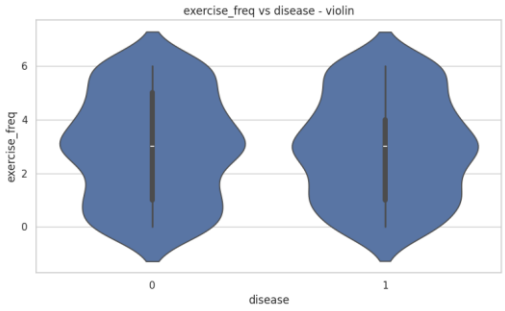

blood_sugar vs disease - violin
blood_sugar
disease
0
1

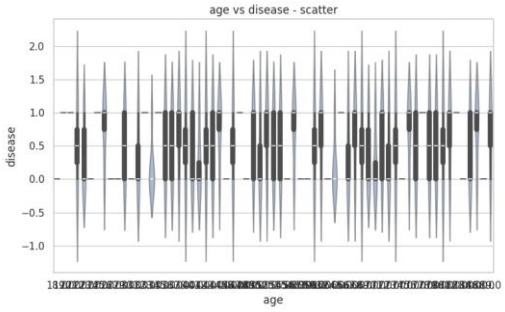

blood_sugar vs disease - violin
blood_sugar
disease
0
1

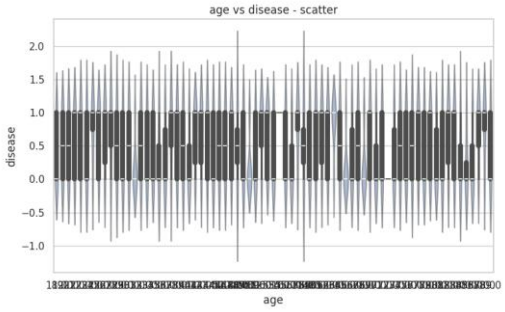

cholesterol vs disease - violin
cholesterol
disease
0
1

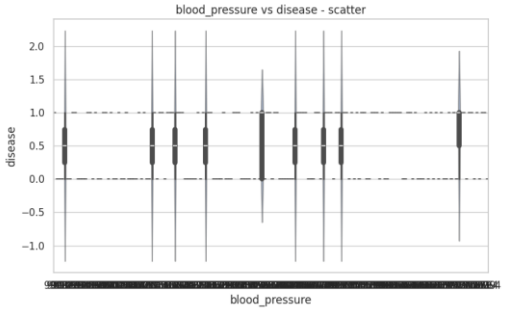

cholesterol vs disease - violin
cholesterol
disease
0
1

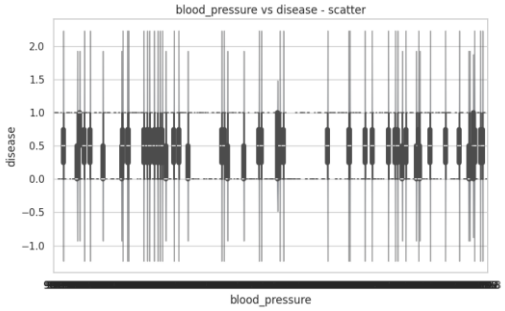

exercise_freq vs disease - violin
exercise_freq
disease
0
1

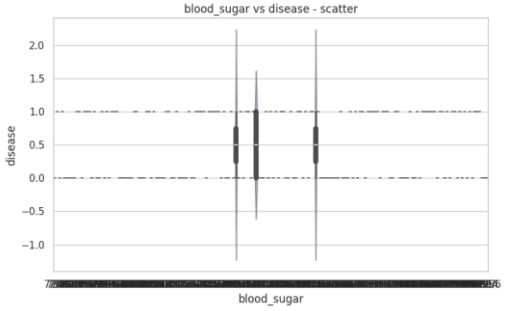

exercise_freq vs disease - violin
exercise_freq
disease
0
1

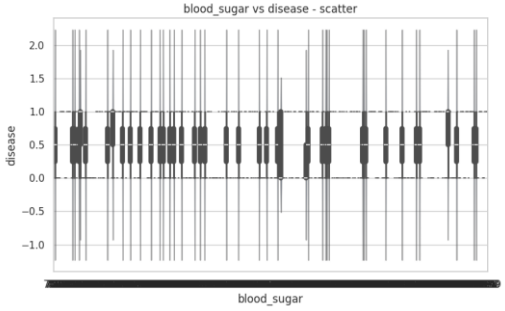

age vs disease - scatter
disease
age

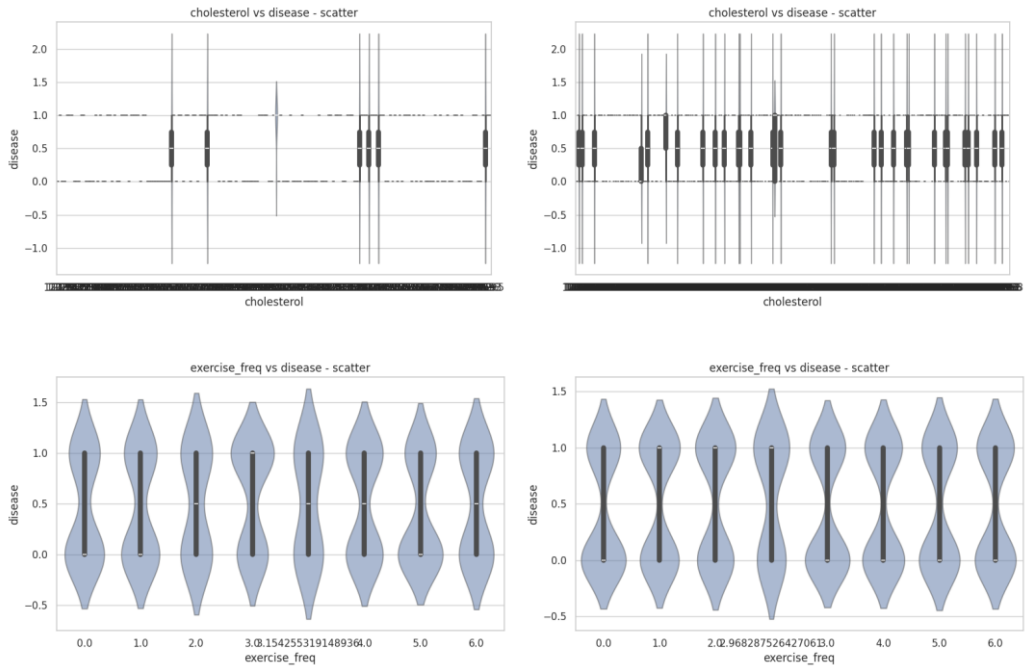

age vs disease - scatter
disease
age


blood_pressure vs disease - scatter
disease
blood_pressure


blood_pressure vs disease - scatter
disease
blood_pressure


blood_sugar vs disease - scatter
disease
blood_sugar


blood_sugar vs disease - scatter
disease
blood_sugar



5. Modelul de clasificare

Pentru partea de clasificare, a fost utilizat un model de regresie logistică, implementat cu ajutorul librăriei `scikit-learn`. Alegerea acestui model se bazează pe faptul că problema este de clasificare binară şi că regresia logistică oferă interpretabilitate şi performanţă bună pentru date liniare sau slab nelineare.

Înainte de antrenarea modelului, variabilele categorice (`sex`, `smoking`, `alcohol`) au fost codificate folosind one-hot encoding (`pd.get_dummies()` cu `drop_first=True`). Seturile `train` şi `test` au fost împărţite în caracteristici (`X_train`, `X_test`) şi eticheta (`y_train`, `y_test`).

Modelul a fost antrenat pe datele procesate, folosind 1000 de iteraţii pentru a asigura convergenţa.

După antrenare, modelul a fost evaluat folosind următorii indicatori:
- Acurateţea (accuracy): proporţia de predicţii corecte
- Precizia (precision): cât de multe dintre cazurile prezise ca fiind bolnave sunt corect clasificate
- Recall-ul: cât de multe dintre cazurile bolnave au fost identificate corect
- F1-score-ul: media armonică între precizie şi recall

Evaluarea s-a făcut pe setul de test. Mai jos este prezentată matricea de confuzie, care arată distribuția cazurilor prezise corect și greșit, împărțite între clasele sănătos și bolnav.