ML Aplicat-Tema2 IA
Minca Ecaterina – Ioana 334CA

1. Descriere generala

Am avut 2 seturi de date si a trebuit sa realizam o analiza a acestora si apoi sa aplicam 2 algoritmi pentru a putea sa prezicem pentru fiecare in functie de un tinta specifica, folosind regresie logistica si MLP.

2. Analiza date AVC

Avem atat date numerice, cat si numerice. Am ales sa fac analiza pe setul de date full. Variabila dupa care se face clasificarea este “cerebrovascular_accident”.

2.1 Date continue

Data Full

	count	mean	std	min	25%	50%	75%	max
mean_blood_sugar_level	5110.0	106.147677	45.283560	55.120000	77.245000	91.885000	114.090000	271.740000
body_mass_indicator	4909.0	28.893237	7.854067	10.300000	23.500000	28.100000	33.100000	97.600000
years_old	5110.0	46.568665	26.593912	0.080000	26.000000	47.000000	63.750000	134.000000
analysis_results	4599.0	323.523446	101.577442	104.829714	254.646209	301.031628	362.822769	756.807975
biological_age_index	5110.0	134.784256	50.399352	-15.109456	96.710581	136.374631	172.507322	266.986321

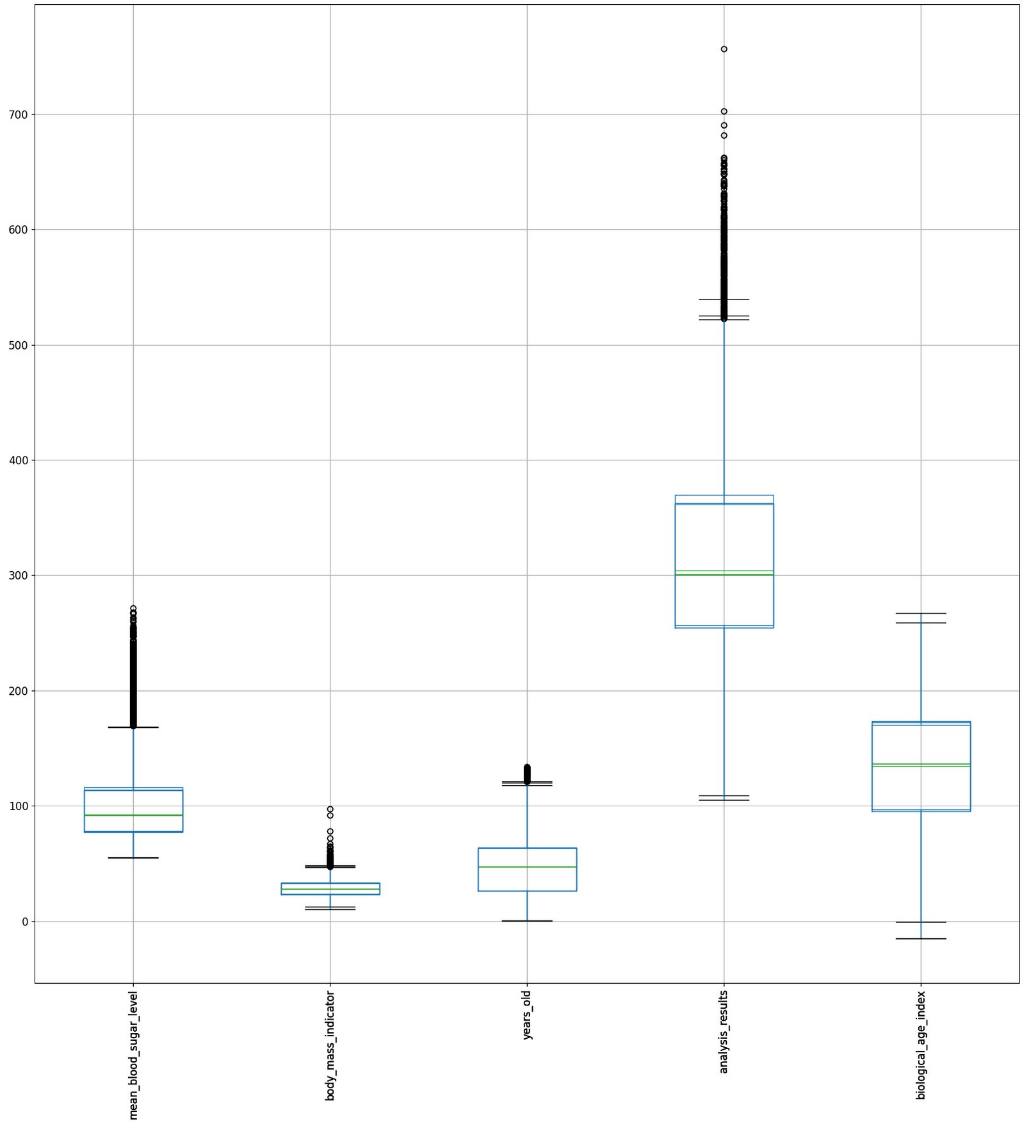
2.2 Date discrete AVC

Se poate observa ca avem lipsa niste valori la 'married'.

```
Data Full
Unique values:
cardiovascular_issues: 2
job_category: 5
sex: 2
tobacco_usage: 4
high_blood_pressure: 2
married: 3
living_area: 2
chaotic_sleep: 2
cerebrovascular_accident: 2

Not missing values:
cardiovascular_issues: 5110
job_category: 5110
sex: 5110
tobacco_usage: 5110
high_blood_pressure: 5110
married: 4599
living_area: 5110
chaotic_sleep: 5110
cerebrovascular_accident: 5110
```

2.3 BoxPlot date continue AVC

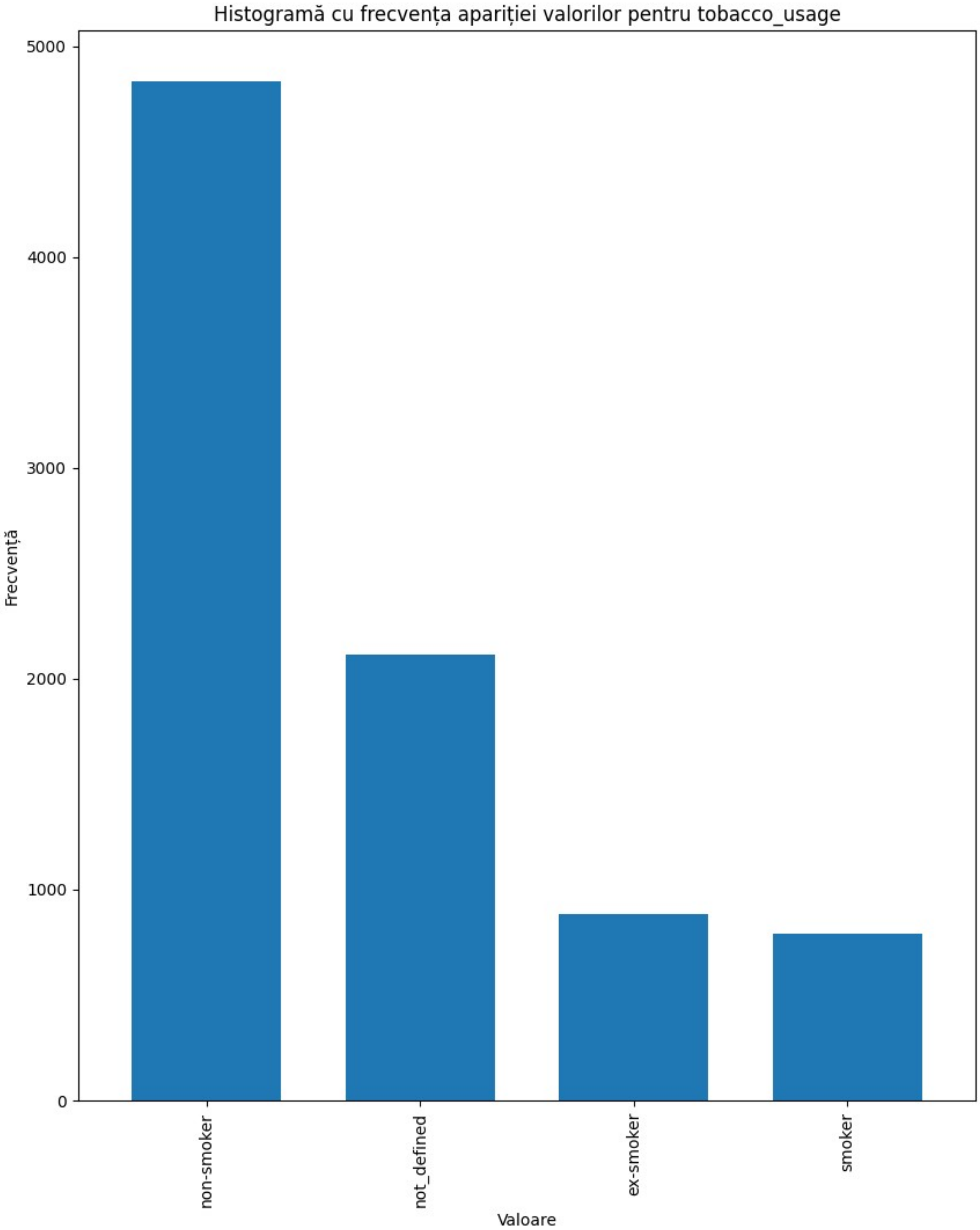


Se poate observa incadrat de acel dreptunghi la fiecare valoarea medie, iar linia de sus si jos sunt valorile minime si maxime, iar restul sunt outliers. Se poate observa de exemplu la "analysis results" ca are multe outliers.

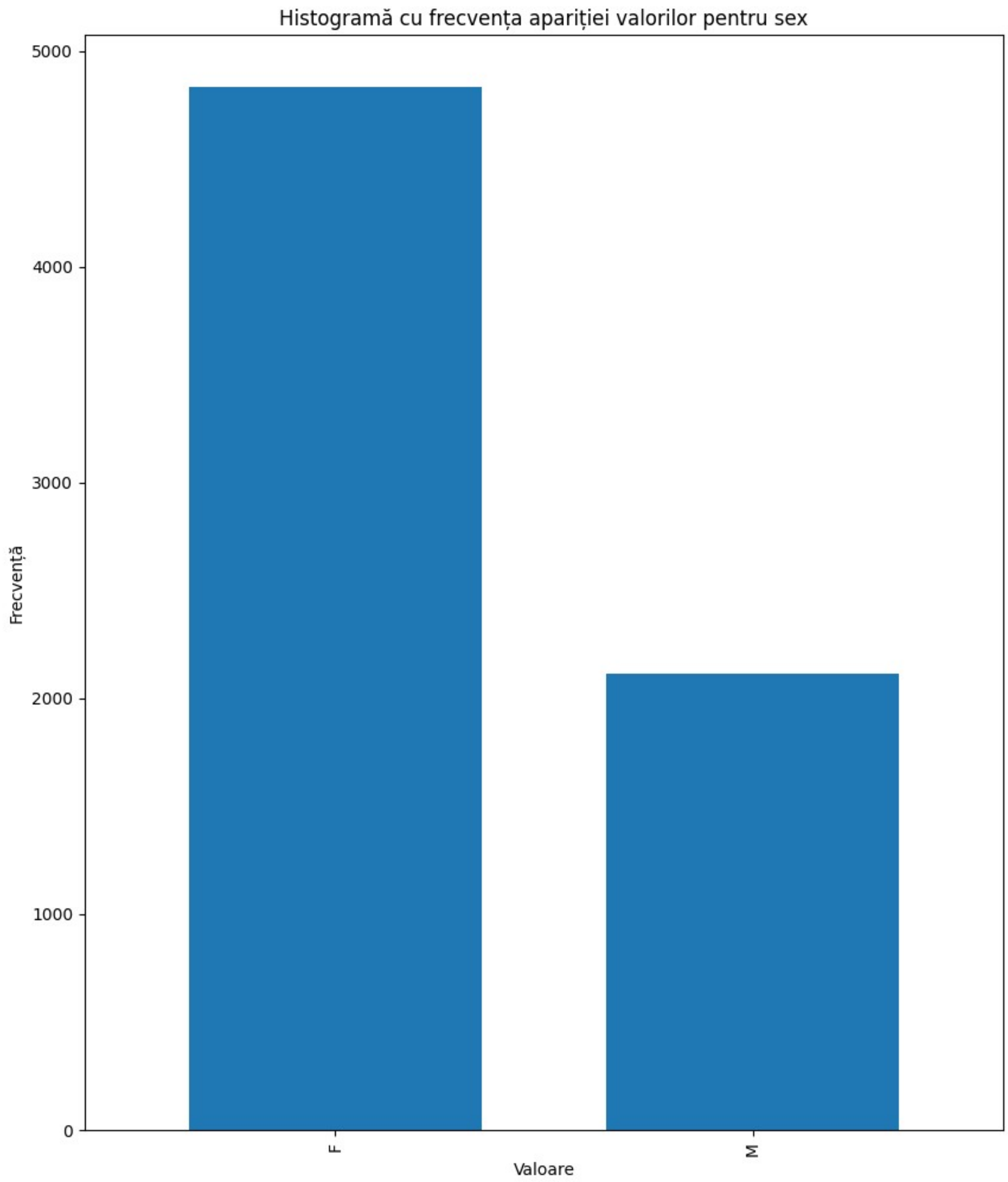
2.4 Histograme date discrete AVC

Pentru fiecare coloana in parte, am relizat cate o histograma in care se va observa cate persoane sunt din fiecare categorie.

2,4,1 Tobacco usage

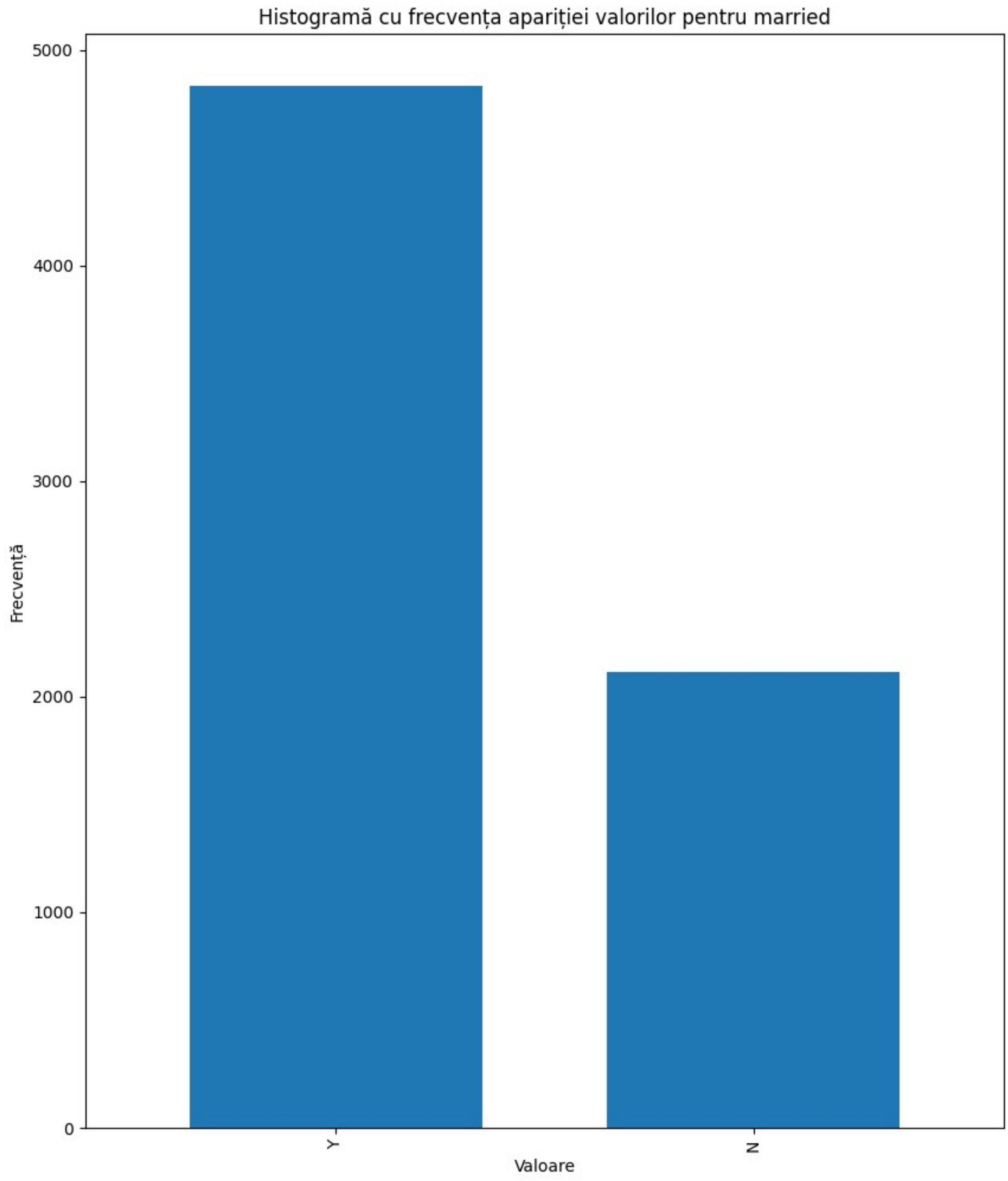


2.4.2 Sex

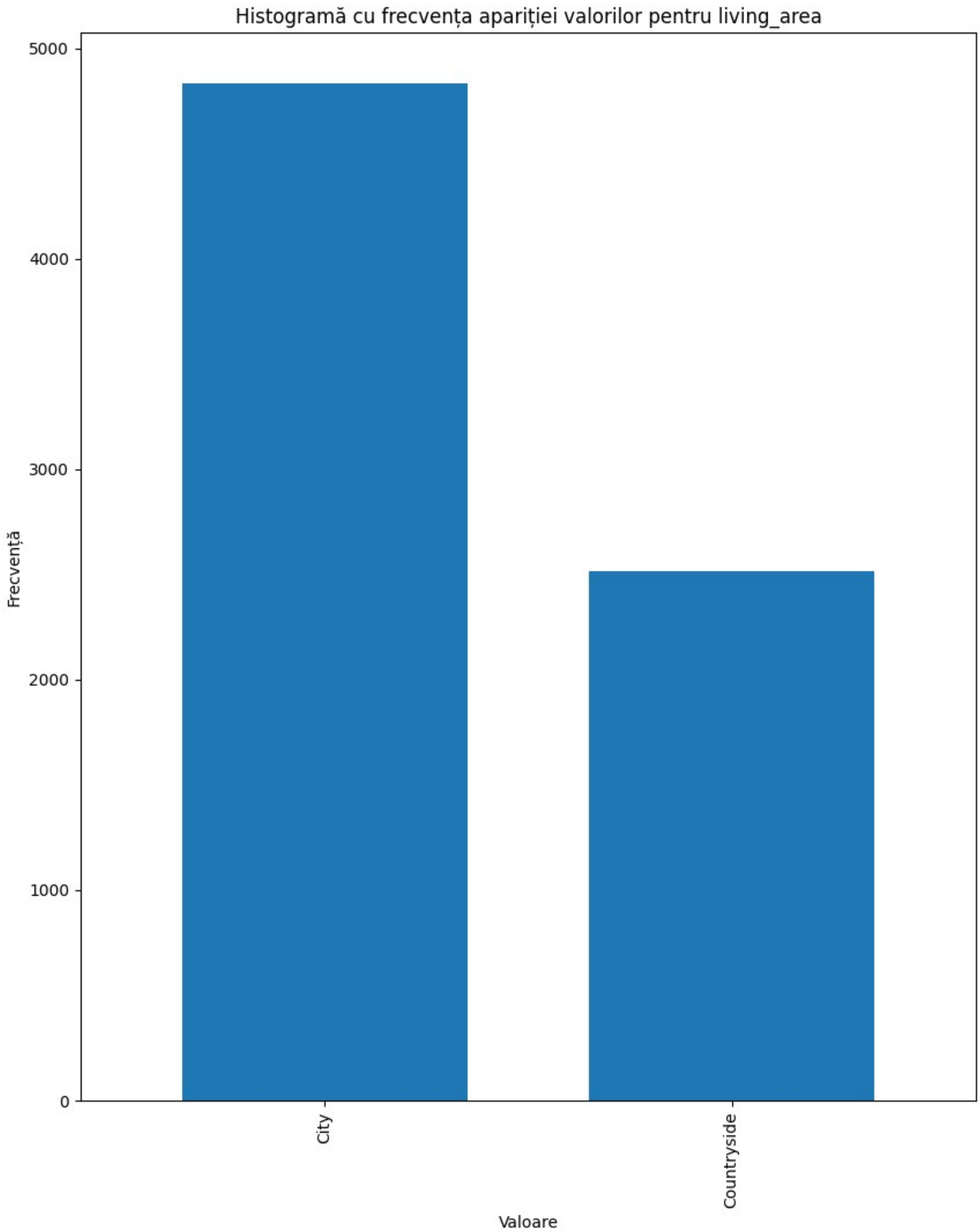


Histogramă cu frecvența apariției valorilor pentru sex
5000
4000
3000
2000
1000
0
Frecvență
F
M
Valoare

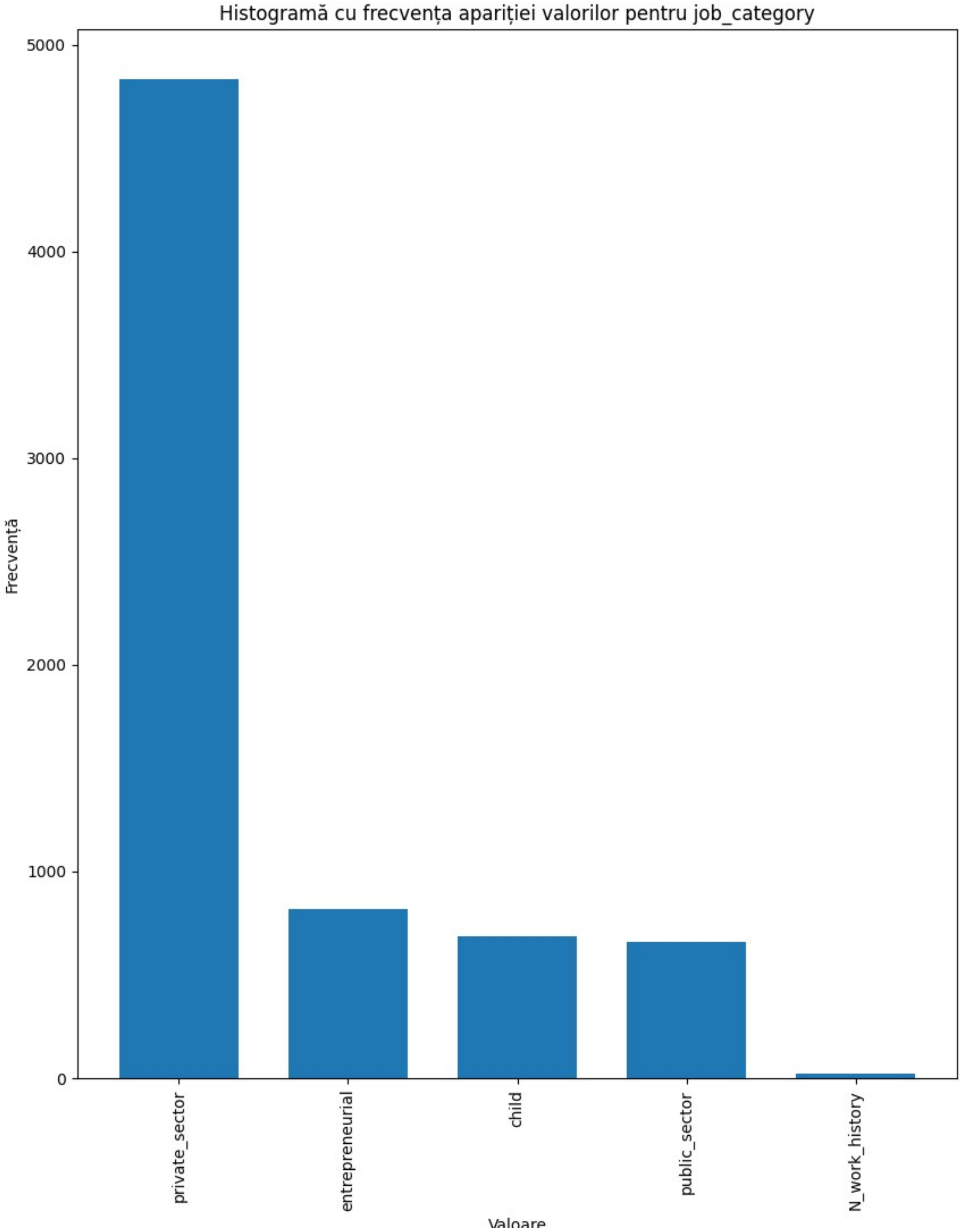
2.4.3 Married


Histogramă cu frecvența apariției valorilor pentru married
Frecvență
5000
4000
3000
2000
1000
0
Y
N
Valoare

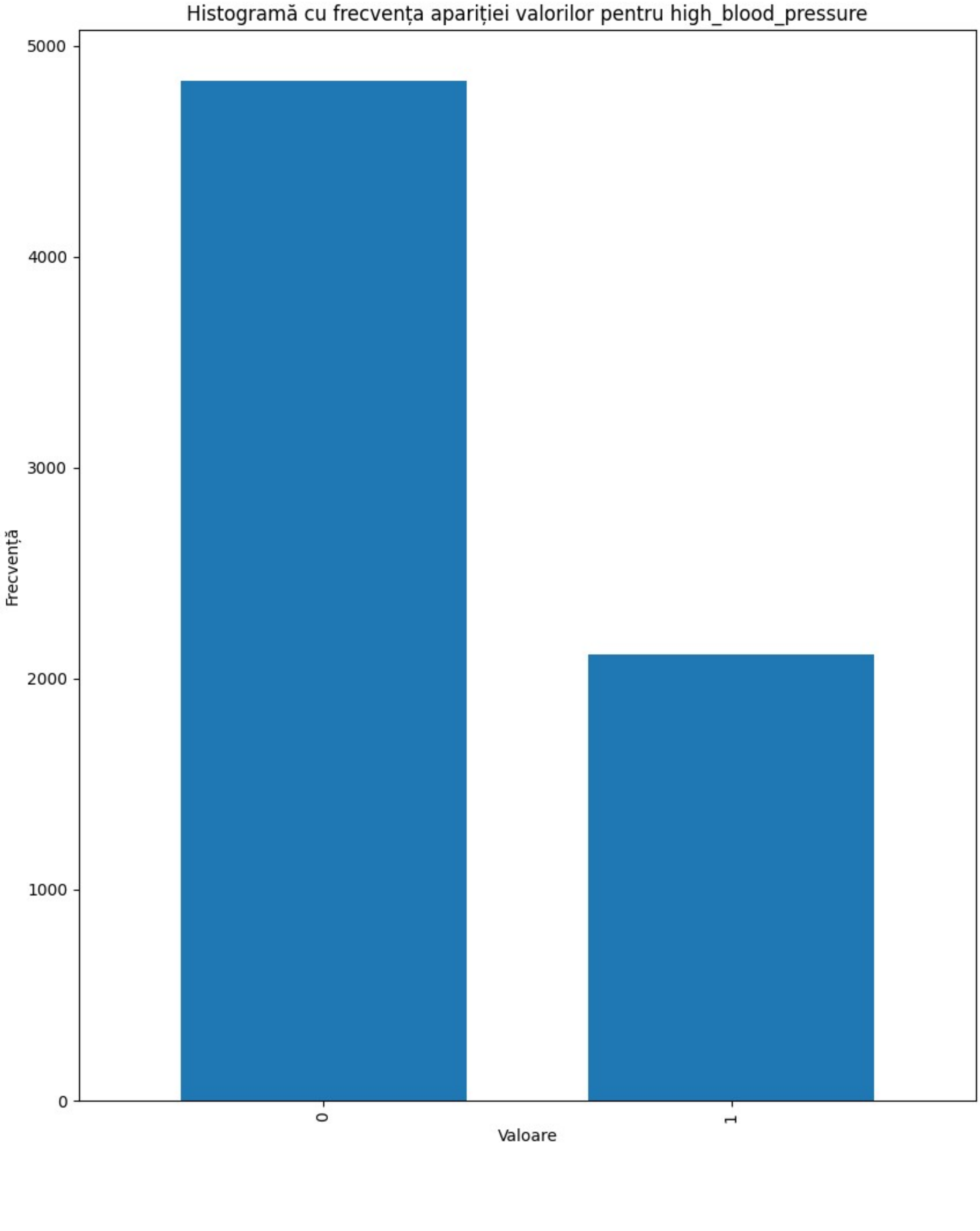
2.4.4 Living area



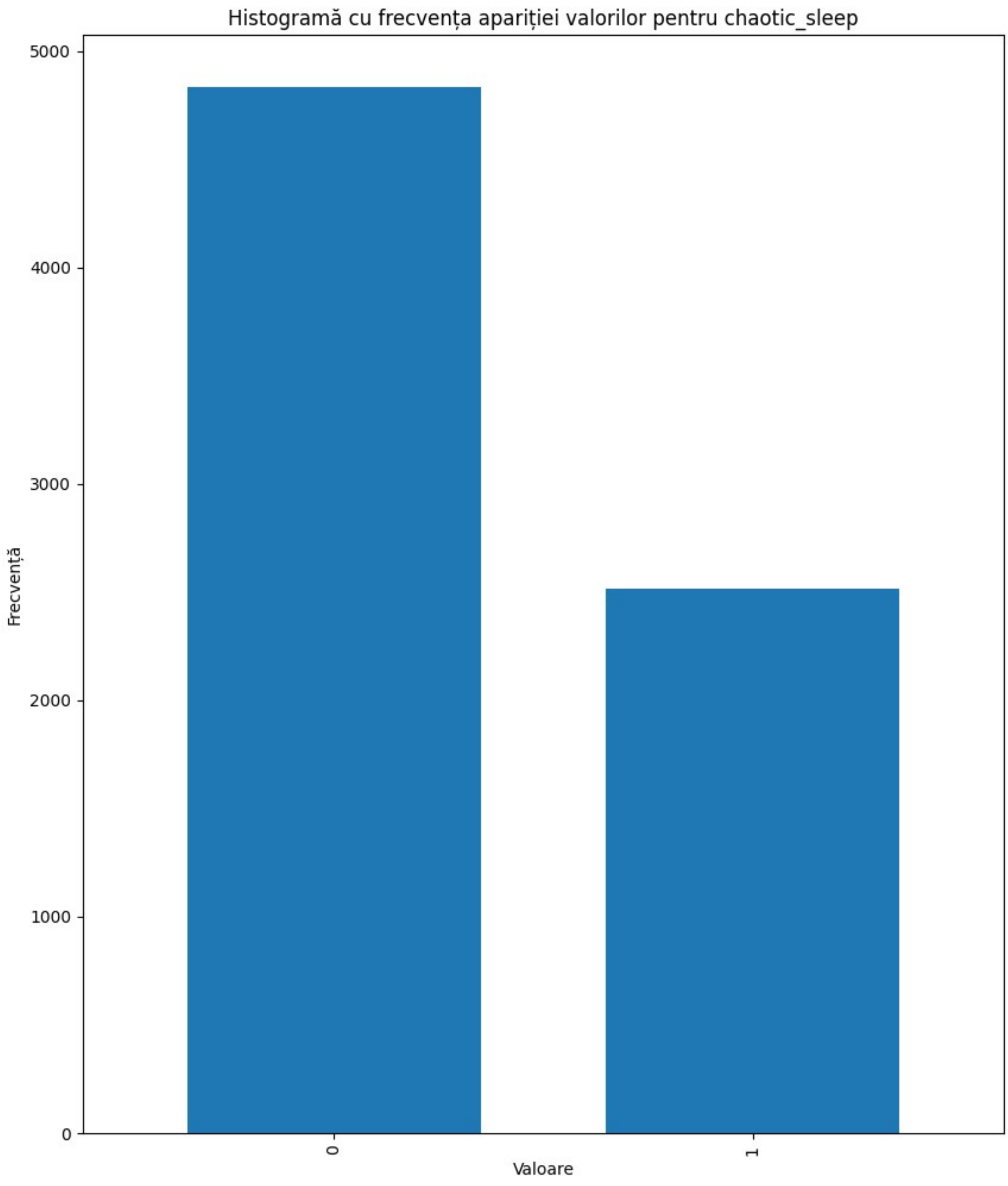
2.4.5 Job category



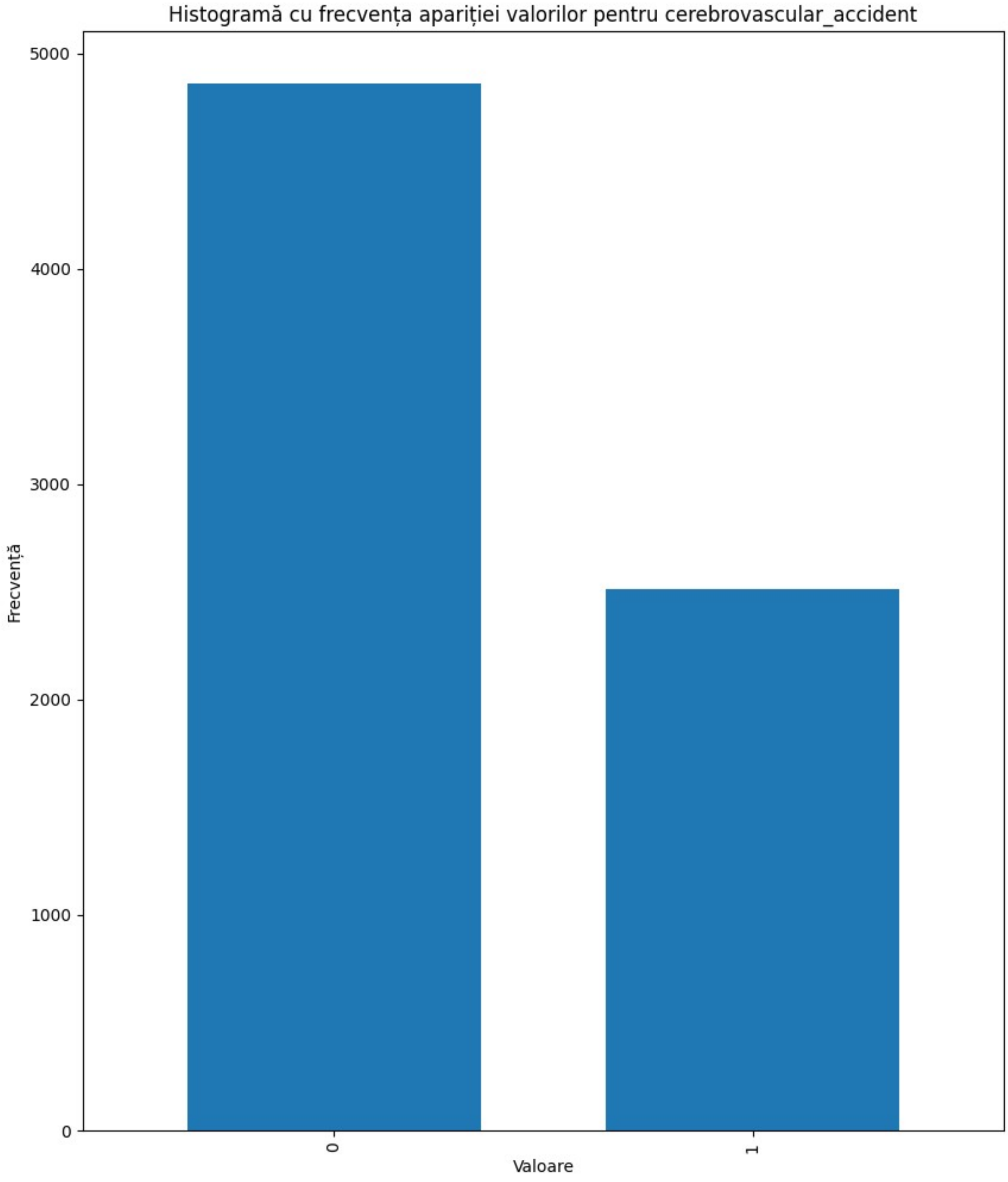
2.4.6 High blood pressure



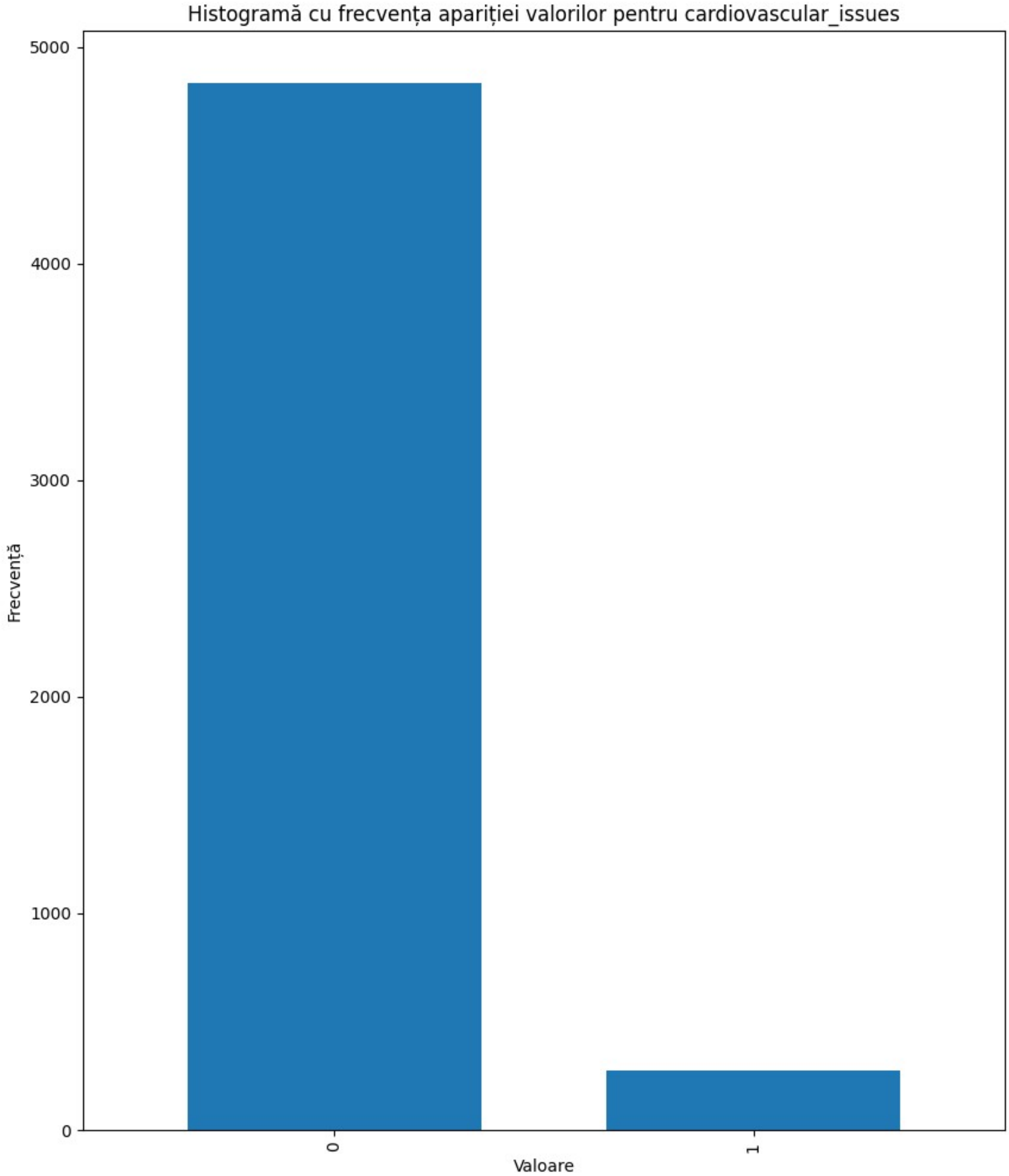
2.4.7 Chaotic sleep



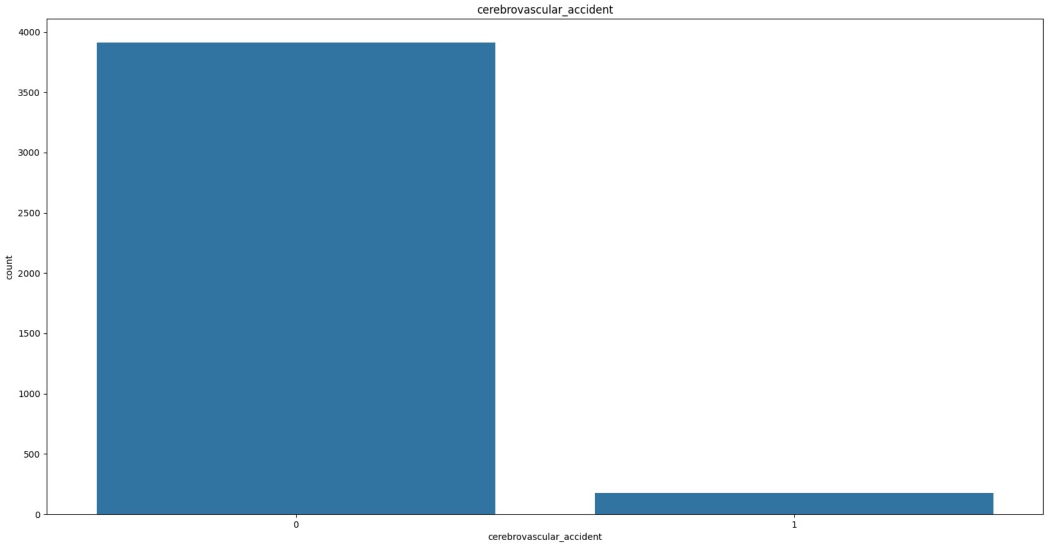
2.4.8 Cerebrovascular accident

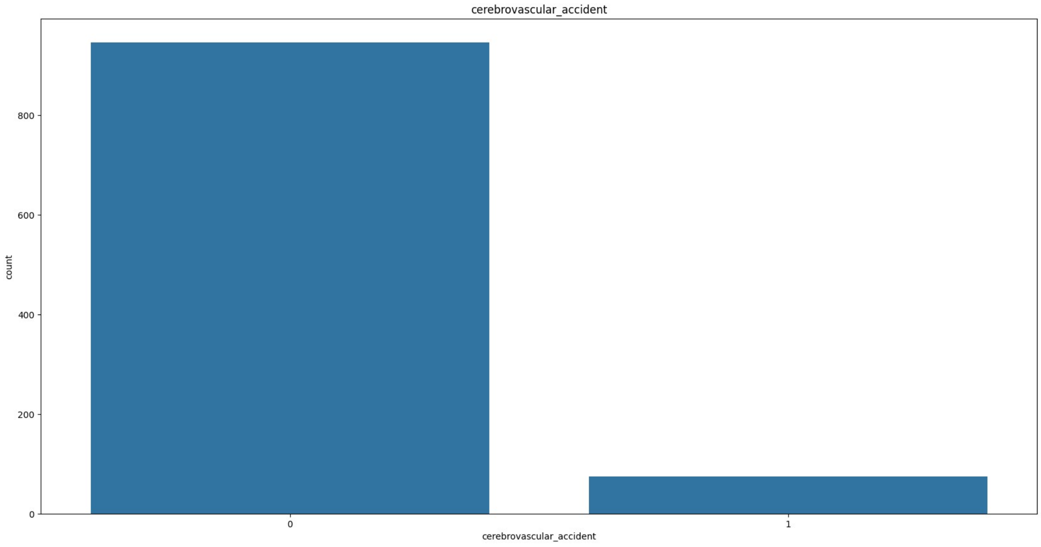


2.4.9 Cardiovascular issues



2.5 Analiza echilibru clasa train vs test AVC

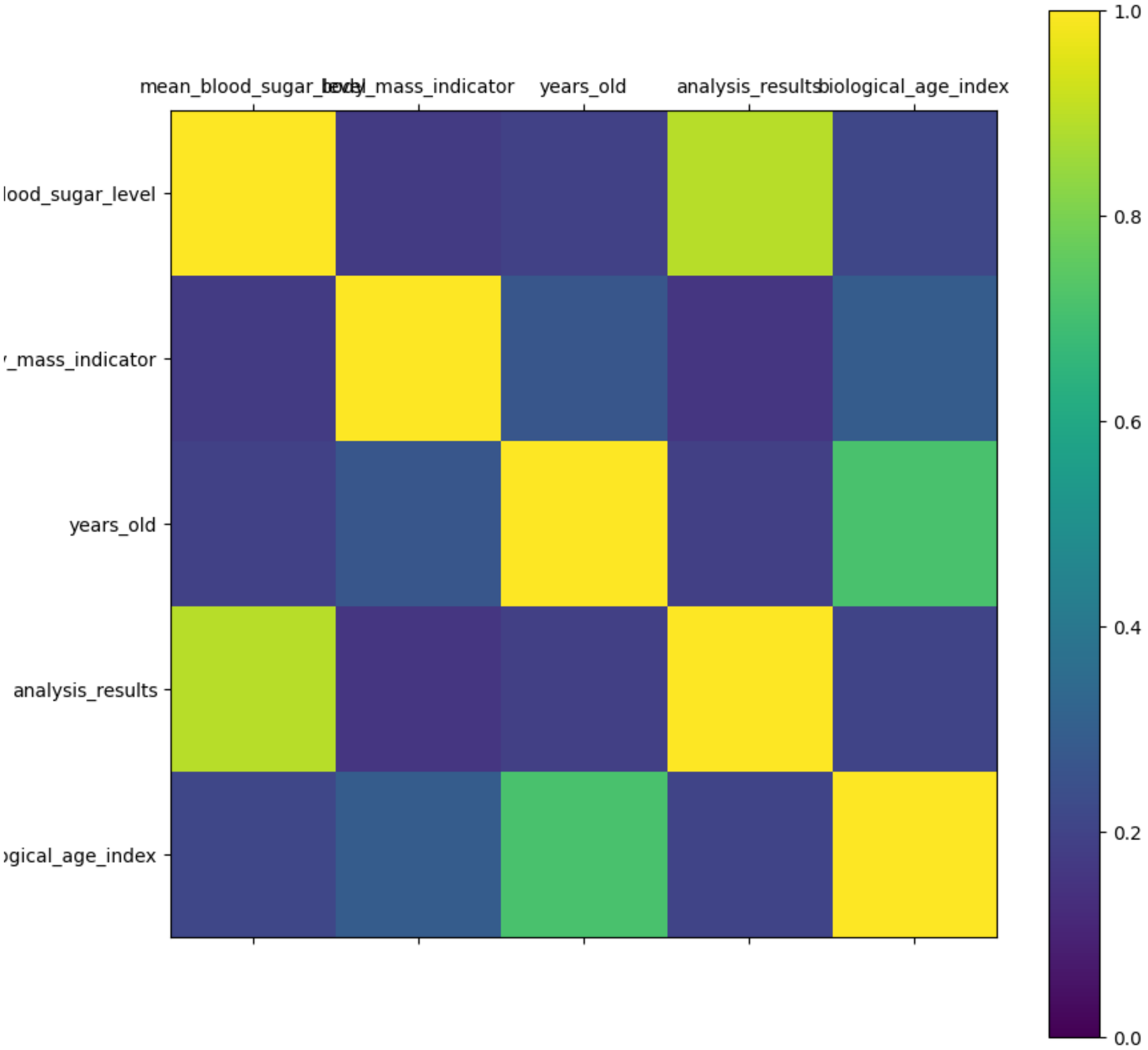




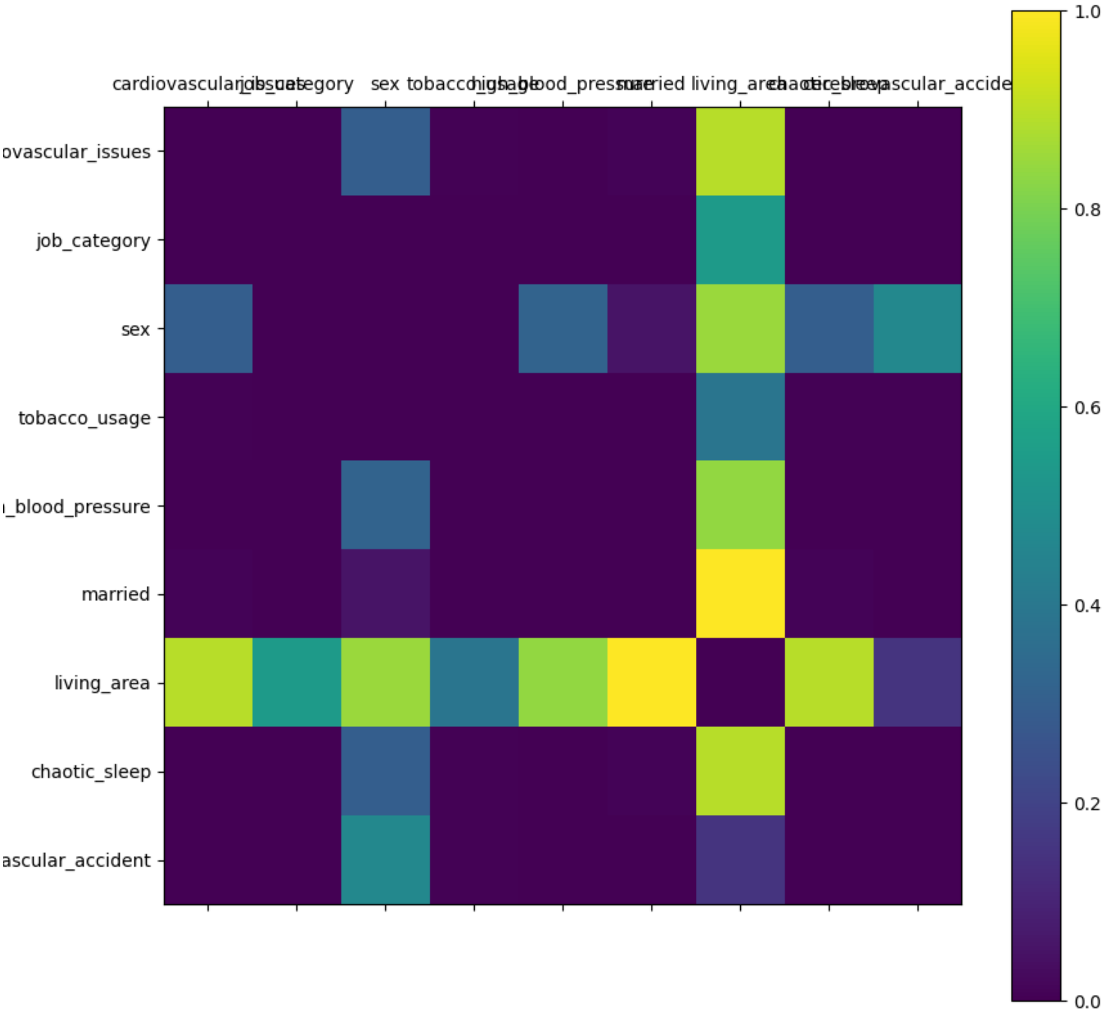
Se poate observa ca este foarte dezechilibrata.

2.6 Analiza corelației între atribute AVC

2.6.1 Date numerce



2.6.2 Date categorice



2.7 Concluzie analiza AVC

Se poate observa ca avem o clasa destul de dezechilibrata si avem multe outliers, valori nule, valori extreme.

3. Analiza date Salary Classification

3.1 Date continue

```
Data Full
          count           mean            std      min       25%       50%       75%        max
fnl      9999.0  190352.902090  106070.862686  19214.0  118282.5  178472.0  237311.0  1455435.0
hpw      9199.0      40.416241      12.517356      1.0      40.0      40.0      45.0       99.0
gain     9999.0     979.853385    7003.795382      0.0       0.0       0.0       0.0    99999.0
edu_int  9999.0      14.262026      24.770835      1.0       9.0      10.0      13.0      206.0
years    9999.0      38.646865      13.745101     17.0      28.0      37.0      48.0       90.0
loss     9999.0      84.111411     394.035484      0.0       0.0       0.0       0.0     3770.0
prod     9999.0    2014.927593   14007.604496    -28.0      42.0      57.0      77.0   200125.0
```

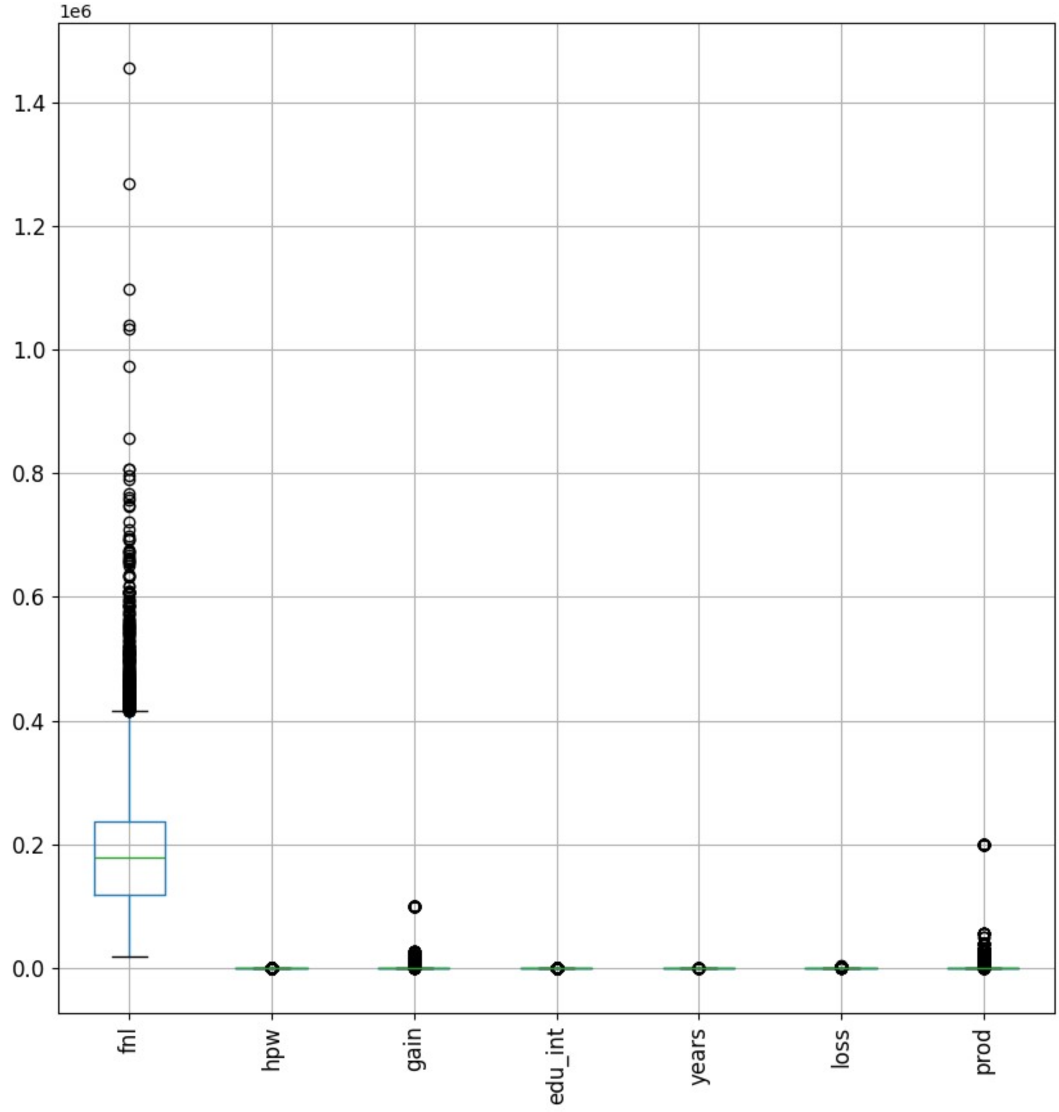
3.2 Date discrete

```
Data Full
Unique values:
relation: 6
country: 41
job: 14
work_type: 9
partner: 7
edu: 16
gender: 3
race: 5
gtype: 2
money: 2

Not missing values:
relation: 9999
country: 9999
job: 9999
work_type: 9999
partner: 9999
edu: 9999
gender: 9199
race: 9999
gtype: 9999
money: 9999
```

Se poate observa ca la ‘gender’ sunt valori lipsa.

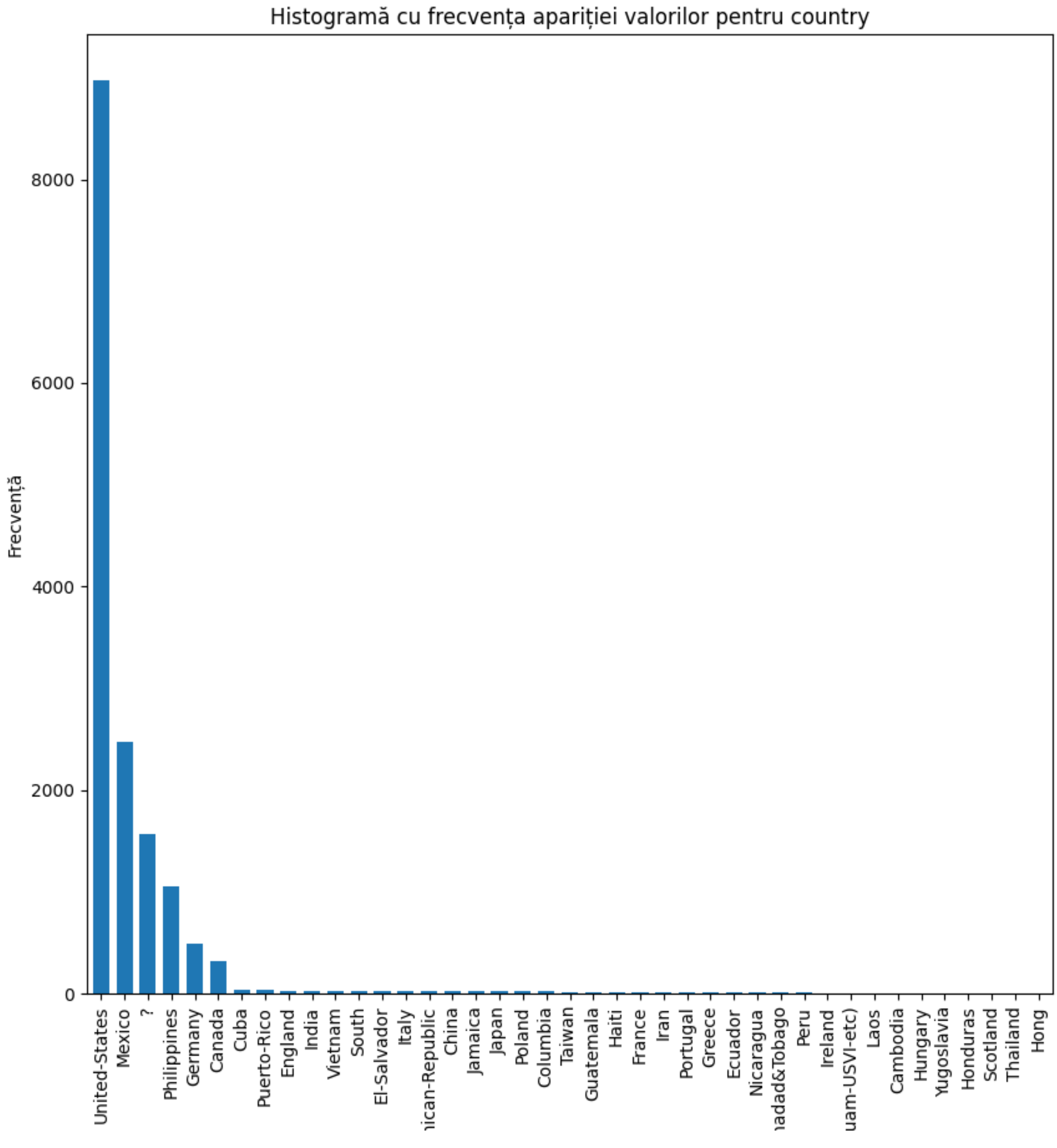
3.3 BoxPlot date continue Salary Classification



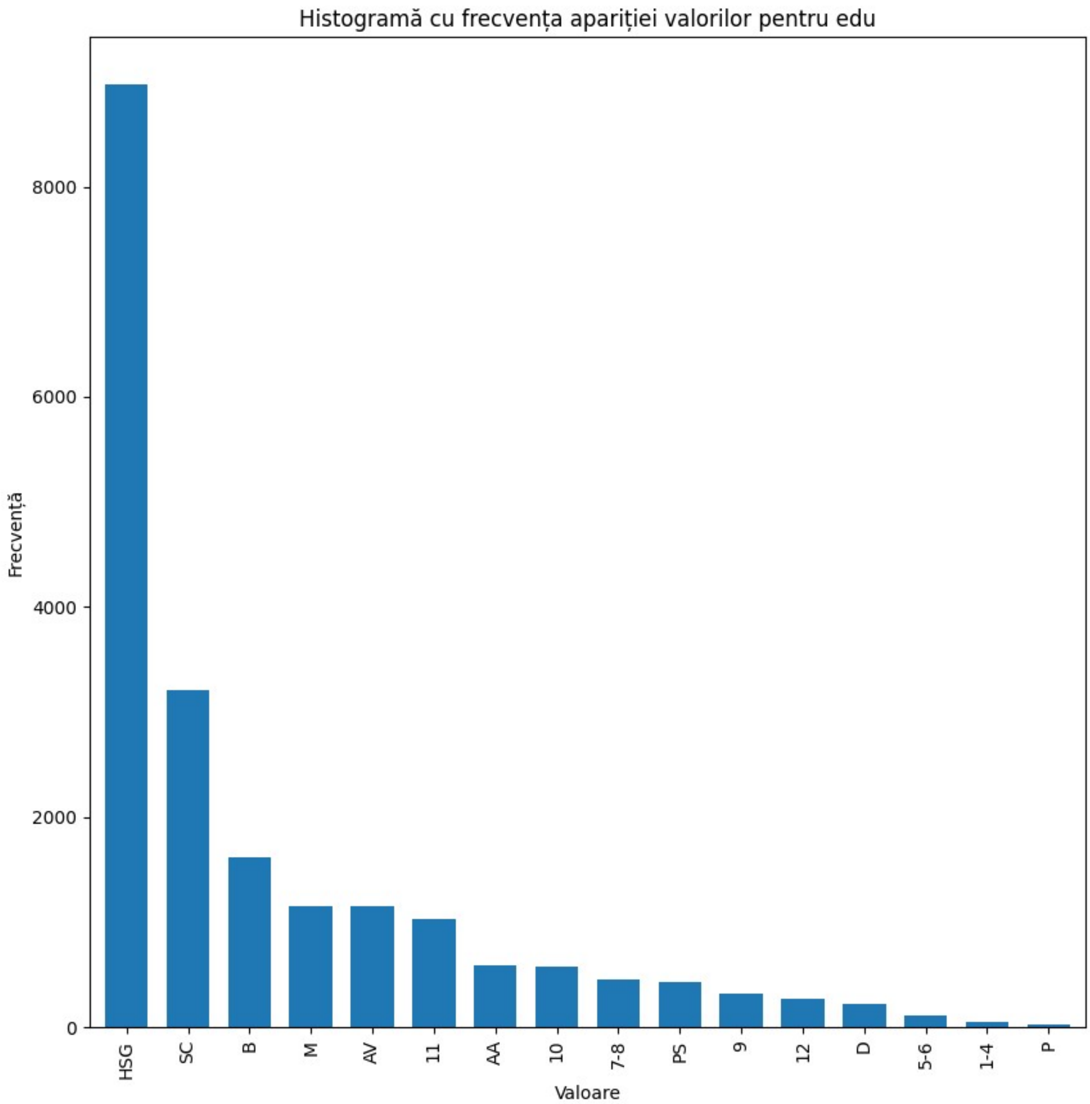
La fel se observa multe outliers mai ales la fnl.

3.4 Histograme set date discrete

3.4.1 Country

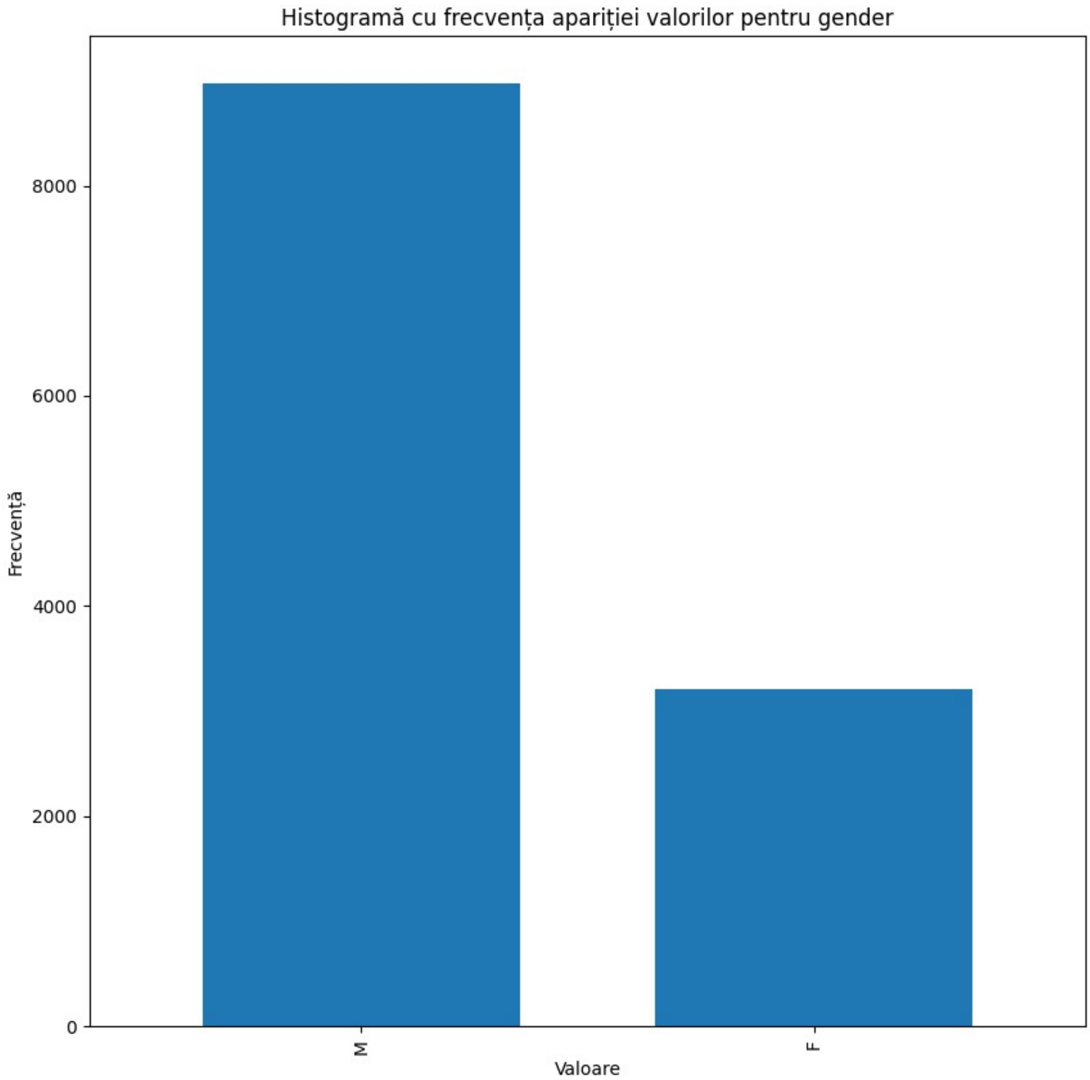


3.4.2 Edu



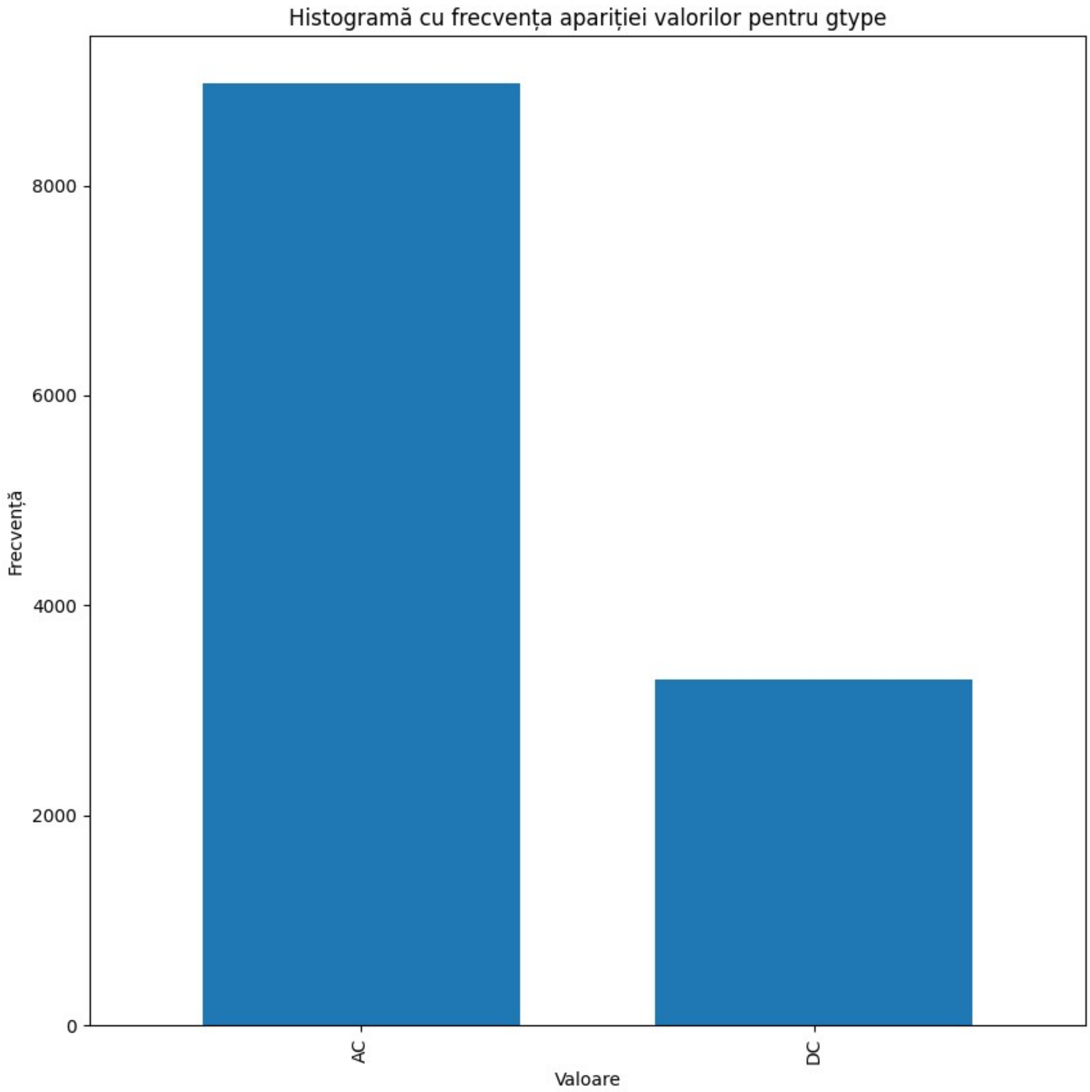
Histogramă cu frecvența apariției valorilor pentru edu
Frecvență
Valoare
0
2000
4000
6000
8000
HSG
SC
B
M
AV
11
AA
10
7-8
PS
9
12
D
5-6
1-4
P

3.4.3 Gender

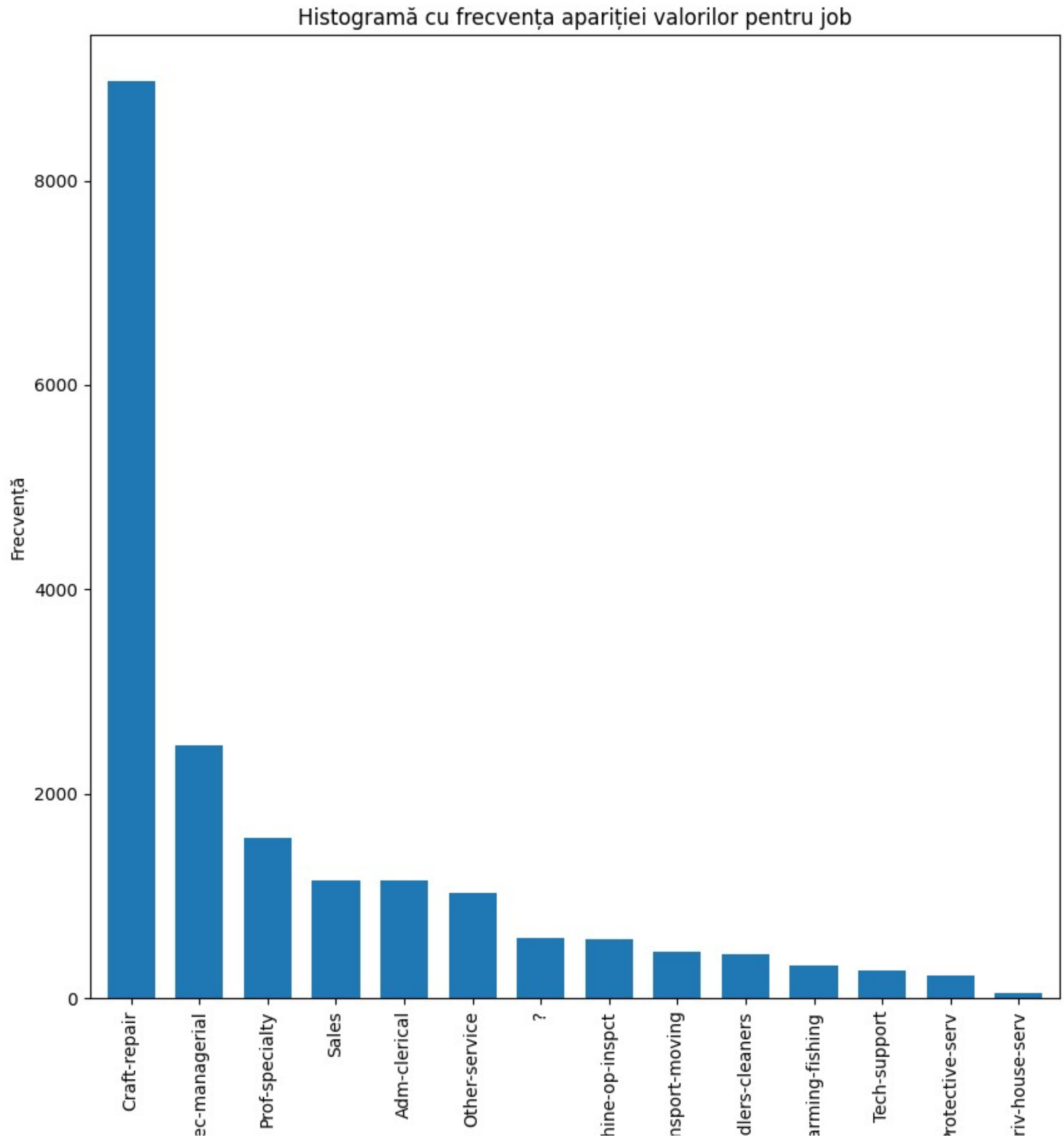


Histogramă cu frecvența apariției valorilor pentru gender
Frecvență
Valoare
0
2000
4000
6000
8000
M
F

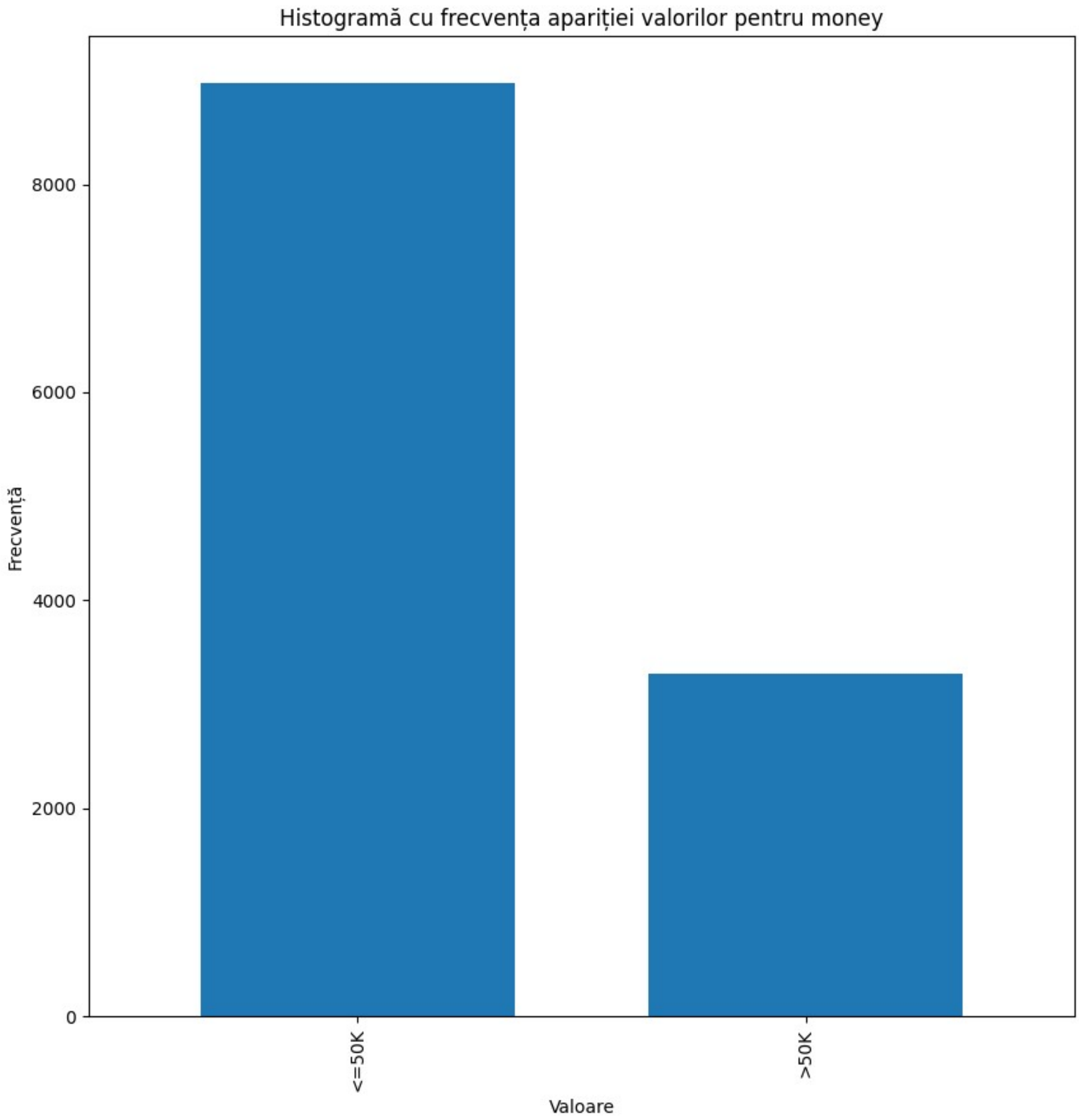
3.4.4 Gtype


Histogramă cu frecvența apariției valorilor pentru gtype
Frecvență
Valoare
0
2000
4000
6000
8000
AC
DC

3.4.5

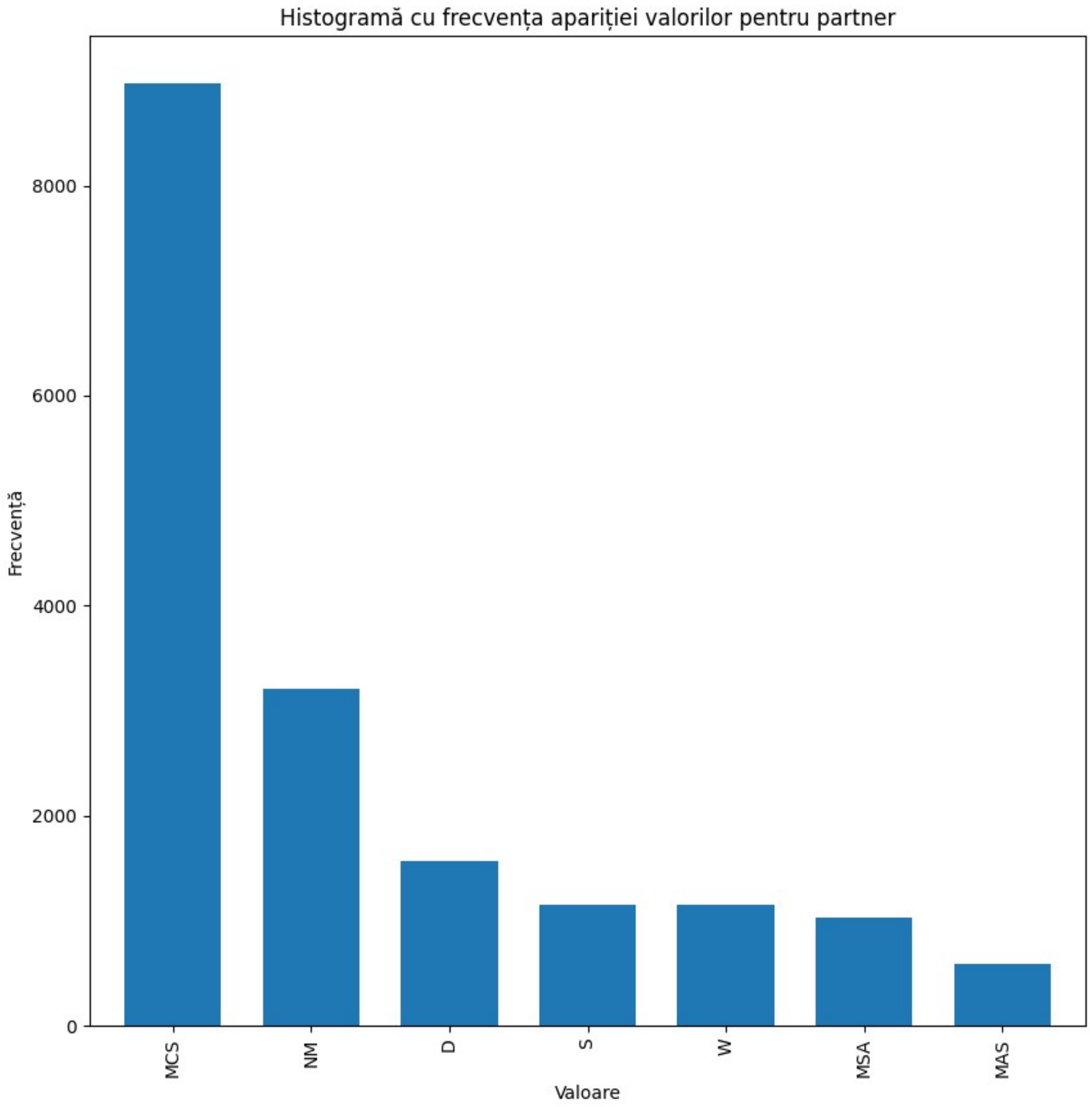

Histogramă cu frecvența apariției valorilor pentru job
Frecvență
0
2000
4000
6000
8000
Craft-repair
ec-managerial
Prof-specialty
Sales
Adm-clerical
Other-service
?
hine-op-inspct
nsport-moving
dlers-cleaners
arming-fishing
Tech-support
rotective-serv
riv-house-serv

3.4.6 Money



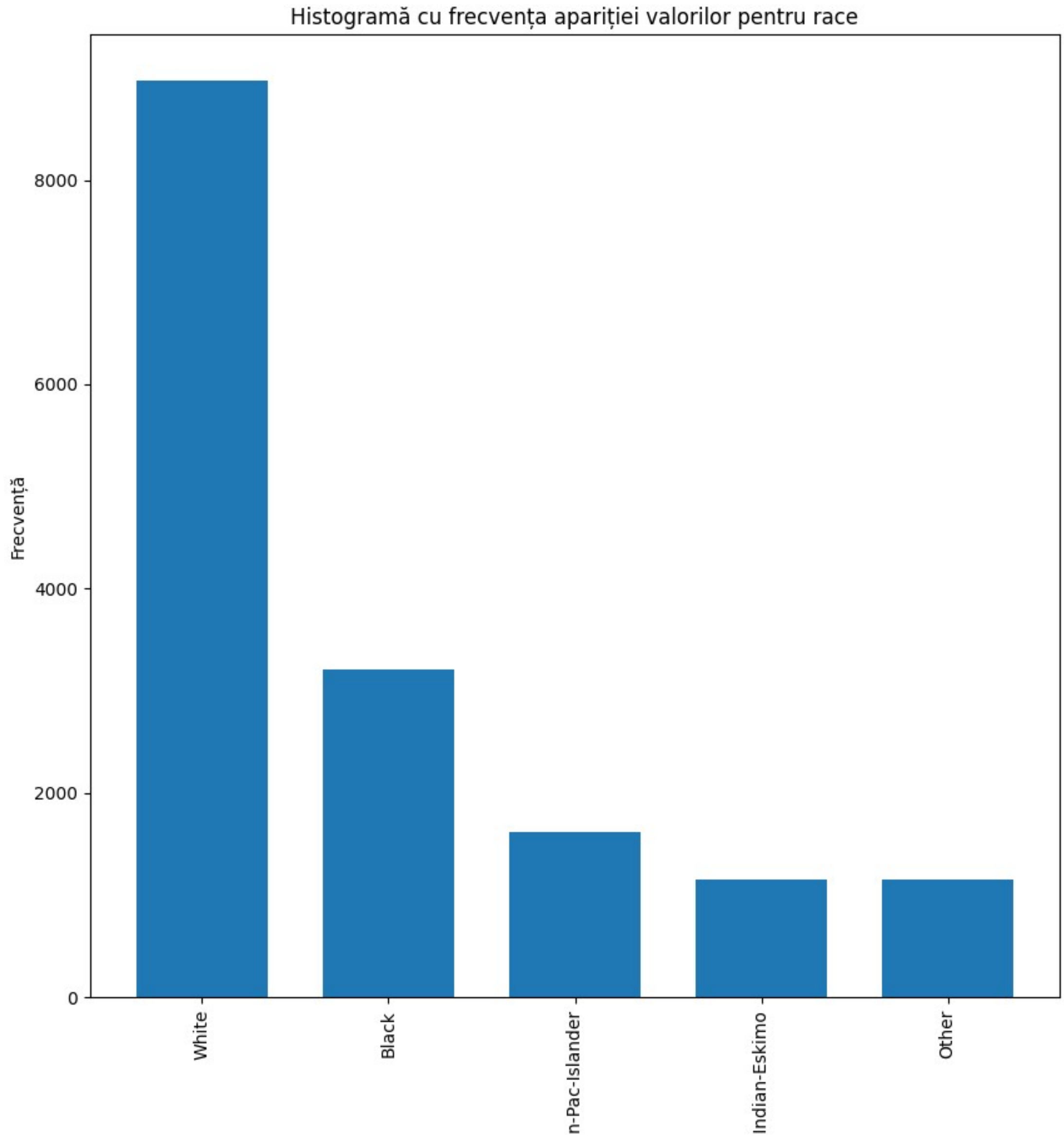
Histogramă cu frecvența apariției valorilor pentru money
Frecvență
8000
6000
4000
2000
0
<=50K
>50K
Valoare

3.4.7 Partner



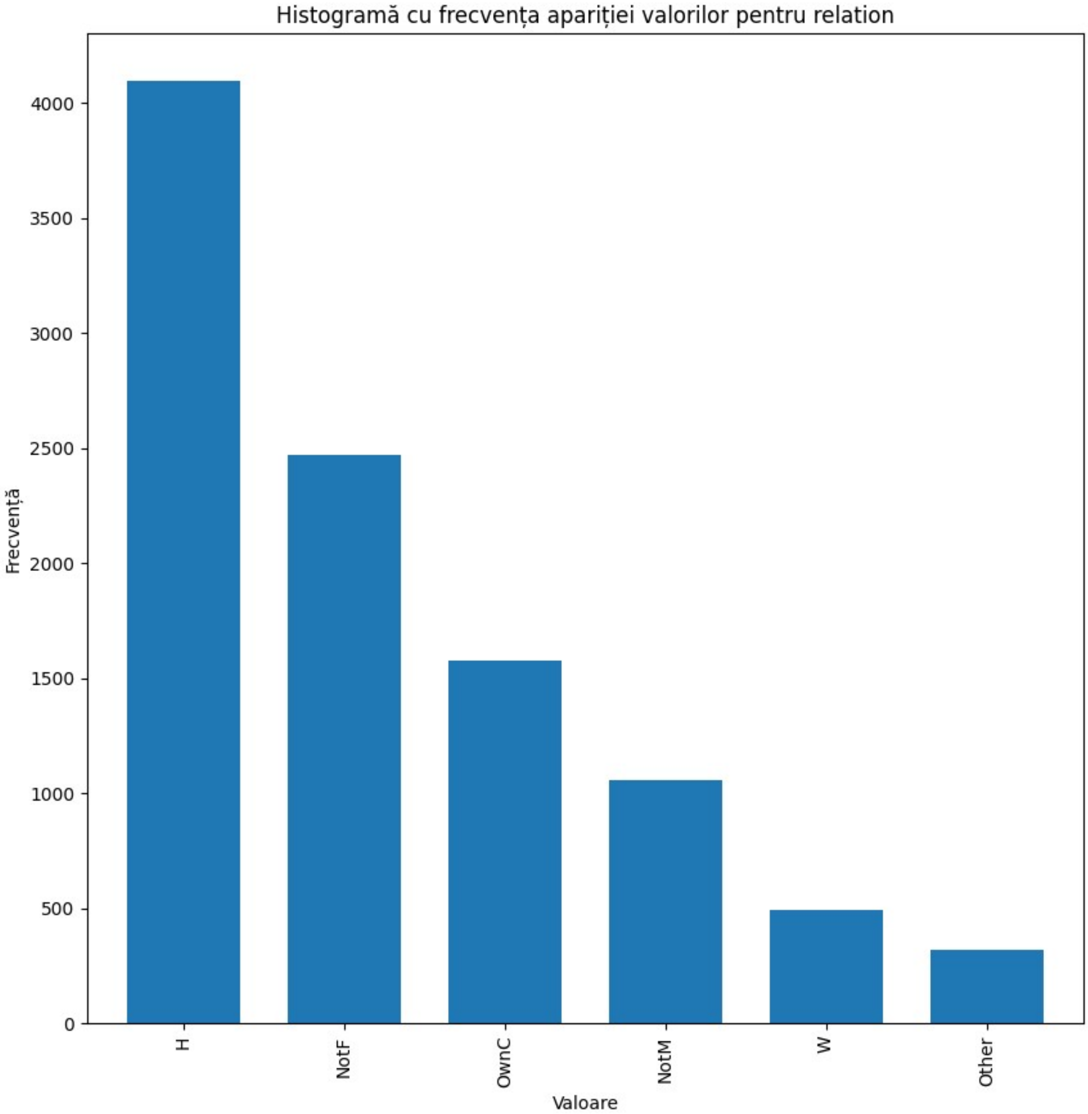
Histogramă cu frecvența apariției valorilor pentru partner
Frecvență
0
2000
4000
6000
8000
MCS
NM
D
S
W
MSA
MAS
Valoare

3.4.8 Race

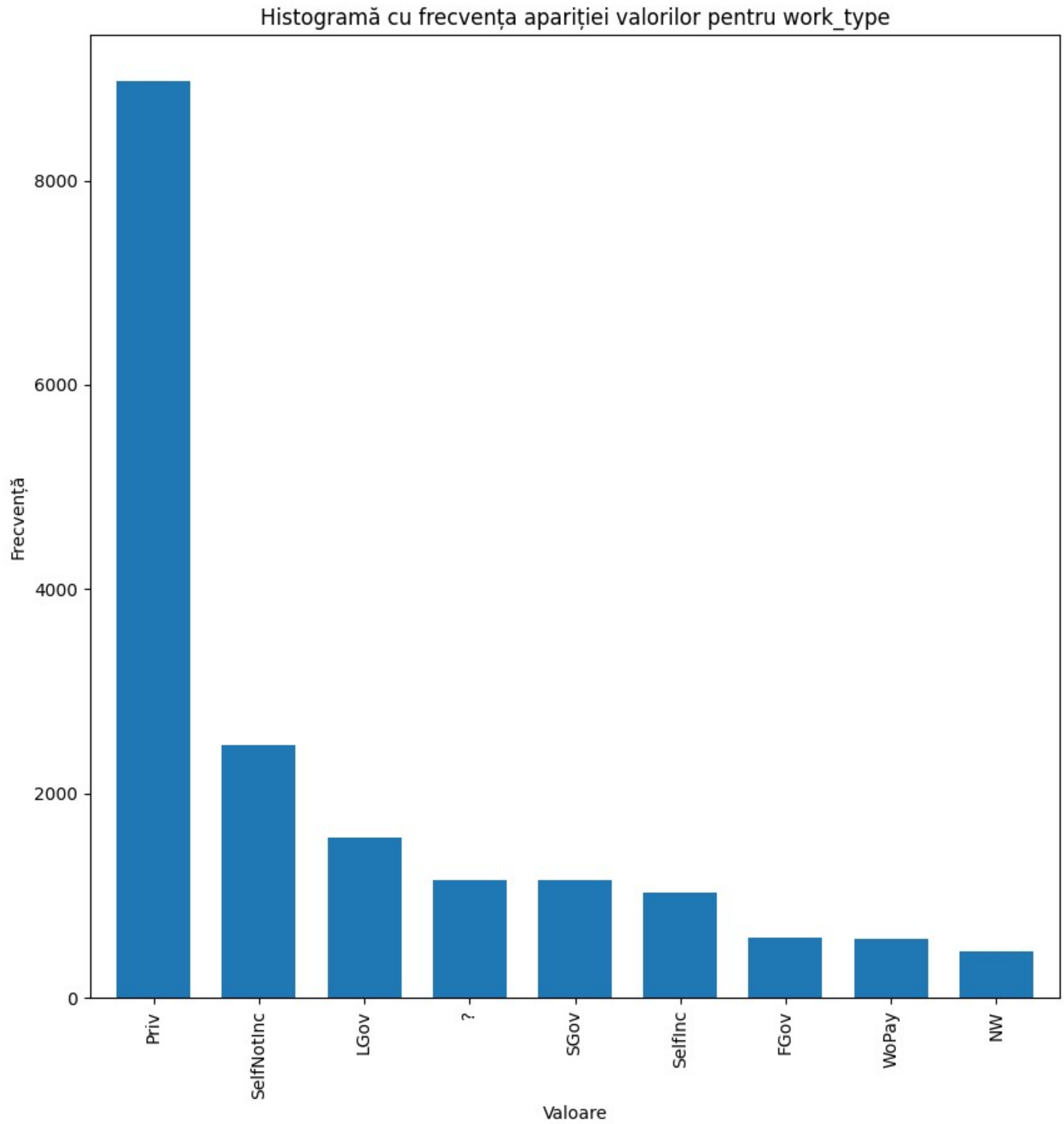


Histogramă cu frecvența apariției valorilor pentru race
Frecvență
0
2000
4000
6000
8000
White
Black
n-Pac-Islander
Indian-Eskimo
Other

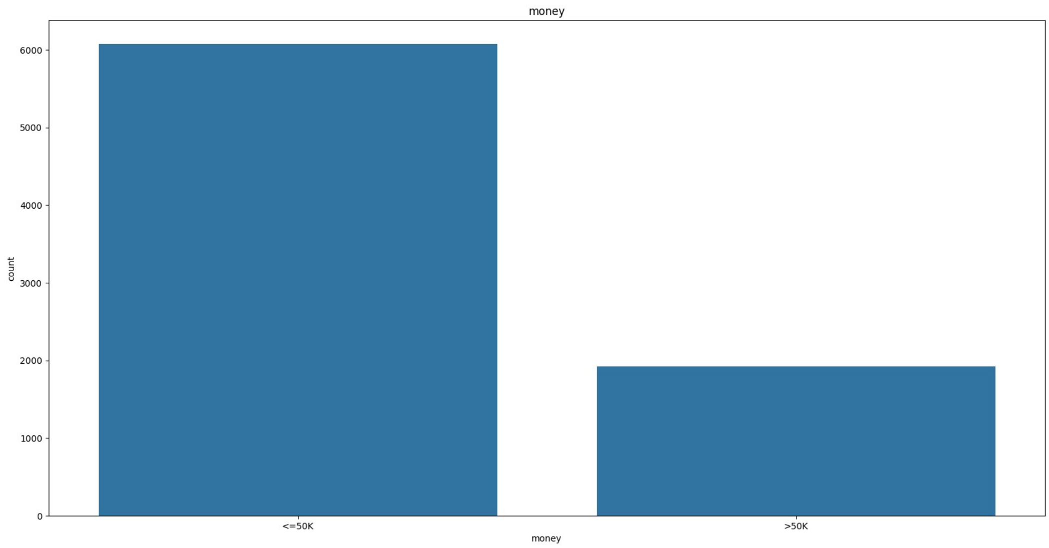
3.4.9 Relation

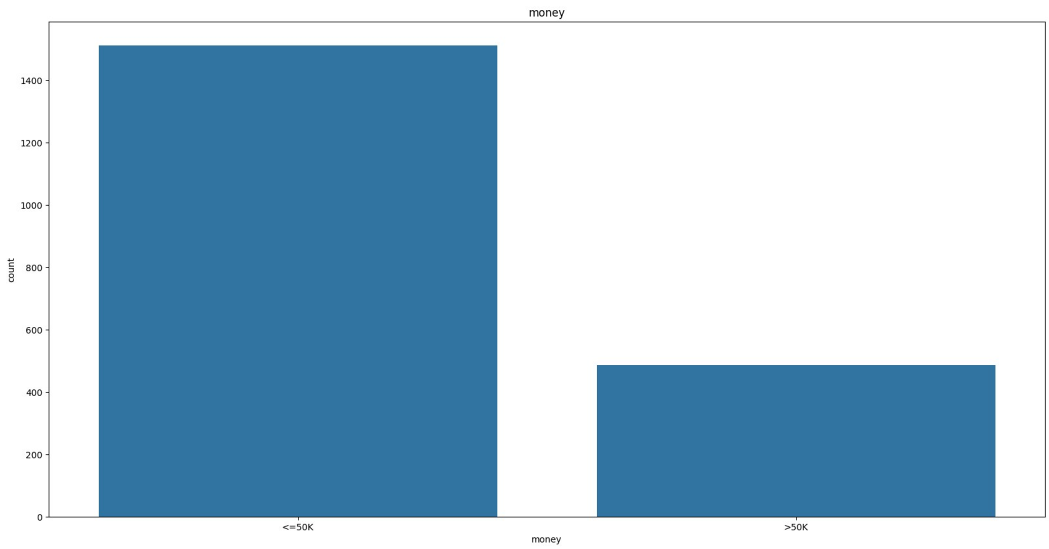


3.4.10 Work type



3.5 Analiza echilibru clasa train vs test Salary

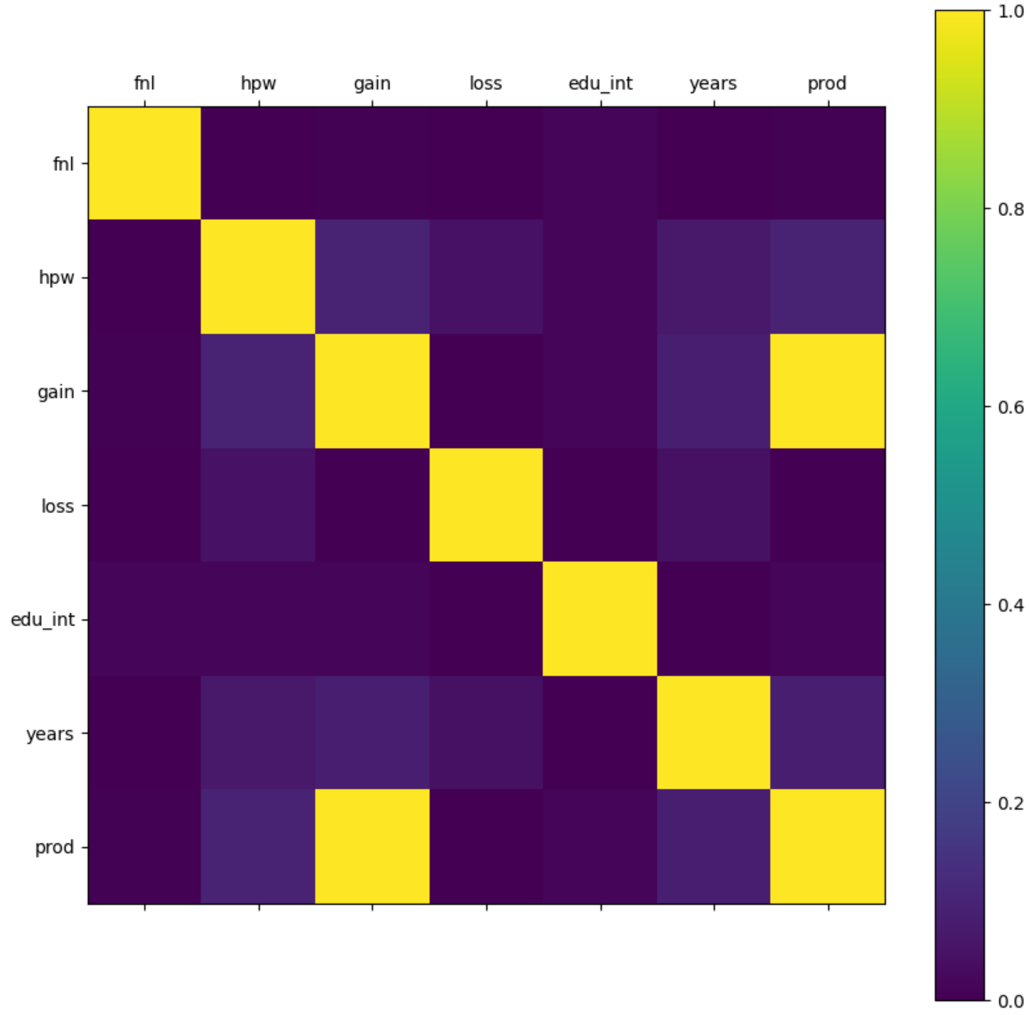




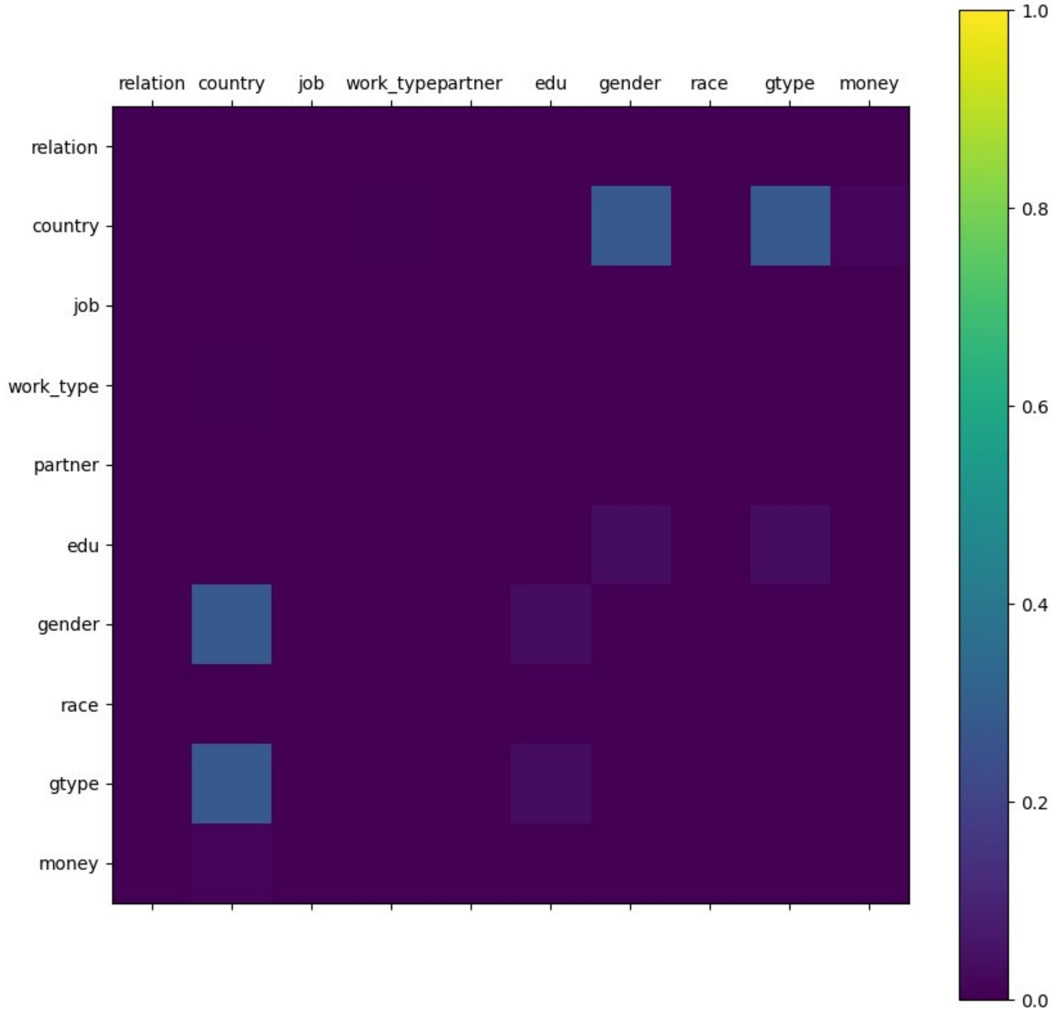
Si pe acest set de date se vede iar o dezechilibrare mare.

3.6 Analiza corelației între atribute Salary

3.6.1 Date continue



2.6.2 Date discrete



3.7 Concluzie
Clasa este dezechilibrata si exista valori extreme si nume in setul de date.

4. Preprocesare
Am inlocuit la ambele seturi de date toate valorile nule si cele extreme si apoi am realizat o standardizare a datelor pentru a putea aplica algoritmii. Am aplicat SimpleImputer pentru a pune valori valori in locul celor nule si am pus la cele categorice valoarea cea mai intalnita per atribut si la cele numerice valoarea medie. Apoi valorile extreme le-am inlocuit cu media. Pentru stanadarizare am folosit StandardScaler.

5. Regresie logistica
- Am folosit ca encoder pentru atributele categorice LabelEncoder
- Setarile algoritmului de optimizare de tip gradient descent: tip optimizator(cel simplu)

6. MLP
- Arhitectura: am folsit la implementarea de mana straturi:
  - ➔ Linear: ca input numarul de coloane primit din setul de date si ca out HIDDEN_UNITS=300(ca in laborator)
  - ➔ activari RELU

- Configurarea optimizatorului:mode= ‘SGB’ , lr = 0.005 (si in cel de la laborator si cel din sklearn)

7.Evaluare algoritmi AVC
7.1 Descriere a setului de hiperparametrii
7.1.1 AVC
Regresie Logistica: LR = 0.01, NUMBER_EPOCH = 100
Regresie Logistica: maxiters = 700, solver = liblinear
MLP: maxiters = 700, solver = SGD
MLP laborator: BATCH_SIZE = 128, HIDDEN_UNITS = 300,
NUMBER_EPOCH = 100, solver = SGD si LR = 0.005

7.1.2 Salary
Regresie Logistica: LR = 0.01, NUMBER_EPOCH = 100
Regresie Logistica: maxiters = 700, solver = liblinear
MLP: maxiters = 700, solver = SGD
MLP laborator: BATCH_SIZE = 128, HIDDEN_UNITS = 300,
NUMBER_EPOCH = 100, solver = SGD si LR = 0.005

7.2 Matrice de confuzie si tabel comparativ AVC

LR laborator

```
[[931  16]
 [ 73   2]]
```

LR SKLEARN

```
[[946   1]
 [ 75   0]]
```

MLP SKLEARN

```
[[947   0]
 [ 75   0]]
```

MLP laborator

```
[[946   1]
 [ 75   0]]
```

Tabel comparativ AVC

LR laborator

	precision	recall	f1-score	support
without avc	0.9273	0.9831	0.9544	947
with avc	0.1111	0.0267	0.0430	75
accuracy			0.9129	1022
macro avg	0.5192	0.5049	0.4987	1022
weighted avg	0.8674	0.9129	0.8875	1022

LR sklearn

	precision	recall	f1-score	support
without avc	0.9265	0.9989	0.9614	947
with avc	0.0000	0.0000	0.0000	75
accuracy			0.9256	1022
macro avg	0.4633	0.4995	0.4807	1022
weighted avg	0.8585	0.9256	0.8908	1022

MLP sklearn

	precision	recall	f1-score	support
without avc	0.9266	1.0000	0.9619	947
with avc	0.0000	0.0000	0.0000	75
accuracy			0.9266	1022
macro avg	0.4633	0.5000	0.4810	1022
weighted avg	0.8586	0.9266	0.8913	1022

MLP laborator

	precision	recall	f1-score	support
without avc	0.9265	0.9989	0.9614	947
with avc	0.0000	0.0000	0.0000	75
accuracy			0.9256	1022
macro avg	0.4633	0.4995	0.4807	1022
weighted avg	0.8585	0.9256	0.8908	1022

7.3 Concluzie AVC

Se poate observa ca din punct de vedere acuratete MLP este mai bun decat regresia logistica. Totusi, doar regresia de la laborator a putut sa-mi detecteze si persoanele cu AVC, in rest pare ca nu prea exista. Acest lucru e destul de firesc sa se intample uitandu-ne pe clasa cine are si nu are “cerebrovascular_accident”, deoarece sunt foarte multe cu nu si putine cu da, atat pe train cat si pe test, ceea ce indica un dezechilibru mare. Totusi se poate observa dupa acuratetea de la ambele ca au prezis destul de bine.

7.4 Matrice de confuzie si tabel comparativ Salary Classification

LR laborator

```
[[1329  184]
 [ 372  115]]
```

LR sklearn

```
[[1329  184]
 [ 372  115]]
```

MLP sklearn

```
[[1329  184]
 [ 372  115]]
```

MLP laborator

```
[[1402  111]
 [ 239  248]]
```

Tabel comparativ

LR laborator

```
              precision    recall  f1-score   support

      <= 50K     0.7813    0.8784    0.8270      1513
       > 50K     0.3846    0.2361    0.2926       487

    accuracy                         0.7220      2000
   macro avg     0.5830    0.5573    0.5598      2000
weighted avg     0.6847    0.7220    0.6969      2000
```

LR sklearn

	precision	recall	f1-score	support
<= 50K	0.8361	0.9405	0.8852	1513
> 50K	0.6980	0.4271	0.5299	487
accuracy			0.8155	2000
macro avg	0.7670	0.6838	0.7076	2000
weighted avg	0.8025	0.8155	0.7987	2000

MLP sklearn

	precision	recall	f1-score	support
<= 50K	0.8426	0.9484	0.8924	1513
> 50K	0.7374	0.4497	0.5587	487
accuracy			0.8270	2000
macro avg	0.7900	0.6991	0.7255	2000
weighted avg	0.8170	0.8270	0.8111	2000

MLP laborator

	precision	recall	f1-score	support
<= 50K	0.8544	0.9266	0.8890	1513
> 50K	0.6908	0.5092	0.5863	487
accuracy			0.8250	2000
macro avg	0.7726	0.7179	0.7377	2000
weighted avg	0.8145	0.8250	0.8153	2000

7.5 Concluzii Salary Classification

Spre deosebire de setul de date trecut, aici avem mult mai multe preziceri, dar care nu sunt neaparat unele bune. Datele sunt mult mai multe decat la AVC si la fel clasele sunt dezechilibrate. Cel mai bine pare ca a prezis MLP implementat la laborator care

are si acuratetea cea mai mare si in matricea de confuzie, nu a gresit la fel de des la negative si are mai multe true negative decat celelalte.

Pentru ambele seturi de date am ales la algoritmi hiperparametrii care dadeau cele mai bune rezultate.