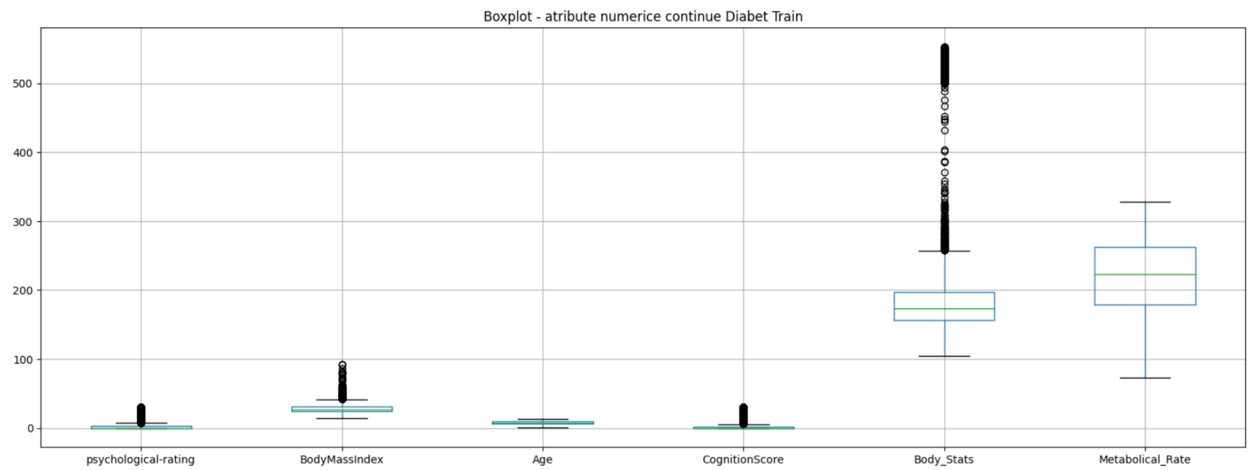
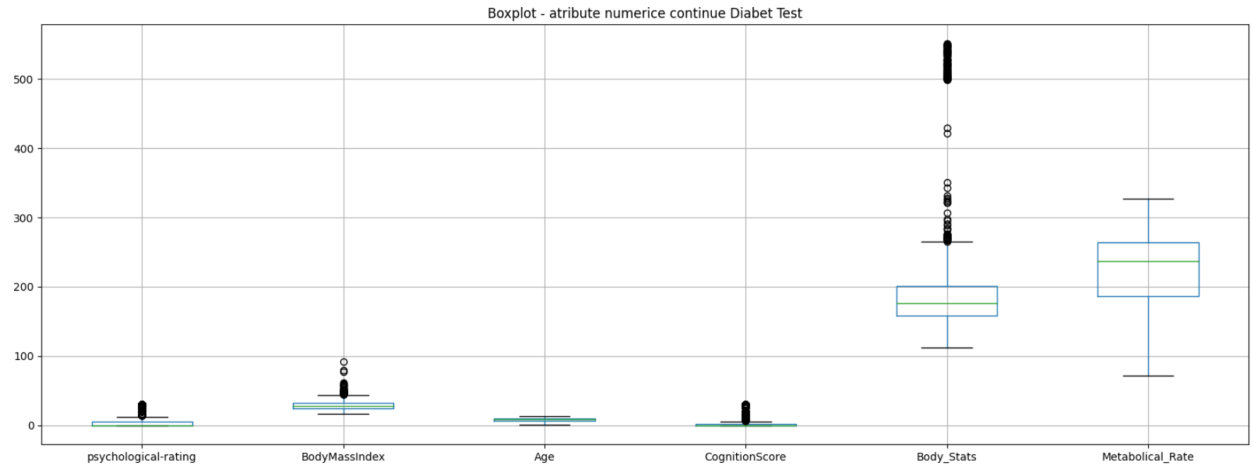
Valori numerice:



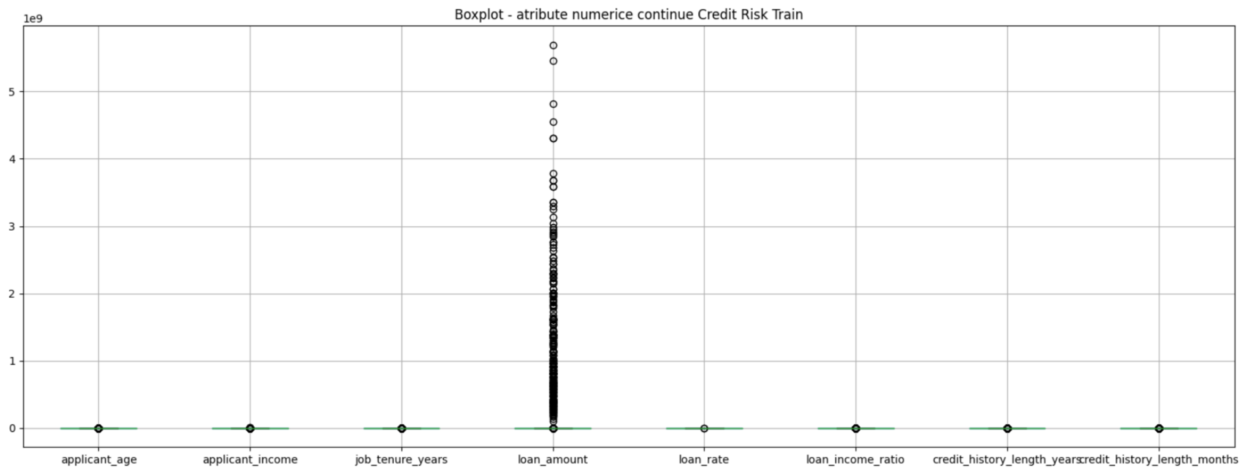

```
                      count        mean        std         min  \
psychological-rating  8000.0    4.150250   8.661253    0.000000
BodyMassIndex         8000.0   28.111250   6.409232   14.000000
Age                   8000.0    8.000000   3.060665    1.000000
CognitionScore        8000.0    3.033875   7.191310    0.000000
Body_Stats            8000.0  194.298697  82.505215  105.063984
Metabolical_Rate      7000.0  220.381400  60.954867   72.774739

                      percentila 25%  percentila 50%  percentila 75%  \
psychological-rating        0.000000        0.000000        3.000000
BodyMassIndex              24.000000       27.000000       31.000000
Age                         6.000000        8.000000       10.000000
CognitionScore              0.000000        0.000000        2.000000
Body_Stats                156.304379      173.402202      196.796633
Metabolical_Rate          179.577803      223.029040      262.350158

                             max
psychological-rating   30.000000
BodyMassIndex          92.000000
Age                    13.000000
CognitionScore         30.000000
Body_Stats            553.000000
Metabolical_Rate      327.936098
```



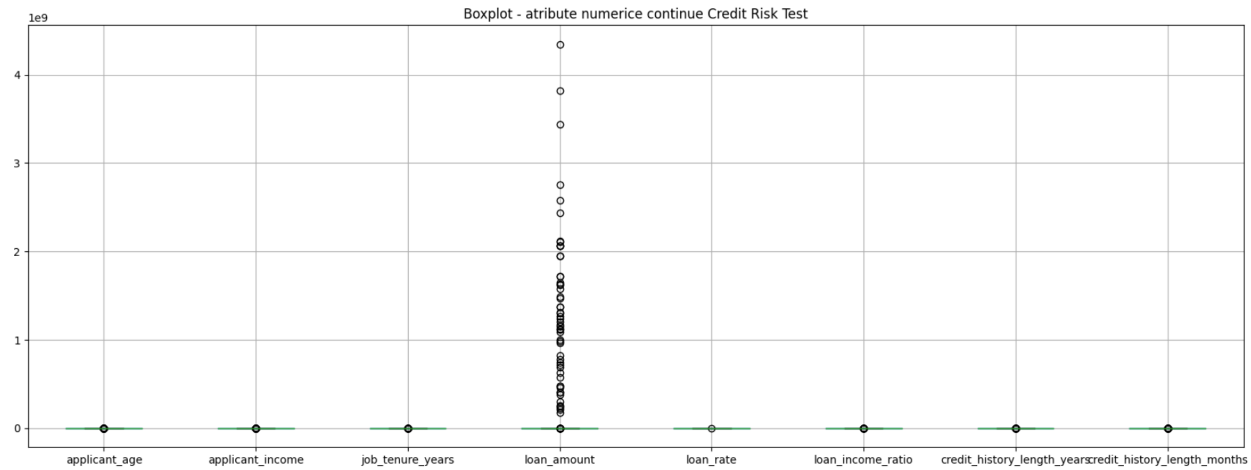
	count	mean	std	min	\
psychological-rating	2000.0	5.224500	9.711398	0.000000	
BodyMassIndex	2000.0	28.787500	6.645783	16.000000	
Age	2000.0	8.287500	2.926282	1.000000	
CognitionScore	2000.0	3.491000	7.751191	0.000000	
Body_Stats	2000.0	197.609132	82.136327	111.704232	
Metabolical_Rate	2000.0	225.831345	58.610273	71.602207	

	percentila 25%	percentila 50%	percentila 75%	\
psychological-rating	0.000000	0.000000	5.000000	
BodyMassIndex	24.000000	28.000000	32.000000	
Age	6.000000	9.000000	10.000000	
CognitionScore	0.000000	0.000000	2.000000	
Body_Stats	158.274650	176.660060	201.001498	
Metabolical_Rate	185.709759	236.422446	263.714385	

	max
psychological-rating	30.000000
BodyMassIndex	92.000000
Age	13.000000
CognitionScore	30.000000
Body_Stats	551.000000
Metabolical_Rate	326.525697



	count	mean	std	min	percentila 25%	percentila 50%	percentila 75%	max
applicant_age	8000.0	2.778912e+01	6.464200e+00	20.00	23.00	26.00	30.00	1.230000e+02
applicant_income	8000.0	6.566213e+04	5.883553e+04	4200.00	38400.00	55000.00	78433.25	2.039784e+06
job_tenure_years	7786.0	4.729001e+00	4.203528e+00	0.00	2.00	4.00	7.00	1.230000e+02
loan_amount	8000.0	3.887819e+07	2.845087e+08	500.00	5000.00	8000.00	13500.00	5.695115e+09
loan_rate	7257.0	1.101108e+01	3.274046e+00	5.42	7.88	10.99	13.47	2.211000e+01
loan_income_ratio	8000.0	1.704438e-01	1.069076e-01	0.00	0.09	0.15	0.23	7.100000e-01
credit_history_length_years	8000.0	5.857500e+00	4.096497e+00	2.00	3.00	4.00	8.00	3.000000e+01
credit_history_length_months	7200.0	7.644986e+01	4.943491e+01	25.00	41.00	57.00	103.00	3.690000e+02



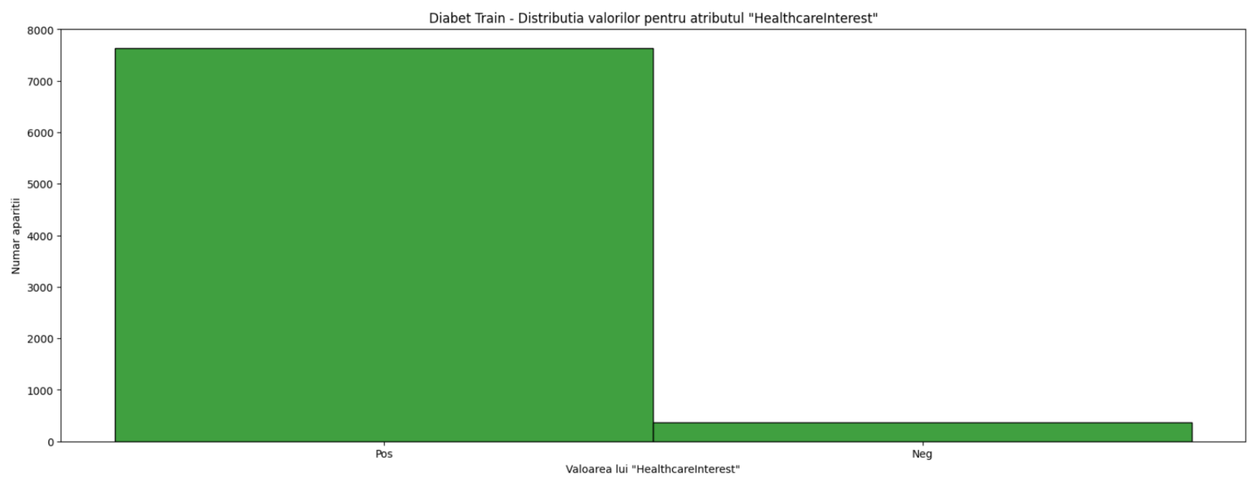
	count	mean	std	min
applicant_age	2000.0	2.756900e+01	5.924001e+00	20.00
applicant_income	2000.0	6.602253e+04	4.866238e+04	8088.00
job_tenure_years	1950.0	5.012308e+00	4.899802e+00	0.00
loan_amount	2000.0	3.660534e+07	2.637011e+08	500.00
loan_rate	1803.0	1.099148e+01	3.236259e+00	5.42
loan_income_ratio	2000.0	1.688750e-01	1.064559e-01	0.01
credit_history_length_years	2000.0	5.625500e+00	3.854955e+00	2.00
credit_history_length_months	2000.0	7.356200e+01	4.633302e+01	25.00

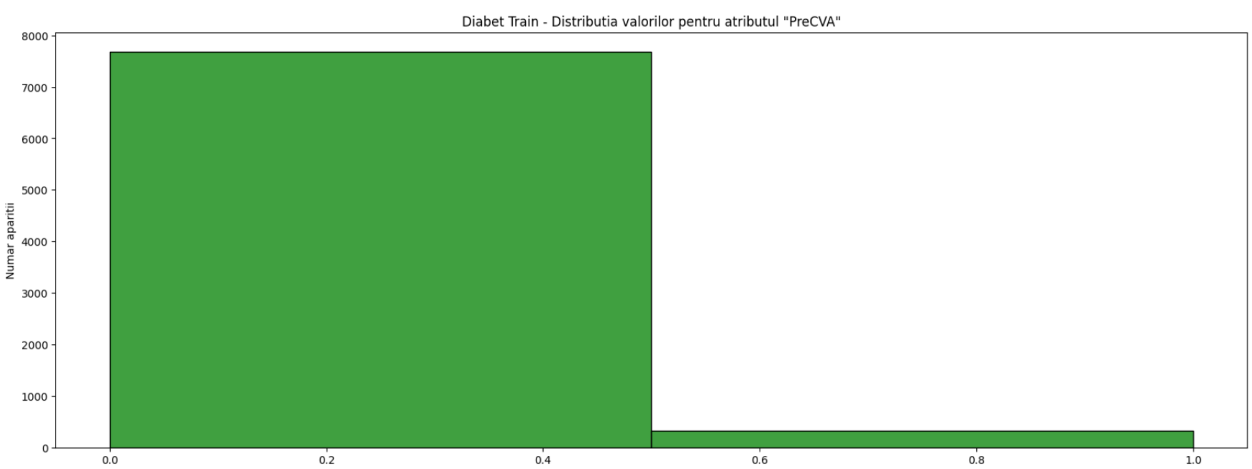
	percentila 25%	percentila 50%	percentila 75%
applicant_age	23.00	26.00	30.00
applicant_income	39540.00	56000.00	80000.00
job_tenure_years	2.00	4.00	7.00
loan_amount	5000.00	8000.00	13200.00
loan_rate	7.90	10.99	13.45
loan_income_ratio	0.09	0.14	0.23
credit_history_length_years	3.00	4.00	8.00
credit_history_length_months	40.00	56.00	98.25

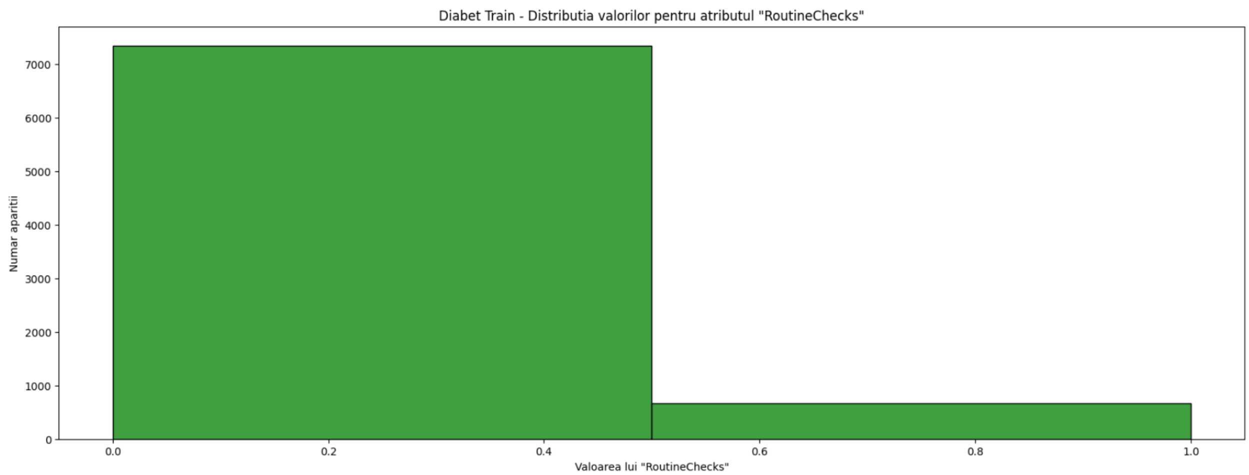
	max
applicant_age	6.900000e+01
applicant_income	7.620000e+05
job_tenure_years	1.230000e+02
loan_amount	4.346934e+09
loan_rate	2.322000e+01
loan_income_ratio	7.600000e-01
credit_history_length_years	2.900000e+01
credit_history_length_months	3.570000e+02

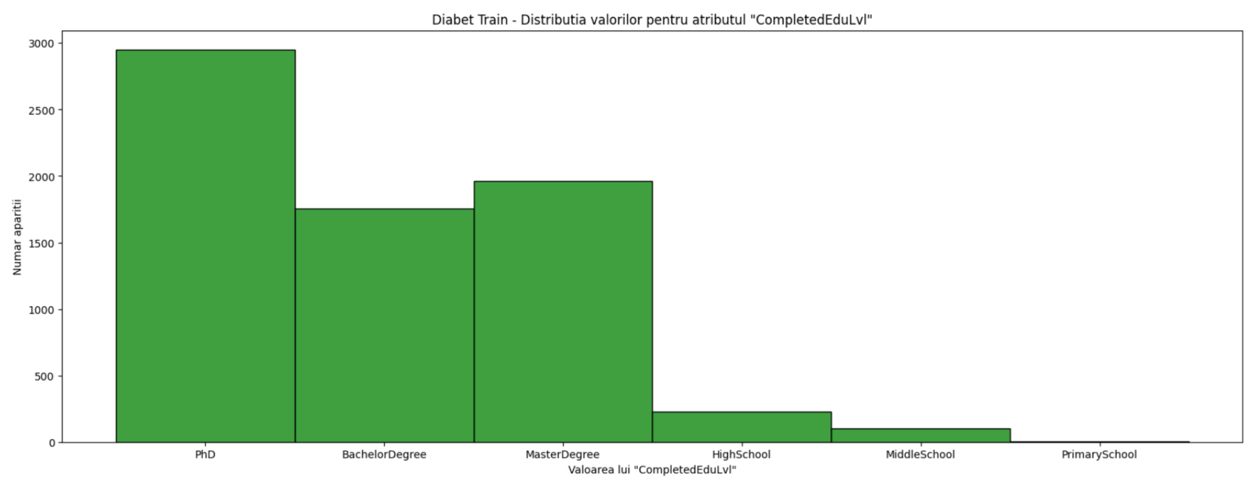
Valori categorice:

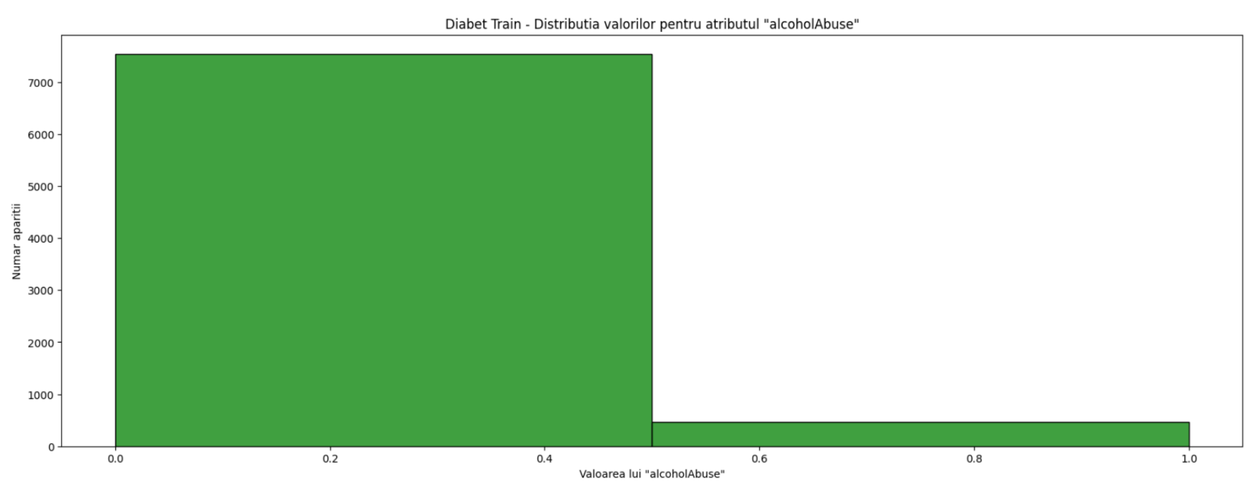
	valori nenule	valori unice
PreCVA	8000	2
RoutineChecks	8000	2
alcoholAbuse	8000	2
cholesterol_ver	8000	2
vegetables	8000	2
HighBP	8000	2
Unprocessed_fructose	8000	2
HealthScore	8000	5
myocardial_infarction	8000	2
SalaryBraket	8000	8
Cardio	8000	2
ImprovedAveragePulmonaryCapacity	8000	2
Diabetes	8000	3

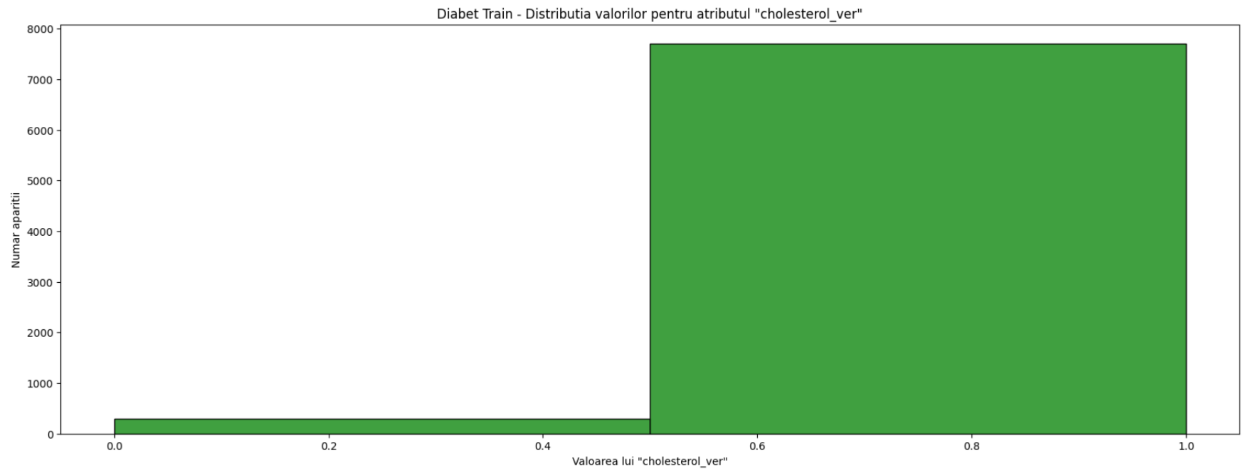


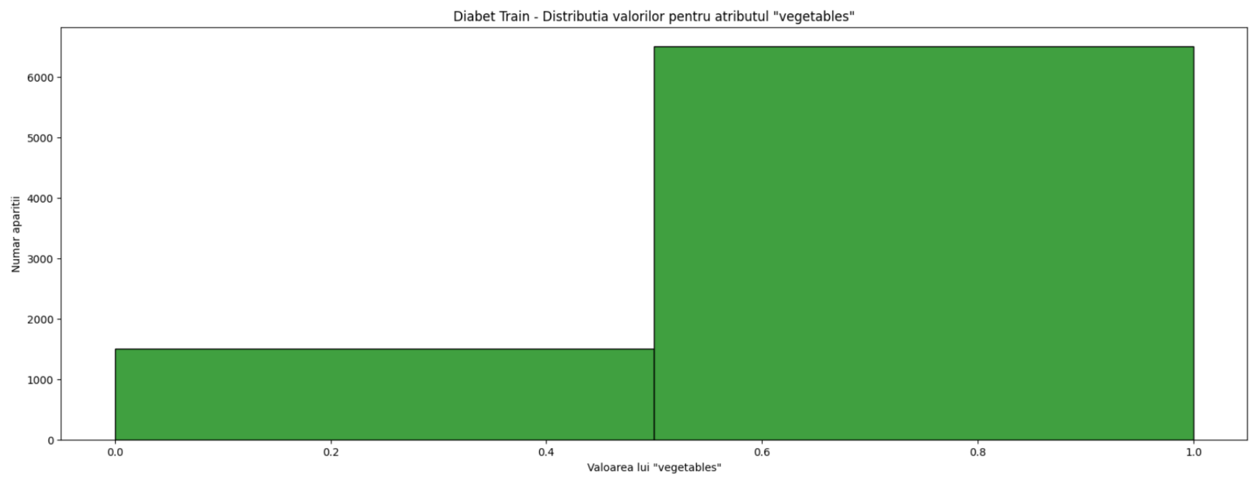


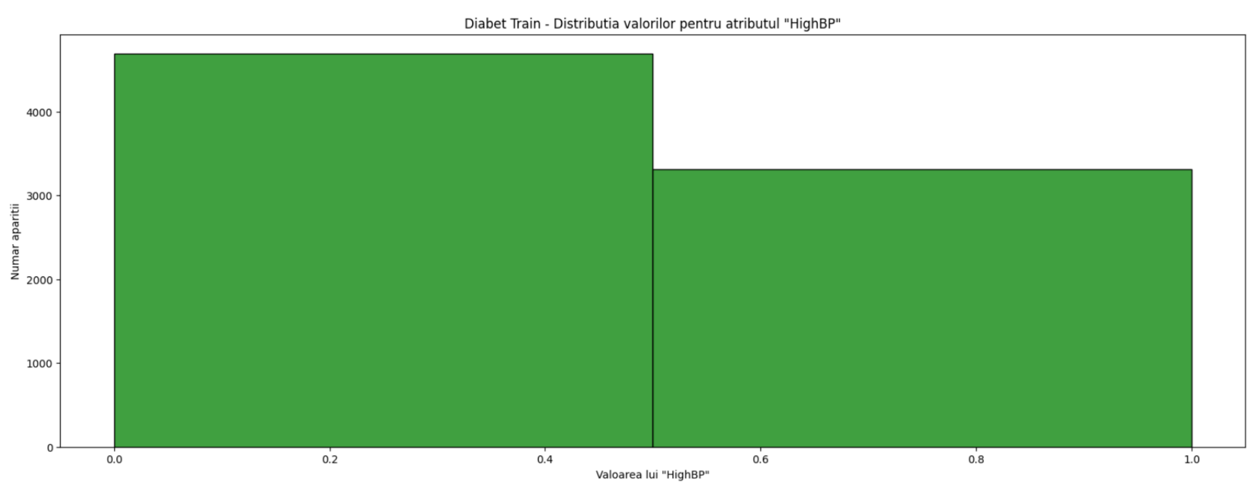

Diabet Train - Distributia valorilor pentru atributul "RoutineChecks"
Numar aparitii
0
1000
2000
3000
4000
5000
6000
7000
0.0
0.2
0.4
0.6
0.8
1.0
Valoarea lui "RoutineChecks"

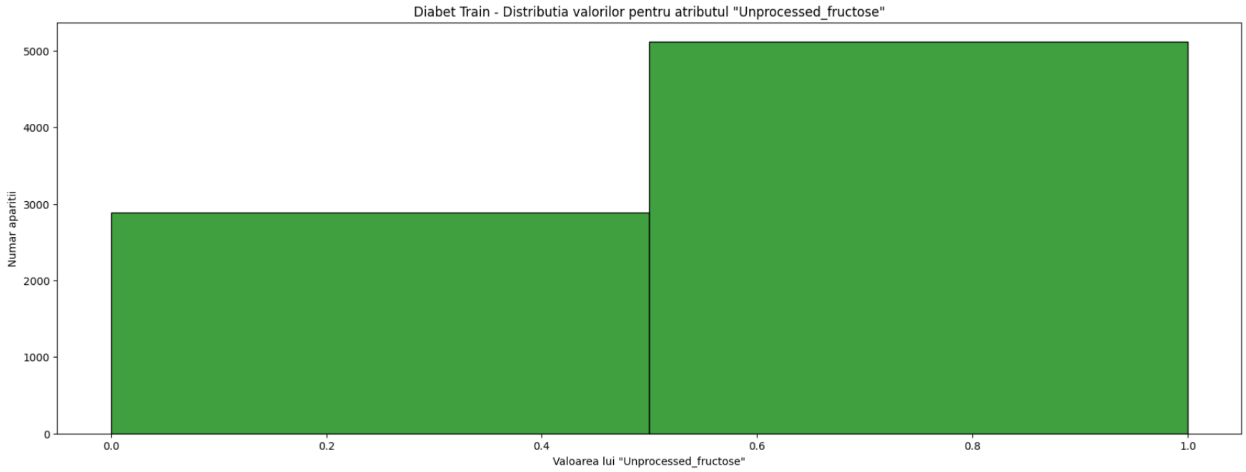

Diabet Train - Distributia valorilor pentru atributul "CompletedEduLvl"
Numar aparitii
0
500
1000
1500
2000
2500
3000
PhD
BachelorDegree
MasterDegree
HighSchool
MiddleSchool
PrimarySchool
Valoarea lui "CompletedEduLvl"

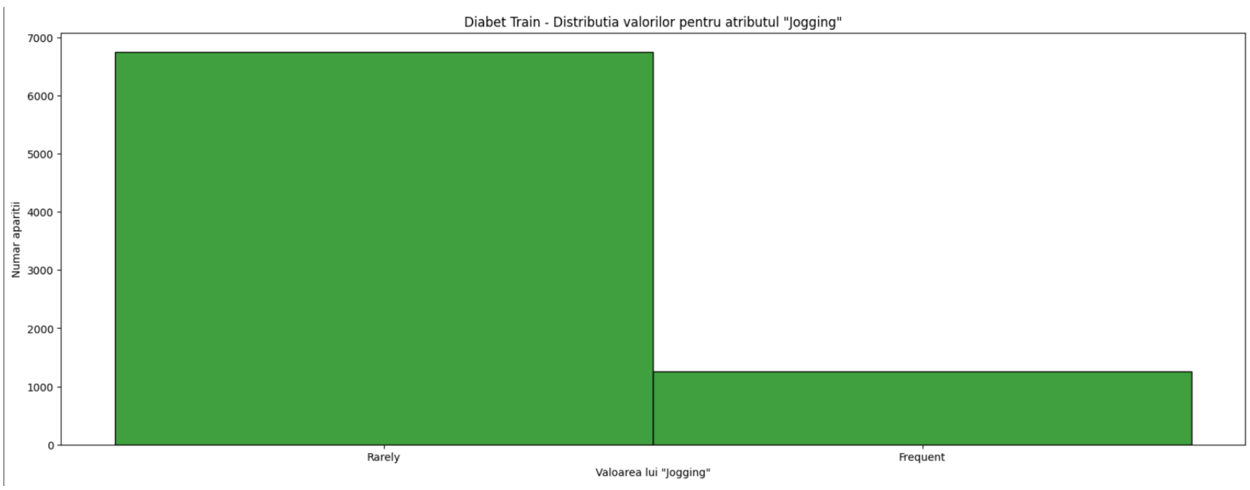

Diabet Train - Distributia valorilor pentru atributul "alcoholAbuse"
Numar aparitii
0
1000
2000
3000
4000
5000
6000
7000
0.0
0.2
0.4
0.6
0.8
1.0
Valoarea lui "alcoholAbuse"

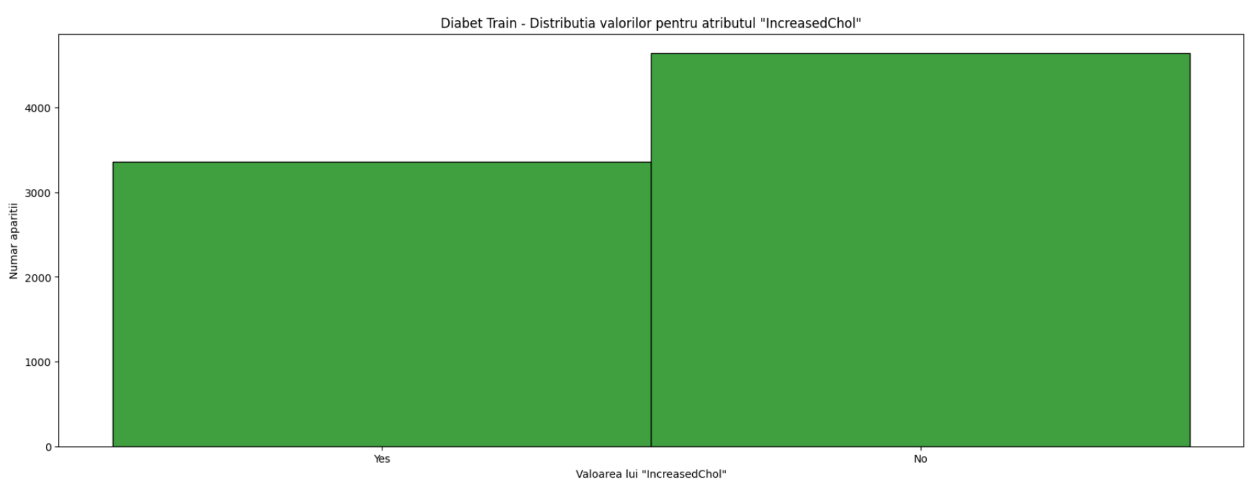

Diabet Train - Distributia valorilor pentru atributul "cholesterol_ver"
Numar aparitii
Valoarea lui "cholesterol_ver"
0
1000
2000
3000
4000
5000
6000
7000
8000
0.0
0.2
0.4
0.6
0.8
1.0

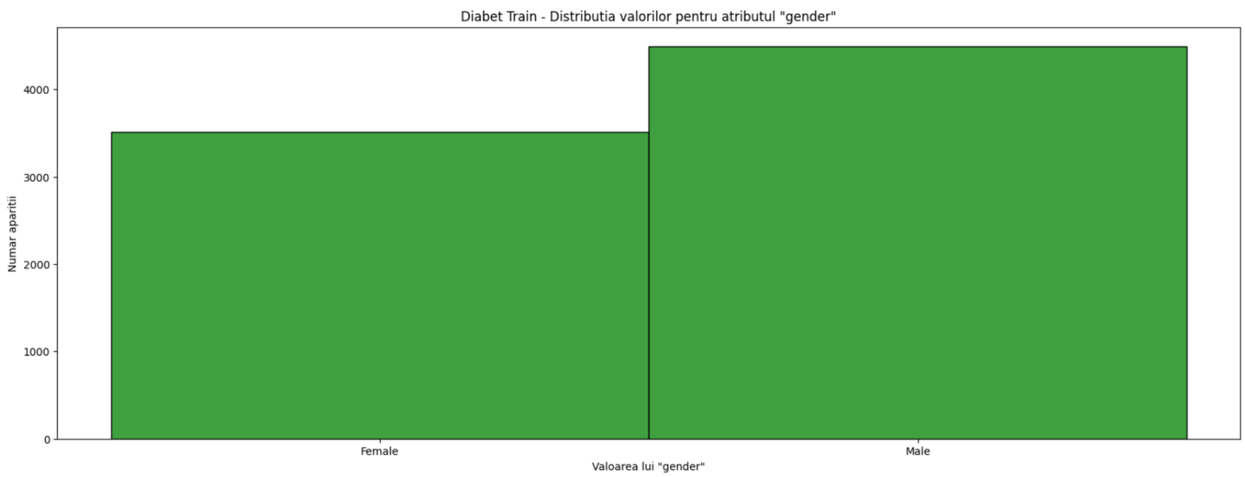

Diabet Train - Distributia valorilor pentru atributul "vegetables"
Numar aparitii
Valoarea lui "vegetables"
0
1000
2000
3000
4000
5000
6000
0.0
0.2
0.4
0.6
0.8
1.0

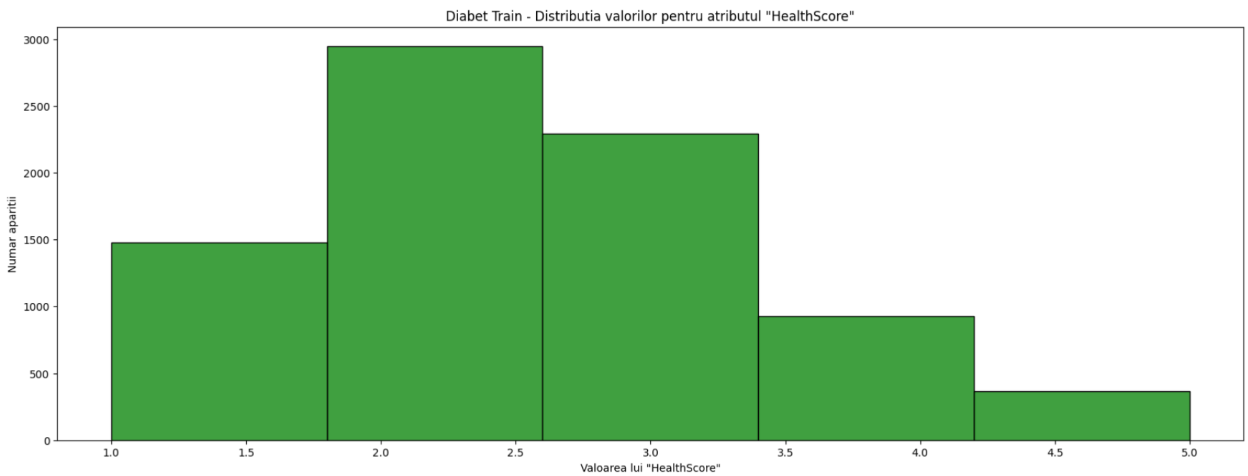

Diabet Train - Distributia valorilor pentru atributul "HighBP"
Numar aparitii
Valoarea lui "HighBP"
0
1000
2000
3000
4000
0.0
0.2
0.4
0.6
0.8
1.0

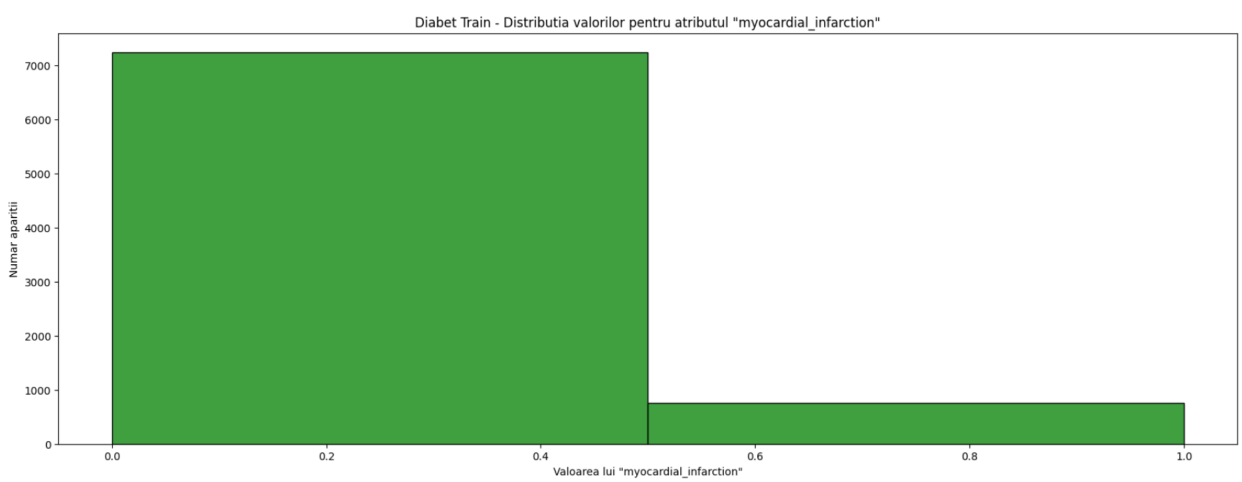

Diabet Train - Distributia valorilor pentru atributul "Unprocessed_fructose"
Numar aparitii
Valoarea lui "Unprocessed_fructose"
0
1000
2000
3000
4000
5000
0.0
0.2
0.4
0.6
0.8
1.0

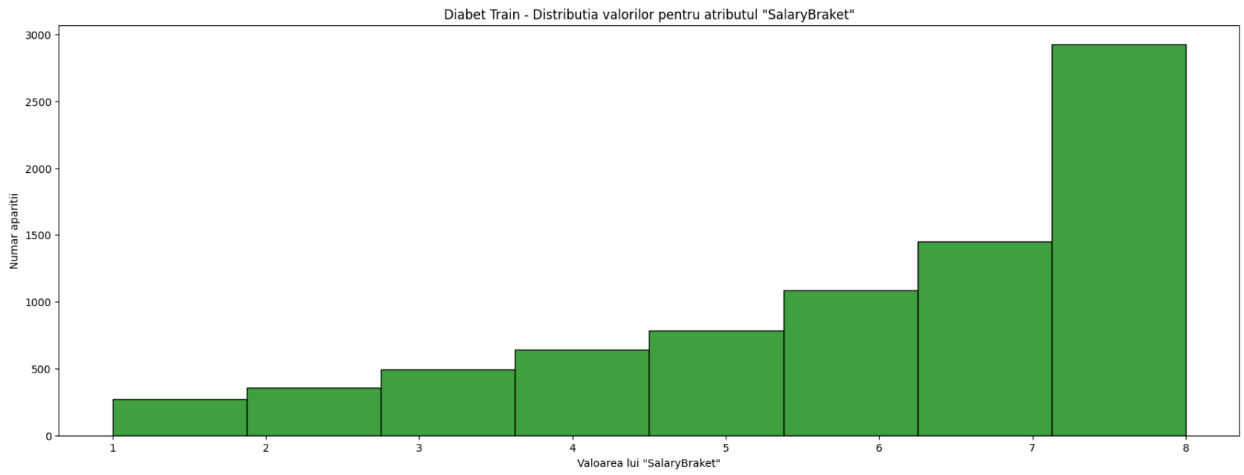

Diabet Train - Distributia valorilor pentru atributul "Jogging"
Numar aparitii
Valoarea lui "Jogging"
0
1000
2000
3000
4000
5000
6000
7000
Rarely
Frequent

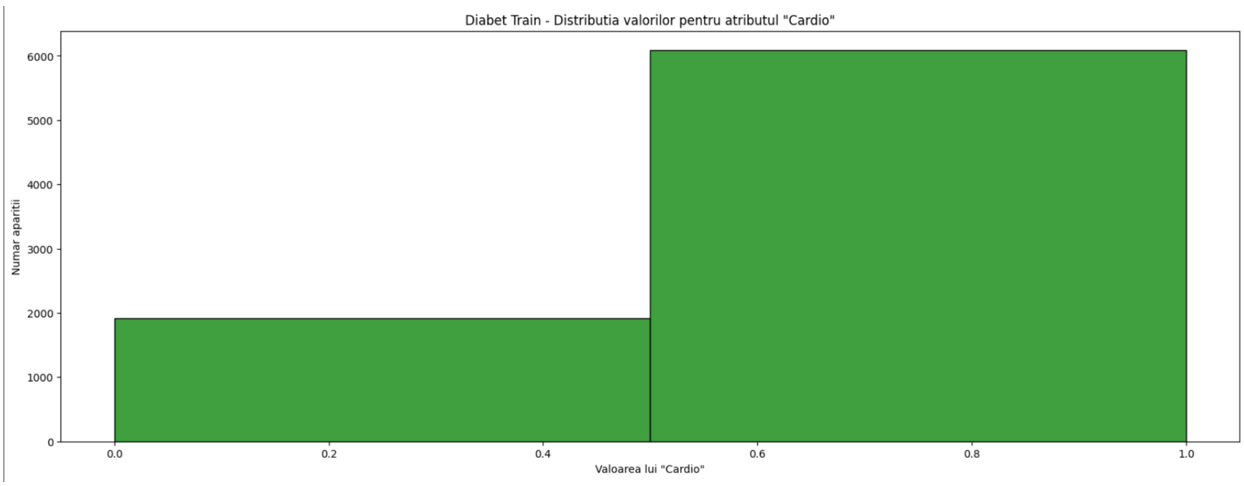

Diabet Train - Distributia valorilor pentru atributul "IncreasedChol"
Numar aparitii
Valoarea lui "IncreasedChol"
0
1000
2000
3000
4000
Yes
No

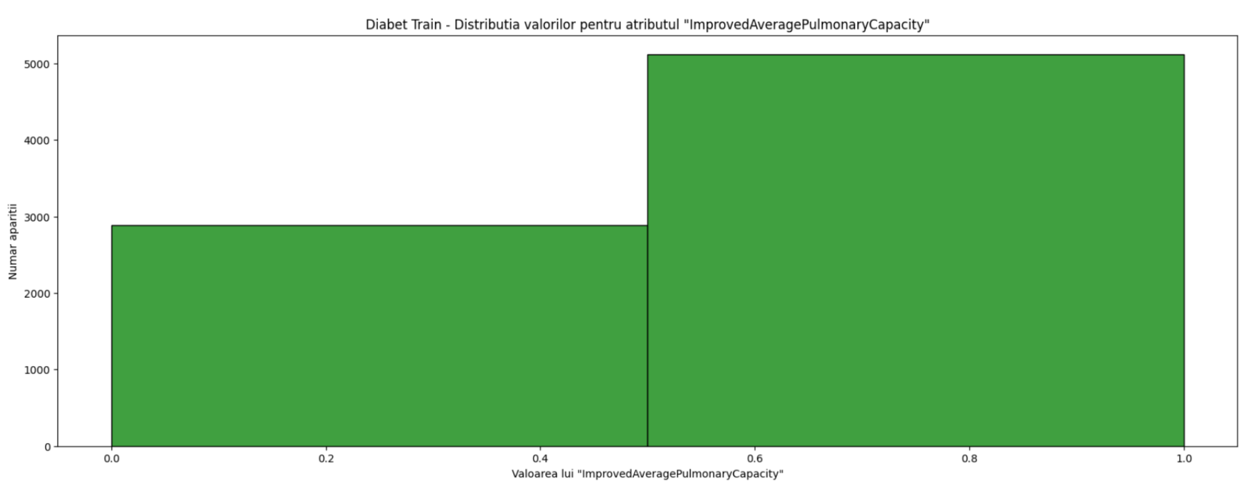

Diabet Train - Distributia valorilor pentru atributul "gender"
Numar aparitii
0
1000
2000
3000
4000
Female
Male
Valoarea lui "gender"

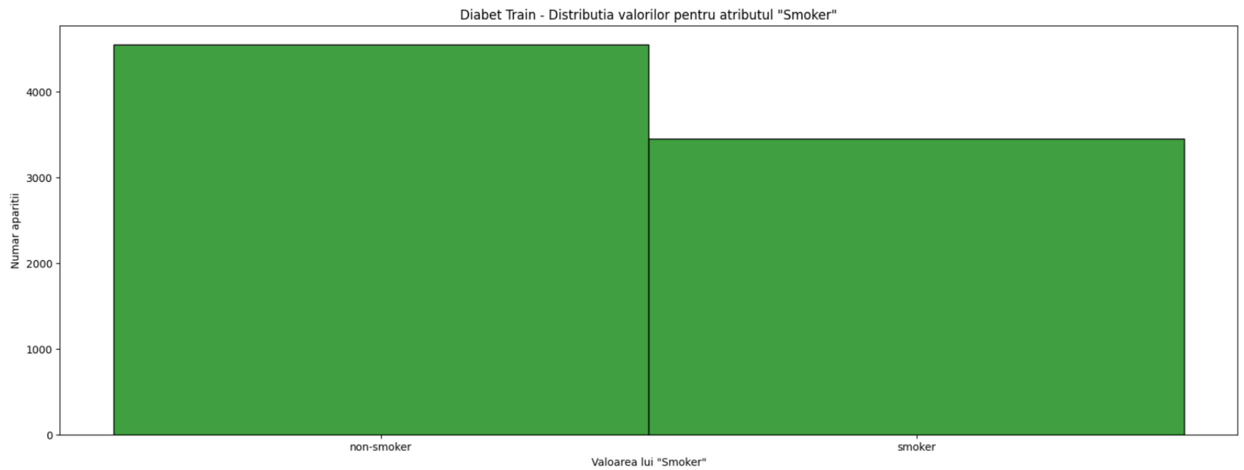

Diabet Train - Distributia valorilor pentru atributul "HealthScore"
Numar aparitii
0
500
1000
1500
2000
2500
3000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Valoarea lui "HealthScore"

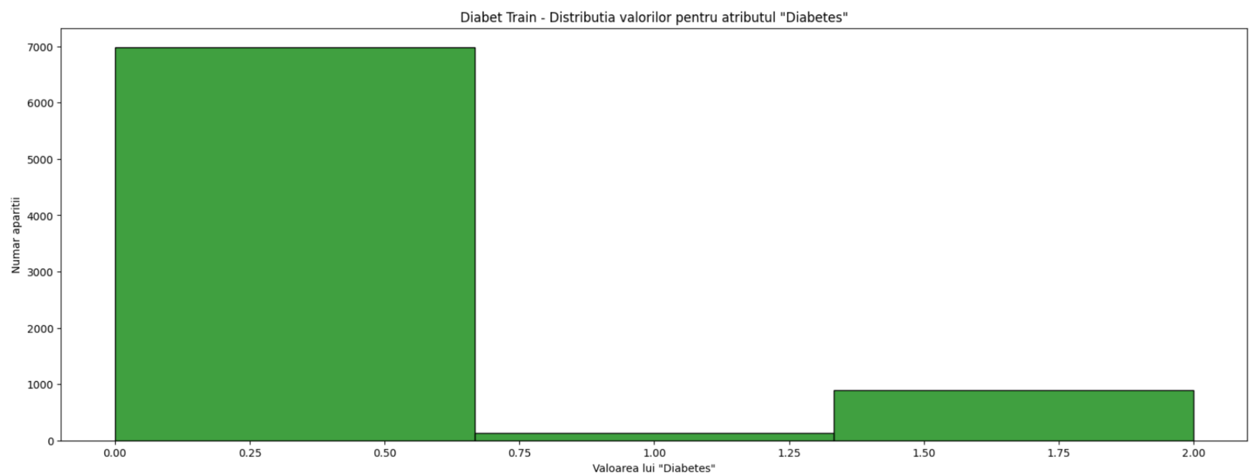

Diabet Train - Distributia valorilor pentru atributul "myocardial_infarction"
Numar aparitii
0
1000
2000
3000
4000
5000
6000
7000
0.0
0.2
0.4
0.6
0.8
1.0
Valoarea lui "myocardial_infarction"

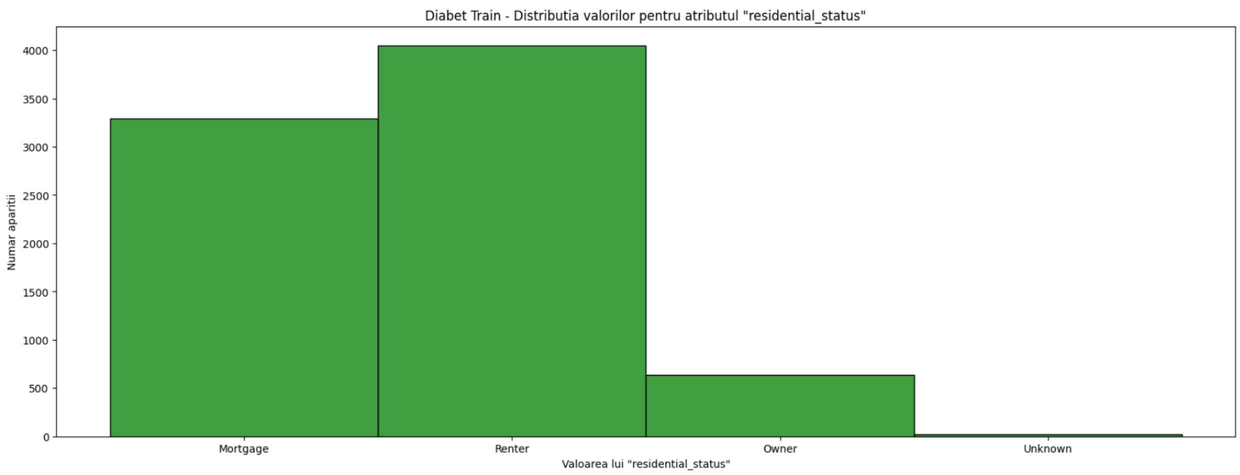

Diabet Train - Distributia valorilor pentru atributul "SalaryBraket"
Numar aparitii
Valoarea lui "SalaryBraket"
0
500
1000
1500
2000
2500
3000
1
2
3
4
5
6
7
8

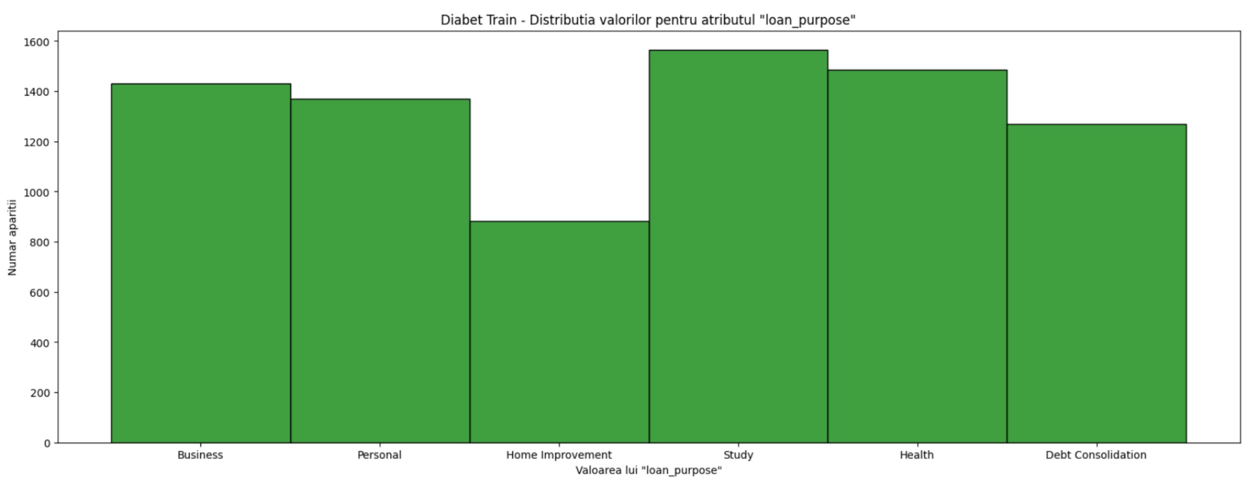

Diabet Train - Distributia valorilor pentru atributul "Cardio"
Numar aparitii
Valoarea lui "Cardio"
0
1000
2000
3000
4000
5000
6000
0.0
0.2
0.4
0.6
0.8
1.0

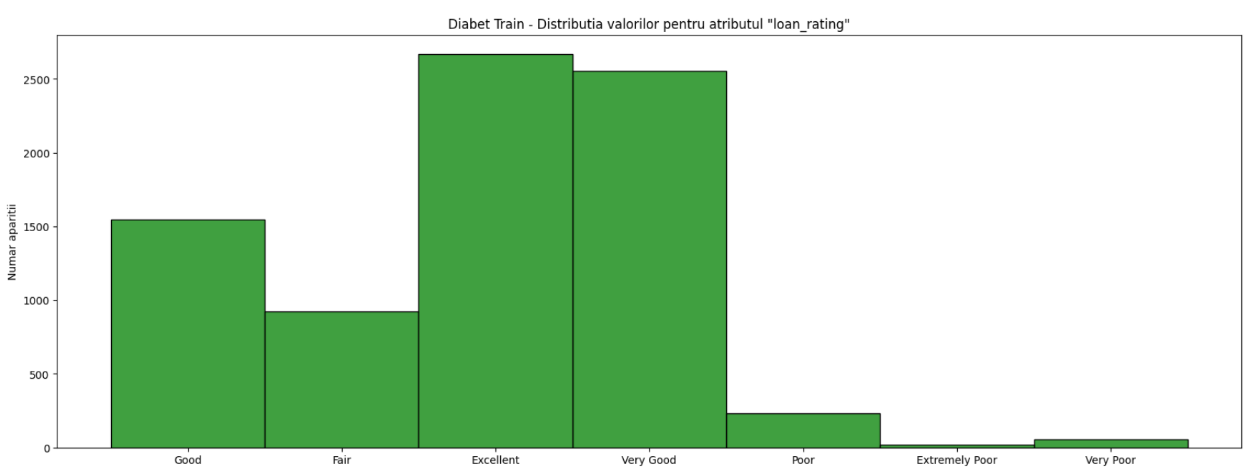

Diabet Train - Distributia valorilor pentru atributul "ImprovedAveragePulmonaryCapacity"
Numar aparitii
Valoarea lui "ImprovedAveragePulmonaryCapacity"
0
1000
2000
3000
4000
5000
0.0
0.2
0.4
0.6
0.8
1.0

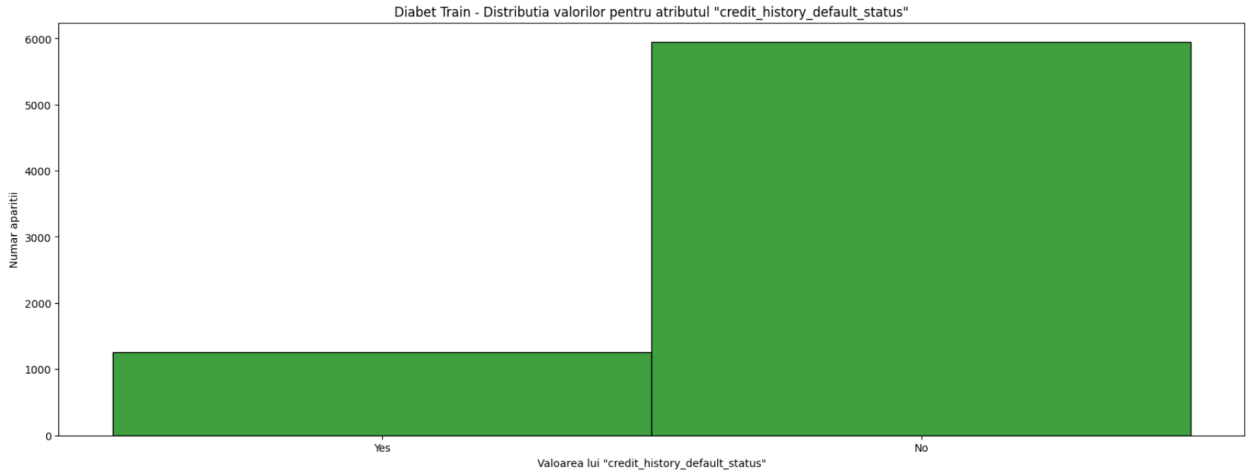

Diabet Train - Distributia valorilor pentru atributul "Smoker"
Numar aparitii
0
1000
2000
3000
4000
non-smoker
smoker
Valoarea lui "Smoker"

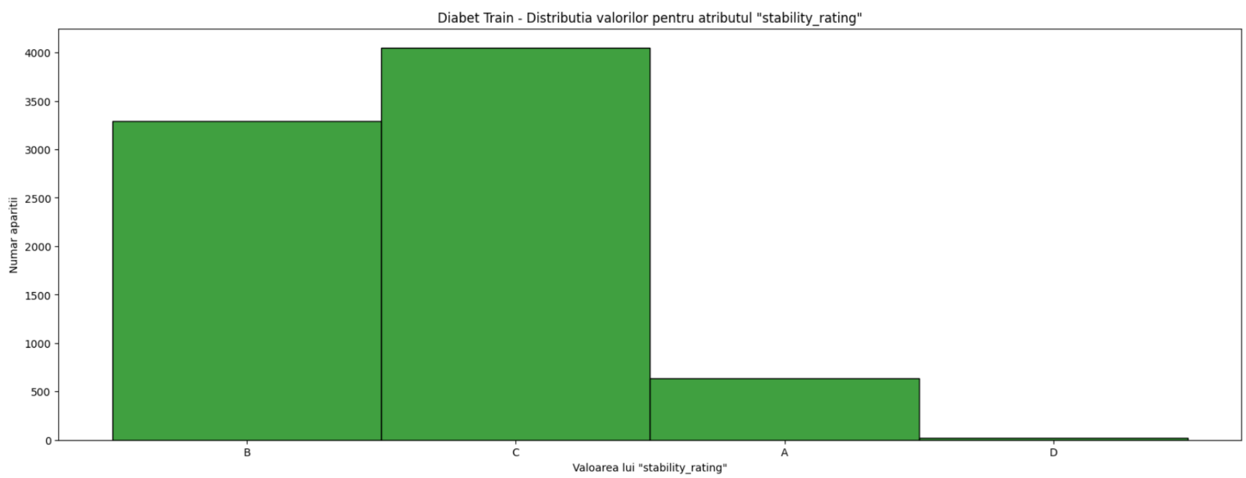

Diabet Train - Distributia valorilor pentru atributul "Diabetes"
Numar aparitii
0
1000
2000
3000
4000
5000
6000
7000
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Valoarea lui "Diabetes"

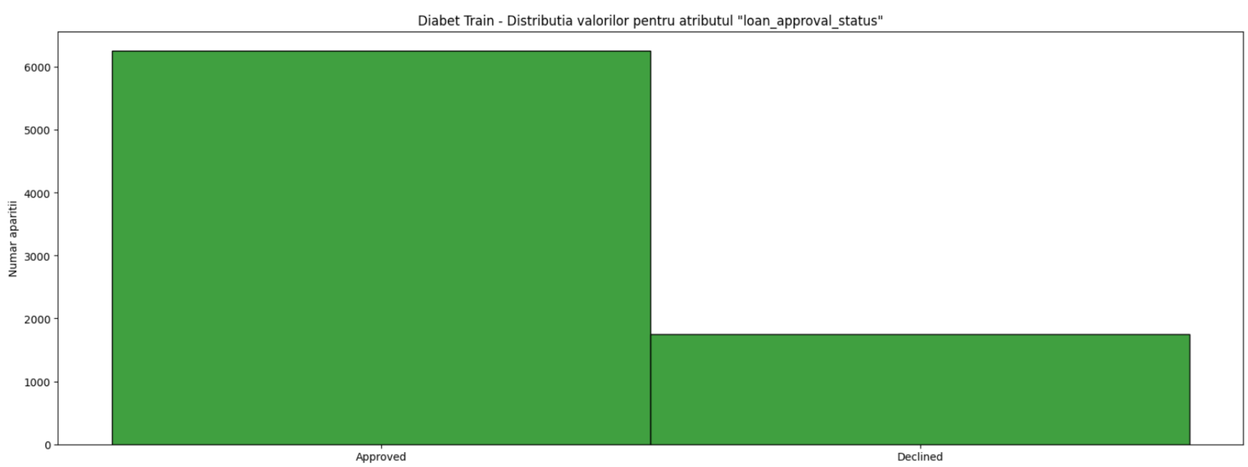

Diabet Train - Distributia valorilor pentru atributul "residential_status"
Numar aparitii
Mortgage
Renter
Owner
Unknown
Valoarea lui "residential_status"


Diabet Train - Distributia valorilor pentru atributul "loan_purpose"
Numar aparitii
Business
Personal
Home Improvement
Study
Health
Debt Consolidation
Valoarea lui "loan_purpose"


Diabet Train - Distributia valorilor pentru atributul "loan_rating"
Numar aparitii
Good
Fair
Excellent
Very Good
Poor
Extremely Poor
Very Poor
Valoarea lui "loan_rating"

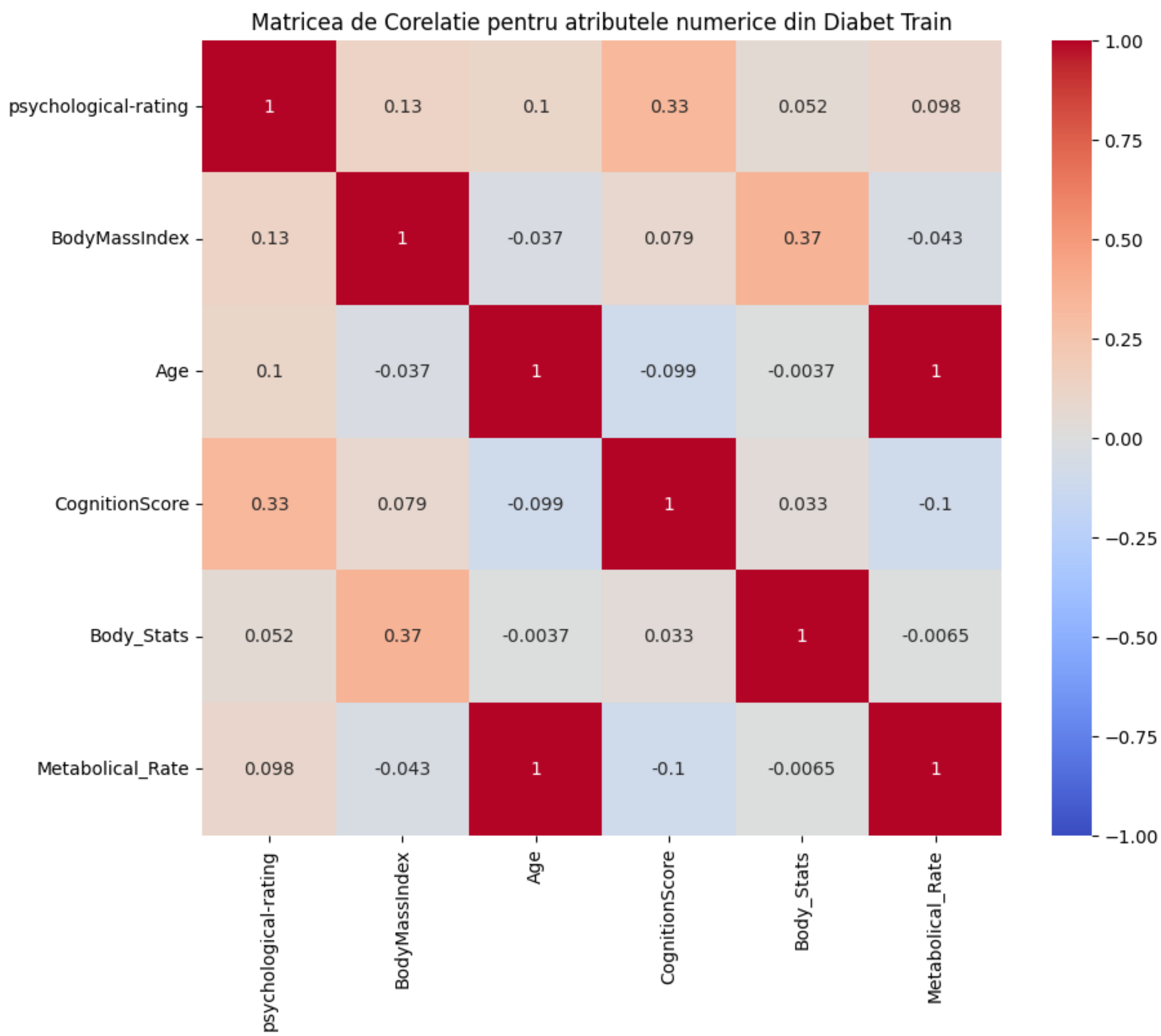


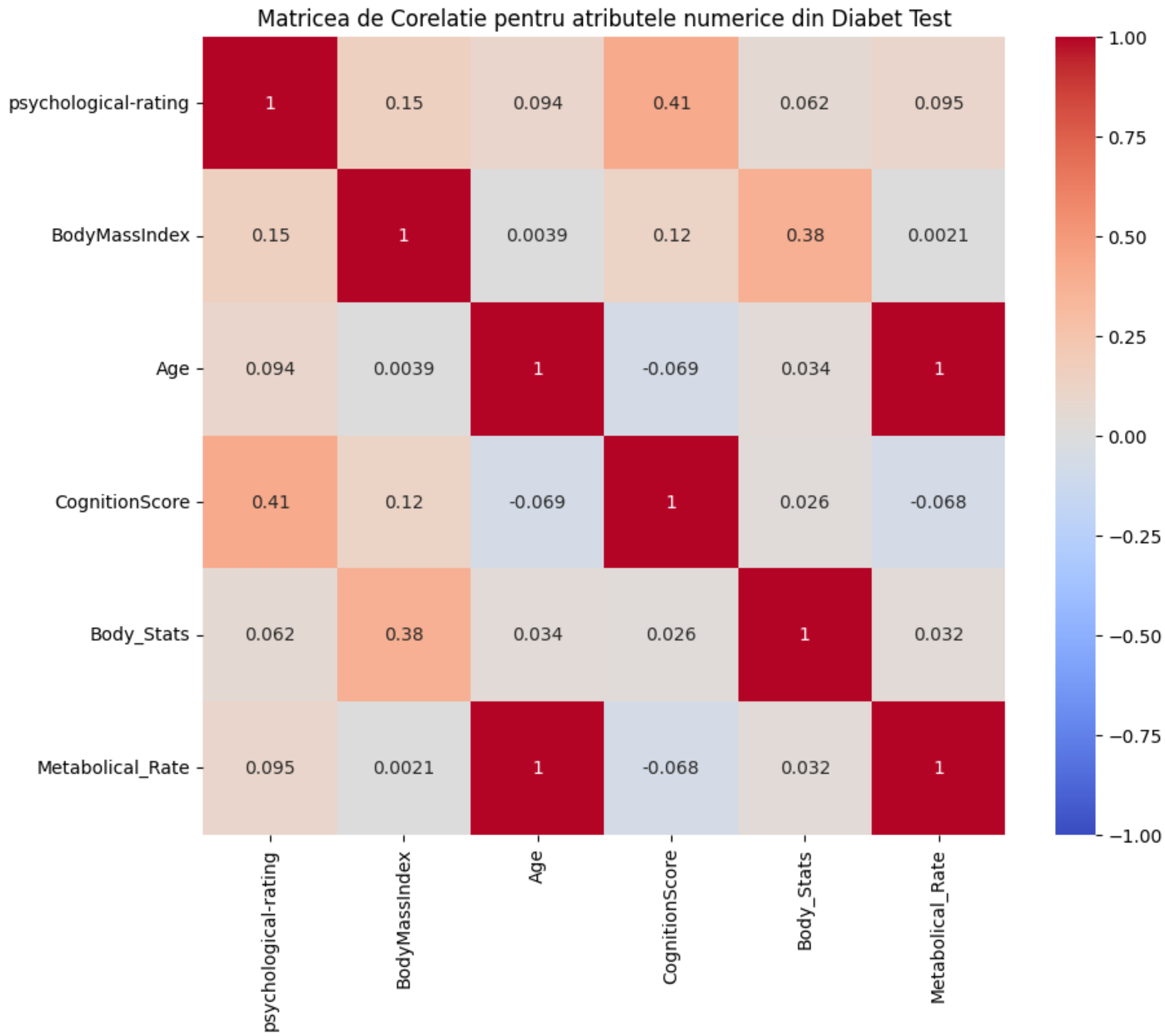




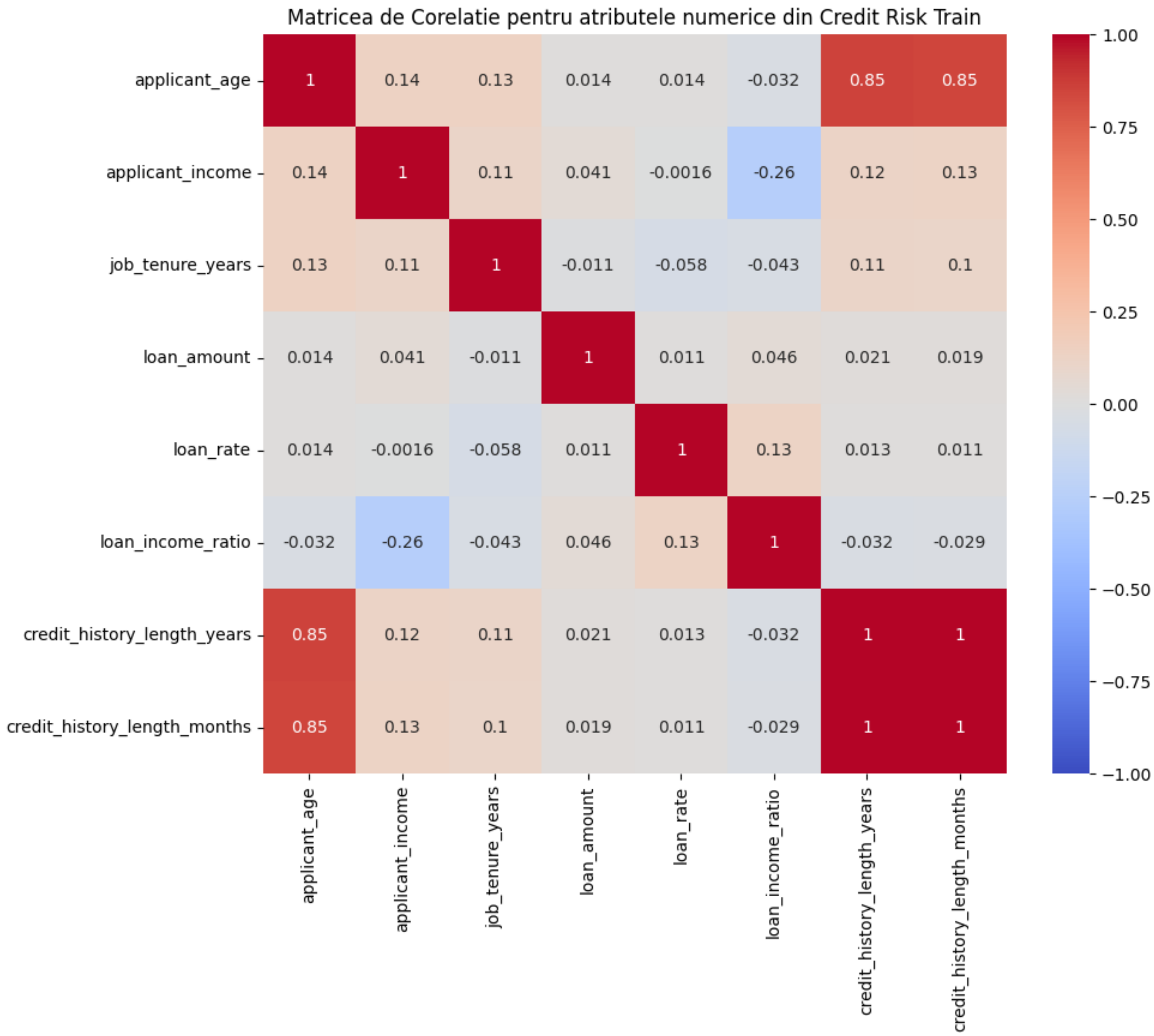
Pentru atributele numerice continue, se poate observa in unele cazuri o variatie mai mare a valorilor, iar in alte cazuri avem unele valori lipsa (valori nule), ceea ce poate produce inexactitatati dupa ce antrenam modelul (pentru ca o sa folosim date imputate).

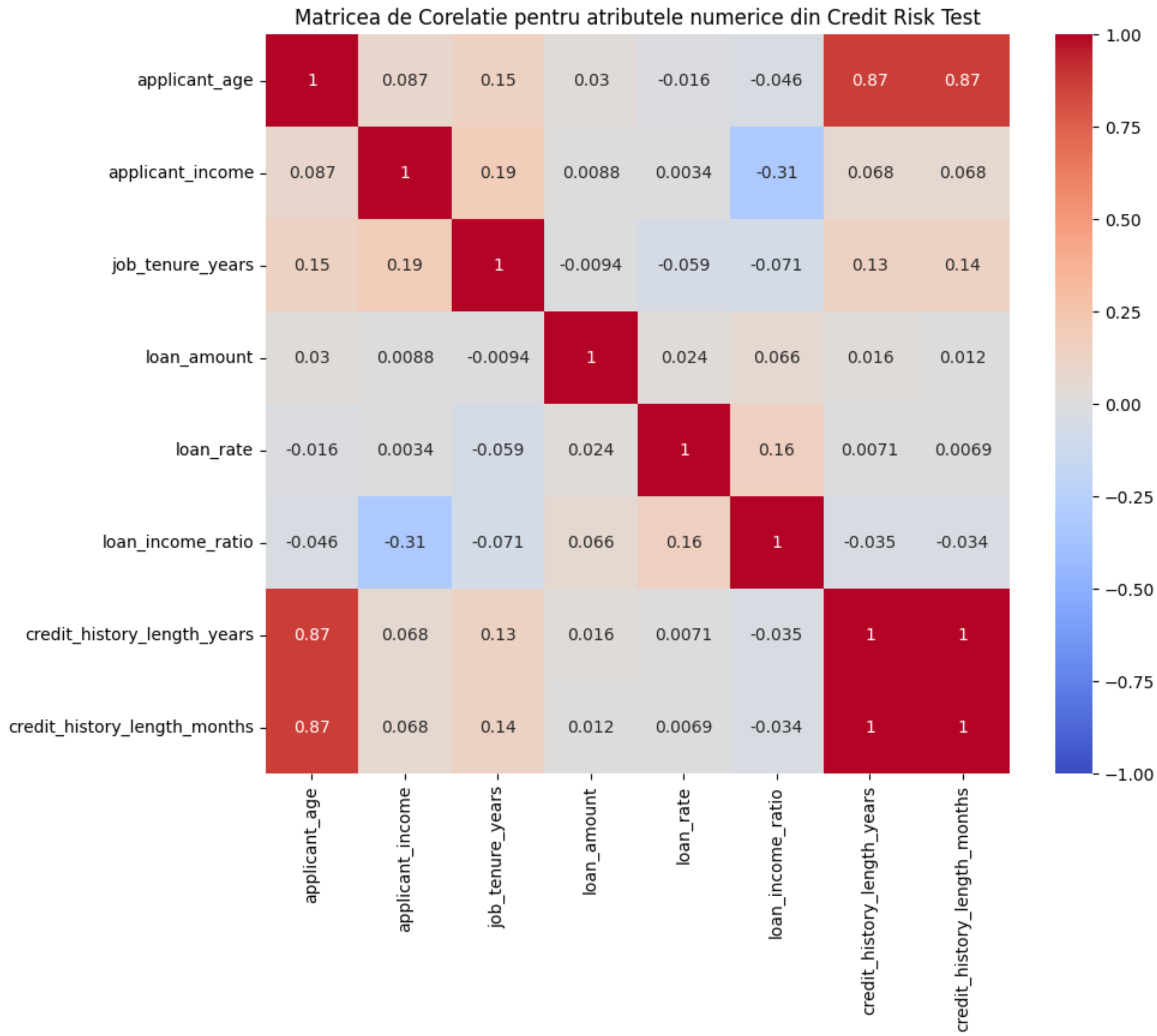
Pentru atributele categorice, atributele care au mai multe valori diferite ajuta mai mult la antrenarea modelului, intrucat se face mai usor distinctia intre intrari. De asemenea, sa zicem ca avem un atribut cu doua valori distincte, daca intrarile ar fi repartizate 50% - 50%, atunci ar fi o distinctie clara intre intrari, dar daca sunt repartizate 90% - 10%, atunci nu te ajuta pentru a distinge intrarile care se afla in cele 90%. In concluzie, pentru a obtine un model cat mai bine antrenat, avem nevoie de o repartitie cat mai uniforma a intrarilor.



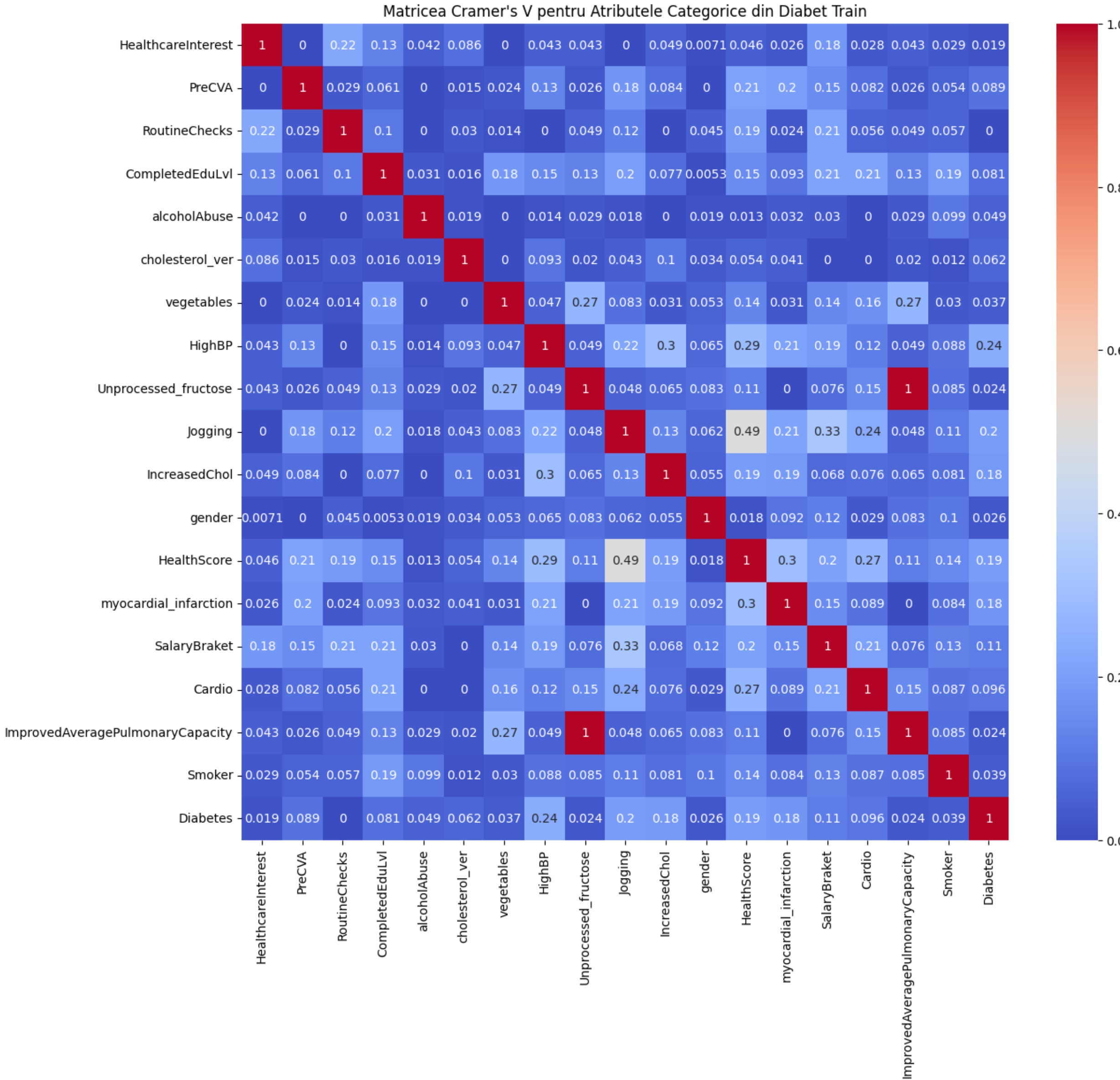


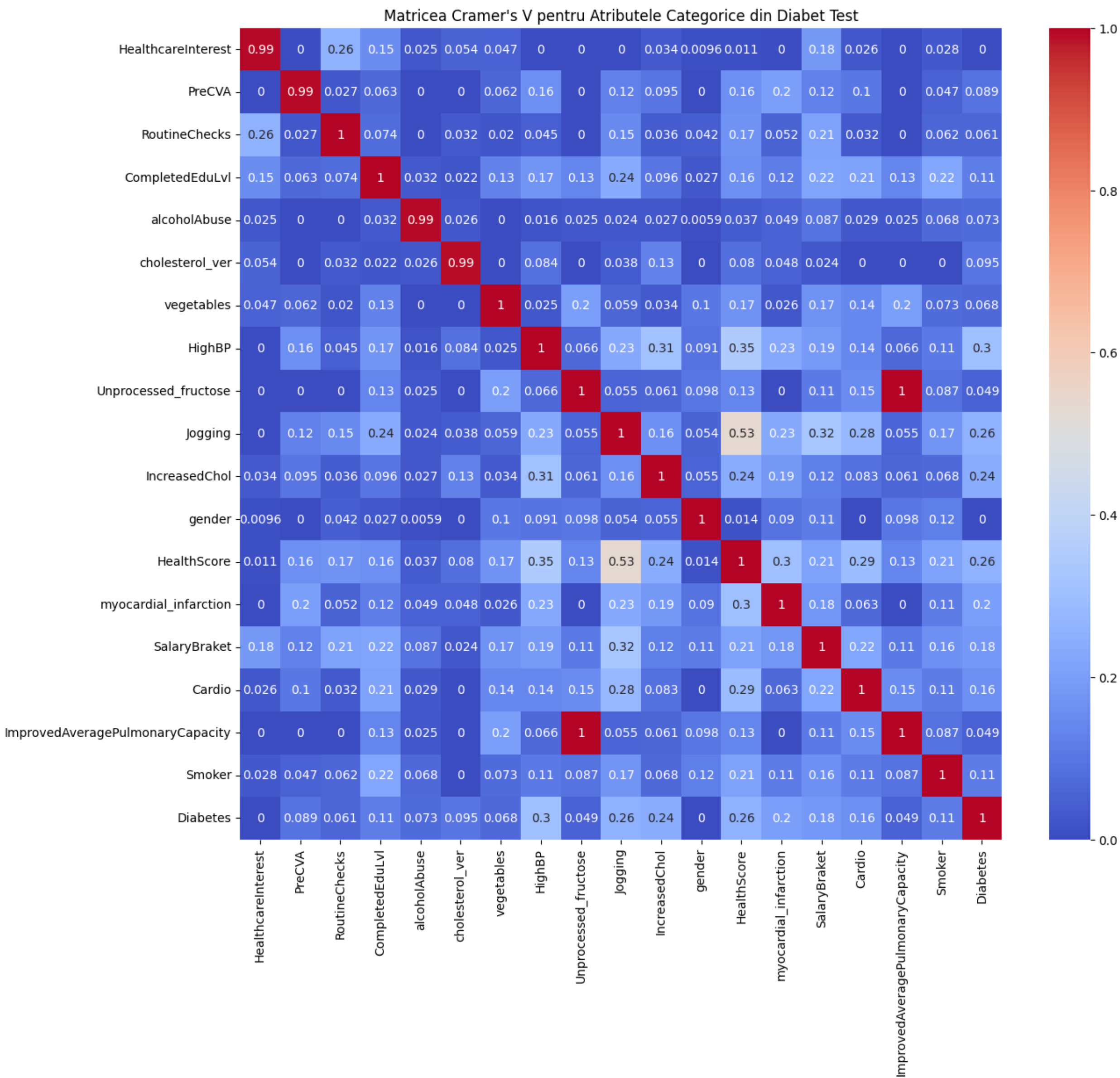
Se poate observa o corelatie mare intre atributele “Age” si “Metabolical_Rate”.


Matricea de Corelatie pentru atributele numerice din Credit Risk Train
applicant_age
applicant_income
job_tenure_years
loan_amount
loan_rate
loan_income_ratio
credit_history_length_years
credit_history_length_months
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

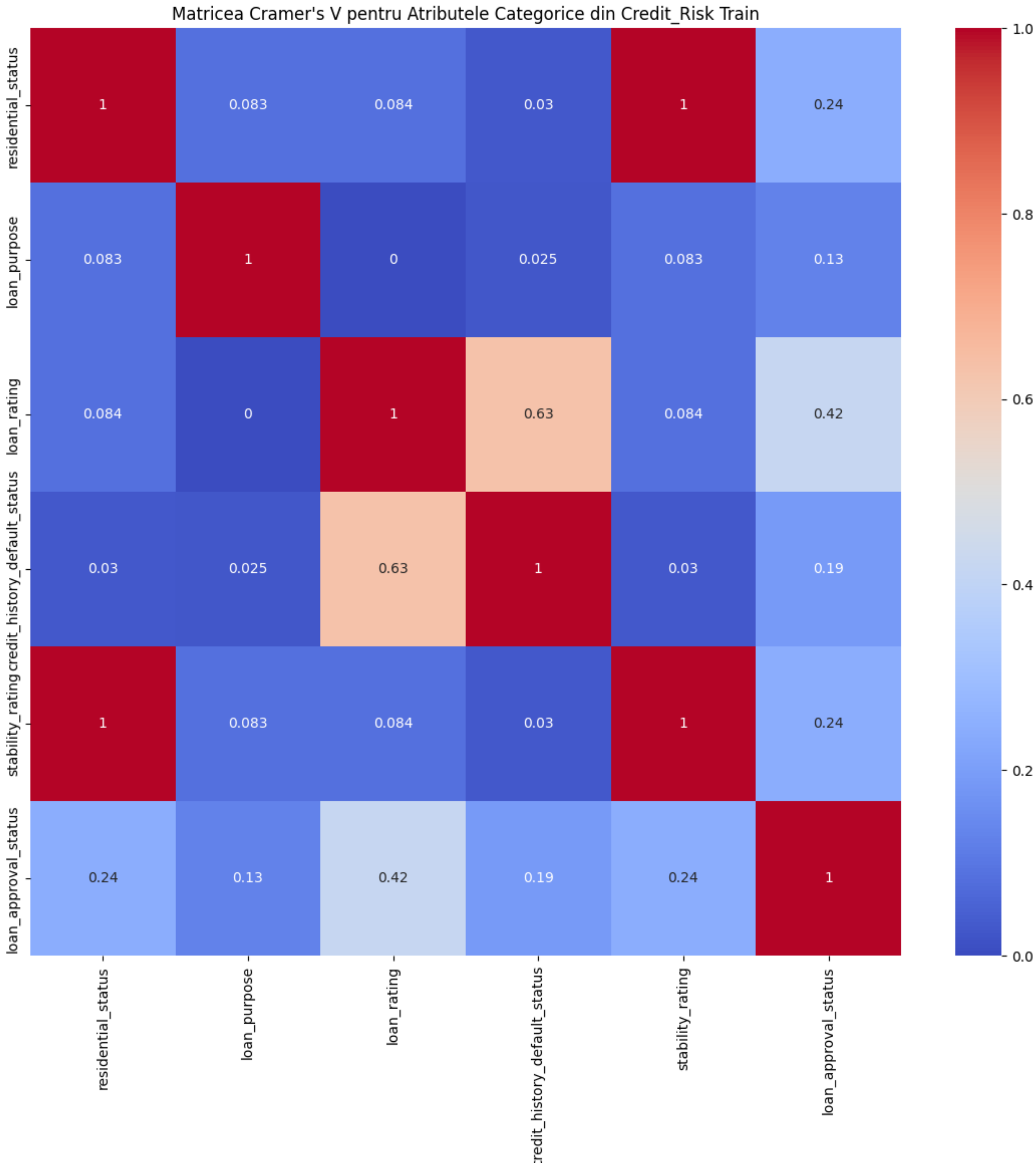


Se poate observa o corelatie mare intre atributele “credit_history_length_years” si “credit_history_length_months” si o corelatie desul de mare, dar nu as spune chiar reduntanta intre “applicant_age” si “credit_history_length_years” si “credit_history_length_months”.

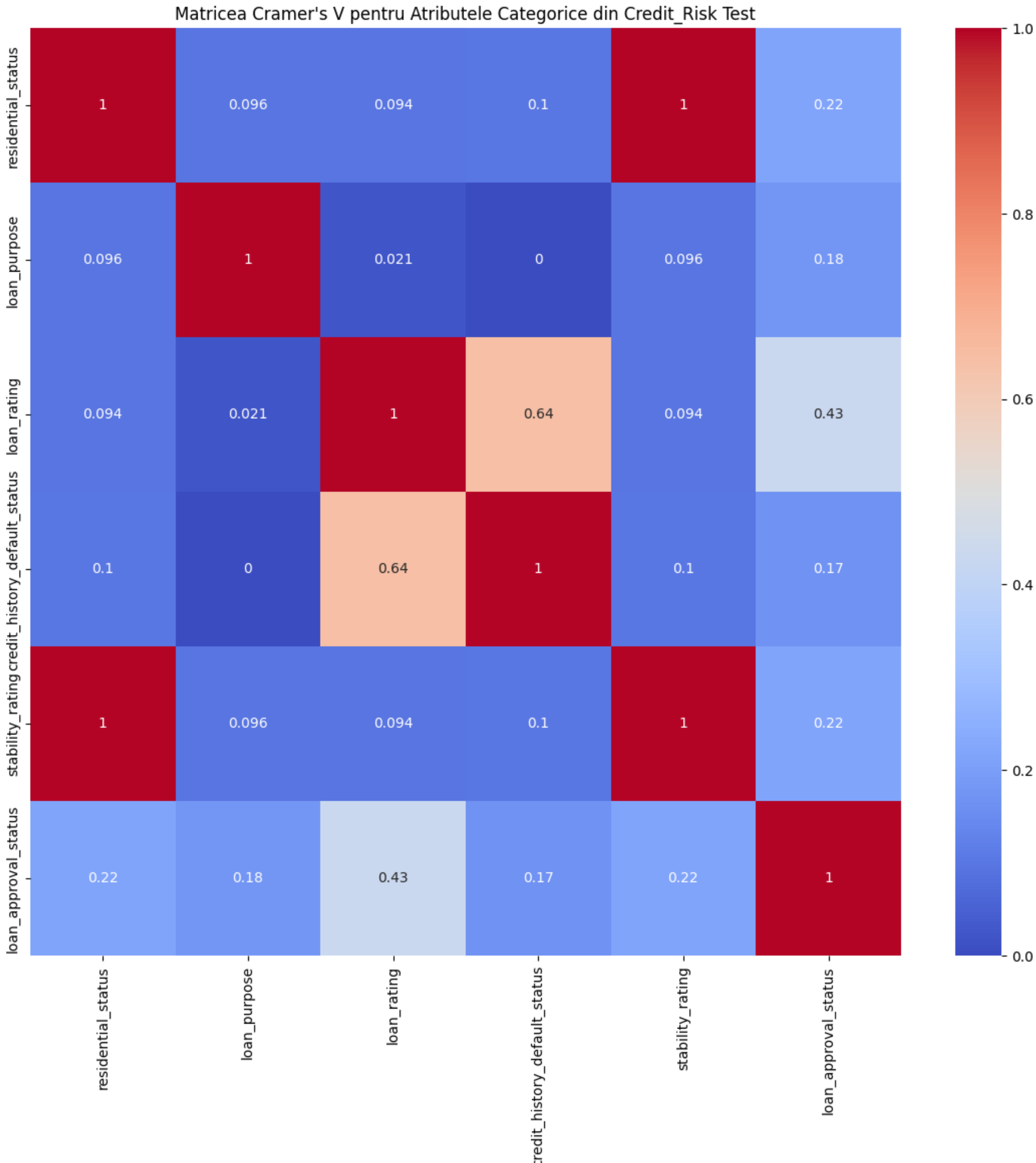

Matricea Cramer's V pentru Atributele Categorice din Diabet Train
HealthcareInterest
PreCVA
RoutineChecks
CompletedEduLvl
alcoholAbuse
cholesterol_ver
vegetables
HighBP
Unprocessed_fructose
Jogging
IncreasedChol
gender
HealthScore
myocardial_infarction
SalaryBraket
Cardio
ImprovedAveragePulmonaryCapacity
Smoker
Diabetes
1.0
0.8
0.6
0.4
0.2
0.0



Se poate observa o corelatie mare intre “Unprocessed_fructose” si “ImprovedAveragePulmonaryCapacity”



Matricea Cramer's V pentru Atributele Categorice din Credit_Risk Train
residential_status
loan_purpose
loan_rating
credit_history_default_status
stability_rating
loan_approval_status
1
0.083
0.084
0.03
1
0.24
0.083
1
0
0.025
0.083
0.13
0.084
0
1
0.63
0.084
0.42
0.03
0.025
0.63
1
0.03
0.19
1
0.083
0.084
0.03
1
0.24
0.24
0.13
0.42
0.19
0.24
1
1.0
0.8
0.6
0.4
0.2
0.0



Se poate observa o corelatie mare intre “residential_status” si “stability_rating”.

Random Forest:

```
Evaluarea modelului pentru Diabetes:

              precision    recall  f1-score   support

           0       0.74      0.99      0.84      1446
           1       0.00      0.00      0.00        54
           2       0.68      0.08      0.14       500

    accuracy                           0.73      2000
   macro avg       0.47      0.36      0.33      2000
weighted avg       0.70      0.73      0.64      2000

Acuratetea: 0.734

Evaluarea modelului pentru loan_approval_status:

              precision    recall  f1-score   support

    Approved       0.92      0.99      0.95      1564
    Declined       0.95      0.68      0.79       436

    accuracy                           0.92      2000
   macro avg       0.93      0.83      0.87      2000
weighted avg       0.92      0.92      0.92      2000

Acuratetea: 0.9215
```

MLP:

```
Evaluarea modelului pentru Diabetes:

              precision    recall  f1-score   support

           0       0.77      0.89      0.83      1446
           1       0.17      0.06      0.08        54
           2       0.50      0.33      0.39       500

    accuracy                           0.72      2000
   macro avg       0.48      0.42      0.43      2000
weighted avg       0.69      0.72      0.70      2000

Acuratetea: 0.7235

Evaluarea modelului pentru loan_approval_status:

              precision    recall  f1-score   support

           0       0.91      0.97      0.94      1564
           1       0.85      0.66      0.74       436

    accuracy                           0.90      2000
   macro avg       0.88      0.82      0.84      2000
weighted avg       0.90      0.90      0.90      2000

Acuratetea: 0.901
```

Dupa cum se poate observa, al doilea set de date, Credit Risk, este cel care obtine o acuratete mai buna. Primul set de date este problematic, sunt multe date eronate care, la antrenare, construieste o logica si niste corelatii nu neaparat adevarate, ceea ce duce la predictii eronate.

Rezultatele sunt asemanatoare intre (Diabetes – Random Forest si Diabetes – MLP) si (Credit Risk – Random Forest si Credit Risk – MLP).

Precision este important in situatiile in care costul unui fals pozitiv este mare si, dupa cum se poate observa, este mai mare la modelul Credit Risk.

Recall este important in situatiile in care costul unui fals negativ este mare si, dupa cum se poate observa, este mai mare la modelul Credit Risk.

F1 score reprezinta media armonica dintre precision si recall. Este util atunci cand trebuie sa gasim un echilibru intre precision si recall.