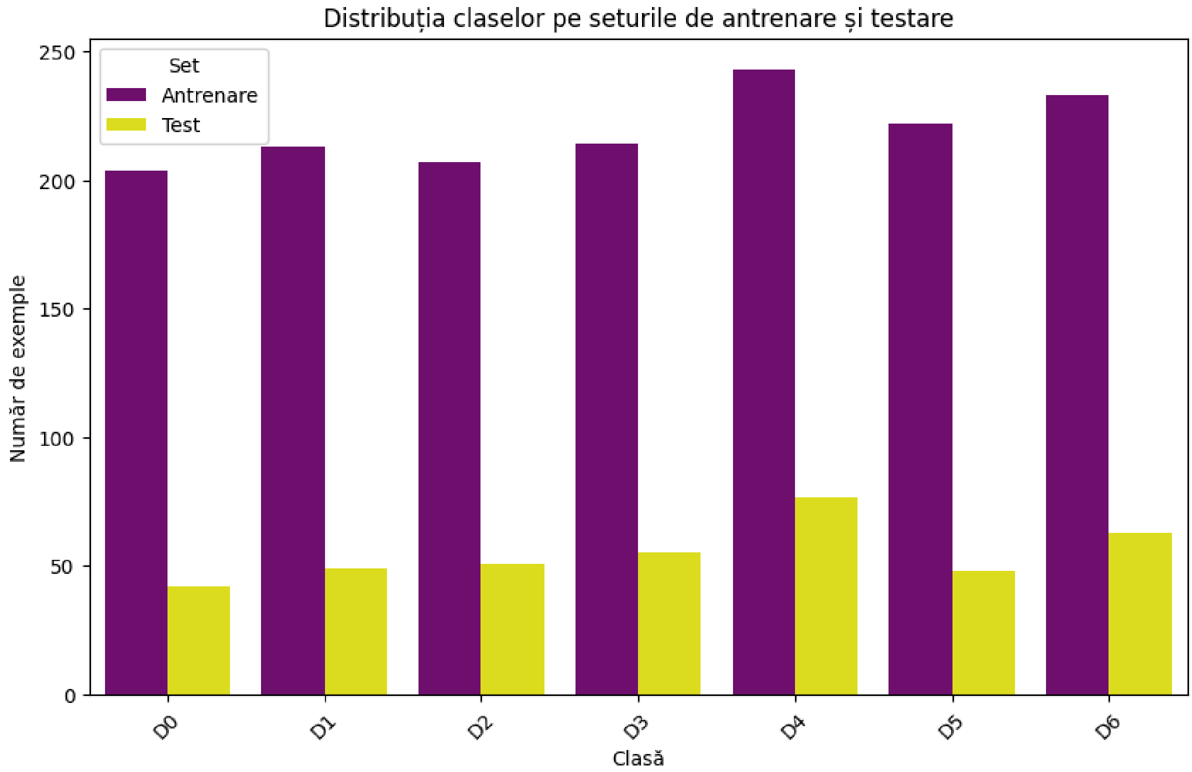
Cotelin Maria-Diana
342C4

Învățare Automată
Tema 1 - 2024

1. Explorarea Datelor

a. Analiza echilibrului de clase

```
Dimensiunea x_train: (1536, 18)
Dimensiunea y_train: (1536,)
Dimensiunea x_test: (385, 18)
Dimensiunea y_test: (385,)
```



b. Vizualizarea datelor

i. Atribute numerice:

<u>Valori Statistice</u>

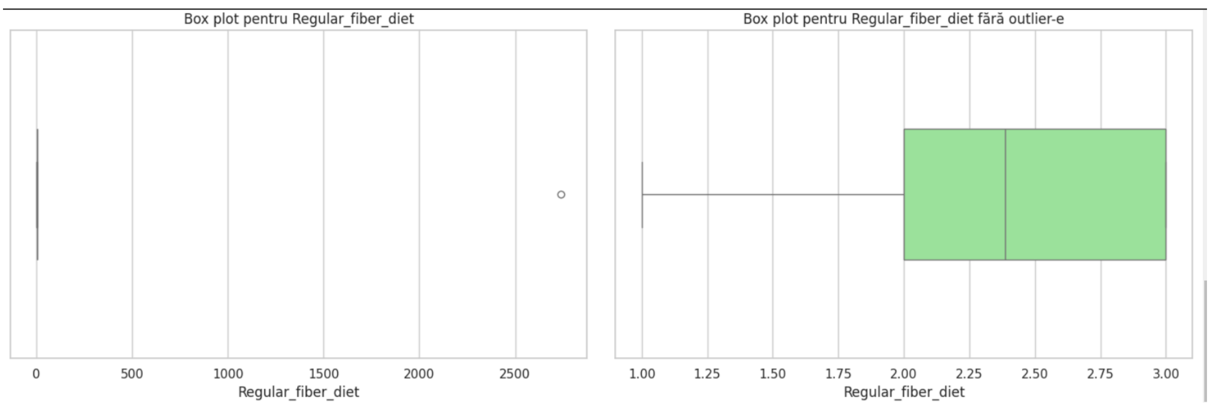
Pentru fiecare atribut numeric au fost calculate urmatoarele:

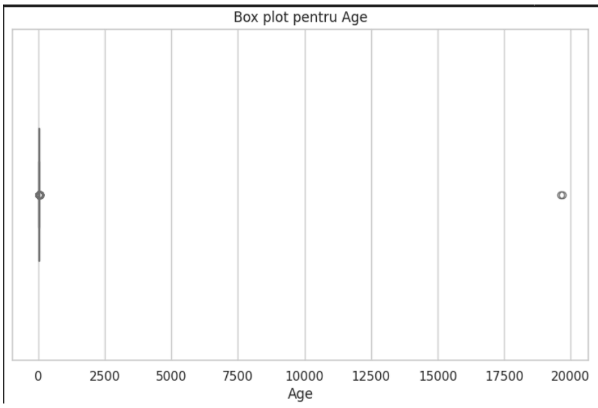
- Medie
- Abaterea standard
- Abaterea medie absolută
- Valoare minimă
- Valoare maximă
- Diferența de valori maxime și minime
- Mediană

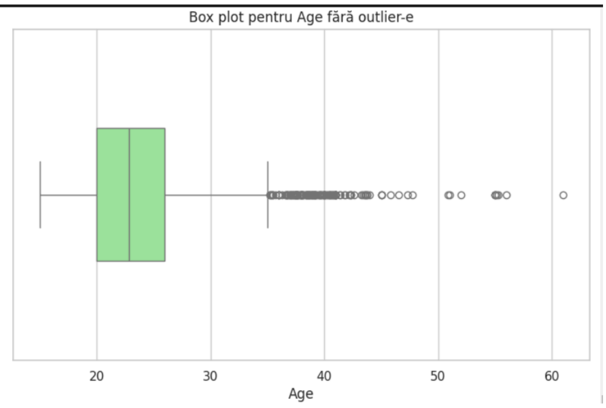
- Abaterea mediană absolută
- Intervalul intercuartil

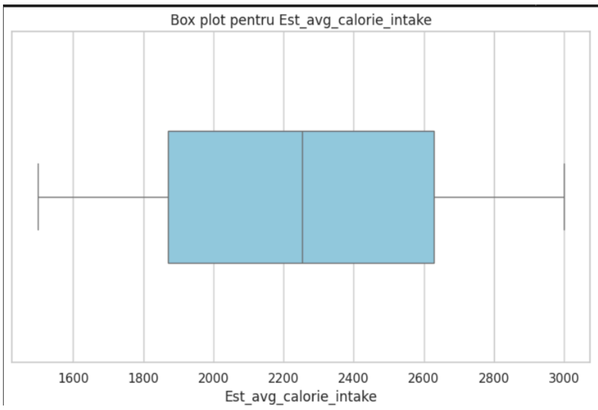
index	Medie	Abatere Medie abs	Abatere Standard	Min	Max	Max - Min	Mediana	Abatere mediana abs	Interval intercuartil
Regular_fiber_diet	3.845	2.848	62.440	1.000	2738.000	2738.000	2.387	0.387	1.000
Sedentary_hours_daily	3.694	1.134	21.759	2.210	965.580	954.370	3.130	0.439	0.870
Age	77.792	40.947	633.312	15.000	19685.000	19670.000	22.830	3.170	6.028
Est_avg_calorie_intake	2253.688	375.362	434.076	1500.000	3000.000	1500.000	2253.000	380.000	757.000
Main_meals_daily	2.683	0.596	0.779	1.000	4.000	3.000	3.000	0.000	0.341
Height	3.573	3.738	58.098	1.450	1915.000	1913.550	1.700	0.070	0.140
Water_daily	2.010	0.470	0.611	1.000	3.000	2.000	2.000	0.445	0.874
Weight	205.637	254.648	3225.654	-1.000	82628.000	82629.000	80.386	24.386	46.205
Physical_activity_level	1.010	0.702	0.855	0.000	3.000	3.000	1.000	0.815	1.568
Technology_time_use	1.345	1.511	29.789	0.000	1306.000	1306.000	1.000	1.000	1.000

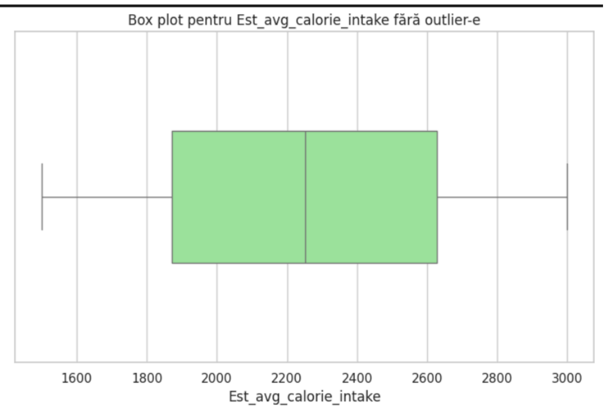
Observand ca in unele cazuri diferenta dintre medie si mediana este semnificativa, am ales sa elimin valorile outleier-e, realizand apoi grafice de tip Box Plot pentru a arata diferenta distributiei datelor inainte si dupa eliminare.

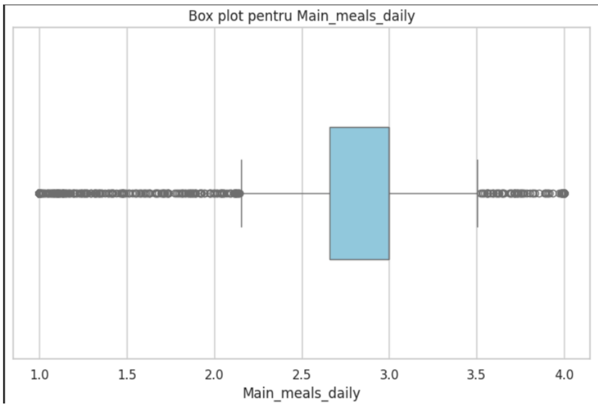


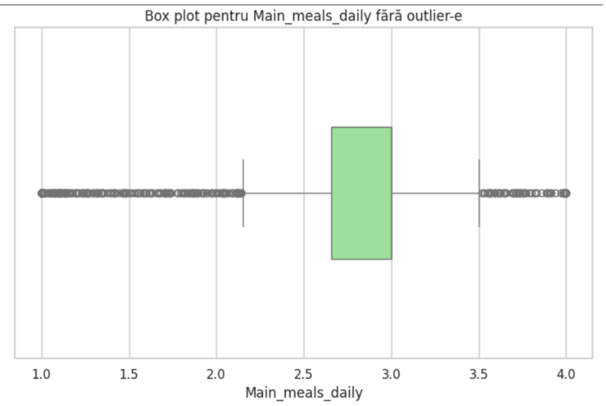

Box plot pentru Age
0
2500
5000
7500
10000
12500
15000
17500
20000
Age

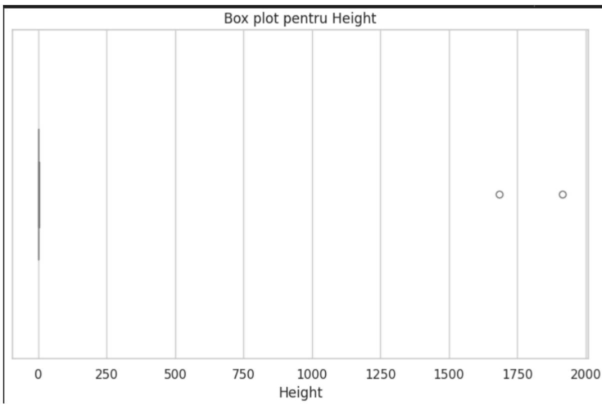

Box plot pentru Age fără outlier-e
20
30
40
50
60
Age

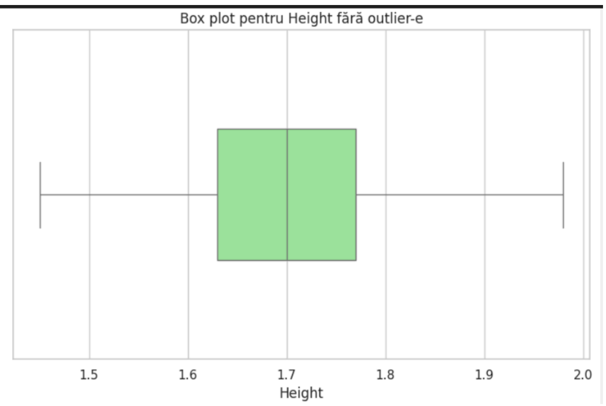

Box plot pentru Est_avg_calorie_intake
1600
1800
2000
2200
2400
2600
2800
3000
Est_avg_calorie_intake

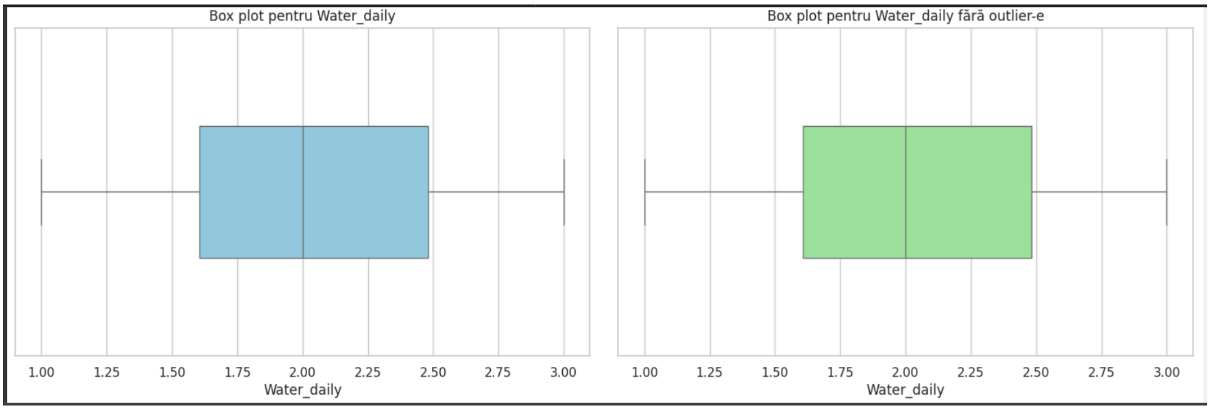

Box plot pentru Est_avg_calorie_intake fără outlier-e
1600
1800
2000
2200
2400
2600
2800
3000
Est_avg_calorie_intake

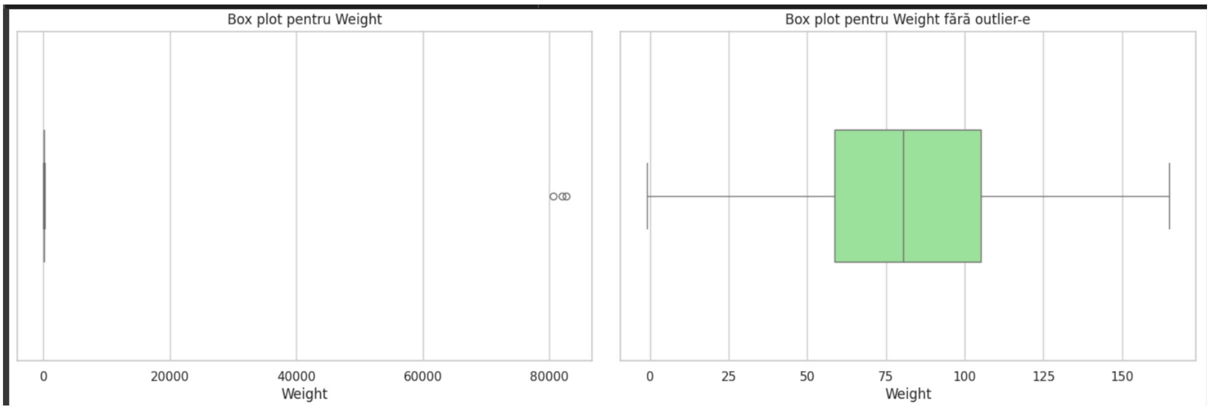

Box plot pentru Main_meals_daily
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Main_meals_daily

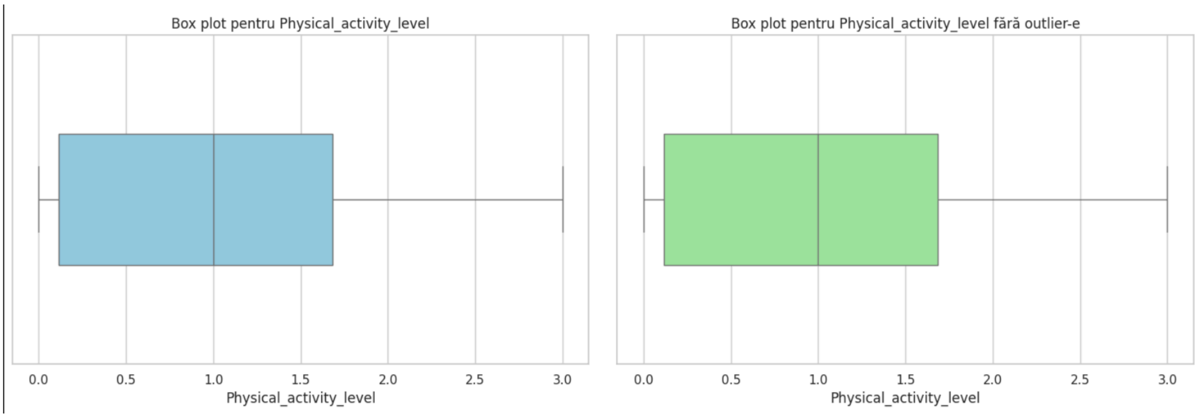

Box plot pentru Main_meals_daily fără outlier-e
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Main_meals_daily

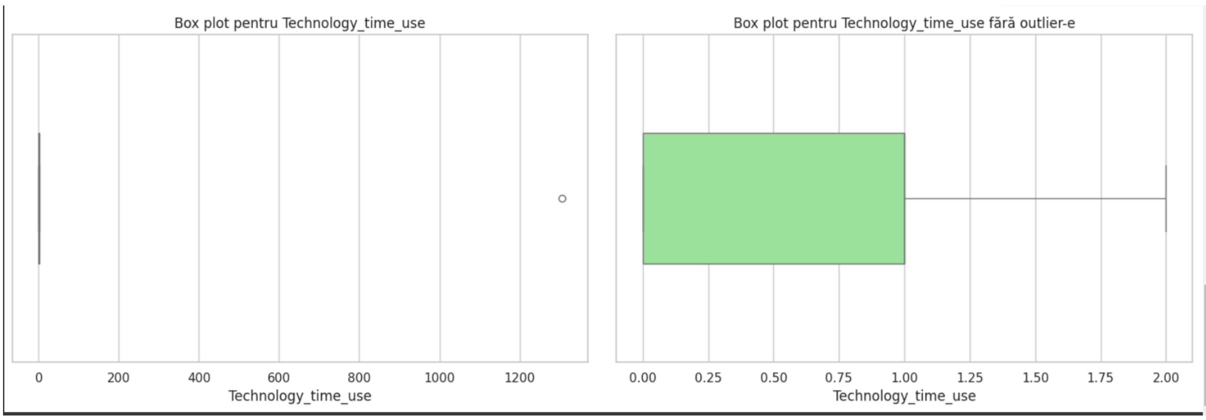

Box plot pentru Height
0
250
500
750
1000
1250
1500
1750
2000
Height


Box plot pentru Height fără outlier-e
1.5
1.6
1.7
1.8
1.9
2.0
Height







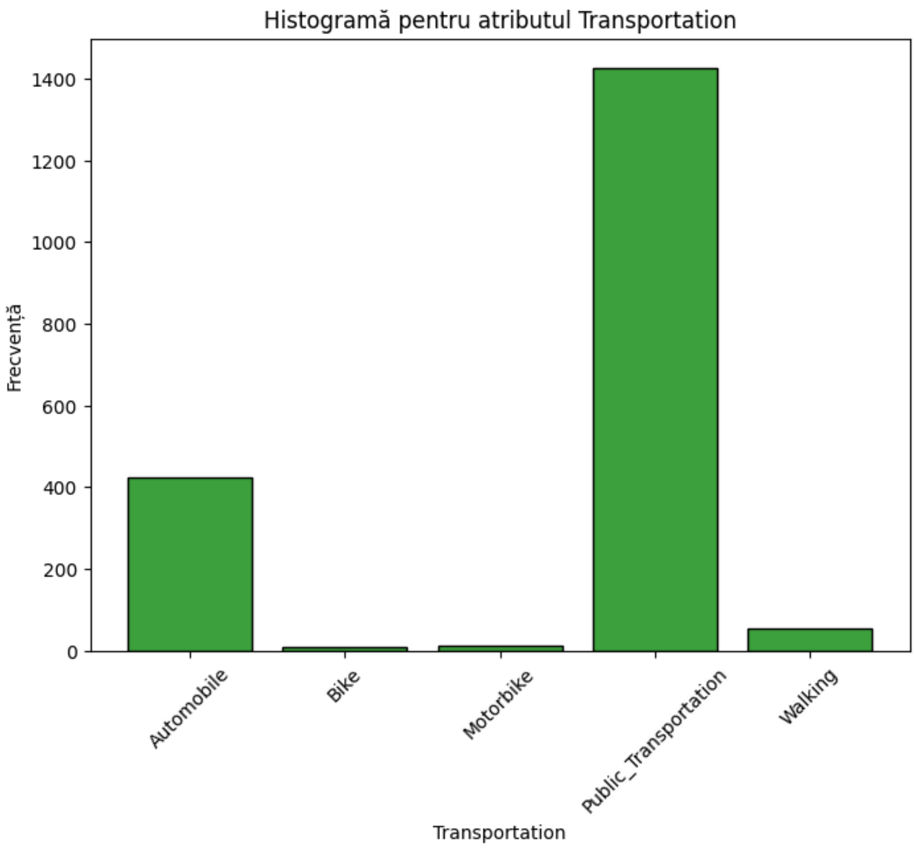


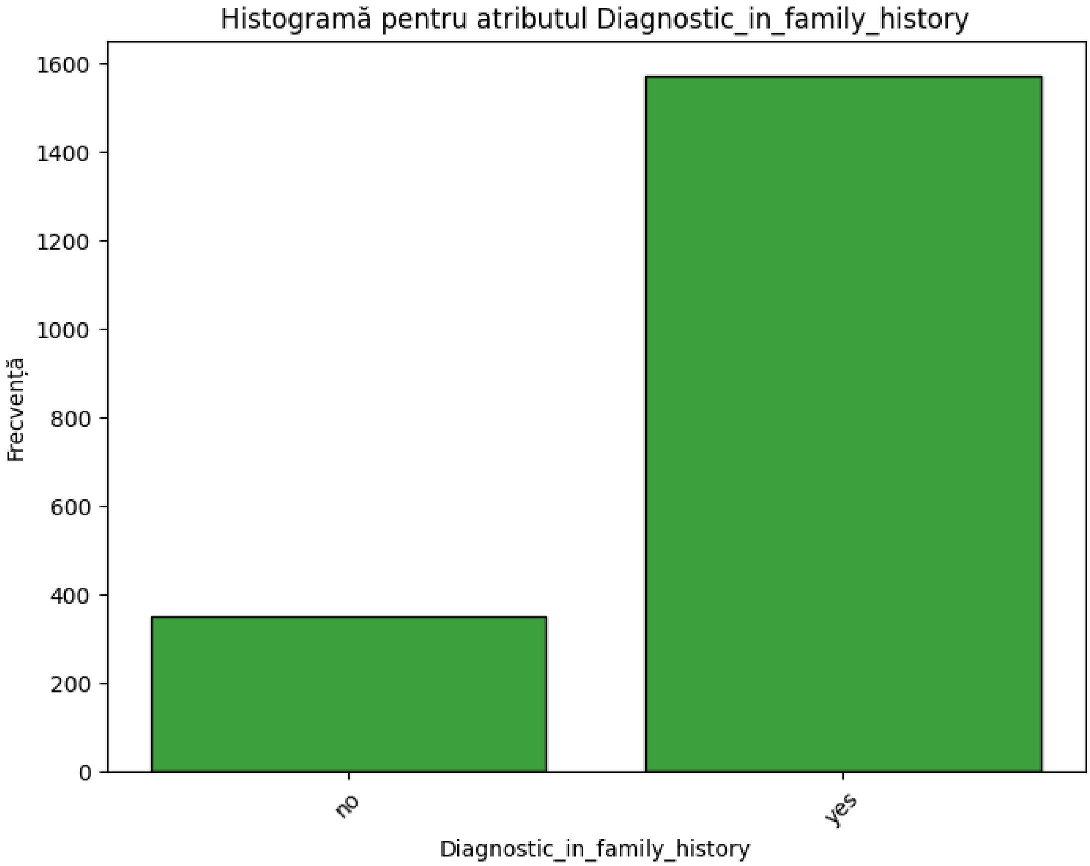
Se observa imbunatatiri semnificative pentru atributele: Regular_fiber_diet, Age, Height, Weight, Technology_time_use.

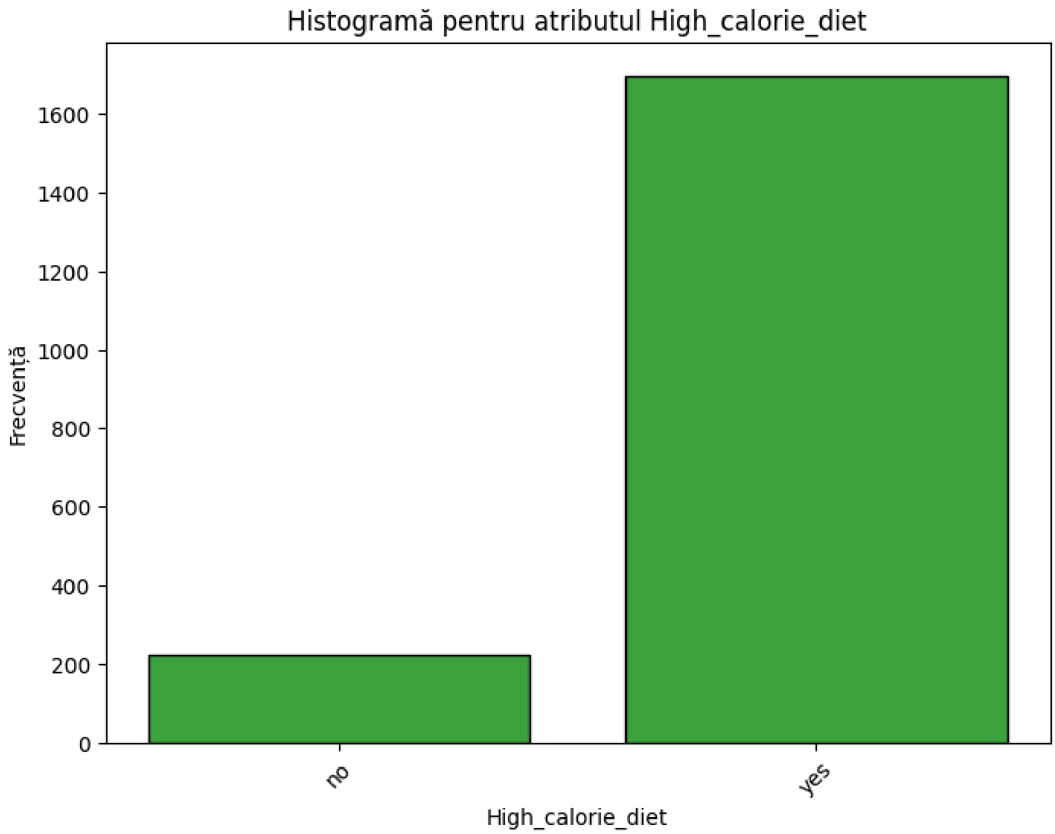
ii. Atribute Categorice

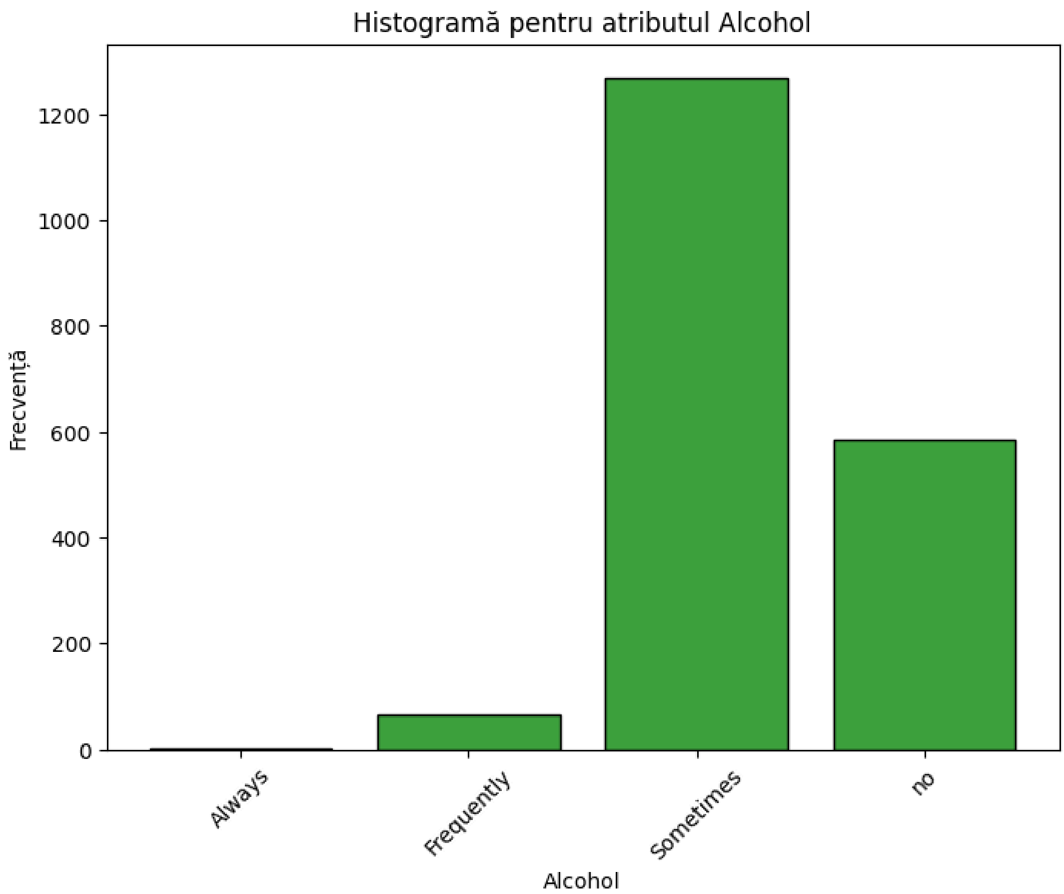
Valori unice pentru atributele categorice:

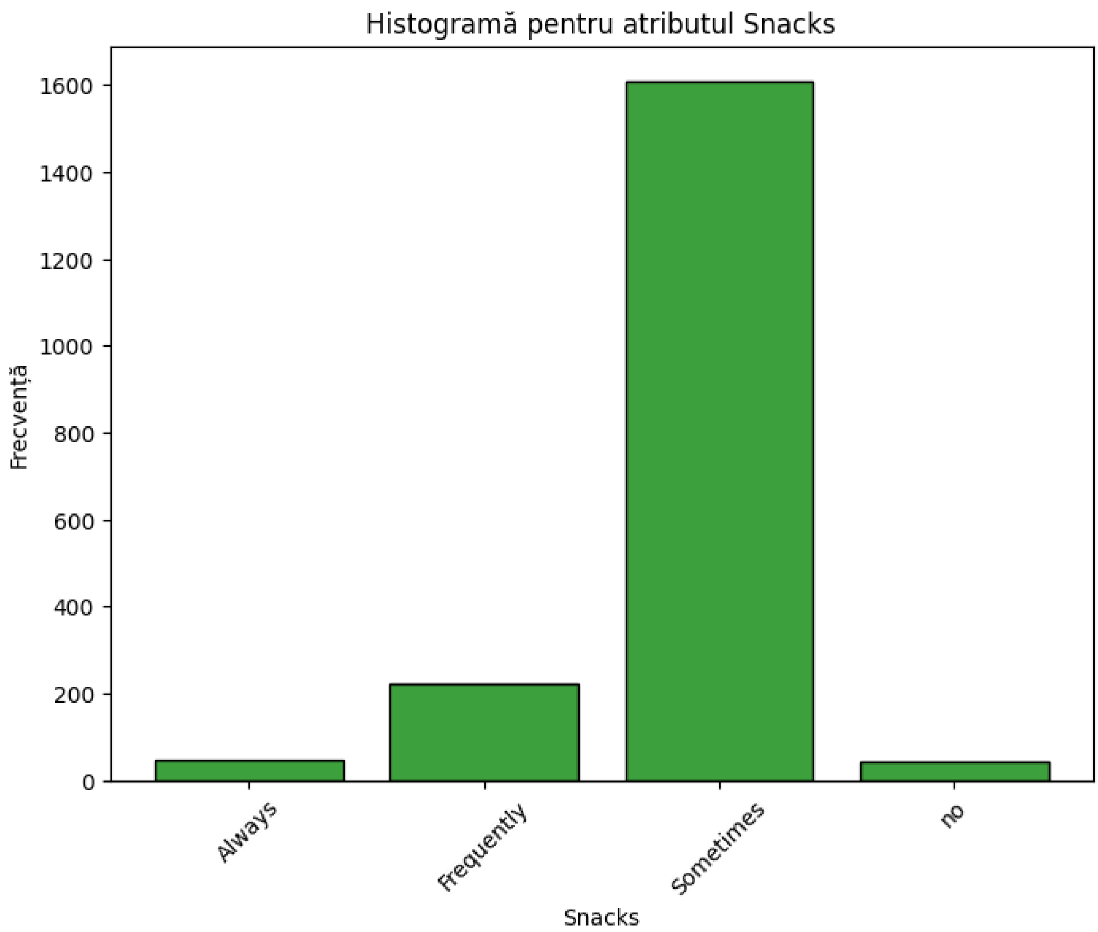
Transportation	5
Diagnostic_in_family_history	2
High_calorie_diet	2
Alcohol	4
Snacks	4
Smoker	2
Calorie_monitoring	2
Gender	2

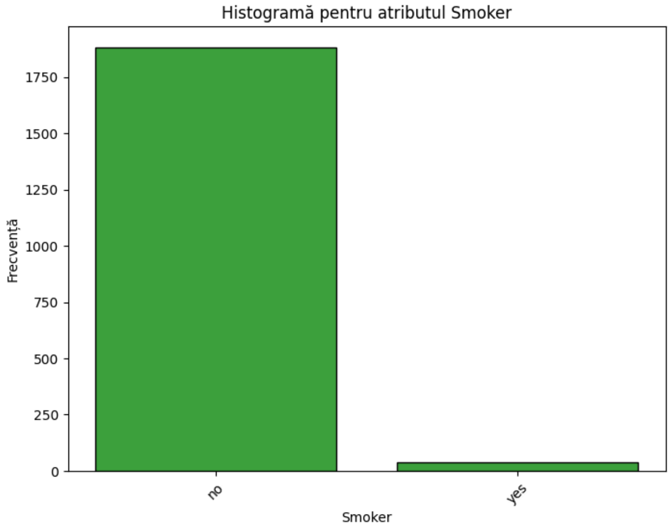


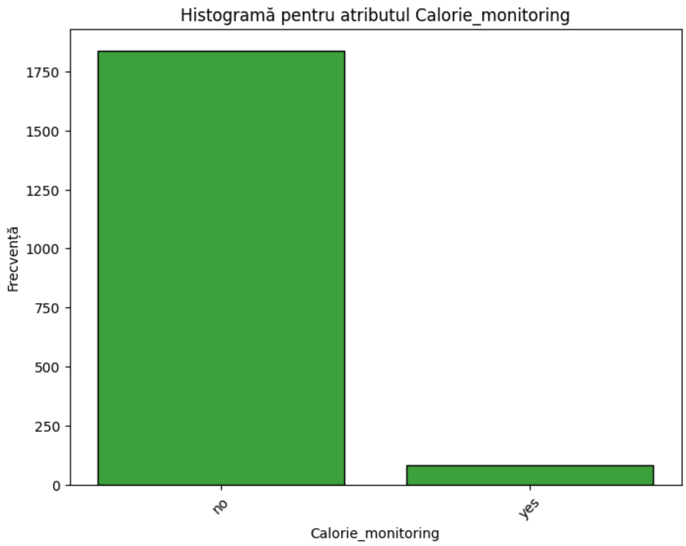

Histogramă pentru atributul Diagnostic_in_family_history
1600
1400
1200
1000
800
600
400
200
0
Frecvență
no
yes
Diagnostic_in_family_history

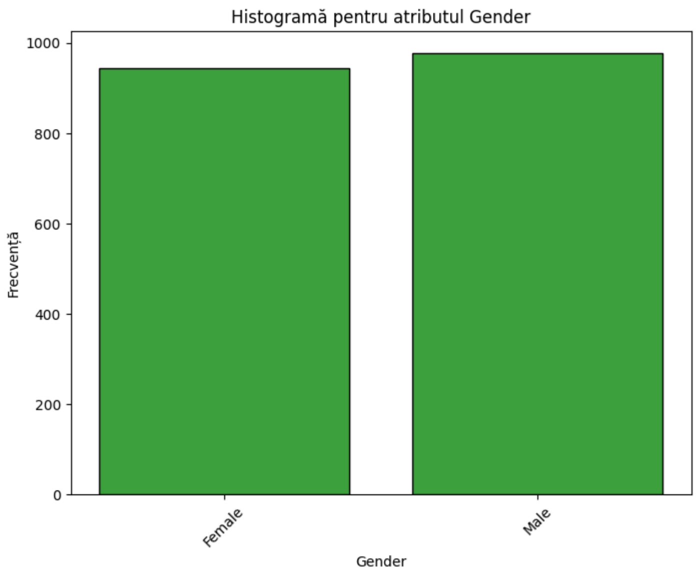

Histogramă pentru atributul High_calorie_diet
1600
1400
1200
1000
800
600
400
200
0
Frecvență
no
yes
High_calorie_diet


Histogramă pentru atributul Alcohol
Frecvență
0
200
400
600
800
1000
1200
Always
Frequently
Sometimes
no
Alcohol


Histogramă pentru atributul Snacks
Frecvență
0
200
400
600
800
1000
1200
1400
1600
Always
Frequently
Sometimes
no
Snacks


Histogramă pentru atributul Smoker
1750
1500
1250
1000
750
500
250
0
Frecvență
no
yes
Smoker


Histogramă pentru atributul Calorie_monitoring
1750
1500
1250
1000
750
500
250
0
Frecvență
no
yes
Calorie_monitoring


Histogramă pentru atributul Gender
1000
800
600
400
200
0
Frecvență
Female
Male
Gender

c. Analize de Covarianta:

<u>Matrice de covarianta</u>

- Ofera informatii despre modul in care doua variabile diferite variaza imreuna.
- Valoare pozitiva -> covariatie pozitiva
- Valoare negativa -> covariatie negativa
- Valori apropiate de 0 -> independente
- Ca interpretate putem lua de exemplu Regular_fiber_diet si Est_avg_calorie_intake, cu cat una creste mai mult cu atat cealalta scade, avand corelatie negativa

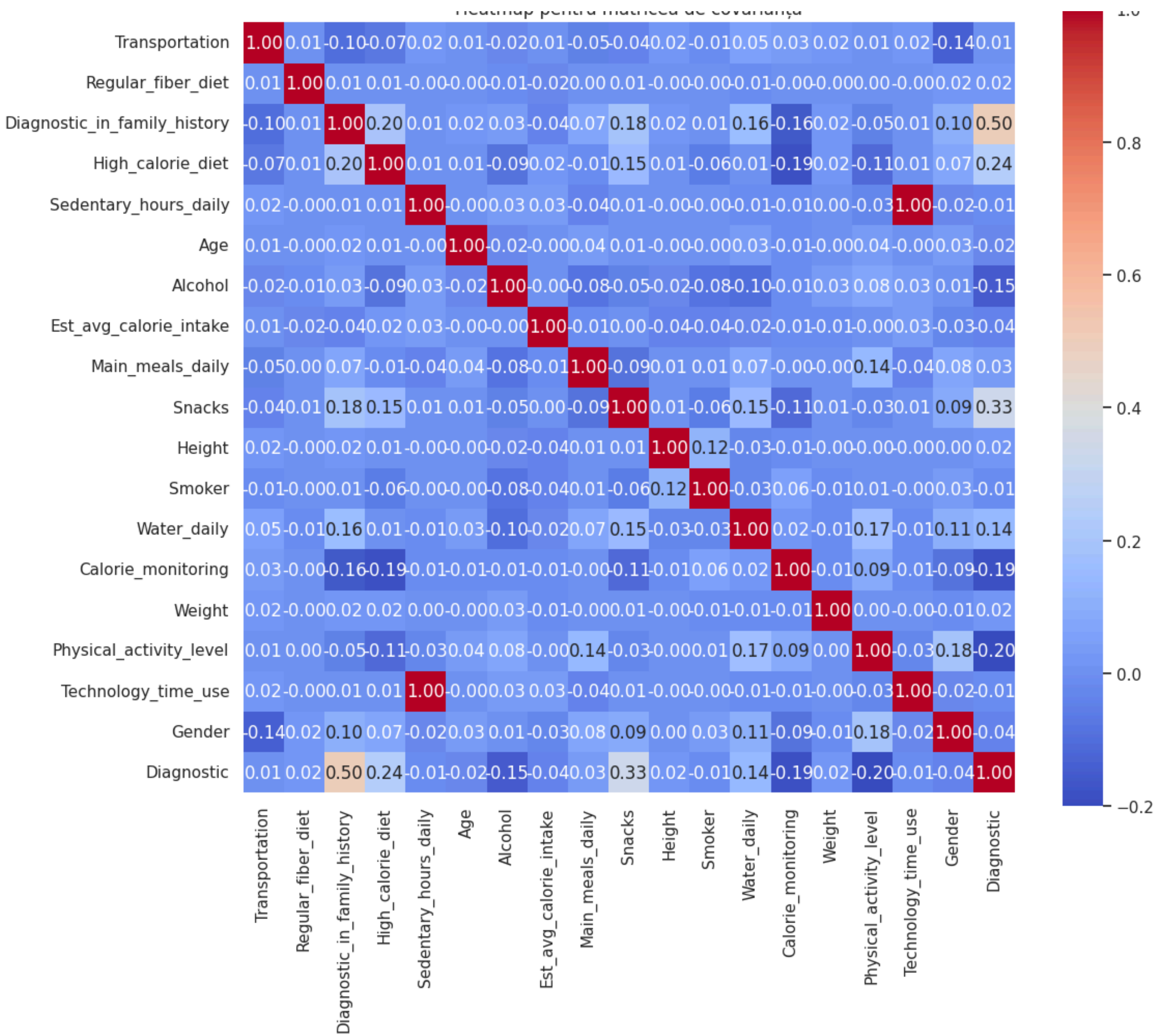
index	Transportation	Regular_fiber_diet	Diagnostic_in_family_history	High_calorie_diet	Sedentary_hours_daily	Age	Alcohol	Est_avg_calorie_intake	Main_meals_daily	Snacks	Height	Smoker	Water_daily	Calorie_monitoring	Weight	Physical_activity_level	Technology_time_use	Gender	Diagnostic
Transportation	1.613	0.966	-0.049	-0.030	0.423	8.340	-0.016	8.080	-0.049	-0.026	1.200	-0.002	0.041	0.009	82.965	0.014	0.585	-0.088	0.031
Regular_fiber_diet	0.966	3898.706	0.267	0.162	-2.044	-28.299	-0.402	-411.186	0.098	0.190	-2.769	-0.028	-0.240	-0.054	-135.811	0.128	-2.234	0.627	2.915
Diagnostic_in_family_history	-0.049	0.267	0.148	0.025	0.095	4.201	0.007	-6.864	0.021	0.031	0.348	0.001	0.037	-0.012	27.775	-0.018	0.130	0.019	0.386
High_calorie_diet	-0.030	0.162	0.025	0.103	0.072	2.514	-0.014	2.786	-0.002	0.022	0.224	-0.003	0.002	-0.012	16.862	-0.029	0.092	0.011	0.155
Sedentary_hours_daily	0.423	-2.044	0.095	0.072	473.490	-13.087	0.376	273.718	-0.612	0.063	-0.588	-0.010	-0.071	-0.024	19.164	-0.476	648.167	-0.252	-0.638
Age	8.340	-28.299	4.201	2.514	-13.087	401083.883	-5.636	-88.785	18.485	3.155	-34.923	-0.343	10.451	-1.016	-2773.866	19.992	-22.877	10.204	-19.336
Alcohol	-0.016	-0.402	0.007	-0.014	0.376	-5.636	0.269	-0.294	-0.030	-0.011	-0.510	-0.006	-0.031	-0.001	49.098	0.036	0.509	0.003	-0.154
Est_avg_calorie_intake	8.080	-411.186	-6.864	2.786	273.718	-88.785	-0.294	188421.795	-4.574	0.611	-930.358	-2.613	-4.249	-1.236	-20457.233	-1.739	371.218	-5.736	-32.981
Main_meals_daily	-0.049	0.098	0.021	-0.002	-0.612	18.485	-0.030	-4.574	0.607	-0.033	0.611	0.001	0.032	0.000	-0.381	0.095	-0.845	0.030	0.040

Snacks	-0.026	0.190	0.031	0.022	0.063	3.155	-0.011	0.611	-0.033	0.217	0.269	-0.004	0.044	-0.011	21.545	-0.010	0.083	0.022	0.301
Height	1.200	-2.769	0.348	0.224	-0.588	-34.923	-0.510	-930.358	0.611	0.269	3375.396	0.958	-1.181	-0.083	-182.428	-0.238	-1.676	0.073	2.671
Smoker	-0.002	-0.028	0.001	-0.003	-0.010	-0.343	-0.006	-2.613	0.001	-0.004	0.958	0.020	-0.003	0.002	-2.570	0.002	-0.013	0.002	-0.001
Water_daily	0.041	-0.240	0.037	0.002	-0.071	10.451	-0.031	-4.249	0.032	0.044	-1.181	-0.003	0.373	0.002	-28.512	0.089	-0.100	0.035	0.166
Calorie_monitoring	0.009	-0.054	-0.012	-0.012	-0.024	-1.016	-0.001	-1.236	0.000	-0.011	-0.083	0.002	0.002	0.041	-6.560	0.016	-0.031	-0.009	-0.076
Weight	82.965	-135.811	27.775	16.862	19.164	-2773.866	49.098	-20457.233	-0.381	21.545	-182.428	-2.570	-28.512	-6.560	10404840.733	7.406	-52.872	-21.290	113.933
Physical_activity_level	0.014	0.128	-0.018	-0.029	-0.476	19.992	0.036	-1.739	0.095	-0.010	-0.238	0.002	0.089	0.016	7.406	0.732	-0.648	0.079	-0.336
Technology_time_use	0.585	-2.234	0.130	0.092	648.167	-22.877	0.509	371.218	-0.845	0.083	-1.676	-0.013	-0.100	-0.031	-52.872	-0.648	887.440	-0.346	-0.873
Gender	-0.088	0.627	0.019	0.011	-0.252	10.204	0.003	-5.736	0.030	0.022	0.073	0.002	0.035	-0.009	-21.290	0.079	-0.346	0.250	-0.036
Diagnostic	0.031	2.915	0.386	0.155	-0.638	-19.336	-0.154	-32.981	0.040	0.301	2.671	-0.001	0.166	-0.076	113.933	-0.336	-0.873	-0.036	3.936

Matrice de convolutie

Matricea de covarianță nu este standardizată și poate fi dificil de interpretat, deoarece valorile sunt sensibile la scala datelor. Prin urmare, este adesea preferată utilizarea coeficientului de corelație.

- Măsoară direcția și forța relațiilor liniare între variabilele continue
- Valorile sunt standardizate între -1 și 1, unde o valoare apropiată de 1 indică o corelație pozitivă perfectă, o valoare apropiată de -1 indică o corelație negativă perfectă
- O valoare apropiată de 0 indică o corelație slabă sau inexistentă



Transportation
Regular_fiber_diet
Diagnostic_in_family_history
High_calorie_diet
Sedentary_hours_daily
Age
Alcohol
Est_avg_calorie_intake
Main_meals_daily
Snacks
Height
Smoker
Water_daily
Calorie_monitoring
Weight
Physical_activity_level
Technology_time_use
Gender
Diagnostic
0.8
0.6
0.4
0.2
0.0
−0.2

2. Utilizarea Algoritmilor de învățare automată

Luand in calcul rezultatele de la pasul anterior am ales sa standardizez datele si sa aplic o tehnica de selectare a atributelor (Variance Threshold)

Numărul total de caracteristici: 18
Numărul de caracteristici selectate folosind **VarianceThreshold**: 15
Numărul de caracteristici selectate folosind **SelectPercentile**: 12
Caracteristicile **eliminate** folosind VarianceThreshold: {'High_calorie_diet', 'Smoker', 'Calorie_monitoring'}
Caracteristicile **eliminate** folosind SelectPercentile: {'Weight', 'Technology_time_use', 'Regular_fiber_diet', 'Est_avg_calorie_intake', 'Height', 'Sedentary_hours_daily'}

VarianceThreshold elimina caracteristicile cu variatie mica, cele trei eliminate fiind categorice, au avut etichetele codificate si pe baza variatiei dintre ele au fost eliminate.
SelectPercentile se foloseste de scorurile de importanta a caracteristicilor in functie de o anumita metrica (in cazul de fata f_score)

Acuratete algoritmi folosind hiperparametrii cei mai buni rezultati in urma Grid-Search:
Cea mai buna acuratete o are **GradientBoostedTrees**

index	Model	Hyperparameters	Accuracy
0	RandomForest	{'max_depth': None, 'max_samples': 1.0, 'n_estimators': 100}	0.912
1	ExtraTrees	{'bootstrap': True, 'max_depth': None, 'max_samples': 1.0, 'n_estimators': 200}	0.888
2	**GradientBoostedTrees**	**{'learning_rate': 0.5, 'max_depth': 5, 'n_estimators': 200}**	**0.945**
3	SVM	{'C': 300, 'kernel': 'rbf'}	0.613

Precizia, Recall-ul si F1-score-ul pentru fiecare clasa in parte encodata in numere:

- Se observa ca valorile maxime sunt intalnite la RandomForest si ExtraTrees pe toate coloanele pentru clasa D6, desi cea mai buna acuratete este la GradientBoostedTrees. Acest lucru inseamna ca desi clasa D6 este clasificata perfect, celelalte clase nu.

index	Model	Class	Precision	Recall	F1
0	RandomForest	0	0.940	0.959	0.949
1	RandomForest	1	0.831	0.891	0.860
2	RandomForest	2	0.885	0.900	0.893
3	RandomForest	3	0.830	0.830	0.830
4	RandomForest	4	0.900	0.844	0.871

5	RandomForest	5	**1.000**	0.956	0.977
6	RandomForest	6	**1.000**	**1.000**	**1.000**
7	ExtraTrees	0	0.904	0.959	0.931
8	ExtraTrees	1	0.837	0.745	0.788
9	ExtraTrees	2	0.797	0.783	0.790
10	ExtraTrees	3	0.851	0.851	0.851
11	ExtraTrees	4	0.841	0.906	0.872
12	ExtraTrees	5	**1.000**	0.978	0.989
13	ExtraTrees	6	**1.000**	**1.000**	**1.000**
14	GradientBoostedTrees	0	**1.000**	0.918	0.957
15	GradientBoostedTrees	1	0.931	0.982	0.956
16	GradientBoostedTrees	2	0.931	0.900	0.915
17	GradientBoostedTrees	3	0.872	0.872	0.872
18	GradientBoostedTrees	4	0.953	0.953	0.953
19	GradientBoostedTrees	5	0.978	0.978	0.978
20	GradientBoostedTrees	6	0.956	1.000	0.977
21	SVM	0	0.621	0.735	0.673
22	SVM	1	0.690	0.364	0.476
23	SVM	2	0.640	0.267	0.376
24	SVM	3	0.677	0.447	0.538
25	SVM	4	0.520	0.609	0.561
26	SVM	5	0.438	0.867	0.582
27	SVM	6	0.833	1.000	0.909

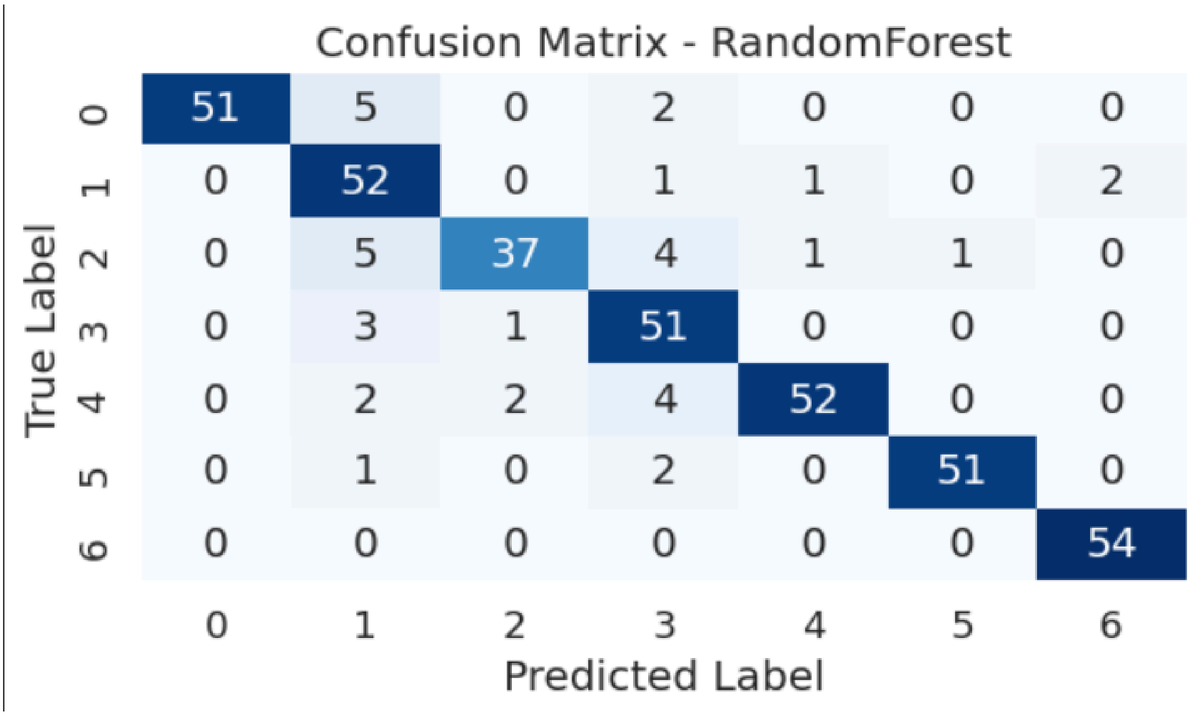
Rezultatele medii pentru scoruri pe fiecare clasa in parte:
Clasele cu cele mai bune predictii sunt D0 si D6

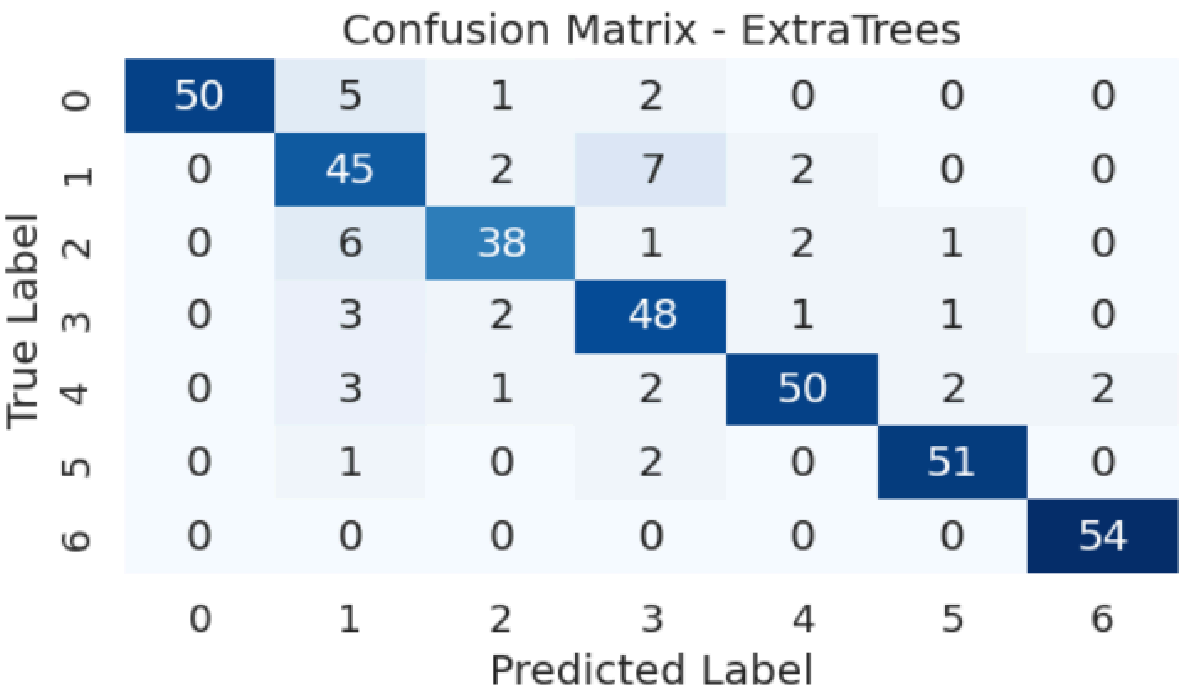
Class	Precision Mean	Precision Variance	Recall Mean	Recall Variance	F1 Mean	F1 Variance
D0	**0.914**	0.030	0.806	**0.007**	0.855	0.015
D1	0.727	**0.009**	0.763	0.070	0.735	0.035

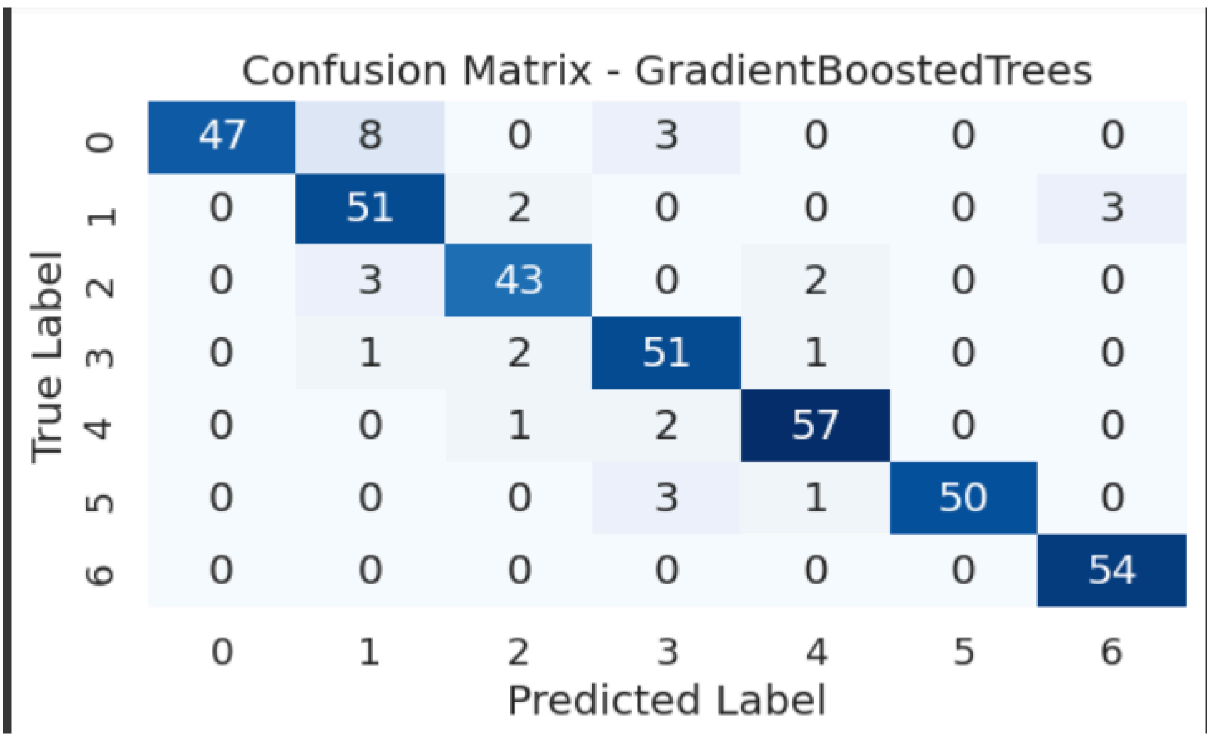
D2	0.786	0.044	0.672	0.098	0.714	0.081
D3	0.739	0.027	0.714	0.144	0.701	0.101
D4	0.772	0.080	0.758	0.064	0.765	0.071
D5	0.844	0.069	0.944	**0.000**	0.871	0.029
D6	**0.918**	0.011	**1.000**	**0.000**	**0.955**	**0.004**

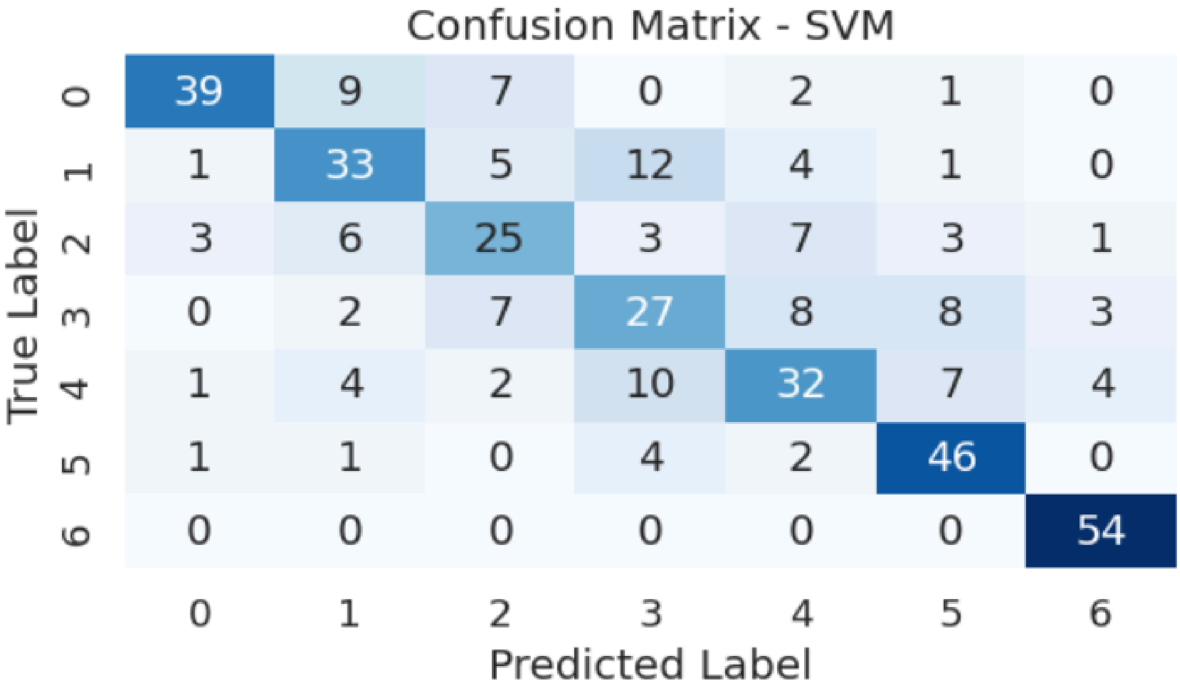
Matricele de confuzie pentru cei mai buni parametrii de la fiecare algoritm:

- Dupa cum se observa si dupa acuratete, ExtraTrees este cel mai echilibrat (diagonala cea mai observabila)




Confusion Matrix - ExtraTrees
True Label
Predicted Label
0 1 2 3 4 5 6
50 5 1 2 0 0 0
0 45 2 7 2 0 0
0 6 38 1 2 1 0
0 3 2 48 1 1 0
0 3 1 2 50 2 2
0 1 0 2 0 51 0
0 0 0 0 0 0 54


Confusion Matrix - GradientBoostedTrees
True Label
Predicted Label
0 1 2 3 4 5 6
47 8 0 3 0 0 0
0 51 2 0 0 0 3
0 3 43 0 2 0 0
0 1 2 51 1 0 0
0 0 1 2 57 0 0
0 0 0 3 1 50 0
0 0 0 0 0 0 54


Confusion Matrix - SVM
True Label
Predicted Label
0
1
2
3
4
5
6
39 9 7 0 2 1 0
1 33 5 12 4 1 0
3 6 25 3 7 3 1
0 2 7 27 8 8 3
1 4 2 10 32 7 4
1 1 0 4 2 46 0
0 0 0 0 0 0 54