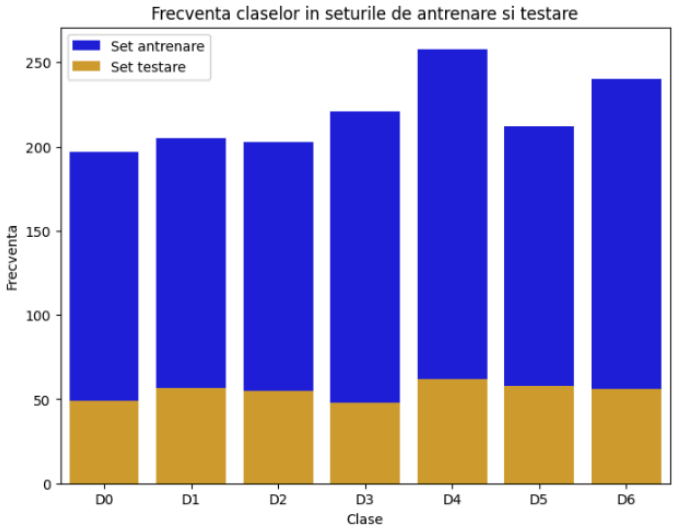
Tema 1 ML - Man Andrei Vlad

Exploratory Data Analysis

1. Echilibrul de clase



Folosind metoda *train_test_split* din sklearn obtinem o impartire echilibrata a claselor.

2. Vizualizarea datelor

Am determinat cele doua categorii de atribute, numerice si categorice:

```
NUMERICAL_COLUMNS = [
    "Age",
    "Est_avg_calorie_intake",
    "Main_meals_daily",
    "Height",
    "Water_daily",
    "Weight",
    "Sedentary_hours_daily",
    "Physical_activity_level",
    "Regular_fiber_diet",
]
```

```
CATEGORICAL_COLUMNS = [
    "Transportation",
    "Diagnostic_in_family_history",
    "High_calorie_diet",
    "Alcohol",
    "Snacks",
    "Smoker",
    "Calorie_monitoring",
    "Technology_time_use",
    "Gender"
]
```

a) Atribute numerice

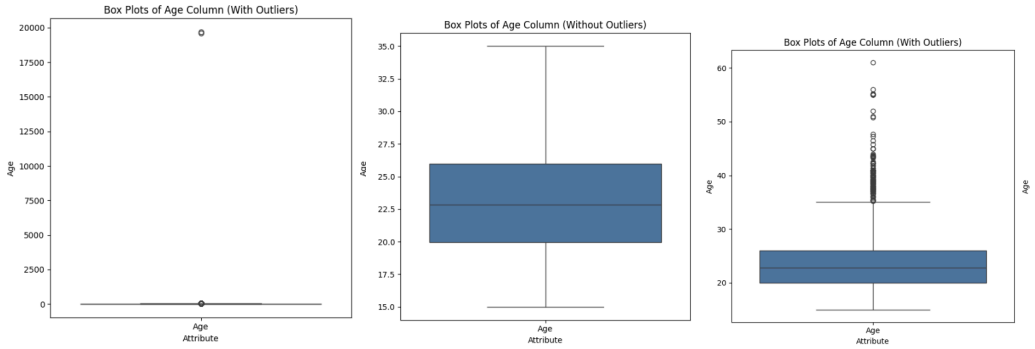
Pentru aceste atribute voi prezenta niste valori statistice insotite de un grafic ce reprezinta valorile minime si maxime, media si intervalul interquartil. Suplimentar, acest grafic afiseaza si valorile outliers. Voi

analiza valorile si voi stabili ce valori sunt intr-un interval “bun” si ce valori sunt cu adevarat daunatoare pentru datele prezente.

Age

```
Statistics for column 'Age':
Mean: 44.79250626392504
Standard Deviation: 633.3118370767136
Min: 15.0
Max: 19685.0
Range: 19670.0
Median: 22.829753
Interquartile Range: 19.97166 - 26.0 (6.02834)
```

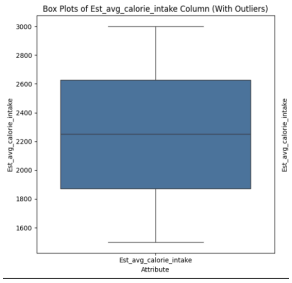
Pe baza graficului de mai jos sunt identificate ca fiind outliers atat valori mici (35+ ani cat si valorile mari din partea superioara ~20k ani). Putem stabili un threshhold pentru varsta ca fiind in intervalul 0-100 pentru valori normale. Eliminand outliers ramase, noul grafic va fi urmatorul:



Est_avg_calorie_intake

```
Mean: 2253.68766267569
Standard Deviation: 434.07579419142866
Min: 1500.0
Max: 3000.0
Range: 1500.0
Median: 2253.0
Interquartile Range: 1871.0 - 2628.0 (757.0)
```

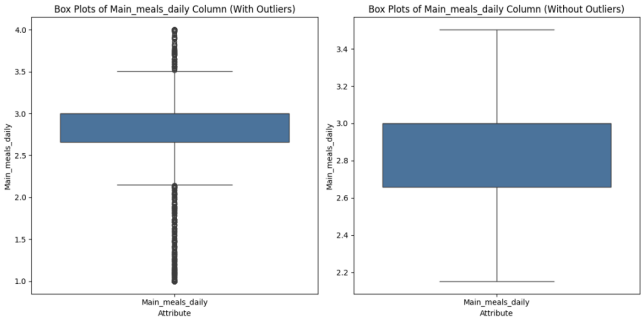
Nu sunt gasite valorile outliers



Main_meals_daily

```
Mean: 2.683471861009891
Standard Deviation: 0.7791790556845525
Min: 1.0
Max: 4.0
Range: 3.0
Median: 3.0
Interquartile Range: 2.658639 - 3.0 (0.341361)
```

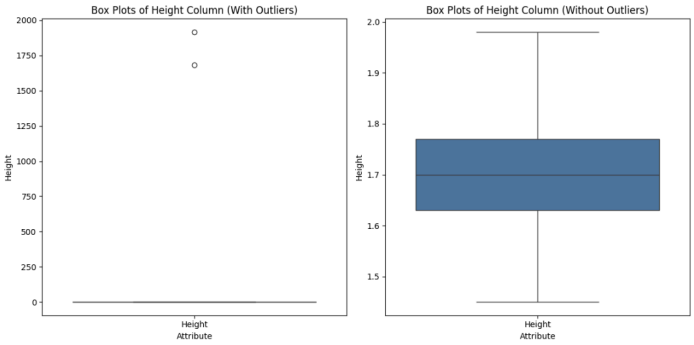
Pe baza graficului vom vedea ce valori outliers suunt identificate. Acestea au valori normale de fapt, insa statistic se afla in intervalele extreme. Le vom considera bune.



Height

```
Mean: 3.5734877667881313
Standard Deviation: 58.09815976912103
Min: 1.45
Max: 1915.0
Range: 1913.55
Median: 1.7
Interquartile Range: 1.63 - 1.77 (0.14000000000000012)
```

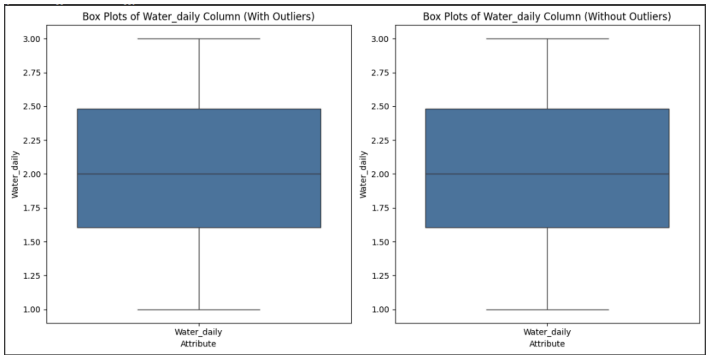
Identificam cateva valori “rele”. Fara acestea restul valorilor sunt normale.



Water_daily

```
Mean: 2.010367264445601
Standard Deviation: 0.6110342044515745
Min: 1.0
Max: 3.0
Range: 2.0
Median: 2.0
Interquartile Range: 1.606076 - 2.480555 (0.8744789999999998)
```

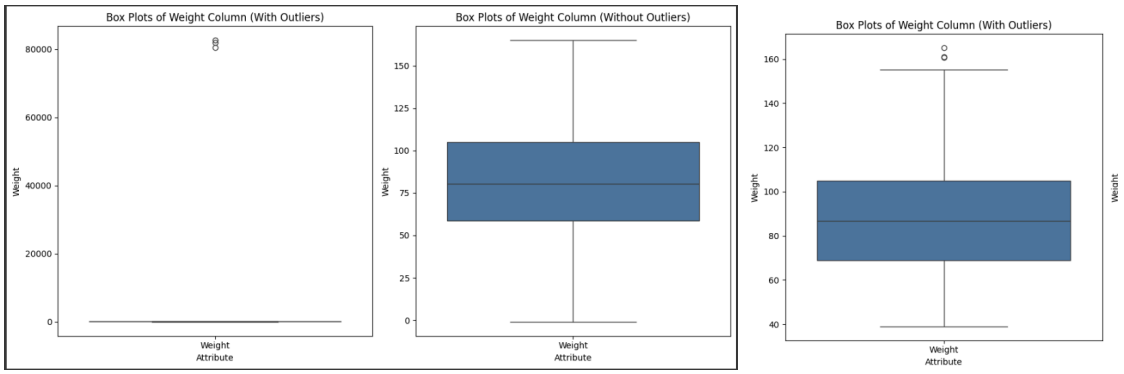
Toate valorile sunt intr-un interval bun.



Weight

```
Mean: 205.63734420249872
Standard Deviation: 3225.6535358208953
Min: -1.0
Max: 82628.0
Range: 82629.0
Median: 80.386078
Interquartile Range: 58.83071 - 105.036075 (46.20536499999999)
```

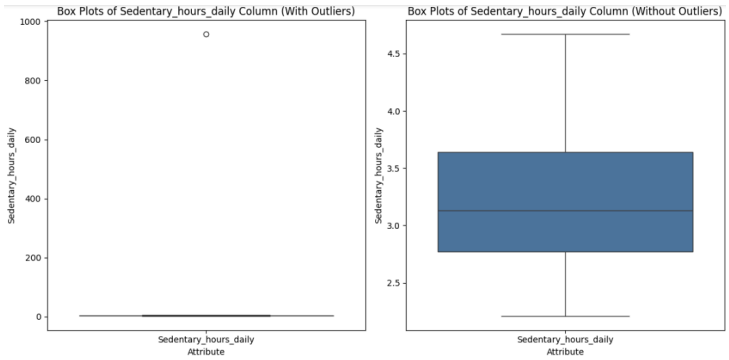
Identificam ca fiind outliers valorile extreme (de ordinul 10k). Alaturi, avem si valorile ce au -1.



Sedentary_hours_daily

```
Mean: 3.693571056741281
Standard Deviation: 21.759834908880748
Min: 2.21
Max: 956.58
Range: 954.37
Median: 3.13
Interquartile Range: 2.77 - 3.64 (0.8700000000000001)
```

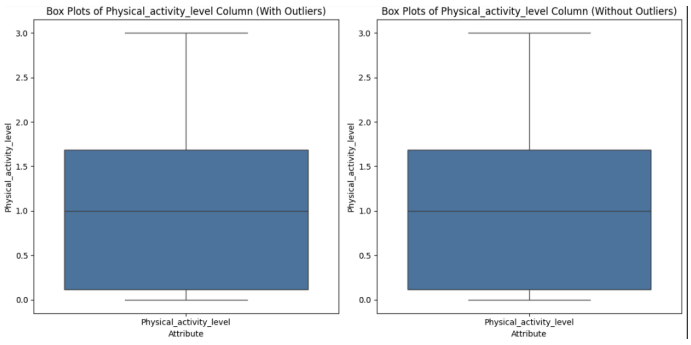
Identificam cateva valori anormale in intervalul extrem.



Physical_activity_level

```
Mean: 1.0126402805830297
Standard Deviation: 0.8555256424802281
Min: 0.0
Max: 3.0
Range: 3.0
Median: 1.0
Interquartile Range: 0.115974 - 1.683497 (1.567523)
```

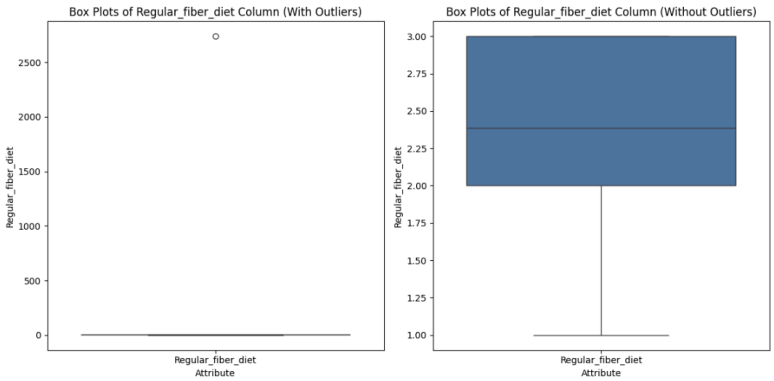
Valorile sunt in regula.



Regular_fiber_diet

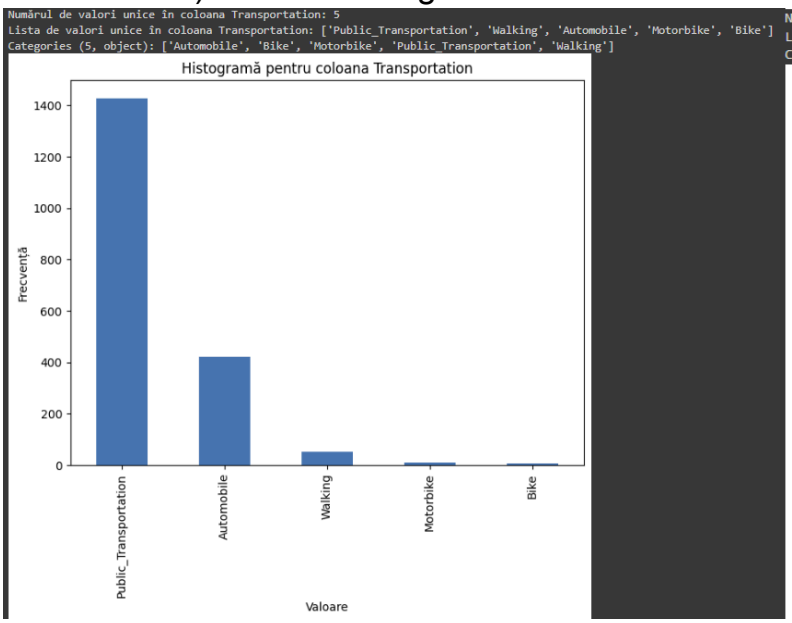
```
Mean: 3.8449373862571576
Standard Deviation: 62.4396174995684
Min: 1.0
Max: 2739.0
Range: 2738.0
Median: 2.387426
Interquartile Range: 2.0 - 3.0 (1.0)
```

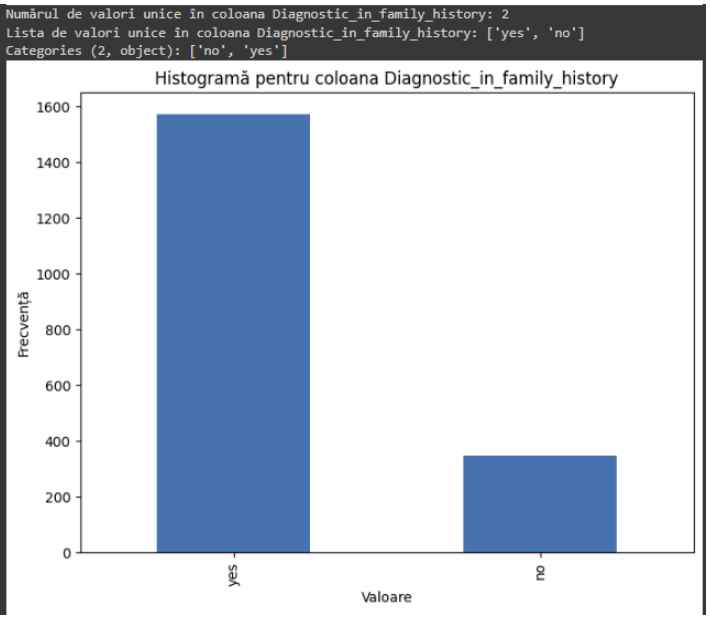
Identificam o valoarea outlier.



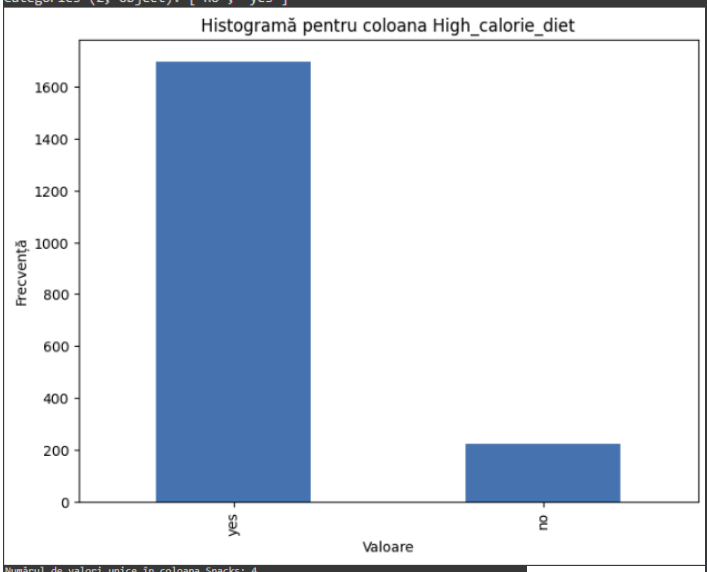
Pentru valorile outlier reale (stabilite de mine) si pentru coloana weight (cu valori -1) am stabilit o strategie de inlocuire a acestora. Vor fi inlocuite cu media valorilor care nu sunt outliers. Putem vedea de exemplu pentru *weight* ca noile valori sunt acum intr-un interval corect (39, 165), initial fiind in intervalul (-1, 82k).

b) Atribute categorice

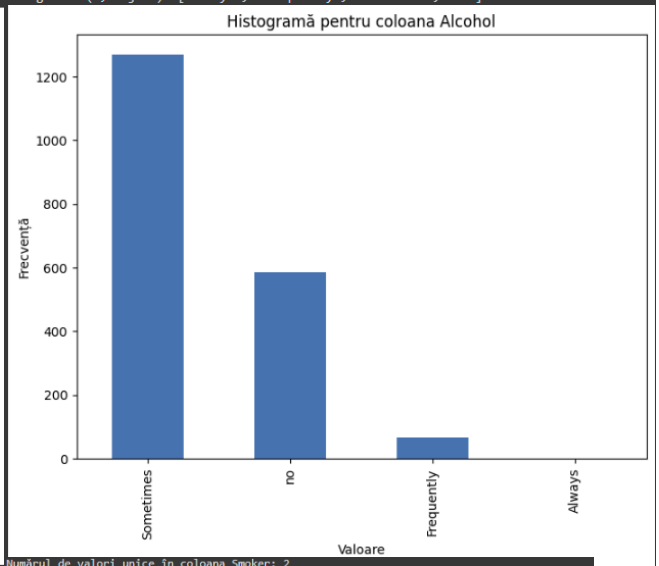




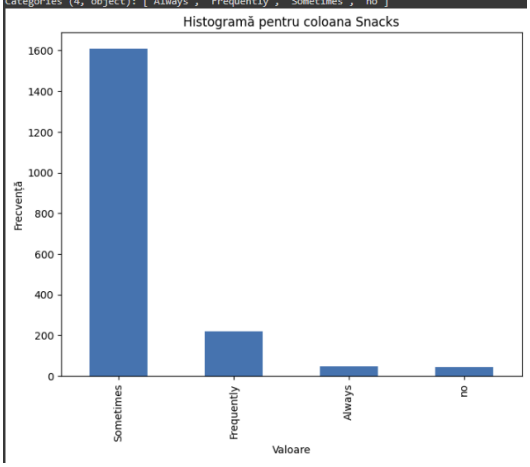
```
Numărul de valori unice în coloana High_calorie_diet: 2
Lista de valori unice în coloana High_calorie_diet: ['no', 'yes']
Categories (2, object): ['no', 'yes']
```



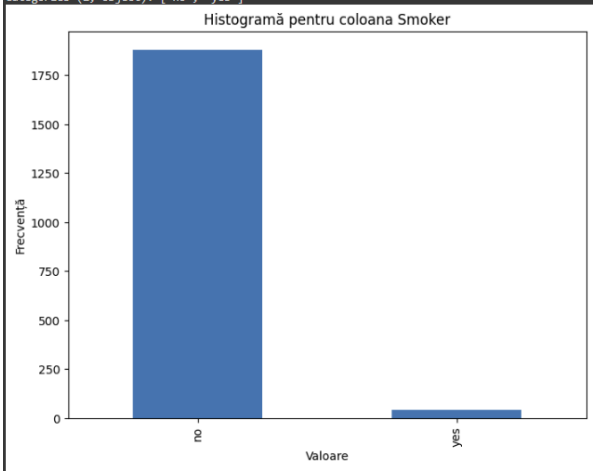

```
Numărul de valori unice în coloana Alcohol: 4
Lista de valori unice în coloana Alcohol: ['no', 'Sometimes', 'Frequently', 'Always']
Categories (4, object): ['Always', 'Frequently', 'Sometimes', 'no']
```



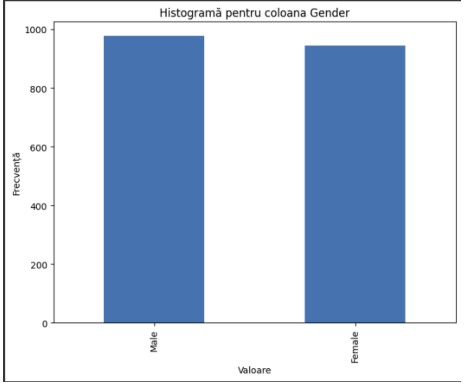

```
Numărul de valori unice în coloana Snacks: 4
Lista de valori unice în coloana Snacks: ['Sometimes', 'Frequently', 'Always', 'no']
Categories (4, object): ['Always', 'Frequently', 'Sometimes', 'no']
```



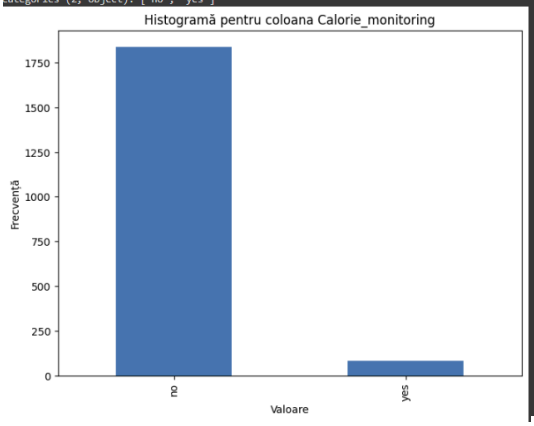

```
Numărul de valori unice în coloana Smoker: 2
Lista de valori unice în coloana Smoker: ['no', 'yes']
Categories (2, object): ['no', 'yes']
```



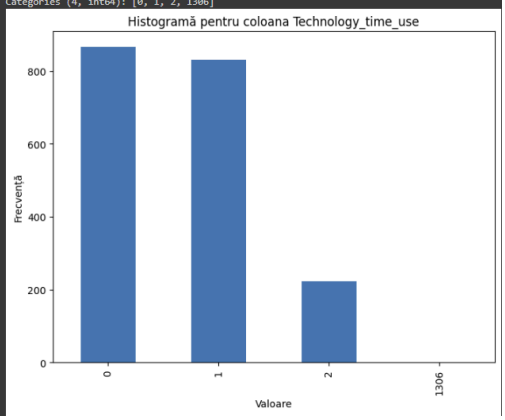

```
Numărul de valori unice în coloana Gender: 2
Lista de valori unice în coloana Gender: ['Female', 'Male']
Categories (2, object): ['Female', 'Male']
```



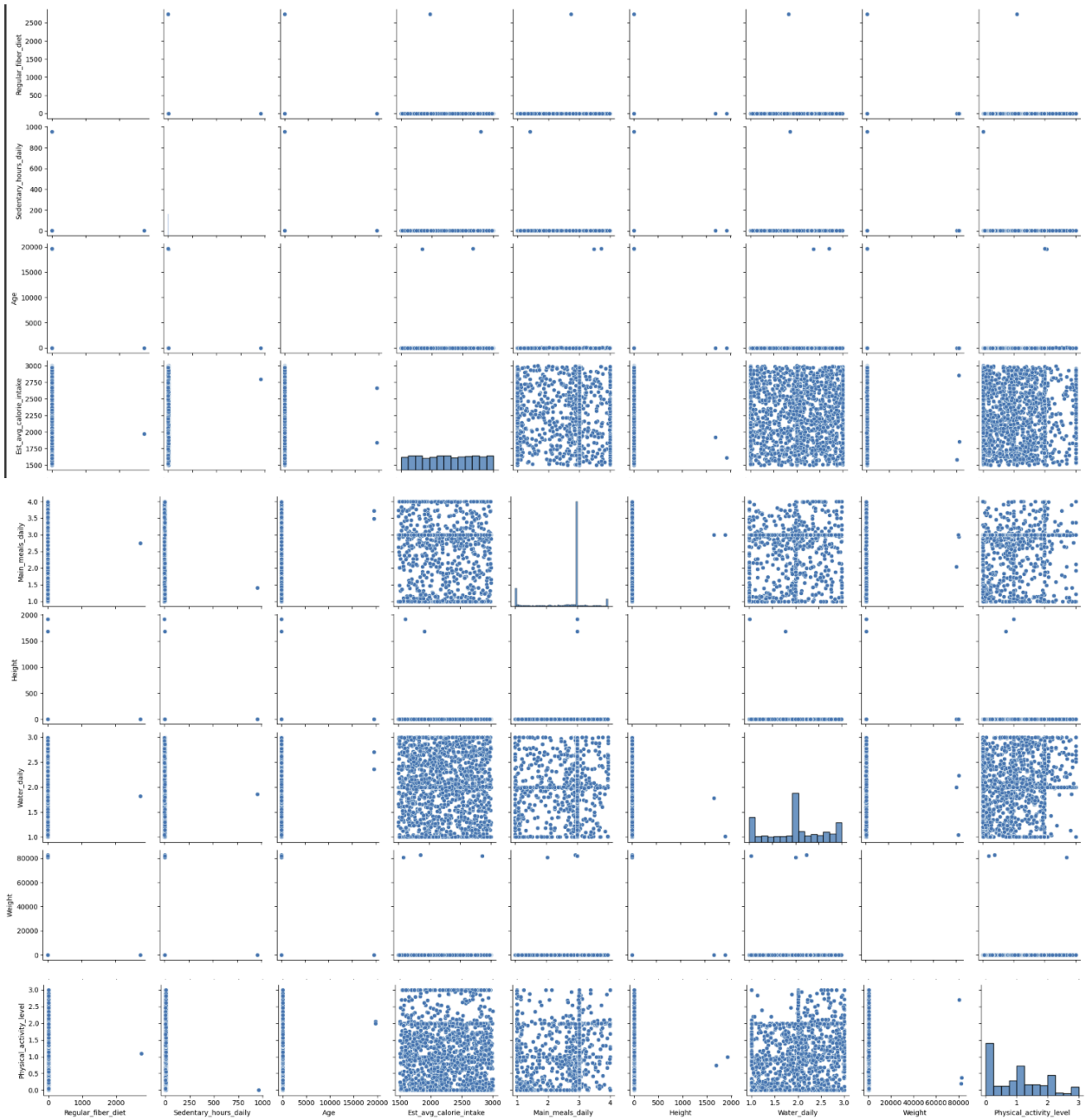

```
Numărul de valori unice în coloana Calorie_monitoring: 2
Lista de valori unice în coloana Calorie_monitoring: ['no', 'yes']
Categories (2, object): ['no', 'yes']
```



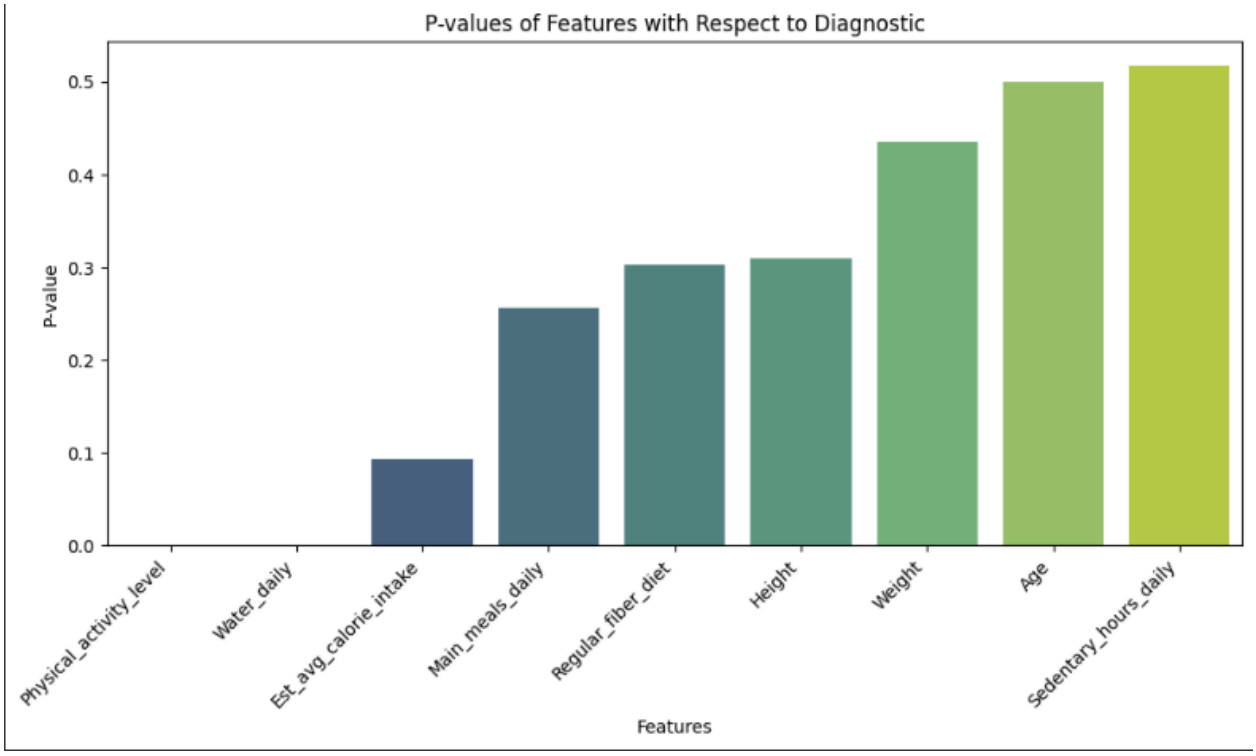

```
Numărul de valori unice în coloana Technology_time_use: 4
Lista de valori unice în coloana Technology_time_use: [1, 0, 2, 1306]
Categories (4, int64): [0, 1, 2, 1306]
```

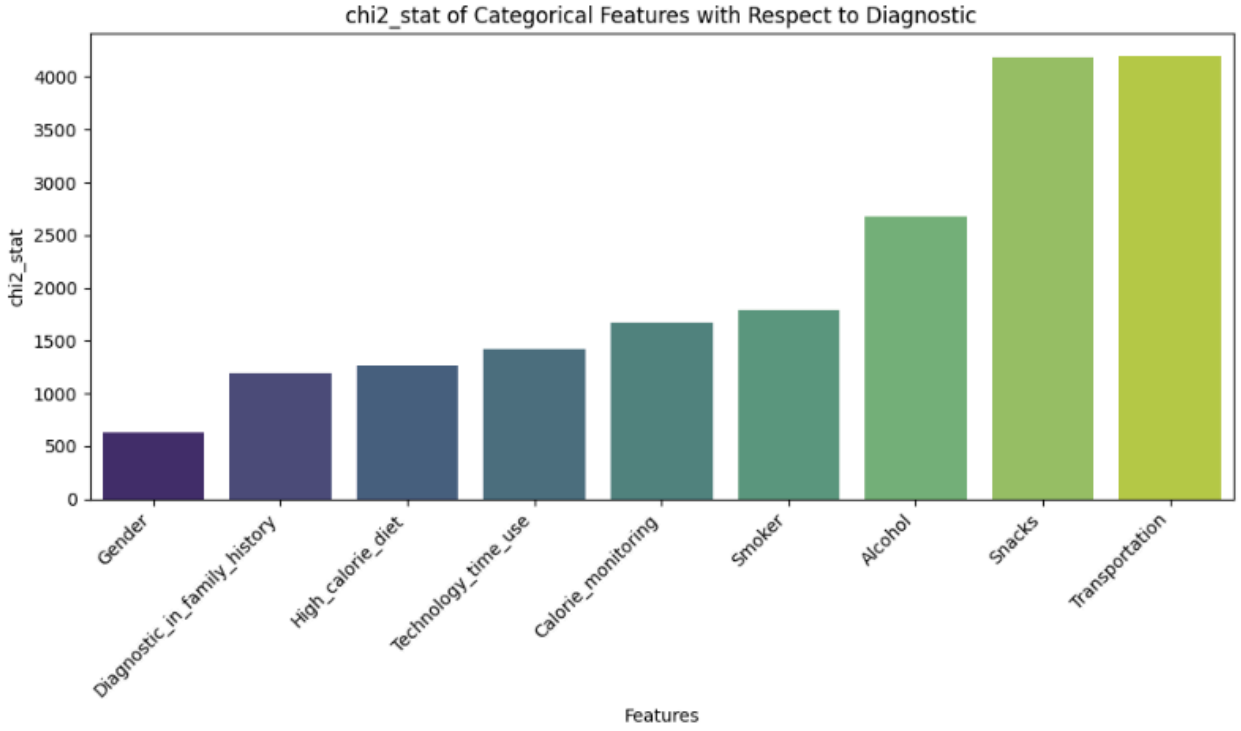


Corelarea valorilor (intre ele)



Corelarea valorilor cu valoarea tinta (Diagnostic)





P-value/Chi-squared ridicat -> corelare puternica

Antrenare si Rezultate

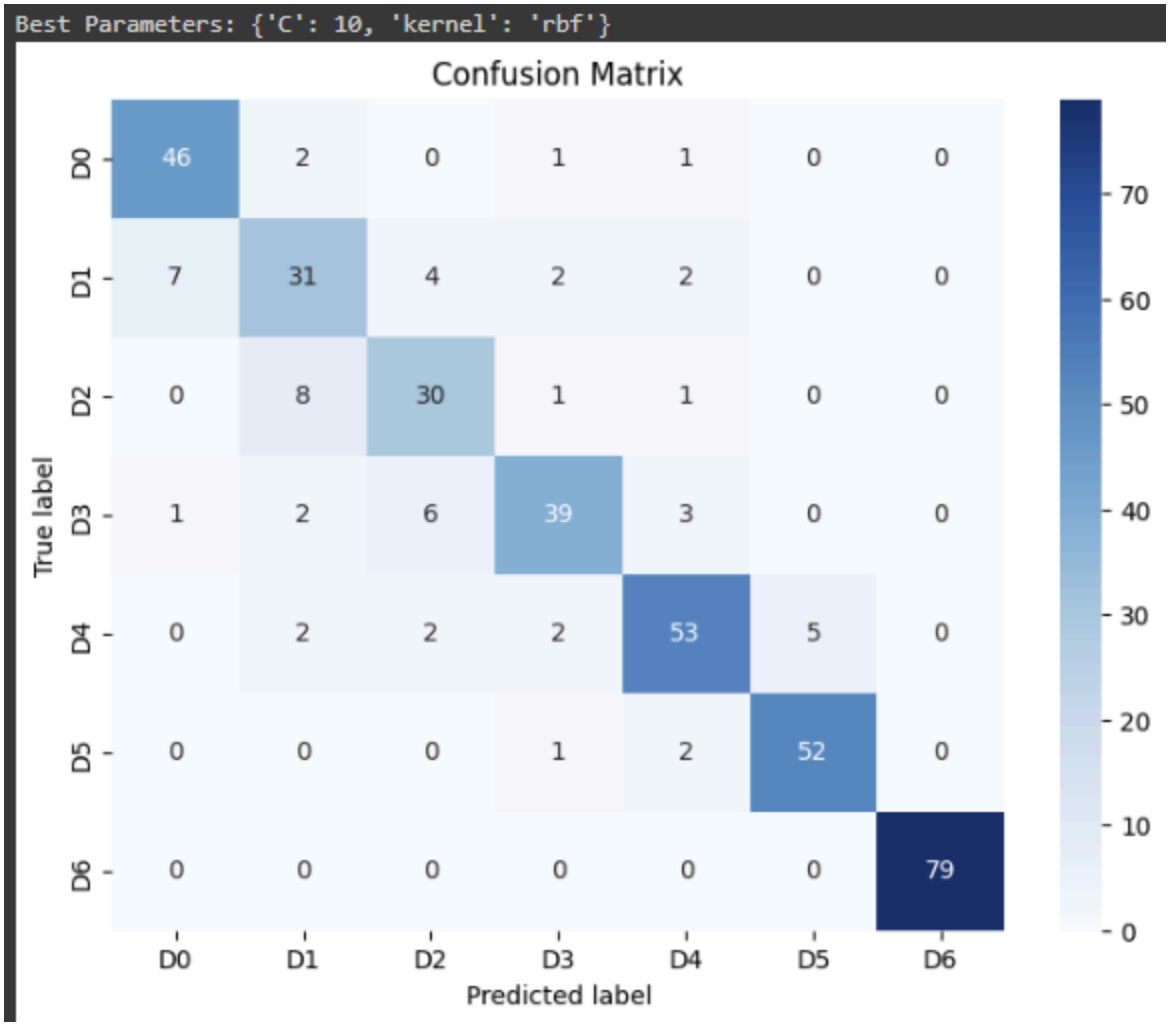
Pentru standardizarea datelor am folosit *StandardScaler*, care are rolul de a transforma valorile numerice prin scaderea mediei si scalare la deviatia standard.

Pentru valorile outliers si cele de la atributul weight (-1), am folosit *SimpleImputer (cu metoda average)* - implementat de mine pentru a putea face handle si la indexii valorilor outliers stabilite de mine.

Pentru selectarea atributelor cele mai relevante, folosesc *VarianceThreshold* cu un threshold p*(1 - p) cu p = 0.8.
Aceasta metoda elimina urmmatoarele atribute: *Diagnostic_in_family_history, High_calorie_diet, Smoker, Calorie_monitoring.* Acestea nu vor mai fi folosite la antrenare/ testare. (de la 18 ajungem la 14)

<u>Support Vector Classifier</u>

		Precision	Precision dev	Recall	Recall dev	F1 Score	F1 Score dev	Accuracy	Accuracy dev
param_C	param_kernel								
0.010000	linear	0.612	0.015	0.613	0.023	0.586	0.02	0.613	0.023
	poly	0.17	0.003	0.232	0.011	0.139	0.011	0.232	0.011
	rbf	0.028	0.0	0.167	0.001	0.048	0.001	0.167	0.001
	sigmoid	0.028	0.0	0.167	0.001	0.048	0.001	0.167	0.001
0.100000	linear	0.773	0.022	0.766	0.02	0.765	0.02	0.766	0.02
	poly	0.671	0.017	0.661	0.015	0.656	0.016	0.661	0.015
	rbf	0.603	0.034	0.561	0.017	0.524	0.016	0.561	0.017
	sigmoid	0.448	0.048	0.465	0.008	0.394	0.008	0.465	0.008
1	linear	0.823	0.012	0.816	0.013	0.815	0.013	0.816	0.013
	poly	0.809	0.016	0.798	0.013	0.799	0.013	0.798	0.013
	rbf	0.79	0.013	0.783	0.013	0.784	0.013	0.783	0.013
	sigmoid	0.477	0.019	0.473	0.025	0.472	0.021	0.473	0.025
10	linear	0.835	0.015	0.828	0.017	0.828	0.017	0.828	0.017
	poly	0.821	0.024	0.816	0.023	0.815	0.023	0.816	0.023
	rbf	0.841	0.011	0.835	0.01	0.836	0.01	0.835	0.01
	sigmoid	0.427	0.02	0.421	0.025	0.416	0.022	0.421	0.025



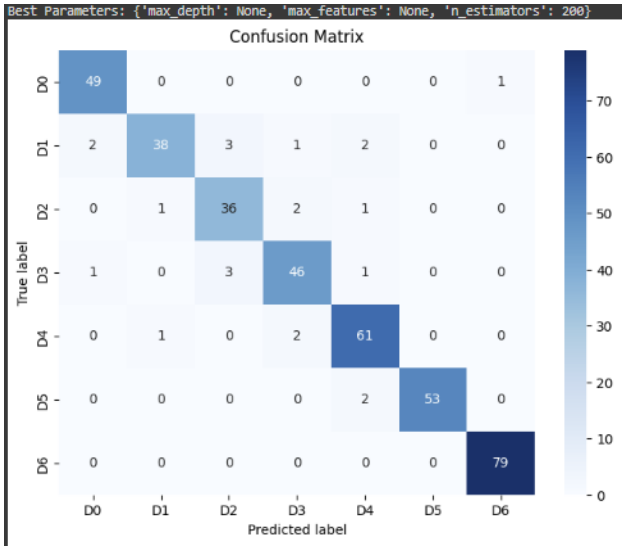
Hiperparametrii influenteaza mult rezultatele. Observam ca parametrul C este cel care influenteaza cel mai mult rezultatele (C mai mare -> acuratete mai mare). La kernel, observam ca sigmoid este in general cel mai slab. Pentru C mare nu sunt diferente foarte mari intre kernele, fata de rezultatele in care avem un C mai mic.

Extra Trees Classifier

param_n_estimators	param_max_depth	param_max_features	Precision	Precision dev	Recall	Recall dev	F1 Score	F1 Score dev	Accuracy	Accuracy dev
50	5	0.500000	0.777	0.022	0.776	0.019	0.773	0.018	0.776	0.019
100	5	0.500000	0.795	0.014	0.79	0.014	0.787	0.013	0.79	0.014
200	5	0.500000	0.791	0.021	0.789	0.02	0.786	0.02	0.789	0.02
50	5	0.750000	0.788	0.024	0.779	0.019	0.779	0.019	0.779	0.019
100	5	0.750000	0.802	0.015	0.794	0.012	0.794	0.012	0.794	0.012
200	5	0.750000	0.801	0.022	0.794	0.021	0.794	0.02	0.794	0.021
50	5	nan	0.816	0.024	0.802	0.025	0.803	0.024	0.802	0.025
100	5	nan	0.807	0.025	0.793	0.022	0.793	0.022	0.793	0.022
200	5	nan	0.807	0.027	0.794	0.029	0.795	0.027	0.794	0.029
50	7	0.500000	0.857	0.013	0.852	0.015	0.852	0.014	0.852	0.015
100	7	0.500000	0.861	0.016	0.855	0.017	0.856	0.017	0.855	0.017
200	7	0.500000	0.857	0.021	0.852	0.022	0.853	0.022	0.852	0.022
50	7	0.750000	0.86	0.024	0.853	0.024	0.853	0.024	0.853	0.024
100	7	0.750000	0.877	0.016	0.867	0.018	0.869	0.017	0.867	0.018
200	7	0.750000	0.873	0.022	0.865	0.023	0.866	0.022	0.865	0.023
50	7	nan	0.886	0.013	0.875	0.015	0.877	0.014	0.875	0.015
100	7	nan	0.883	0.016	0.873	0.021	0.874	0.02	0.873	0.021
200	7	nan	0.881	0.018	0.872	0.022	0.873	0.021	0.872	0.022
50	nan	0.500000	0.91	0.01	0.908	0.009	0.908	0.009	0.908	0.009
100	nan	0.500000	0.913	0.015	0.911	0.015	0.911	0.015	0.911	0.015
200	nan	0.500000	0.914	0.009	0.912	0.009	0.912	0.009	0.912	0.009
50	nan	0.750000	0.917	0.016	0.916	0.017	0.916	0.017	0.916	0.017
100	nan	0.750000	0.923	0.016	0.921	0.016	0.921	0.016	0.921	0.016
200	nan	0.750000	0.923	0.016	0.921	0.017	0.92	0.017	0.921	0.017
50	nan	nan	0.92	0.013	0.919	0.014	0.919	0.014	0.919	0.014
100	nan	nan	0.924	0.016	0.923	0.016	0.923	0.016	0.923	0.016
200	nan	nan	0.925	0.011	0.924	0.011	0.924	0.011	0.924	0.011

Numarul de estimatori nu influenteaza mult rezultatele. Se pot atinge rezultate similare si cu estimatori mai putini. Atat max_depth cat si max_features, cu cat sunt mai crescute cu atat obtin rezultate mai bune. Rezultatele cele mai bune au fost obtinute cu optiunea ‘None’, care de fapt reprezinta maximul la fiecare (adancime maxima si toate features luate in considerare)

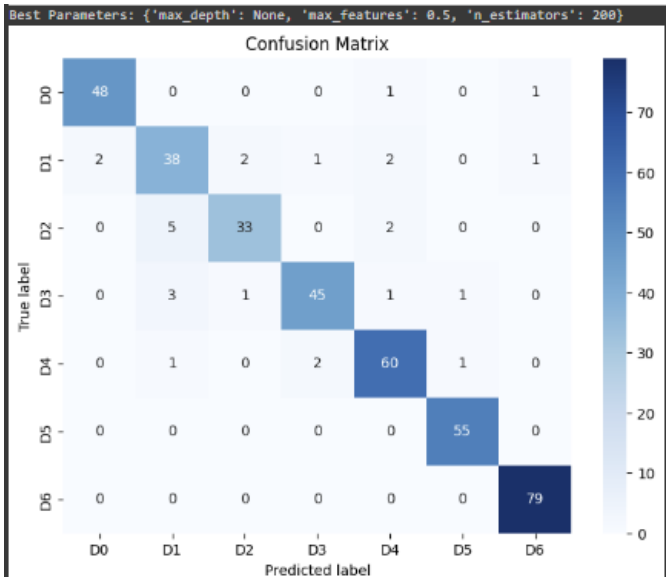
Best Parameters: {'max_depth': None, 'max_features': None, 'n_estimators': 200}



Random Forest Classifier

param_n_estimators	param_max_depth	param_max_features	Precision	Precision dev	Recall	Recall dev	F1 Score	F1 Score dev	Accuracy	Accuracy dev
50	5	0.500000	0.81	0.017	0.8	0.019	0.803	0.018	0.8	0.019
100	5	0.500000	0.808	0.025	0.792	0.026	0.797	0.026	0.792	0.026
200	5	0.500000	0.807	0.025	0.794	0.025	0.798	0.027	0.794	0.025
50	5	0.750000	0.799	0.028	0.773	0.034	0.78	0.032	0.773	0.034
100	5	0.750000	0.809	0.034	0.788	0.039	0.797	0.037	0.788	0.039
200	5	0.750000	0.807	0.034	0.782	0.037	0.788	0.036	0.782	0.037
50	5	nan	0.797	0.031	0.78	0.032	0.786	0.031	0.78	0.032
100	5	nan	0.791	0.032	0.75	0.039	0.757	0.035	0.75	0.039
200	5	nan	0.788	0.03	0.747	0.035	0.754	0.032	0.747	0.035
50	7	0.500000	0.862	0.022	0.854	0.023	0.855	0.022	0.854	0.023
100	7	0.500000	0.867	0.02	0.862	0.021	0.86	0.02	0.862	0.021
200	7	0.500000	0.871	0.02	0.862	0.02	0.864	0.019	0.862	0.02
50	7	0.750000	0.867	0.019	0.856	0.02	0.859	0.019	0.856	0.02
100	7	0.750000	0.865	0.026	0.854	0.028	0.856	0.027	0.854	0.028
200	7	0.750000	0.864	0.026	0.853	0.027	0.858	0.026	0.853	0.026
50	7	nan	0.861	0.022	0.845	0.025	0.849	0.023	0.845	0.025
100	7	nan	0.864	0.024	0.849	0.03	0.853	0.028	0.849	0.03
200	7	nan	0.865	0.023	0.85	0.027	0.854	0.025	0.85	0.027
50	nan	0.500000	0.903	0.017	0.901	0.017	0.901	0.017	0.901	0.017
100	nan	0.500000	0.903	0.013	0.9	0.014	0.9	0.014	0.9	0.014
200	nan	0.500000	0.908	0.016	0.905	0.016	0.905	0.016	0.905	0.016
50	nan	0.750000	0.905	0.016	0.903	0.017	0.903	0.017	0.903	0.017
100	nan	0.750000	0.903	0.027	0.902	0.027	0.902	0.027	0.902	0.027
200	nan	0.750000	0.904	0.028	0.902	0.029	0.902	0.028	0.902	0.029
50	nan	nan	0.899	0.024	0.898	0.025	0.897	0.025	0.898	0.025
100	nan	nan	0.9	0.024	0.898	0.024	0.898	0.024	0.898	0.024
200	nan	nan	0.895	0.029	0.894	0.028	0.893	0.028	0.894	0.028

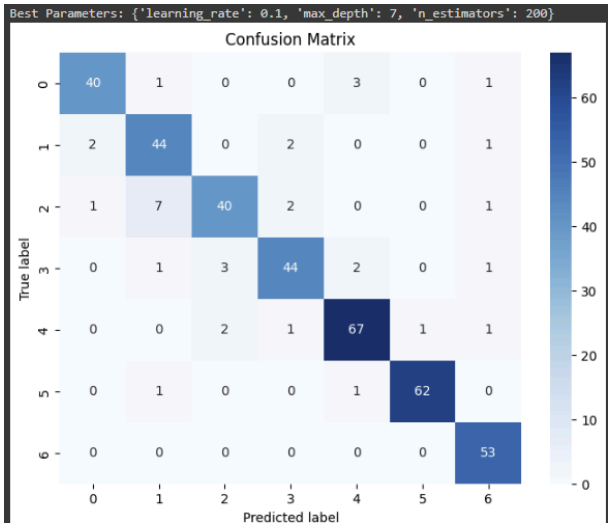
Numarul de estimatori nu influenteaza mult rezultatele. Se pot atinge rezultate similare si cu estimatori mai putini. Max_features nu influenteaza cu mult acuratetea. Parammetrul cel mai immportant este max_depth, care da cele mai bune rezultate ca fiind maxim.



Boosted Trees Classifier

			Precision	Precision dev	Recall	Recall dev	F1 Score	F1 Score dev	Accuracy	Accuracy dev
param_n_estimators	param_max_depth	param_learning_rate								
50	5	0.100000	0.891	0.017	0.889	0.015	0.888	0.018	0.889	0.019
100	5	0.100000	0.907	0.009	0.905	0.01	0.905	0.01	0.905	0.01
200	5	0.100000	0.913	0.012	0.91	0.013	0.91	0.013	0.91	0.013
50	7	0.100000	0.897	0.015	0.894	0.016	0.894	0.015	0.894	0.016
100	7	0.100000	0.905	0.011	0.903	0.012	0.903	0.012	0.903	0.012
200	7	0.100000	0.915	0.011	0.913	0.011	0.913	0.011	0.913	0.011
50	nan	0.100000	0.897	0.011	0.895	0.013	0.894	0.012	0.895	0.013
100	nan	0.100000	0.909	0.011	0.906	0.011	0.906	0.011	0.906	0.011
200	nan	0.100000	0.914	0.013	0.911	0.014	0.911	0.013	0.911	0.014
50	5	0.010000	0.823	0.018	0.816	0.02	0.817	0.019	0.816	0.02
100	5	0.010000	0.845	0.013	0.84	0.014	0.84	0.013	0.84	0.014
200	5	0.010000	0.857	0.013	0.854	0.015	0.854	0.014	0.854	0.015
50	7	0.010000	0.869	0.016	0.865	0.017	0.864	0.016	0.865	0.017
100	7	0.010000	0.879	0.018	0.876	0.018	0.876	0.018	0.876	0.018
200	7	0.010000	0.89	0.014	0.887	0.014	0.886	0.013	0.887	0.014
50	nan	0.010000	0.845	0.024	0.842	0.025	0.842	0.024	0.842	0.025
100	nan	0.010000	0.859	0.019	0.856	0.02	0.856	0.019	0.856	0.02
200	nan	0.010000	0.877	0.018	0.874	0.019	0.874	0.018	0.874	0.019
50	5	0.001000	0.806	0.012	0.796	0.017	0.797	0.016	0.796	0.017
100	5	0.001000	0.815	0.016	0.807	0.02	0.807	0.019	0.807	0.02
200	5	0.001000	0.822	0.02	0.814	0.024	0.814	0.023	0.814	0.024
50	7	0.001000	0.848	0.021	0.844	0.022	0.844	0.022	0.844	0.022
100	7	0.001000	0.857	0.016	0.853	0.017	0.852	0.016	0.853	0.017
200	7	0.001000	0.862	0.021	0.858	0.023	0.857	0.022	0.858	0.023
50	nan	0.001000	0.829	0.018	0.826	0.019	0.824	0.02	0.826	0.019
100	nan	0.001000	0.833	0.015	0.83	0.016	0.829	0.016	0.83	0.016
200	nan	0.001000	0.836	0.019	0.833	0.02	0.832	0.019	0.833	0.02

Best Parameters: {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 200}



Numarul de estimatori influenteaza incremmental acuratetea (la fel ca la restul algoritmilor). Learning rate-ul este parametrul care influenteaza cel mai mult rezultatele. Fata de restul algoritmilor, max_depth-ul nu mai este mai bun cu cat e mai mare.