Inteligență Artificială Tema 2 - ML Aplicat

Lache Alexandra Florentina Georgiana, 332CD

20.05 - 25.05.2025

1 Descriere

Scopul acestui proiect este de a trece prin diferitele stagii ale dezvoltării unei soluții în domeniul inteligenței artificiale. Prin etapele de dezvoltare se numără:

- EDA - etapa de analiză și explorare a datelor
- preprocesarea datelor
- utilizarea propriu-zisă a unui algoritm de învățare automată

În cadrul acestui proiect ne vom folosi de două seturi de date deja separate în train și test. Primul reprezintă riscul de dezvoltare boală coronariană având mai mulți factori, în timp ce cel de-al doilea conține datele înregistrate de niște senzori cu infraroșu.

2 Explorarea și analizarea datelor

Heart Disease Dataset

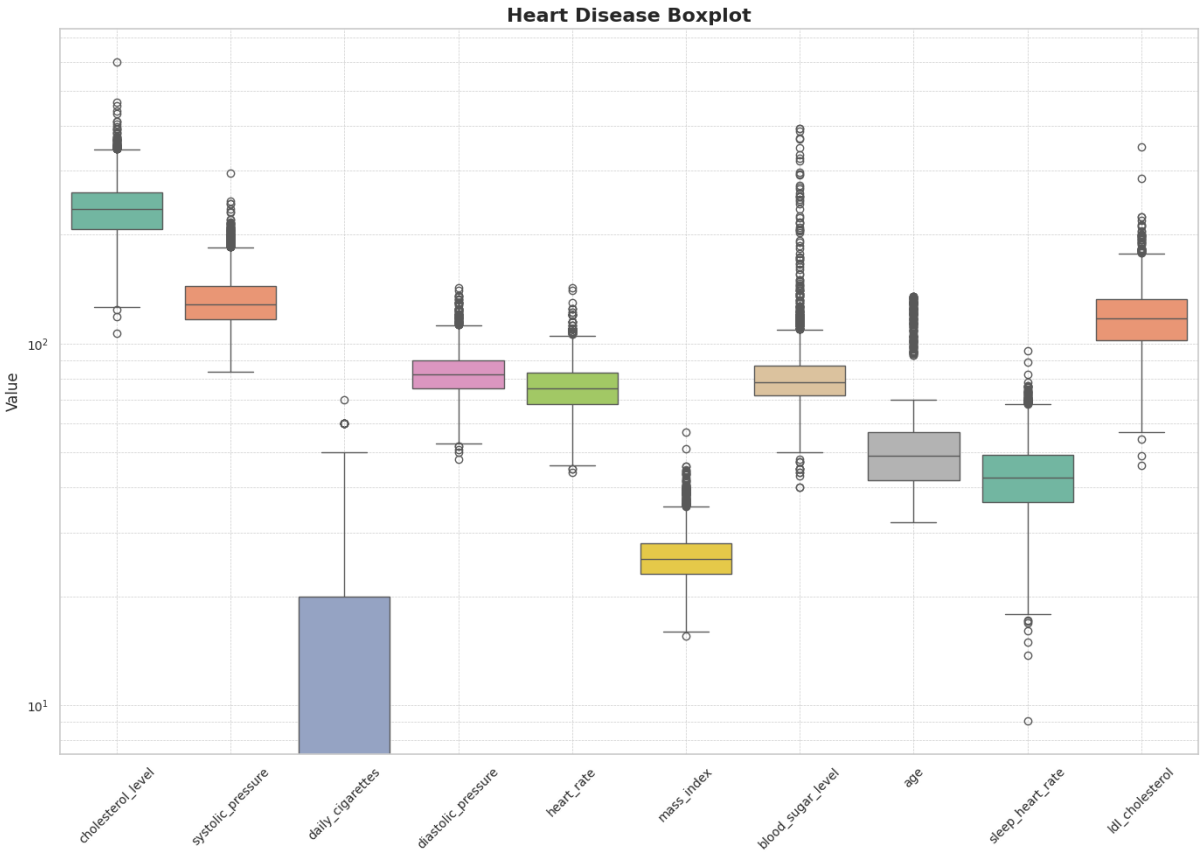
- **Analiza tipului de atribute și a plajei de valori a acestora**
  - Pentru setul Heart Disease am identificat două tipuri de atribute: atribute cu valori continue și atribute cu valori discrete.
  - Atributele cu valori continue sunt: 'cholesterol_level', 'systolic_pressure', 'daily_cigarettes', 'diastolic_pressure', 'heart_rate', 'mass_index', 'blood_sugar_level', 'age', 'sleep_heart_rate', 'ldl_cholesterol'.

Heart Disease Continuous Variables

index	count	mean	std	min	25%	50%	75%	max
cholesterol_level	3878.00000	236.61822	43.48108	107.00000	207.00000	235.00000	262.00000	600.00000
systolic_pressure	4240.00000	132.35460	22.03330	83.50000	117.00000	128.00000	144.00000	295.00000
daily_cigarettes	4217.00000	9.00586	11.91398	0.00000	0.00000	0.00000	20.00000	70.00000
diastolic_pressure	4240.00000	82.89776	11.91039	48.00000	75.00000	82.00000	90.00000	142.50000
heart_rate	4239.00000	75.87898	12.02535	44.00000	68.00000	75.00000	83.00000	143.00000
mass_index	4222.00000	25.80078	4.07936	15.54000	23.07000	25.40000	28.04000	56.80000
blood_sugar_level	3919.00000	81.96366	23.74864	40.00000	72.00000	78.00000	87.00000	394.00000
age	4240.00000	51.57283	14.19855	32.00000	42.00000	49.00000	57.00000	134.79749
sleep_heart_rate	4239.00000	43.16371	9.76962	9.06045	36.48940	42.57163	49.10418	95.25031
ldl_cholesterol	4203.00000	118.28012	23.31328	45.99315	102.08380	117.13560	132.36507	349.27374

  - Câteva observații pe care le putem face pe baza datelor din tabelul de mai sus:
    - există intrări în tabelul de training de unde lipsesc valorile pentru anumite caracteristici, după cum observăm în coloana **count**(*dintr-un total de 4240 intrări în cele două tabele combinate, există atribute pentru care avem mai puține valori, de exemplu 'cholesterol_level' unde avem aproximativ cu 10% mai puține valori decât numărul de intrări*)

* pentru anumite caracteristici nu avem o distribuție simetrică(*de exemplu, media pentru 'daily_cigarettes' este de 9.02018, dar mai mult de 50% din cei chestionați nu sunt fumători*)



– Câteva observații pe care le putem face pe baza boxplot-ului de mai sus:
  * prin reprezentarea pătratelor colorate se observă unde se află 50% din valorile centrate în mediană pentru fiecare caracteristică. Din această cauză, feature-urile care variază mai mult sunt reprezentate prin dreptunghiuri cu lățime mai mare.
  * se observă și anomaliile pentru fiecare caracteristică sub formă de cerculețe care vor avea nevoie să fie tratate separat(spre exemplu, 'blood_sugar_level' are un număr mare de anomalii)

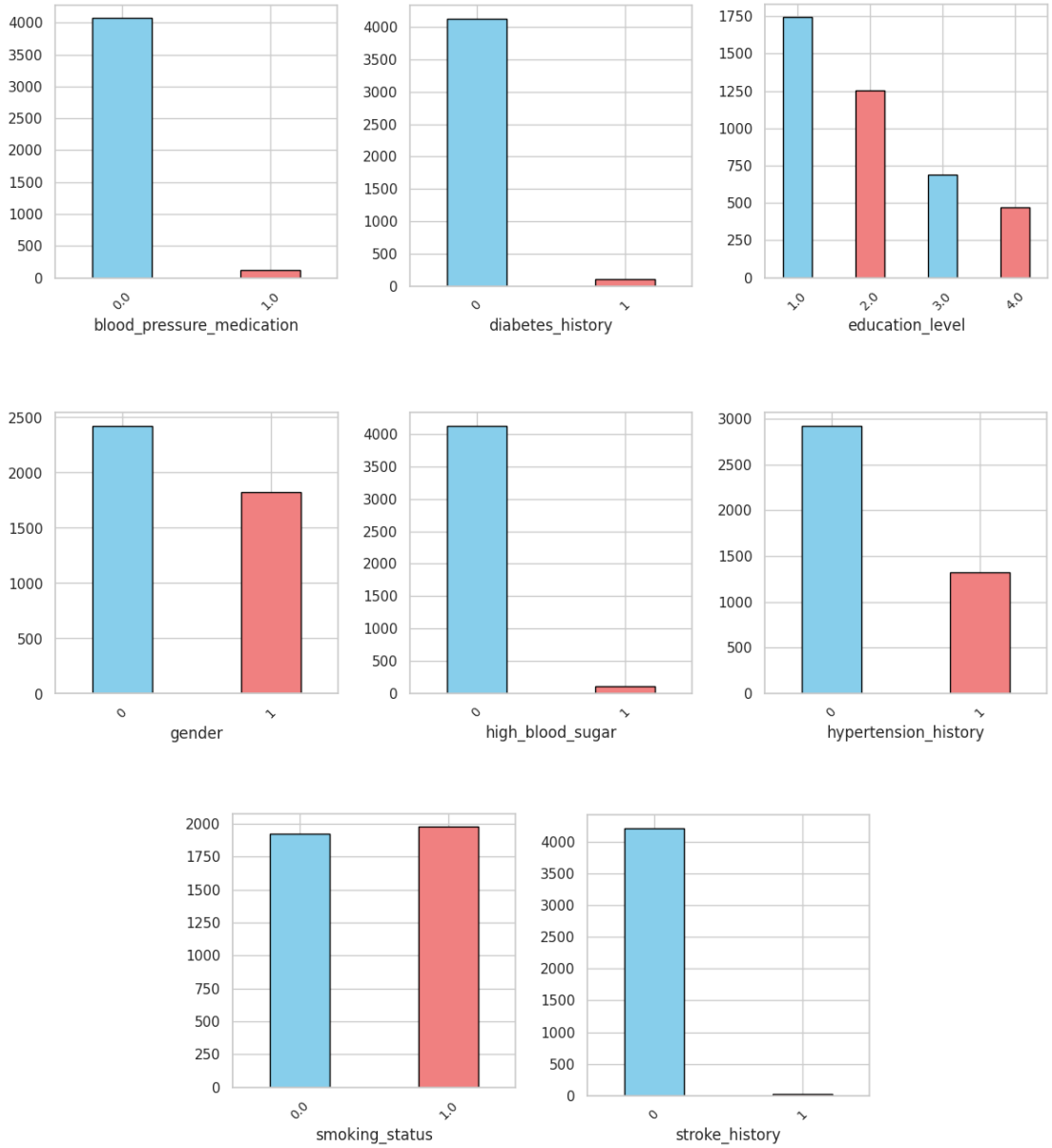
– Atributele cu valori discrete sunt: 'blood_pressure_medication', 'stroke_history', 'hypertension_history', 'smoking_status', 'diabetes_history', 'education_level', 'gender', 'high_blood_sugar'.

Heart Disease Discrete Variables

index	count	no_unique_values	value_counts
blood_pressure_medication	4196	2	{0.0: 4072, 1.0: 124}
stroke_history	4240	2	{0: 4215, 1: 25}
hypertension_history	4240	2	{0: 2923, 1: 1317}
smoking_status	3901	2	{1.0: 1978, 0.0: 1923}
diabetes_history	4240	2	{0: 4131, 1: 109}
education_level	4158	4	{1.0: 1743, 2.0: 1253, 3.0: 689, 4.0: 473}
gender	4240	2	{0: 2420, 1: 1820}
high_blood_sugar	4240	2	{0: 4131, 1: 109}

– Câteva observații pe care le putem face pe baza tabelului de mai sus:

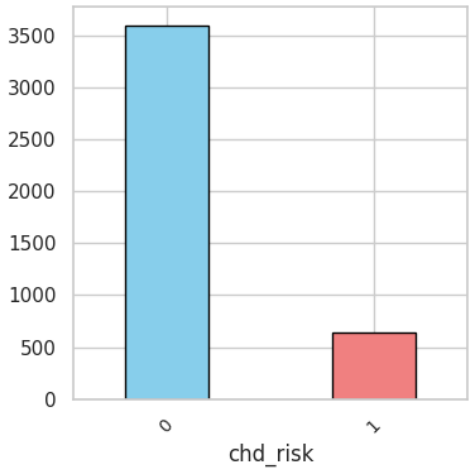
    * Și pentru anumite atribute discrete există valori există intrări lipsă.
    * Majoritatea atributelor studiate pentru acest dataset au valori binare, excepție făcând 'education_level'.
  - Repariția valorilor pentru aceste atribute cu valori discrete poate fi vizualizată cu ajutorul următoarelor grafice:



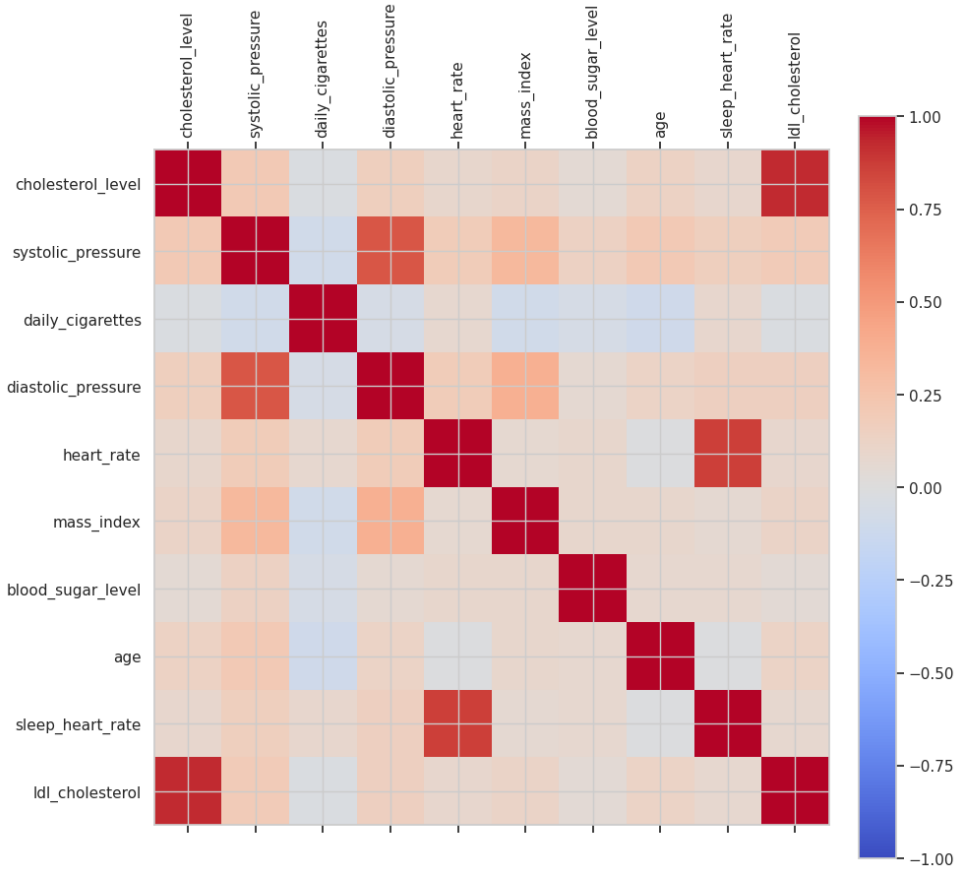
  - Pot fi trase următoarele concluzii:
    * Anumite caracteristici au o distribuție dezechilibrată de valori, cum ar fi 'stroke_history' sau 'diabetes_history', unde majoritatea covârșitoare a observațiilor sunt negative (nu au istoric). Această distribuție mult dezechilibrată poate duce la probleme de clasificare pentru valorile subreprezentate.
    * Atribute precum 'hypertension_history' au o distribuție mai echilibrată, ceea ce le poate face mai ușor de interpretat de către modele și mai robuste în clasificare.

- **Analiza echilibrului de clase**
  - Pentru setul de date referitor la bolile cardiovasculare, scopul este de a prezice incidența dezvoltării unei astfel de patologii în următorii 10 ani. Analizând distribuția valorilor pentru eticheta 'chd_risk', se observă și aici un dezechilibru între cele două valori, numărul de cazuri

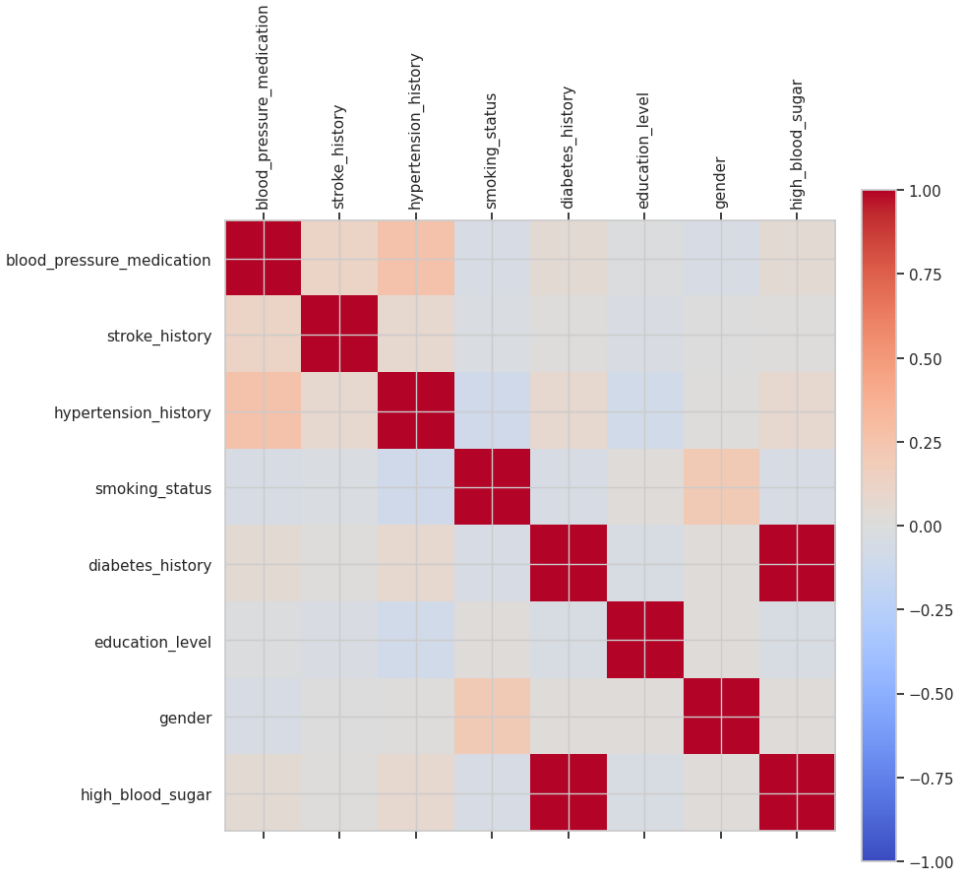
pozitive reprezentând aproximativ 15% din numărul total de valori. Apare riscul ca modelul antrenat să încline mai des înspre clasa majoritară.



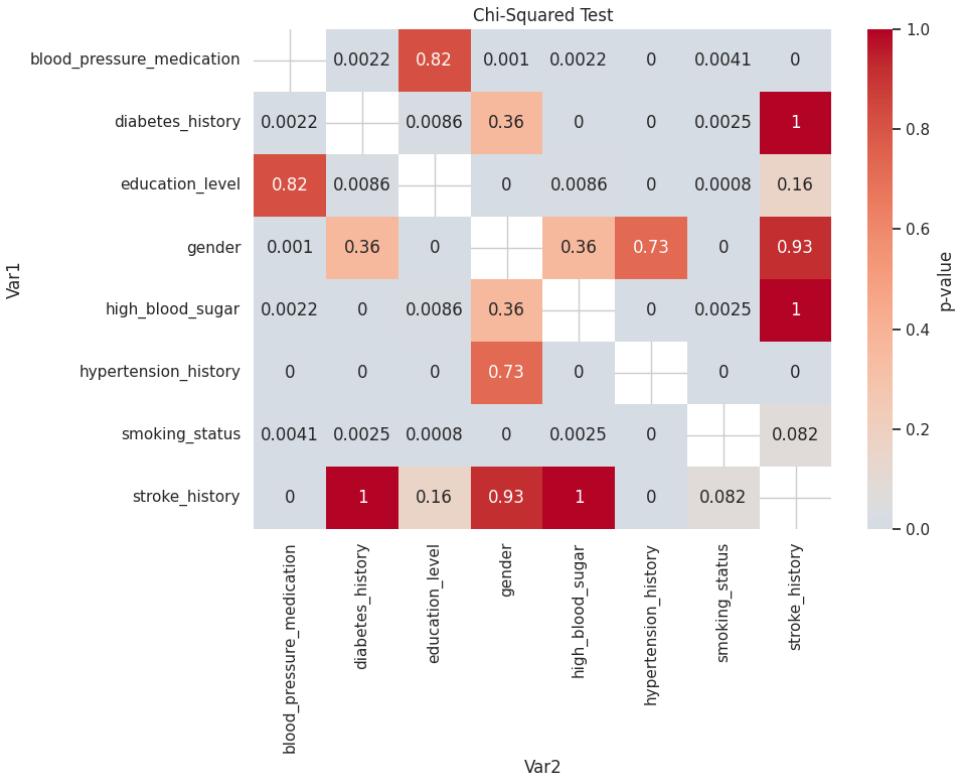
- **Analiza corelației între atribute**
  - Pentru caracteristicile continue, se observă corelații între următoarele perechi: ('sleep_heart_rate', 'heart_rate'), ('cholesterol_level','ldl_cholesterol'), ('systolic_pressure','diastolic_pressure').
  - Pe de altă parte, atributul 'daily_cigarettes' nu pare a fi direct corelat cu niciun alt atribut.



  - În cazul caracteristicilor discrete, se observă o corelație puternică între 'high_blood_sugar' și 'diabetes_history'



– Pentru testul static Chi-Pătrat efectuat asupra valorilor discrete, se obțin următoarele rezultate:



– Conform tabelului, perechide de atribute care au un p-value peste 0.05 sunt independente.

Pirvision Dataset

- **Analiza tipului de atribute și a plajei de valori a acestora**

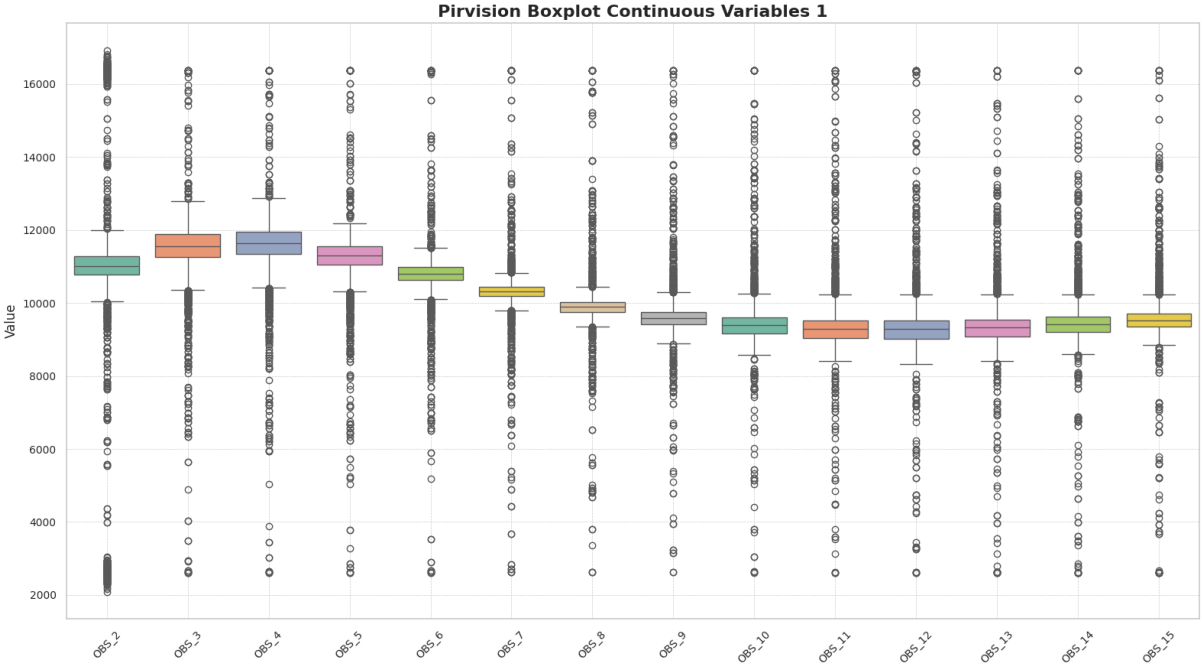
Pirvision Continuous Variables

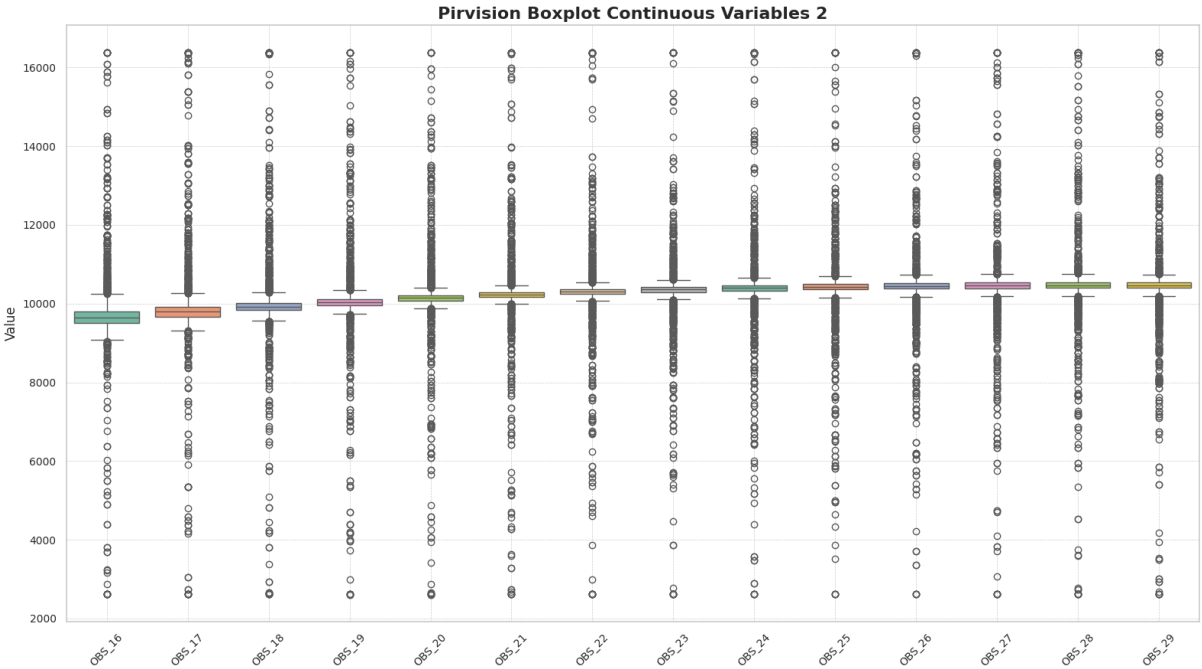
index	count	mean	std	min	25%	50%	75%	max
Temp (F)	10000.00000	80.39260	22.85846	0.00000	86.00000	86.00000	88.00000	89.00000
Temp (C)	10000.00000	26.69850	12.42325	-17.00000	30.00000	30.00000	31.00000	31.00000
OBS_1	9000.00000	293891.30611	4662186.35897	2613.00000	10335.00000	10433.00000	10564.25000	111602625.00000
OBS_2	10000.00000	10962.18660	1363.58794	2092.00000	10779.00000	11001.00000	11281.00000	16928.00000
OBS_3	10000.00000	11521.63370	669.18525	2614.00000	11273.00000	11560.00000	11889.00000	16383.00000
OBS_4	10000.00000	11583.09270	673.74107	2611.00000	11341.00000	11634.00000	11959.00000	16383.00000
OBS_5	10000.00000	11273.43060	617.61260	2612.00000	11056.00000	11303.00000	11553.00000	16383.00000
OBS_6	10000.00000	10799.28120	550.86507	2613.00000	10631.00000	10794.00000	10983.00000	16383.00000
OBS_7	10000.00000	10315.35770	515.67934	2618.00000	10185.00000	10312.00000	10443.00000	16383.00000
OBS_8	10000.00000	9909.64570	511.57918	2620.00000	9763.00000	9908.00000	10035.00000	16383.00000
OBS_9	10000.00000	9612.18290	529.90015	2621.00000	9408.00000	9594.00000	9762.00000	16383.00000
OBS_10	10000.00000	9423.81250	556.64488	2614.00000	9160.00000	9399.00000	9600.00000	16383.00000
OBS_11	10000.00000	9330.88990	577.17340	2611.00000	9041.00000	9299.00000	9524.00000	16383.00000
OBS_12	10000.00000	9313.99110	588.84145	2605.00000	9027.00000	9287.00000	9514.00000	16383.00000
OBS_13	10000.00000	9356.15500	572.80409	2602.00000	9086.00000	9329.00000	9548.00000	16383.00000
OBS_14	10000.00000	9443.04900	554.93783	2603.00000	9204.00000	9421.00000	9618.00000	16383.00000
OBS_15	10000.00000	9557.34860	546.06658	2611.00000	9354.00000	9528.00000	9707.00000	16383.00000
OBS_16	10000.00000	9683.59290	546.87557	2613.00000	9513.00000	9655.00000	9807.00000	16383.00000
OBS_17	10000.00000	9813.07220	547.92866	2614.00000	9674.00000	9792.00000	9913.00000	16383.00000
OBS_18	10000.00000	9936.63850	533.46778	2613.00000	9832.00000	9922.00000	10014.00000	16383.00000
OBS_19	10000.00000	10049.63910	507.31918	2612.00000	9965.00000	10040.00000	10119.00000	16383.00000
OBS_20	10000.00000	10146.63340	488.60989	2612.00000	10077.00000	10143.50000	10209.00000	16383.00000
OBS_21	10000.00000	10228.62070	490.50518	2613.00000	10172.00000	10232.00000	10291.00000	16383.00000
OBS_22	10000.00000	10296.55700	485.51089	2614.00000	10243.00000	10303.00000	10363.00000	16383.00000
OBS_23	10000.00000	10348.93260	481.94117	2614.00000	10294.00000	10358.00000	10420.00000	16383.00000
OBS_24	10000.00000	10391.50300	481.20215	2618.00000	10335.00000	10402.00000	10466.00000	16383.00000
OBS_25	10000.00000	10423.54890	463.08903	2619.00000	10365.00000	10431.00000	10501.00000	16383.00000
OBS_26	10000.00000	10444.62690	466.05026	2620.00000	10381.00000	10452.00000	10521.00000	16383.00000
OBS_27	10000.00000	10458.13890	482.16052	2616.00000	10393.00000	10464.00000	10535.00000	16383.00000
OBS_28	10000.00000	10463.24300	496.69354	2616.00000	10399.00000	10470.00000	10541.00000	16383.00000
OBS_29	10000.00000	10457.74030	501.87860	2616.00000	10396.00000	10470.00000	10535.00000	16383.00000
OBS_30	10000.00000	10449.53550	489.41928	2615.00000	10393.00000	10461.00000	10529.00000	16383.00000
OBS_31	10000.00000	10439.53660	478.12431	2614.00000	10383.00000	10453.00000	10519.00000	16383.00000
OBS_32	10000.00000	10427.43790	484.41934	2613.00000	10373.00000	10441.00000	10505.00000	16383.00000
OBS_33	10000.00000	10416.05280	499.88155	2614.00000	10365.00000	10430.00000	10494.00000	16383.00000
OBS_34	10000.00000	10404.56740	502.76852	2614.00000	10355.00000	10419.00000	10481.00000	16383.00000
OBS_35	10000.00000	10393.95240	498.32257	2612.00000	10343.00000	10407.00000	10470.00000	16383.00000
OBS_36	10000.00000	10381.69290	487.66884	2614.00000	10331.00000	10393.00000	10456.00000	16383.00000
OBS_37	10000.00000	10239.91840	489.49923	2609.00000	10153.00000	10265.50000	10353.00000	16383.00000
OBS_38	10000.00000	9642.69900	582.75269	2611.00000	9400.00000	9672.00000	9887.00000	16383.00000
OBS_39	10000.00000	8955.73640	680.06805	2612.00000	8562.00000	8929.00000	9254.00000	16383.00000
OBS_40	10000.00000	8712.99610	724.63694	2612.00000	8235.00000	8681.50000	9002.00000	16383.00000
OBS_41	10000.00000	8981.48840	673.17819	2612.00000	8615.00000	8941.00000	9253.25000	16383.00000
OBS_42	10000.00000	9556.98750	569.13805	2611.00000	9321.00000	9563.00000	9777.00000	16383.00000
OBS_43	10000.00000	10212.36730	501.12329	2611.00000	10047.00000	10215.00000	10381.00000	16383.00000
OBS_44	10000.00000	10790.04390	511.95681	2613.00000	10607.00000	10787.00000	10997.00000	16383.00000
OBS_45	10000.00000	11210.75980	558.19620	2610.00000	10994.00000	11240.00000	11504.00000	16383.00000
OBS_46	10000.00000	11462.78340	596.63551	2606.00000	11223.00000	11491.00000	11839.00000	16383.00000
OBS_47	10000.00000	11562.52970	620.51011	2606.00000	11315.00000	11588.00000	11962.00000	16383.00000
OBS_48	10000.00000	11544.71600	616.34430	2607.00000	11305.00000	11570.00000	11928.00000	16383.00000
OBS_49	10000.00000	11442.94070	597.22353	2613.00000	11214.00000	11478.00000	11783.00000	16383.00000
OBS_50	10000.00000	11286.81120	566.48574	2605.00000	11084.00000	11323.00000	11574.00000	16383.00000
OBS_51	10000.00000	11105.03350	538.63875	2609.00000	10939.00000	11137.00000	11336.00000	16383.00000
OBS_52	10000.00000	10920.67450	510.25748	2605.00000	10795.75000	10949.00000	11098.00000	16383.00000
OBS_53	10000.00000	10743.87790	479.47776	2604.00000	10650.00000	10761.50000	10877.00000	16383.00000
OBS_54	10000.00000	10585.93500	453.28252	2603.00000	10514.00000	10596.00000	10684.00000	16383.00000
OBS_55	10000.00000	10449.01570	437.93765	2602.00000	10388.00000	10453.00000	10523.00000	16383.00000
OBS_56	10000.00000	11014.59020	480.73706	2547.00000	10929.00000	11026.00000	11129.00000	17157.00000
OBS_57	10000.00000	10026.85050	420.56931	2530.00000	9957.00000	10033.00000	10109.00000	15713.00000

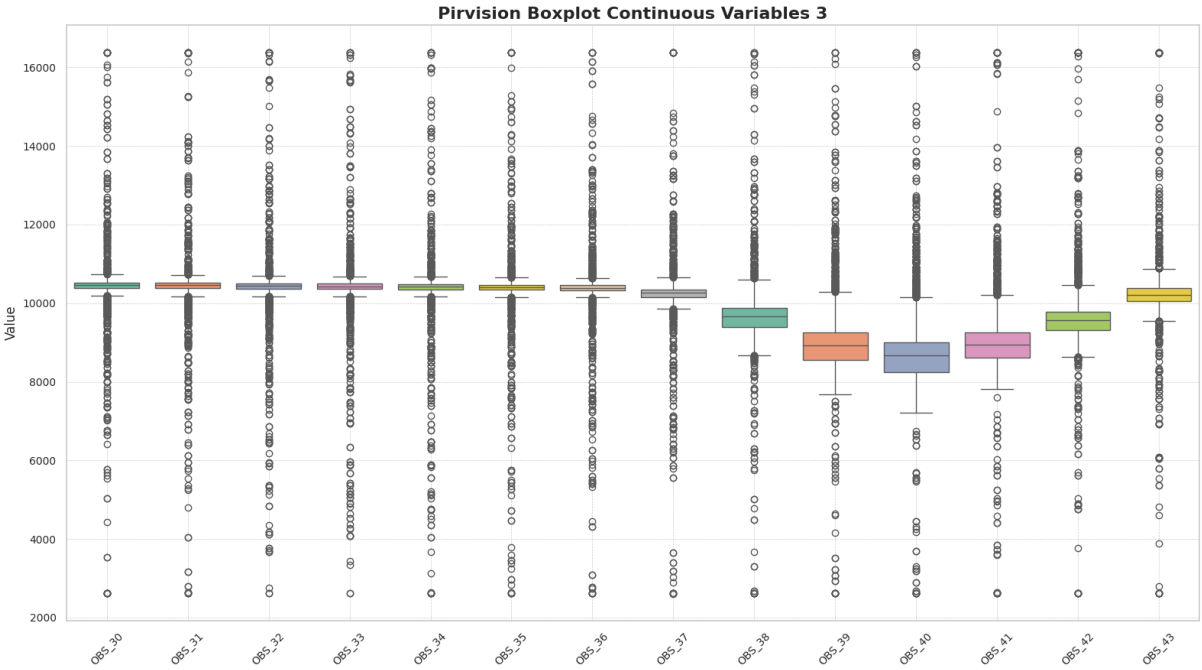
  - Observând valorile înregistrate pentru setul de date Pirvision, pot reieși următoarele concluzii:
    * Cu excepția 'OBS_1', care are 9000 de valori înregistrate, toate celelalte au tot atâtea valori câte înregistrări în setul de date.

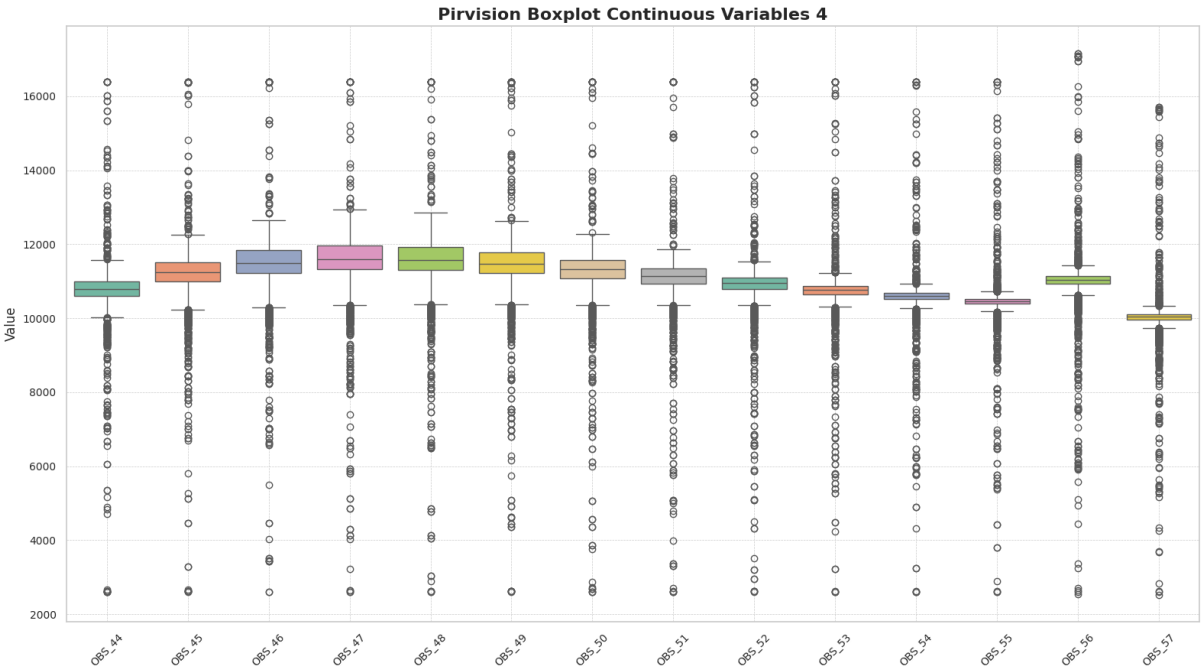
* 'OBS_2' - 'OBS_57' au valori apropiate ca rang de mărime, în timp ce la 'OBS_1' pare a se fi produs o anomalie.
* 'Temp (F)' și 'Temp (C)' au distribuții dezechilibrate, majoritatea valorilor fiind în intervalul 86-89, respetiv 30-31, lucru care poate indica o temperatură constantă în mediul înregistrat. Valorile minime sunt anormale, putând fii un indicator al unei valori rare sau a unei erori de înregistrare.

– Pentru a observa repartiția de valori și anomalii ale caracteristicilor am realizat o reprezentare de tip boxplot a valorilor.

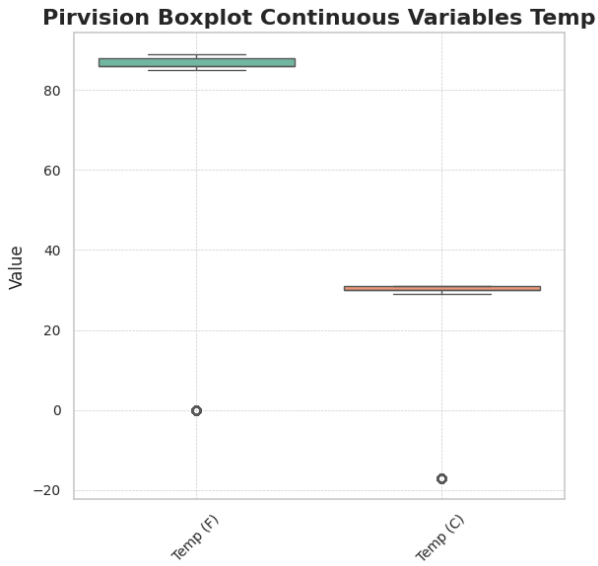






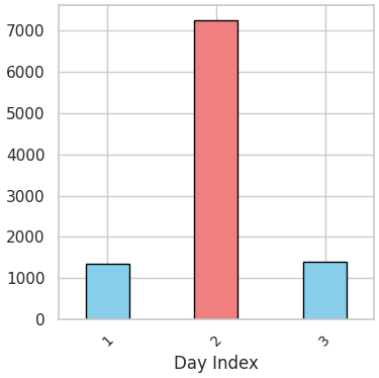


* Distribuția valorilor de tip OBS între percentilele 25% și 75% indică, în general, o variabilitate moderată, cu intervale relativ compacte în raport cu întreaga gamă de valori posibile. De exemplu, pentru caracteristici precum OBS_18 până la OBS_37, diferența dintre cele două percentile este de aproximativ 200 de unități, în timp ce pentru OBS_40, această diferență crește până la aproximativ 700 de unități. Această observație sugerează că majoritatea valorilor sunt concentrate într-un interval relativ restrâns.
* Totuși, există variații destul de mari pentru valorile în afara acestui interval, observându-se minime de aproximativ 2.000 de unități și maxime de 16.000 de unități.



∗ În ceea ce privește repartiția pentru caracteristicile de temperatură, se păstrează observația anterioară, conform cărora valorile acestora se află predominant într-un interval compact, cu existența unor valori rare.

– În cadrul acestui set de date, am identificat o singură variabilă discretă, anume DayIndex, care poate lua trei valori distincte, majoritatea intrărilor având valoarea 2 pentru această caracteristică.

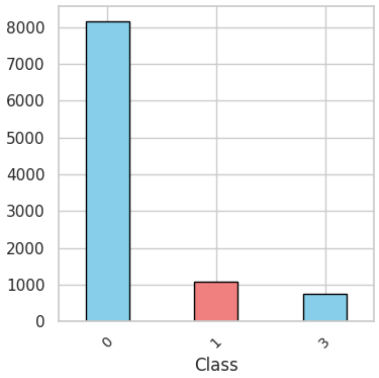


Pirvision Discrete Variables

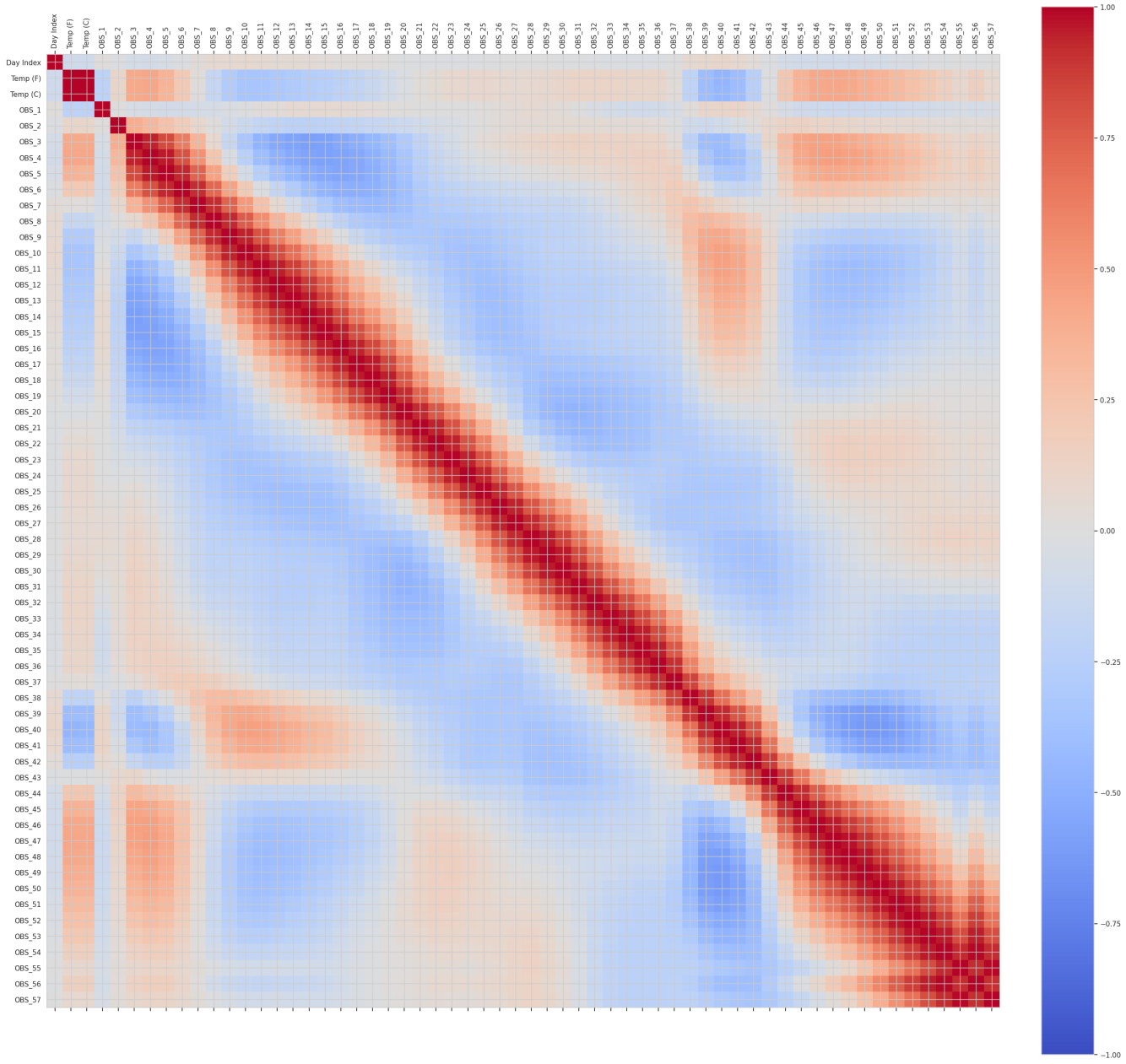
index	count	no_unique_values	value_counts
Day Index	10000	3	{2: 7251, 3: 1402, 1: 1347}

- **Analiza echilibrului de clase**

- Se observă că, în cazul etichetei pentru fiecare intrare, avem un număr considerabil de înregistrări care fac parte din clasa 0, aproximativ de 8 ori mai multe decât fiecare din celelalte clase. Așadar, apare riscul ca modelul antrenat să aibă performanțe mai proaste pentru aceste etichete, neavând același număr de intrări pentru învățare.



- **Analiza corelației între atribute**



3 Preprocesarea Datelor

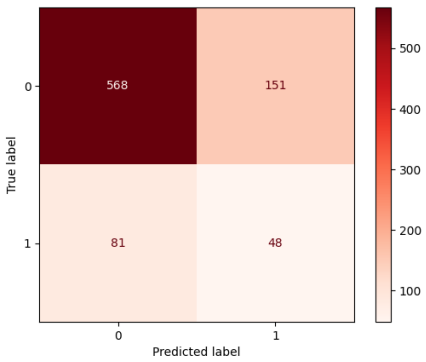
Pentru preprocesarea datelor, am fost realizate următoarele operații:

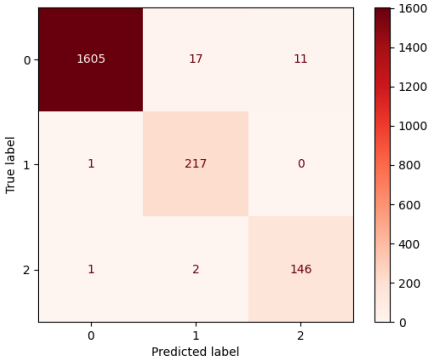
- Eliminarea valorilor outlier pentru date folosind algoritmult IQR(cu quantila1 = 25% și quantila3 =75% pentru setul de date cu riscul de boală coroidară, respectiv 10% și 90% pentru setul de date pirvision)
- Imputarea valorilor lipsă — pentru datele despre boala coroidară am folosit metoda 'most_frequent' (completăm cu valoarea cea mai comună), deoarece multe atribute erau categorice. Pentru datele pirvision am folosit 'median', care este mai robustă la outlieri și potrivită pentru date numerice.
- Eliminarea atributelor puternic corelate(corelație ¿ 0.9) — pentru coloanele care conțin cam aceeași informație.
- Scalarea valorilor — pentru datele despre boala coroidară am folosit scalarea 'robust'. Pentru datele pirvision am folosit 'standard'.

4 Utilizarea algoritmilor de ML

Arbori de decizie

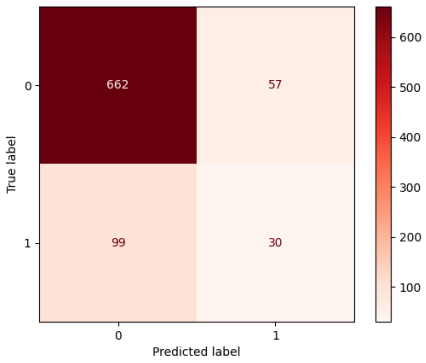
- Atribute max_depth=20, min_samples_leaf=5, criterion='gini'(funcția de impuritate definită în cazul spliturilor), class_weight adaugat(în ambele cazuri exista cel puțin o clasă cu o pondere covârșitoare)
- Matrice de confuzie

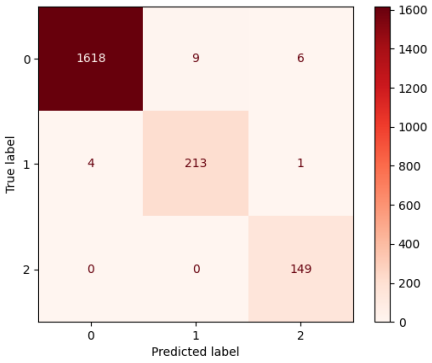




Păduri aleatoare

- Atribute heart_dataset n_estimators=155, max_depth=15,min_samples_leaf=3,criterion='gini', max_samples=0.7,ma class_weight adaugat
- Matrice de confuzie



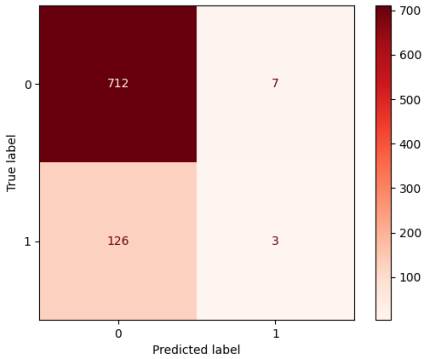


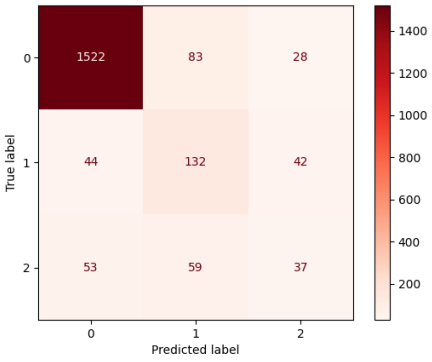
Regresie logistică

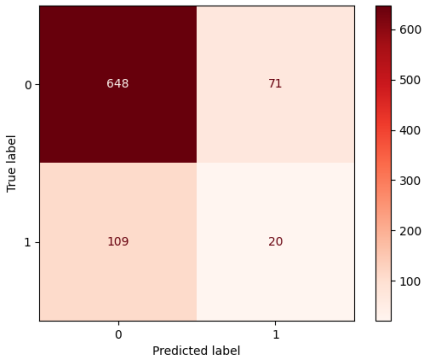
- Atributele categorice au fost codificate cu OneHotEncoder, iar variabila țintă cu LabelEncoder.
- Modelul de regresie logistică a fost antrenat folosind Gradient Descent cu rata de învățare 0.01 și 1000 de epoci.
- Nu a fost aplicată regularizare în această implementare de bază.

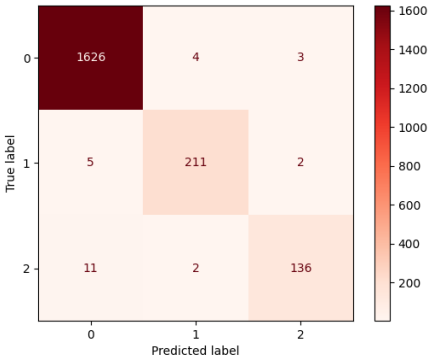
MLP

- Atribute folosite:hidden_layer_sizes=(128, 64), activation='relu', solver='adam', learning_rate_init=0.001, max_iter=5000, batch_size=128, alpha=0.001.









Pentru setul de date pirvision

	accuracy	macro_precision	macro_recall	macro_f1	weighted_precision	weighted_recall	weighted_f1
decision_tree	0.984000	0.949000	0.986000	0.967000	0.985000	0.984000	0.984000
random_forest	0.990000	0.971000	0.989000	0.980000	0.990000	0.990000	0.990000
logistic_regression	0.823000	0.536000	0.555000	0.535000	0.833000	0.823000	0.825000
mlp	0.986000	0.976000	0.959000	0.967000	0.986000	0.986000	0.986000

Pentru setul de date referitor la bolile coroidare:

	accuracy	macro_precision	macro_recall	macro_f1	weighted_precision	weighted_recall	weighted_f1
decision_tree	0.726000	0.558000	0.581000	0.562000	0.779000	0.726000	0.749000
random_forest	0.816000	0.607000	0.577000	0.586000	0.790000	0.816000	0.801000
logistic_regression	0.796000	0.532000	0.520000	0.520000	0.756000	0.796000	0.773000
mlp	0.788000	0.538000	0.528000	0.530000	0.759000	0.788000	0.772000

5 Concluzii

- Pentru acest set de date, toate modelele au obținut scoruri bune. Cel mai bun rezultat l-a avut Random Forest, cu o acuratețe de 0.990, urmat de MLP și Decision Tree. Regresia logistică a avut rezultate mai slabe, cu o acuratețe de 0.823. Metricele macro și weighted confirmă că Random Forest și MLP se descurcă bine pentru toate clasele, nu doar cele frecvente.

- Pe acest set, scorurile sunt mai mici. Random Forest are din nou cele mai bune rezultate (0.816), iar Decision Tree cele mai slabe (0.726). Diferențele dintre metricele macro și weighted arată că modelele se descurcă mai bine pe clasele frecvente și mai slab pe cele rare. Analizînd matricele de confuzie pentru acest set de date, se observă o predispoziție a modelelor de a clasifica o intrare ca făcând parte din clasa majoritară. În ciuda faptului ca acuratețea este ridicată, modelele se descurcă prost la a identifica intrări din clasa nedominantă, având un scor mic per macro.

- Implementarea proprie de Logistic Regression nu este foarte eficientă :¡