Predicția unor diagnostice medicale folosind clasificatori bazați pe arbori și de tip SVM

Radu Mihai Breazu

20 aprilie 2024

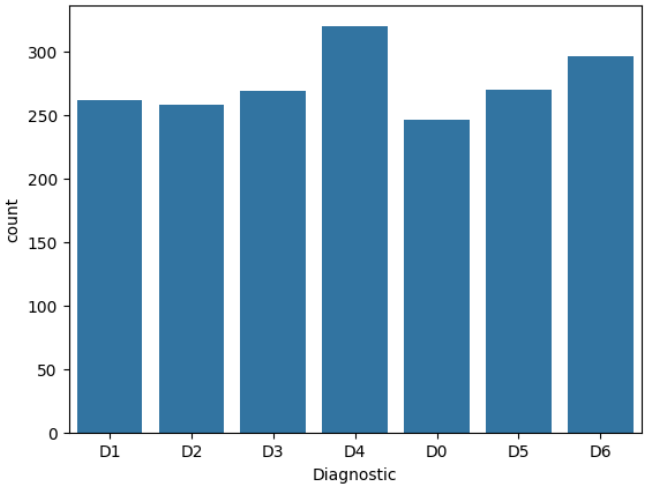
În acest document sunt prezentate rezultatele obținute în urma antrenării și a testării unor clasificatori pe un set de date ce conține informații medicale despre pacienți, în vederea prezicerii diagnosticului unui pacient pe baza unui set de atribute înregistrate în setul de date.

1 Setul de date

Setul de date conține informații despre un număr de 1921 de pacienți, cărora li s-au măsurat (și înregistrat) 18 atribute, între care: tipul de transport folosit, cât de des consumă alcool, dacă este fumător sau nu, nivelul activității fizice, înălțimea, greutatea, vârsta și sexul. De asemenea, s-a reținut, pentru fiecare pacient, diagnosticul ce i-a fost atribuit, sub forma unei reprezentări codificate, astfel încât diagnosticele prezente în setul de date sunt desemnate drept D0, D1, D2, D3, D4, D5 și D6.

1.1 Analiza echilibrului de clase

În acest set de date, clasele sunt date de diagnosticele pacienților. Astfel, am numărat câte persoane au avut fiecare diagnostic, și am construit o histogramă cu rezultatele astfel obținute. Rezultatul este prezentat mai jos:



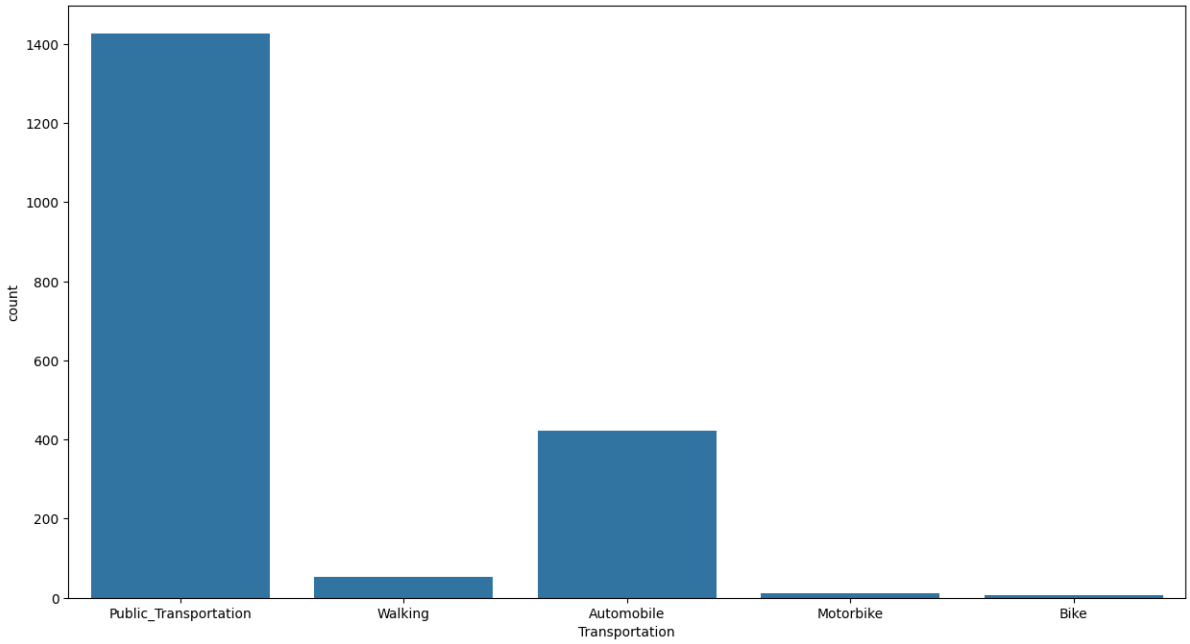
După cum se observă din diagramă, distribuția claselor este aproape uniformă: diagnosticele D4 și D6 sunt cele mai frecvente, pacienții cu D4 fiind în număr de 320. Numărul relativ mare de pacienți cu D4 ”trage în sus” atât media, cât și abaterea standard la nivelul întregii distribuții.

1.2 Analiza atributelor

Atributele categorice

Pentru fiecare atribut categoric, am construit o histogramă, astfel încât să poată fi vizualizat mai ușor numărul de persoane care prezintă o anumită valoare a acelui atribut. Rezultatele sunt prezentate în continuare.

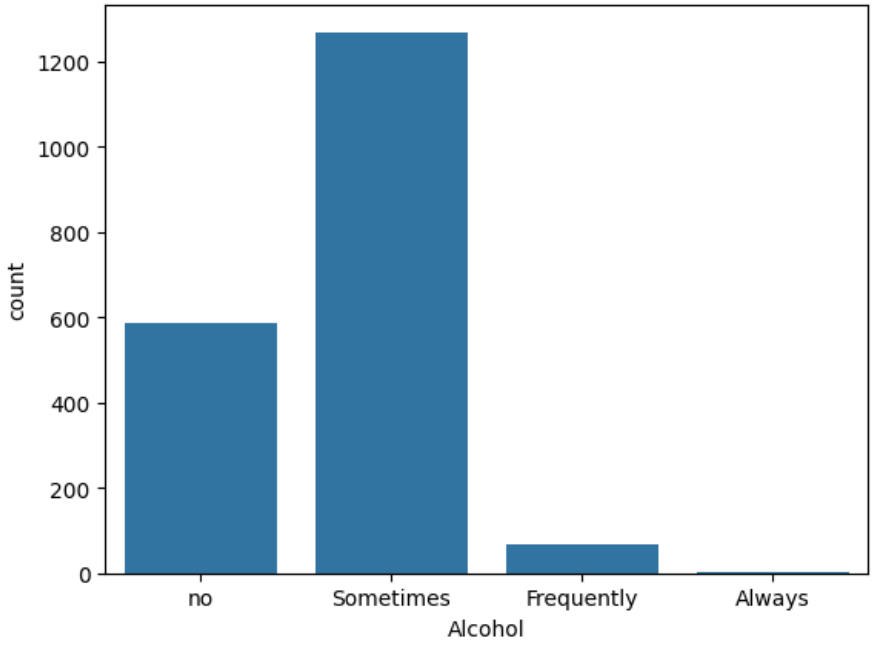
Tipul de transport



Din diagramă reiese că transportul în comun este metoda preferată de deplasare a persoanelor selectate pentru studiu. Mai exact, 1427 de oameni preferă să folosească în mod regulat această formă de transport, comparativ cu automobilul, care este folosit de 423 de oameni. De asemenea, doar 53 de persoane preferă să meargă pe jos, iar 7 persoane se deplasează în mod obișnuit cu bicicleta. După părerea mea, acest atribut este un slab predictor (per total) al diagnosticului unui pacient (a se vedea analiza de covarianță)[1], din cauză că procentul de oameni care folosesc transportul public este foarte mare (comparativ cu cele ale celorlalte forme de transport).

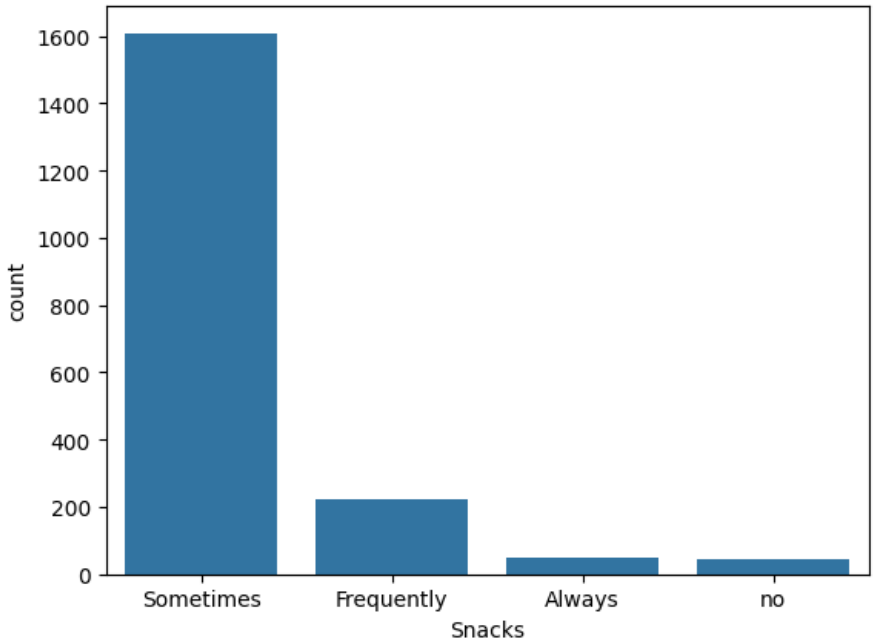
[1] Toate inferențele făcute în această secțiune referitor la gradul de corelare între atribute și clasă sunt făcute în mod intuitiv, și nu formal, deoarece se cunoaște (din statistică) faptul că două variabile aleatoare pot avea coeficientul de corelație 0 și să nu fie independente.

Consumul de alcool



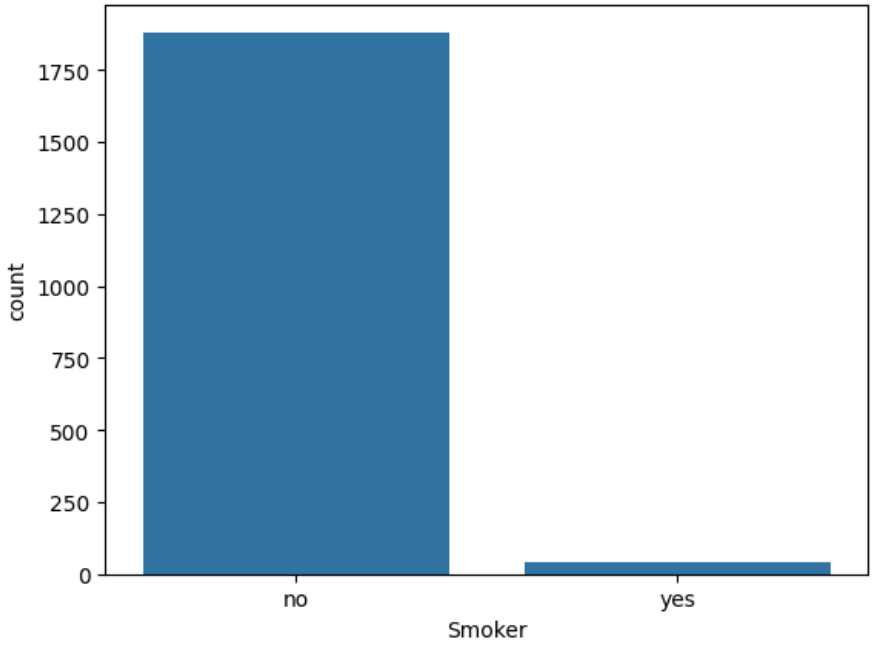
Din diagramă rezultă că deși majoritatea persoanelor analizate (1269 din 1921) consumă uneori alcool, un număr semnificativ dintre acestea (585) nu consumă deloc alcool, fapt care explică puterea mai mare de prezicere a diagnosticului pe baza acestui atribut.

Consumul de snacks-uri



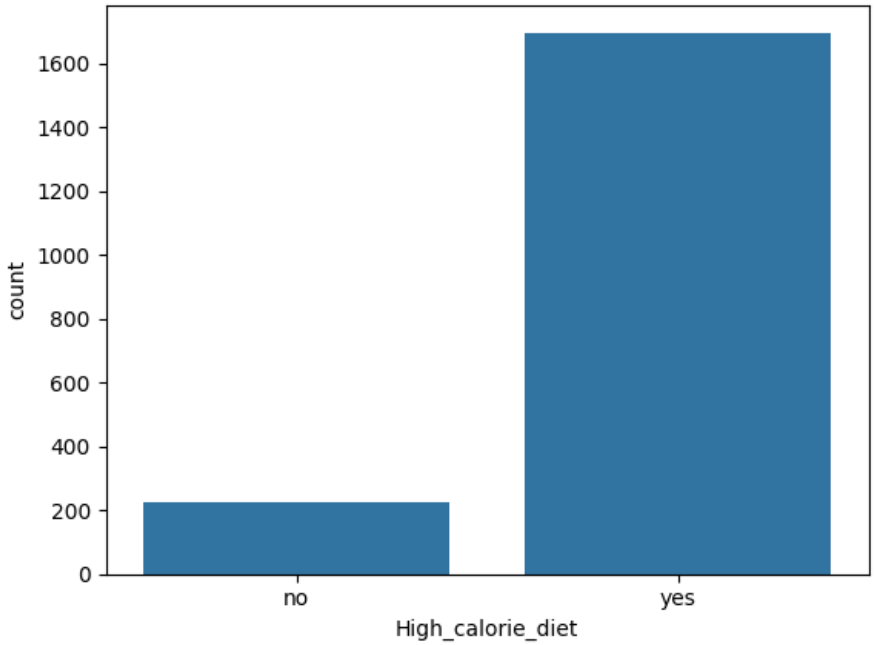
Se observă că distribuția este de tip exponențial. Mai exact, majoritatea pacienților (1609, mai concret) consumă ocazional snacks-uri. De asemenea, aproximativ 200 de pacienți (221, mai precis) consumă frecvent snacks-uri, iar aproximativ 50 de persoane consumă snacks-uri foarte frecvent, respectiv nu consumă deloc astfel de gustări. Cu toate acestea, coeficientul de corelație a consumului de snacks-uri cu diagnosticul este de -0.3, ceea ce duce la concluzia că acest atribut are o oarecare putere de prezicere a diagnosticului.

Fumatul



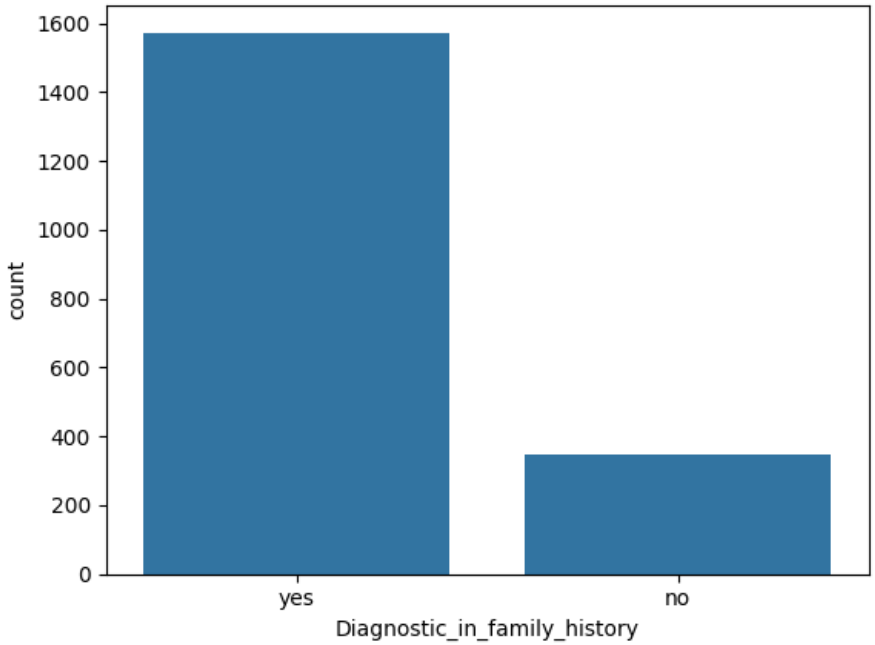
Încă o dată, distribuția este foarte debalansată: aproximativ 1875 de pacienți (adică aproape toți) sunt nefumători. Acest lucru, combinat cu faptul că nu a fost înregistrată frecvența fumatului, cauzează, după părerea mea, puterea predictivă scăzută a acestui atribut.

Dieta bogată în calorii



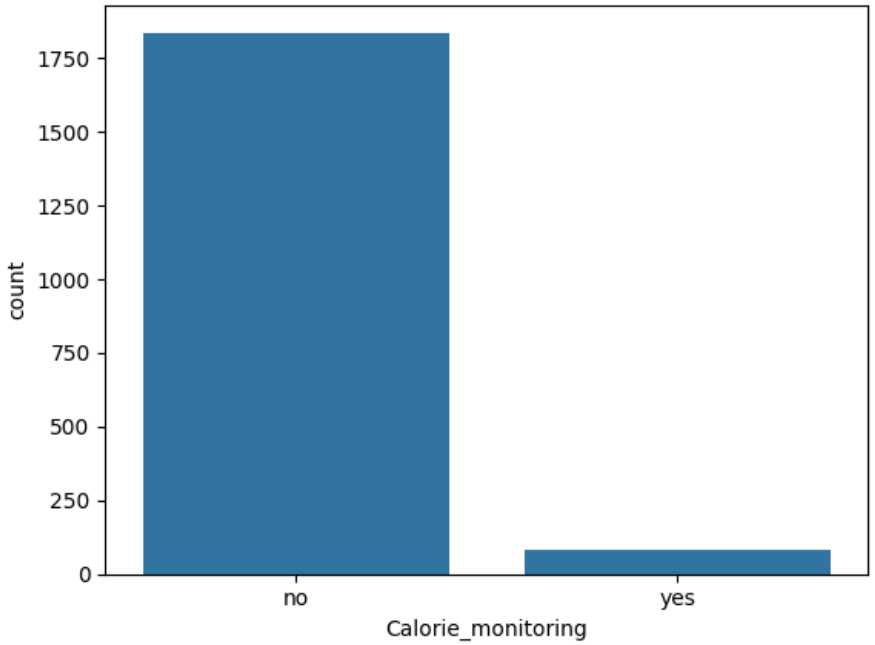
De această dată, distribuția este simetrică uneia exponențiale: aproximativ 1700 de pacienți (mai exact, 1697 de persoane) au o dietă bogată în calorii. Totuși, atributul reușește să prezică într-o anumită măsură diagnosticul unui pacient, întrucât coeficientul de corelație este 0.25.

Istoricul familial



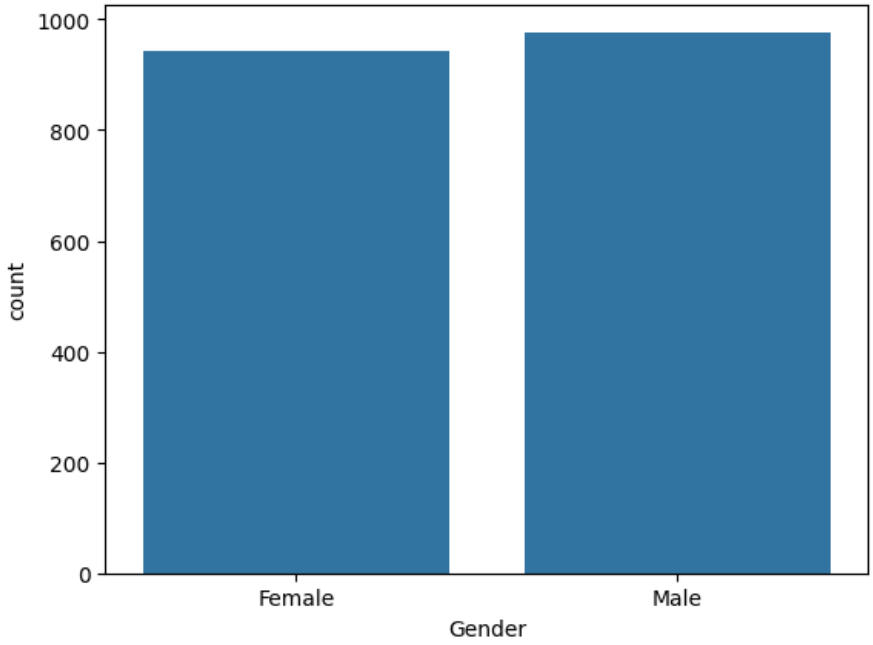
Distribuția este de tip exponențial: aproape 1600 de persoane (1573, mai exact) au un istoric familial ce le predispune la a dezvolta afecțiunea pe care o manifestă. Istoricul medical este corelat într-o măsură semnificativă cu diagnosticul primit de pacient (coeficient de corelație: 0.5), ceea ce conferă atributului o putere crescută de predicție.

Monitorizarea caloriilor



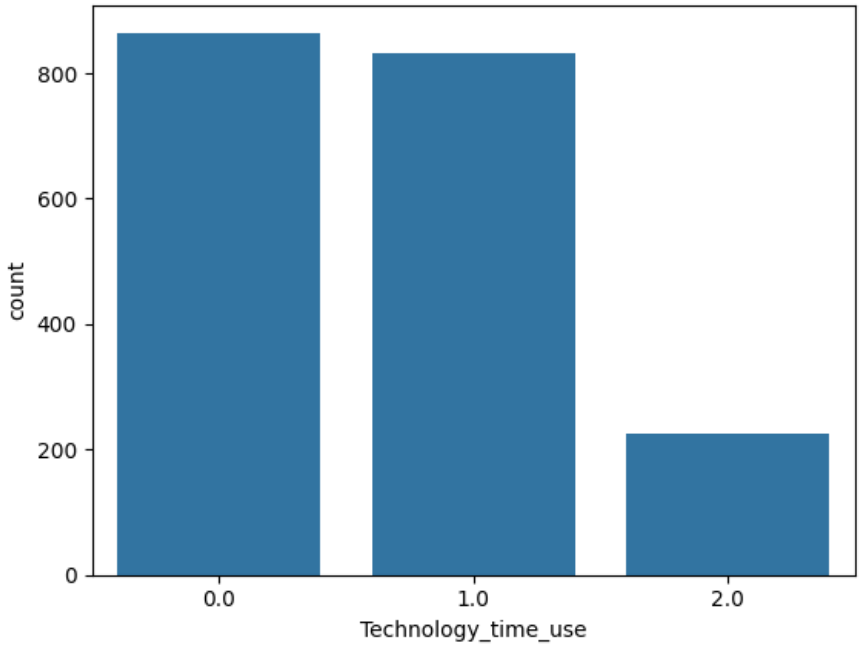
Încă o dată, distribuția este dezechilibrată: în jur de 1800 de pacienți (1838, mai precis) se aflau într-un program de monitorizare a caloriilor, iar restul (83) nu se aflau într-un astfel de program. Coeficientul de corelație cu diagnosticul este de -0.19, ceea ce arată că prezența unui astfel de program este un slab predictor al diagnosticului.

Sexul



De această dată, distribuția este echilibrată: la studiu au participat 977 de bărbați și 944 de femei. Coeficientul de corelație cu diagnosticul este de -0.042, lucru care sugerează o putere predictivă aproape inexistentă a acestui atribut. La nivel intuitiv, acest lucru mai sugerează faptul că niciuna dintre bolile a căror prezicere se încearcă nu afectează preponderent unul dintre sexe.

Timpul de utilizare a tehnologiei



Se observă că procentul celor care au scorul 0 și cel al celor care au scorul 1 în dreptul acestui atribut sunt aproape egale: sunt 865 de persoane cu scorul 0, 831 cu scorul 1, și 224 cu scorul 2. Coeficientul de corelație cu diagnosticul este -0.064, ceea ce arată puterea predictivă aproape inexistentă a acestui atribut.

Atributele numerice

În urma analizei exploratorii a datelor am observat că setul de date conține valori lipsă (de exemplu valori de -1 pe coloana ”Greutate”) sau chiar valori aberante (vârste de ordinul zecilor de mii), drept care, pentru a putea extrage valorile statisticilor uzuale, am filtrat aceste valori și le-am înlocuit cu NaN, astfel încât ele să fie ignorate în momentul calculării statisticilor. Valorile astfel obținute ale statisticilor[2] sunt redate în următorul tabel[3]:

Statistică	Atribut								
	Fibre	Sedentar	Vârstă	Calorii	Mese/zi	Înălțime	Hidratare	Greutate	Sport
Medie	2.4204	3.1973	24.3535	2253.6877	2.6835	1.7023	2.0104	86.8133	1.0126
std	0.5332	0.5758	6.4092	434.0758	0.7792	0.0933	0.611	26.2413	0.8555
min	1.00	2.21	15.00	1500.00	1.00	1.45	1.00	39.00	0.00
max	3.00	4.67	61.00	3000.00	4.00	1.98	3.00	165.0573	3.00
25%	2.00	2.77	19.9671	1871.00	2.6586	1.63	1.6061	65.8152	0.116
50%	2.3865	3.13	22.8296	2253.00	3.00	1.70	2.00	83.32	1.00
75%	3.00	3.64	26.00	2628.00	3.00	1.77	2.4806	108.0143	1.6835
MAD	0.4782	0.4708	4.8129	375.3623	0.5954	0.0771	0.4708	21.8847	0.7022
MedAD	0.4204	0.4327	3.8008	380.3123	0.3165	0.0677	0.4419	21.2026	0.8227
IQR	1.00	0.87	6.0329	757.00	0.3414	0.14	0.8745	42.1991	1.5675

Tabelul 1: Statistici referitoare la distribuția atributelor numerice

Din tabel reiese că persoanele analizate sunt în cea mai mare parte tinere (cuantila 0.75 este egală cu 26, ceea ce înseamnă că 75% dintre pacienți au cel mult 26 de ani), consumă multe fibre (cuantila 0.25 este egală cu 2, de unde rezultă că 75% dintre persoane consumă fibre la cel puțin 2 mese pe zi) și se hidratează corespunzător (mediana consumului zilnic de apă este de 2 l, iar cuantilele 0.25 și 0.75 sunt aproximativ egale cu 1.61 l, respectiv 2.48 l). Totodată, nivelele sedentarismului sunt scăzute: mediana este de 3.13 ore/zi, în timp ce cuantila 0.75 este de 3.64 ore/zi (ceea ce înseamnă că 50% dintre pacienți stau jos, în medie, cel mult 3.13 ore/zi, și 75% dintre persoane stau jos, în medie, sub 3.64 ore într-o zi).

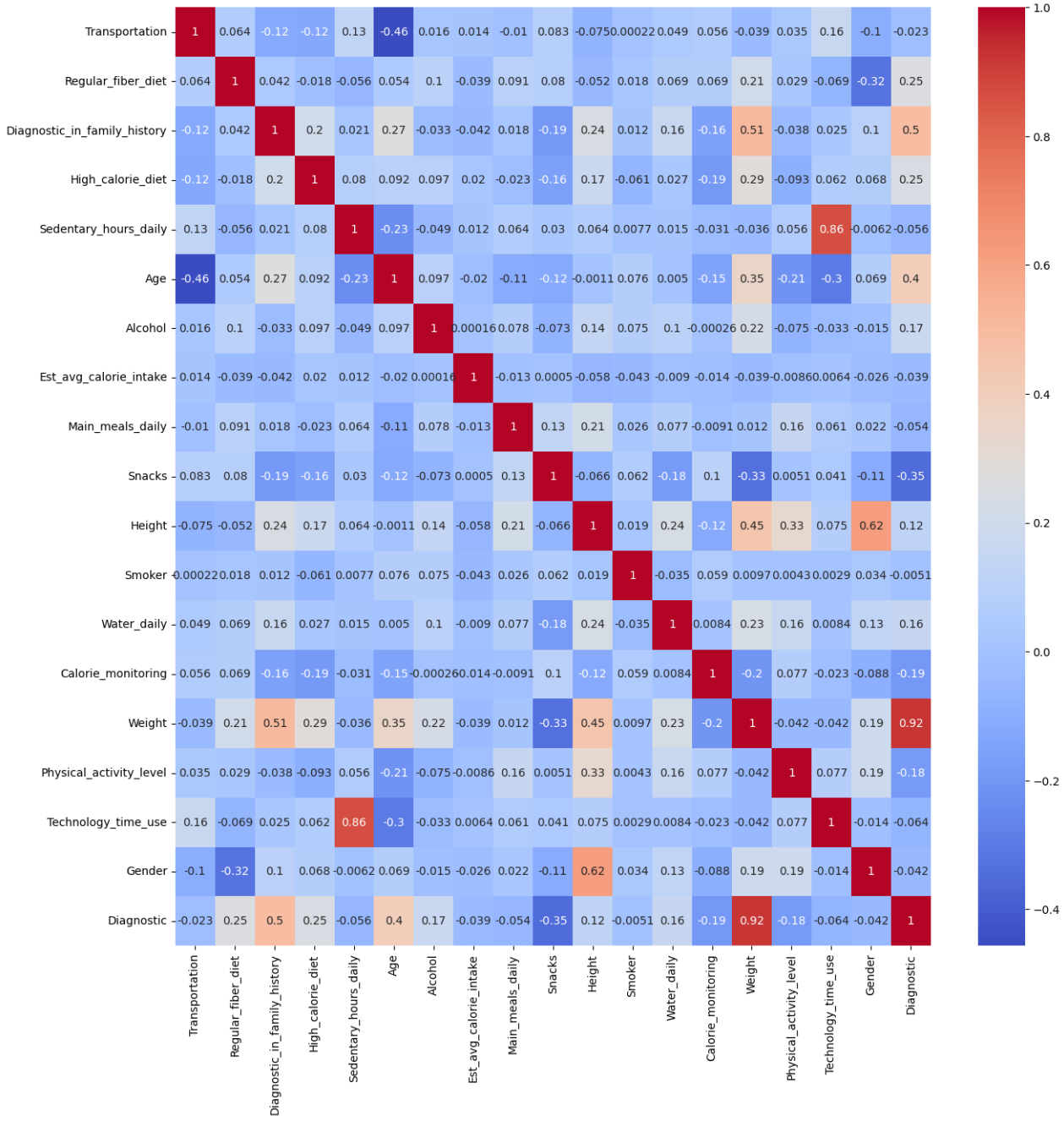
În schimb, pacienții au o greutate mediană de aproximativ 83.32 kg, iar cuantila 0.75 este de aproximativ 108.01 kg, ceea ce indică o masă corporală mare, iar nivelul activității fizice are mediana egală cu 1 oră/săptămână și cuantila 0.25 egală cu 0.116 ore/săptămână, ceea ce indică o slabă antrenare a corpului.

1.3 Analiza de covarianță

După convertirea atributelor categorice la numerice cu valori discrete, am calculat (și am afișat) matricea de covarianță atât între atribute, cât și între atribute și clasă. Pentu calculul matricei am folosit metoda Spearman de evaluare a coeficientului de corelație (coeficientul Pearson detectează numai corelațiile liniare, pe când cel Spearman detectează orice fel de corelație monotonă – liniară sau neliniară – între variabilele analizate). Rezultatul obținut este imaginea de mai jos:

[2]Abrevierile folosite sunt aceleași cu cele care apar în mesajul întors de funcția `describe()` din `pandas`, în afară de următoarele: MAD – abaterea medie absolută, MedAD – abaterea mediană absolută

[3]Valorile sunt rotunjite la a patra zecimală



Din această analiză se desprind un număr de concluzii (pe lângă cele deja menționate în dreptul fiecărui atribut categoric), dintre care câteva sunt prezentate mai jos:

– Există o corelație foarte puternică între greutate și diagnosticul primit de un pacient (coeficient de corelație: 0.92). Acest fapt sugerează că pacienții cu o greutate peste cea medie (a întregii populații) tind să dezvolte boli precum cele indexate cu un număr mare (mai apropiat de 6) în lista de diagnostice din setul de date furnizat.

– Există o corelație foarte puternică (și pozitivă) între timpul de utilizare a tehnologiei și sedentarism (coeficient de corelație: 0.86). Corelația poate fi explicată intuitiv prin faptul că dispozitive precum computerele și televizoarele, pentru a fi folosite, impun o poziție staționară.

– Există o corelație puternică între înălțime și sex (coeficient de corelație: 0.62). Acest

lucru poate fi explicat intuitiv prin aceea că bărbații tind să fie mai înalți decât femeile. Sensul corelației (pozitivă, și nu negativă) este o consecință directă a indexării în vederea conversiei la valori numerice: bărbații au indexul 1, pe când femeile au indexul 0.

– Mijlocul de transport folosit și vârsta sunt corelate negativ cu un coeficient de 0.46 (în modul). Sensul negativ al corelației este dat de modul în care este făcută indexarea mijloacelor de transport, și arată faptul că pacienții mai în vârstă tind să prefere automobilul și transportul în comun, în detrimentul mersului pe jos și al bicicletei.

2 Normalizarea datelor și selecția atributelor

2.1 Normalizarea datelor

Analiza atributelor numerice a revelat faptul că valorile atributului ”Calorii” (reprezentând numărul de calorii ingerate în medie pe zi) sunt de ordinul miilor, pe când celelalte atribute (cu excepția greutății) au valori de ordinul unităților. Astfel, am normalizat valorile de pe această coloană prin împărțire la 1000, astfel încât dispersia datelor să nu se modifice, dar valorile lor să fie de ordinul unităților. Nu am normalizat atributul ”Greutate” deoarece el este puternic corelat cu diagnosticul, fapt care sugerează o putere predictivă foarte mare, astfel că valorile de ordinul zecilor din coloana asociată reflectă acest lucru.

De asemenea, pentru a spori performanțele clasificatorilor antrenați, am înlocuit valorile NaN din pașii anteriori cu valori valide, folosind un `IterativeImputer`, cu strategia inițială de înlocuire fiind cea cu media valorilor existente, pe care l-am lăsat să ruleze maximum 20 de iterații.

2.2 Selecția atributelor

După cum rezultă din analiza de covarianță, nu toate atributele au aceeași putere de predicție a clasei. Astfel, pentru a reduce dimensionalitatea spațiului de lucru, am păstrat numai primele 80% dintre atribute, ordonate după puterea predictivă. Concret, acest pas a dus la eliminarea a patru coloane din setul de date, și anume: ”Sedentary_hours_daily” (sedentarismul), ”Est_avg_calorie_intake” (numărul de calorii ingerate într-o zi), ”Technology_time_use” (timpul de utilizare a tehnologiei) și ”Smoker” (fumător sau nefumător). Într-adevăr, acestea sunt unele dintre atributele cu cei mai mici (în modul) coeficienți de corelație cu diagnosticul, dar nu au *cei mai mici* patru astfel de coeficienți: în valoare absolută, atributul ”Transportation” are coeficientul de corelație egal cu 0.023, pe când ”Sedentary_hours_daily” și ”Smoker” au 0.056, respectiv 0.0051, iar ”Transportation” nu a fost eliminat din lista de atribute (pe când celelalte coloane au fost filtrate).

3 Antrenarea și testarea clasificatorilor

3.1 Clasificatorii folosiți și hiperparametrii folosiți

Pentru a prezice diagnosticele pacienților, am folosit patru clasificatori: unul bazat pe SVM (varianta multi-clasă), unul de tip pădure aleatoare clasică (`Random Forest`), unul de tip pădure aleatoare bazată pe arbori extrem de aleatori (`Extremely Random Trees -- Extra Trees`) și unul de tip `GradientBoosted Trees`. Întrucât combinația de hiperparametri folosită are un efect semnificativ asupra performanțelor clasificatorilor (a se vedea paragrafele următoare), am folosit metoda Grid Search cu Cross Validation pentru a căuta cea mai

bună combinație de hiperparametri pentru fiecare clasificator. Programatic, am folosit un `GridSearchCV` pentru fiecare clasificator, cu un număr de fold-uri egal cu 5 pentru validarea încrucișată, și am căutat cea mai bună combinație de hiperparametri dintr-un grid pe care l-am definit.

3.2 Performanțele clasificatorilor

Am evaluat performanțele clasificatorilor (folosind, în fiecare caz, combinația optimă de hiperparametri) pe baza a patru metrici: acuratețea, precizia, recall-ul, și scorul F1. Dintre acestea, acuratețea a fost calculată la nivelul tuturor claselor, iar celelalte metrici au fost calculate pentru fiecare clasă. Astfel, în tabelele de mai jos, sunt raportate doar media și deviația standard (ambele exprimate în procente) pentru precizie, pentru recall și pentru scorul F1, iar pentru acuratețe, este raportată valoarea sa pentru toate clasele. Toate numerele sunt rotunjite prin trunchiere la 2 zecimale.

Setul redus de atribute
Pentru ambele seturi de date, am căutat cea mai bună combinație de hiperparametri pentru fiecare clasificator. Hiperparametrii considerați au fost:

- pentru SVM: parametrul de regularizare C, parametrul γ, și tipul de kernel folosit
- pentru Random Forest: numărul de arbori, adâncimea maximă a fiecărui arbore, și procentul de eșantionare a datelor
- pentru Extra Trees: numărul de arbori, adâncimea maximă a fiecărui arbore, procentul de eșantionare a datelor, și dacă se folosește sau nu bootstrapping
- pentru clasificatorul XGBoost: numărul de arbori și rata de învățare

Pentru setul de date redus la primele 80% dintre atribute (după puterea predictivă), am obținut următoarele rezultate:

- pentru SVM: `SVC(C=100, gamma=0.1, kernel="rbf")`
- pentru Random Forest: `RandomForestClassifier(n_estimators=500, max_depth=14, max_samples=0.8)`
- pentru Extra Trees: `ExtraTreesClassifier(n_estimators=500, max_depth=14, max_samples=0.8, bootstrap=True)`
- pentru XGBoost: `XGBClassifier(n_estimators=200, learning_rate=0.2)`

Performanțele algoritmilor pe setul de date redus sunt prezentate în tabelele următoare:

Clasă	Acuratețe	Precizie	Recall	Scor F1
0	88.05	88.37	77.55	82.60
1	88.05	75.86	83.01	79.27
2	88.05	84.31	82.69	83.49
3	88.05	81.81	83.33	82.56
4	88.05	90.62	90.62	90.62
5	88.05	**100.00**	96.29	**98.11**
6	88.05	95.16	**100.00**	97.52
Medie	88.05	88.04	87.72	87.80
Abatere standard	0.00	6.67	6.69	6.27

Tabelul 2: Performanțele SVM pe setul redus de date

Clasă	Acuratețe	Precizie	Recall	Scor F1
0	91.94	93.75	91.83	92.78
1	91.94	77.41	90.56	83.47
2	91.94	93.33	80.76	86.59
3	91.94	90.74	90.74	90.74
4	91.94	93.65	92.18	92.91
5	91.94	**100.00**	96.29	98.11
6	91.94	96.72	**100.00**	**98.33**
Medie	91.94	92.21	91.80	91.87
Abatere standard	0.00	5.84	4.85	4.49

Tabelul 3: Performanțele Random Forest pe setul redus de date

Clasă	Acuratețe	Precizie	Recall	Scor F1
0	89.09	87.75	87.75	87.75
1	89.09	72.88	81.13	76.78
2	89.09	83.67	78.84	81.18
3	89.09	90.38	87.03	88.67
4	89.09	90.47	89.06	89.76
5	89.09	98.14	98.14	98.14
6	89.09	**100.00**	**100.00**	**100.00**
Medie	89.09	89.07	88.90	88.94
Abatere standard	0.00	7.44	6.45	6.80

Tabelul 4: Performanțele ExtraTrees pe setul redus de date

Clasă	Acuratețe	Precizie	Recall	Scor F1
0	89.35	87.75	87.75	87.75
1	89.35	85.71	79.24	82.35
2	89.35	82.35	80.76	81.55
3	89.35	89.09	90.74	89.90
4	89.35	87.87	90.62	89.23
5	89.35	**98.07**	94.44	96.22
6	89.35	93.65	**100.00**	**96.72**
Medie	89.35	89.23	89.13	89.14
Abatere standard	0.00	4.23	5.97	4.87

Tabelul 5: Performanțele GradientBoostedTrees pe setul redus de date

Se observă că Random Forest oferă performanțele medii cele mai bune, pentru toate metricile, fiind singurul algoritm pentru care valorile medii ale tuturor metricilor sunt mai mari de 90%. De asemenea, Random Forest obține valorile cele mai bune ale scorului F1 pentru fiecare clasă în parte. Totuși, clasificatorul Extra Trees are scorul F1 de 100% pentru clasa 6 (depășind, astfel, Random Forest). Random Forest prezice cel mai bine diagnosticele D0, D1, D2, D3 și D4, în timp ce Extra Trees prezic cel mai bine diagnosticul D6. Pentru diagnosticul D5, cei doi clasificatori au performanțe practic identice – diferența este de doar 0.03% (din moment ce sunt aproximativ 275 de pacienți cu diagnosticul D5, o diferență de 0.03%

înseamnă că cel mult un singur pacient a fost clasificat greșit de Random Forest). Mai mult, SVM obține aceleași performanțe cu Random Forest pentru diagnosticul D5. Diagnosticele prezise cel mai bine sunt D5 și D6, cu scoruri F1 maxime de 98.14%, respectiv 100%. În schimb, diagnosticul D1 este cel mai prost prezis, cu un scor F1 maxim de 83.47%.

Impactul hiperparametrilor poate fi de la destul de puțin semnificativ până la critic, funcție de algoritm. Astfel:

a) pentru SVM, folosirea kernelului sigmoid în locul celui radial duce la un scoruri (întoarse de Grid Search) semnificativ mai mici față de folosirea kernelului radial. De exemplu, combinația de hiperparametri $C = 100$, $\gamma = 0.1$ și kernel = 'rbf' duce la un scor de 0.8632[4] (acesta fiind și scorul maxim pentru SVM), în timp ce combinația $C = 100$, $\gamma = 0.1$ și kernel = 'sigmoid' duce la un scor de 0.1668. Impactul lui γ este mai puțin semnificativ (dar nu de neglijat), întrucât combinația $C = 100$, $\gamma = 0.01$ și kernel = 'rbf' duce la un scor de 0.841, iar combinația $C = 100$, $\gamma = 1$ și kernel = 'rbf' duce la un scor de 0.7462. Impactul lui C este de asemenea semnificativ, căci o valoare prea mică a lui poate compromite performanțele clasificatorului: de exemplu, combinația $C = 0.01$, $\gamma = 0.1$ și kernel = 'rbf' duce la un scor de 0.167.

b) pentru Random Forest, scorul maxim este de 0.9054, și a fost obținut folsind hiperparametrii n_estimators $= 500$, max_depth $= 14$ și max_samples $= 0.8$. Cât timp adâncimea maximă a unui arbore și procentul de selecție a exemplelor sunt ambele mari, numărul de estimatori are o influență aproape inexistentă asupra scorului: de exemplu, combinația de hiperparametri n_estimators $= 50$, max_depth $= 14$ și max_samples $= 0.8$ duce la un scor de 0.8984. Alegerea hiperparametrilor influențează performanțele algoritmului doar în cazul în care cel puțin doi dintre aceștia sunt setați la valori mici: combinația n_estimators $= 500$, max_depth $= 6$ și max_samples $= 0.3$ duce la un scor de 0.846, iar combinația n_estimators $= 50$, max_depth $= 6$ și max_samples $= 0.3$ duce la un scor de 0.837. Clasificatorul Random Forest este, deci, robust la variația hiperparametrilor.

c) pentru ExtraTrees, scorul maxim este de 0.8888, și a fost obținut folosind hiperparametrii n_estimators $= 500$, max_depth $= 14$ și max_samples $= 0.8$[5]. Spre deosebire de Random Forest, acest algoritm este mai sensibil la alegerea hiperparametrilor, pentru că adâncimea maximă a unui arbore influențează, de una singură, scorul obținut mai mult decât o face alegerea întregii combinații de hiperparametri în cazul Random Forest: pentru Extra Trees, combinația n_estimators $= 500$, max_depth $= 6$ și max_samples $= 0.8$ conduce la un scor de 0.7444. În plus, algoritmul este robust la alegerea procentului de selecție a exemplelor și a numărului de estimatori: combinația n_estimators $= 50$, max_depth $= 14$ și max_samples $= 0.3$ duce la un scor de 0.8726.

d) pentru GradientBoostedTrees, scorul maxim este de 0.9012, și este obținut cu combinația de hiperparametri n_estimators $= 200$ și learning_rate $= 0.2$. Clasificatorul este robust la alegerea hiperparametrilor, deoarece scorul cel mai mic obținut este de 0.8464 (pentru combinația n_estimators $= 50$ și learning_rate $= 0.01$).

Setul complet de atribute

Pentru setul de date complet, următoarele combinații de hiperparametri au fost întoarse de către procedura de Grid Search:

– pentru SVM: `SVC(C=10, gamma=0.1, kernel="rbf")`

[4] Toate scorurile din această secțiune sunt scoruri medii peste toate cele 5 iterații ale validării încrucișate

[5] Pentru toate testele acestui clasificator, am folosit bootstrap = True

– pentru Random Forest: `RandomForestClassifier(n_estimators=500, max_depth=12, max_samples=0.8)`

– pentru Extra Trees: `ExtraTreesClassifier(n_estimators=500, max_depth=14, max_samples=0.8, bootstrap=True)`

– pentru XGBoost: `XGBClassifier(n_estimators=100, learning_rate=0.5)`

Performanțele algoritmilor sunt următoarele:

Clasă	Acuratețe	Precizie	Recall	Scor F1
0	88.57	90.90	81.63	86.02
1	88.57	75.00	84.90	79.64
2	88.57	88.88	76.92	82.47
3	88.57	85.45	87.03	86.23
4	88.57	86.56	90.62	88.54
5	88.57	**100.00**	96.29	**98.11**
6	88.57	95.16	**100.00**	97.52
Medie	88.57	88.82	88.27	88.40
Abatere standard	0.00	6.47	6.61	5.77

Tabelul 6: Performanțele SVM pe setul integral de date

Clasă	Acuratețe	Precizie	Recall	Scor F1
0	91.94	91.66	89.79	90.72
1	91.94	75.80	88.67	81.73
2	91.94	95.34	78.84	86.31
3	91.94	92.59	92.59	92.59
4	91.94	95.31	95.31	95.31
5	91.94	**98.11**	96.29	97.19
6	91.94	96.72	**100.00**	**98.33**
Medie	91.94	92.21	91.70	91.78
Abatere standard	0.00	6.18	5.59	4.91

Tabelul 7: Performanțele Random Forest pe setul integral de date

Clasă	Acuratețe	Precizie	Recall	Scor F1
0	90.64	86.27	89.79	88.00
1	90.64	75.86	83.01	79.27
2	90.64	91.30	80.76	85.71
3	90.64	90.74	90.74	90.74
4	90.64	93.54	90.62	92.06
5	90.64	96.36	98.14	97.24
6	90.64	**100.00**	**100.00**	**100.00**
Medie	90.64	90.62	90.48	90.48
Abatere standard	0.00	6.38	5.77	5.70

Tabelul 8: Performanțele ExtraTrees pe setul integral de date

Clasă	Acuratețe	Precizie	Recall	Scor F1
0	89.87	84.61	89.79	87.12
1	89.87	87.23	77.35	82.00
2	89.87	85.71	80.76	83.16
3	89.87	86.20	92.59	89.28
4	89.87	89.23	90.62	89.92
5	89.87	**98.11**	96.29	97.19
6	89.87	96.72	**100.00**	**98.33**
Medie	89.87	89.72	89.68	89.62
Abatere standard	0.00	0.47	6.59	5.15

Tabelul 9: Performanțele GradientBoostedTrees pe setul integral de date

Se observă că pe setul de date ce conține toate atributele, performanțele (precizie, recall, scor F1) medii ale tuturor algoritmilor se îmbunătățesc cu până la 2%, în cazul ExtraTrees. Cum fiecare clasă are, în medie, 264 de pacienți, acest lucru înseamnă că până la doi pacienți în plus sunt clasificați corect. Aceste tabele evidențiază tradeoff-ul dintre viteza de calcul și corectitudinea clasificării: setul de date prelucrat prin eliminarea a 20% dintre atribute oferă performanțe mai scăzute, dar comparabile, cu cele ale setului integral de date, dar algoritmii ce-l prelucrează oferă avantajul de a găsi mai repede o clasificare a pacienților (la rularea pe calculatorul propriu, antrenarea și rularea tuturor celor patru clasificatori pe setul redus de date și afișarea metricilor de performanță au durat 2 minute și 2.5 secunde, pe când aceleași operațiuni pe setul integral de date au durat 2 minute și 25.9 secunde).

Alegerea hiperparametrilor influențează performanțele algoritmilor în același mod cu cel din cazul setului filtrat de date.

3.3 Matricele de confuzie

Pentru fiecare algoritm, pentru configurația optimă a hiperparametrilor, am realizat câte o matrice de confuzie. Matricele astfel obținute sunt următoarele:

Setul redus de date

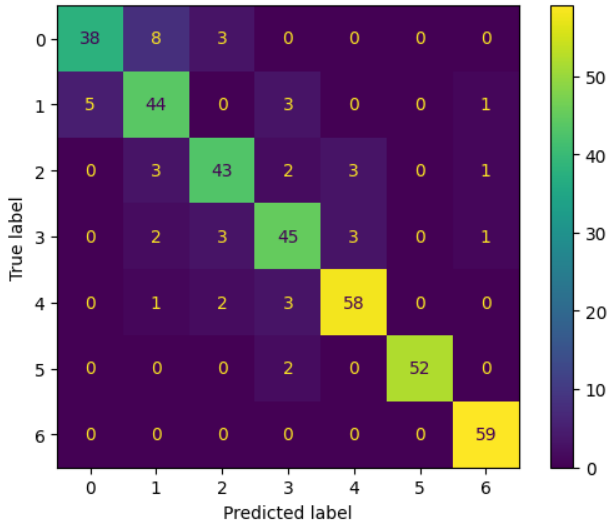


Figura 1: SVM

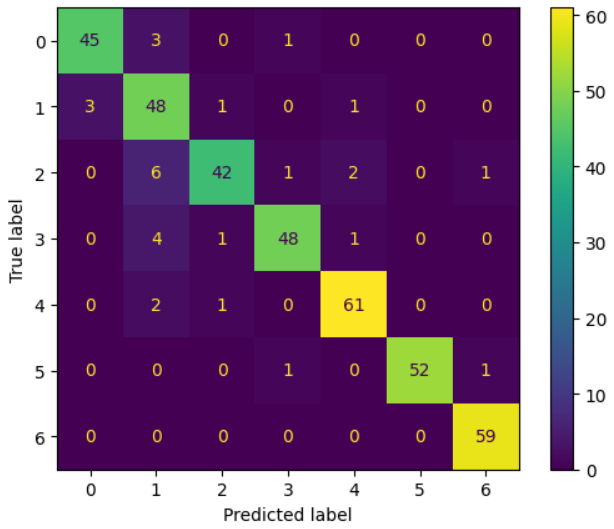


Figura 2: Random Forest

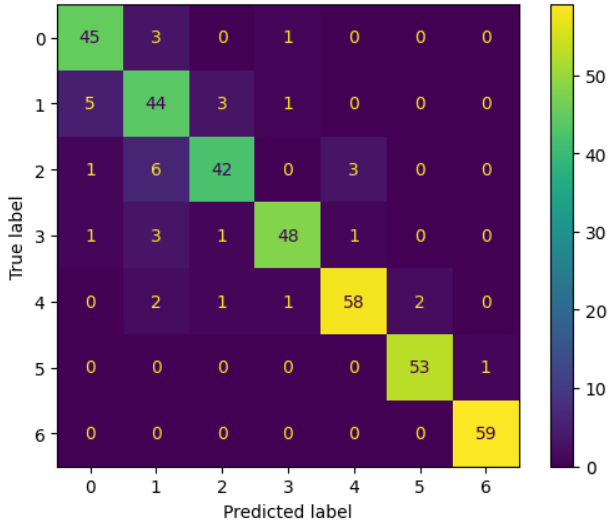


Figura 3: Extra Trees

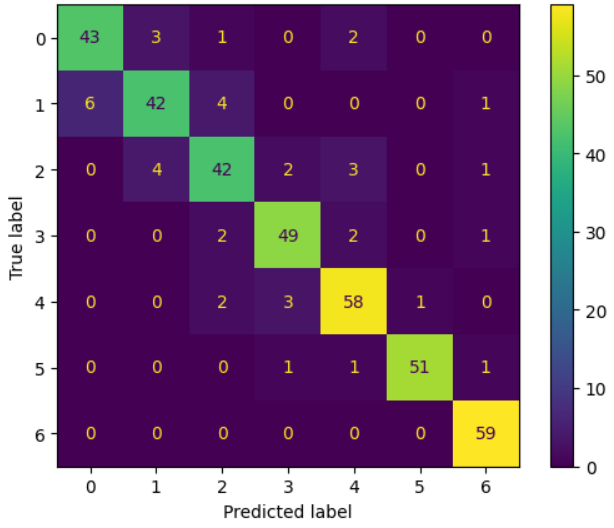


Figura 4: Gradient Boosted Trees

Setul complet de date

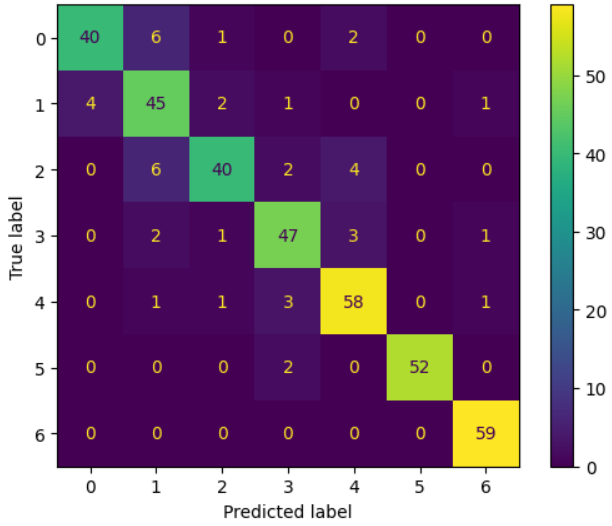


Figura 5: SVM

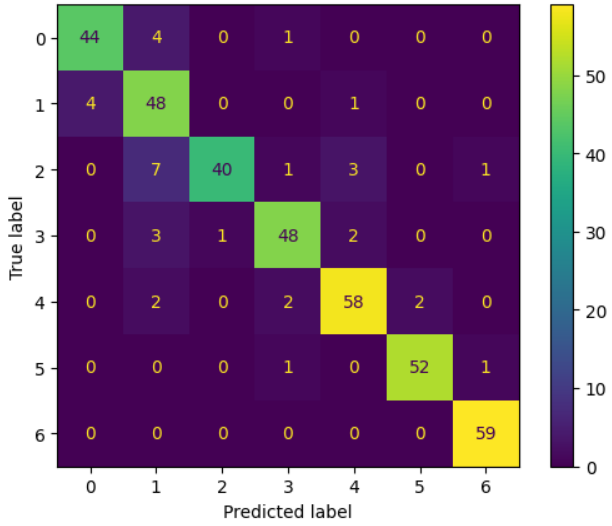


Figura 6: Random Forest

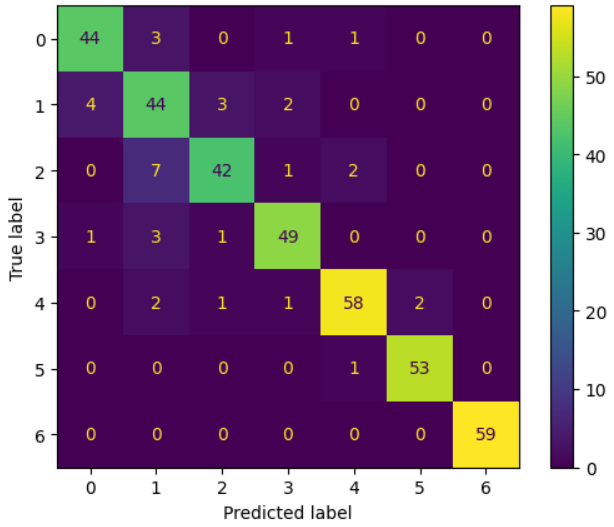


Figura 7: Extra Trees

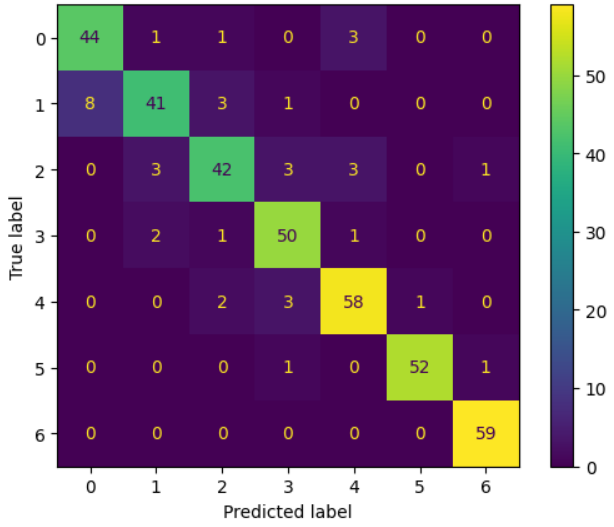


Figura 8: Gradient Boosted Trees