Tema 1 Invățare Automată

Văideanu Renata - Georgia

November 30, 2024

Abstract

Tema 1 la materia Învațare Automata, anul 4 - Universitatea Națională de Știință și Tehnologie Politehnică București

Introducere

Acest document explică procesul de implementare a sarcinilor din Tema 1 de Învățare Automată, cu accent pe extragerea atributelor utilizând diferite metode pentru două seturi de date diferite: Fashion-MNIST și Fruits-360. Obiectivul principal este clasificarea imaginilor în categoriile corecte pe baza atributelor extrase.

1 Extragerea de atribute (Feature Extraction)

Pentru rezolvarea cerintelor, am optat pentru urmatoarele doua metode: PCA si HOG.

Motivele utilizării PCA și HOG:

1. PCA (Principal Component Analysis):
   - Fashion-MNIST: Setul de date include imagini grayscale cu rezoluție scăzută, unde PCA ajută la reducerea dimensionalității prin captarea variațiilor esențiale din date, cum ar fi contururi generale și diferențe de textură.
   - Fruits-360: Având imagini RGB cu rezoluție mai mare și complexitate sporită, PCA simplifică reprezentarea prin evidențierea atributelor dominante, cum ar fi combinațiile de culori și formele mari.
2. HOG (Histogram of Oriented Gradients):
   - Fashion-MNIST: HOG este ideal pentru identificarea caracteristicilor geometrice ale obiectelor (e.g., orientările marginilor), utile pentru clasificarea articolelor vestimentare care au forme distincte.
   - Fruits-360: Metoda HOG este potrivită pentru captarea detaliilor fine ale texturilor și formelor, cum ar fi contururile fructelor, care sunt esențiale pentru diferențierea între clase.

In cazul de fata, PCA este aplicat pe seturile de date pentru reducerea dimensionalității cu 20 de componente principale, reținând 78% din varianța datelor. In ambele dataseturi, valorile cumulativ explicate pentru variantă sunt afișate, ceea ce indică faptul că aceste componente sunt eficiente în reducerea complexității datelor.

Fashion-MNIST: Primele 20 de componente explică aproape 78% din varianța totală. Prima componentă explică 29% din variantă, iar primele 5 componente explică împreună aproape 50%.

Fruits-360: Primele 20 de componente explică aproape 80% din varianța totală. Prima componentă explică 30%, iar primele 5 componente împreună explică aproape 57%.

În ambele cazuri, PCA reduce semnificativ dimensiunea datelor, păstrând în același timp majoritatea informației esențiale pentru clasificare. Aceasta ajută la îmbunătățirea eficienței și performanței modelelor de învățare automată.

Totodata, HOG a fost aplicat pe imagini utilizând parametrii: Pixels per cell: (8, 8) pentru Fashion-MNIST și (16, 16) pentru Fruits-360. Cells per block: (2, 2). Block normalization: L2-Hys. Dimensiunea vectorilor HOG pentru fiecare imagine de antrenament din seturile de date Fashion-MNIST și Fruits-360.

Fashion-MNIST: Dimensiunea vectorilor HOG pentru imaginile de antrenament este 144. Acest lucru înseamnă că pentru fiecare imagine de antrenament, descriptorul HOG are 144 de caracteristici care sunt extrase pe baza orientărilor gradientului și ale structurilor locale ale imaginii (margini, texturi etc.).

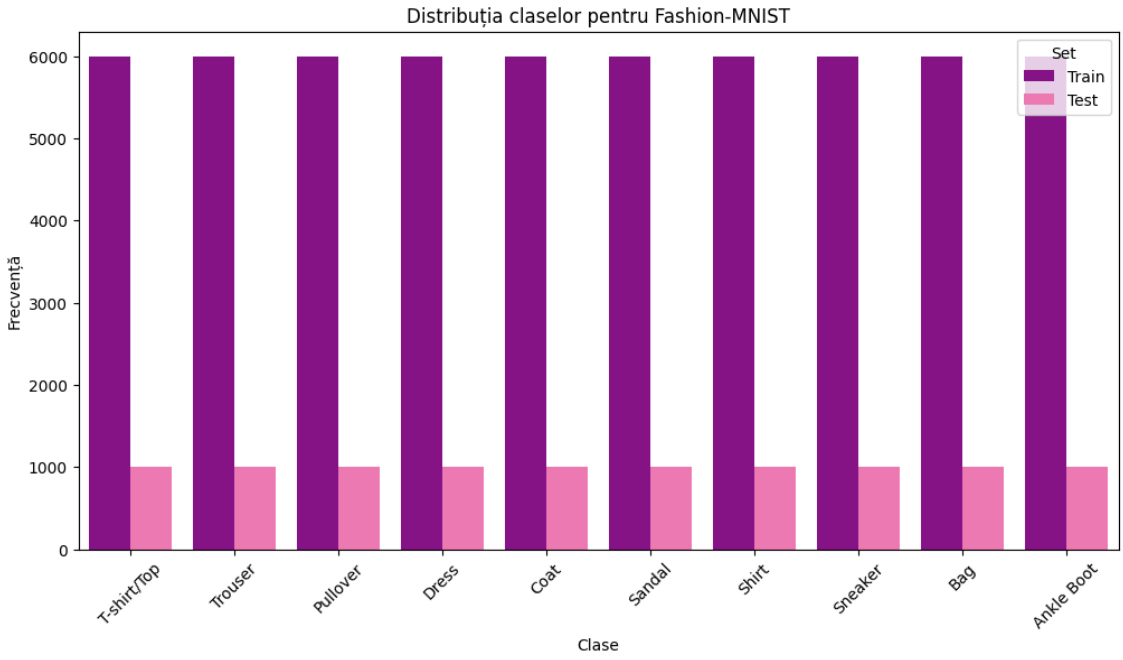
Fruits-360: Dimensiunea vectorilor HOG pentru imaginile de antrenament este 900. Aici, fiecare imagine de antrenament are un descriptor HOG cu 900 de caracteristici, fiind mai detaliat datorită dimensiunii mai mari a imaginilor și a complexității lor (100x100 pixeli RGB).

Diferenta principala de valori dintre celee doua seturi de date este datorata dimensiunii imaginilor din fiecare set. Astfel, Fashion-MNIST are vectori HOG mai mici (144 caracteristici), deoarece imaginile sunt mai mici și grayscale (28x28), in timp ce Fruits-360 are vectori HOG mai mari (900 caracteristici), având imagini mai mari și cu mai multe detalii (100x100 RGB).

2 Vizualizarea atributelor extrase

2.1 Analiza echilibrului de clase

Fashion-MNIST

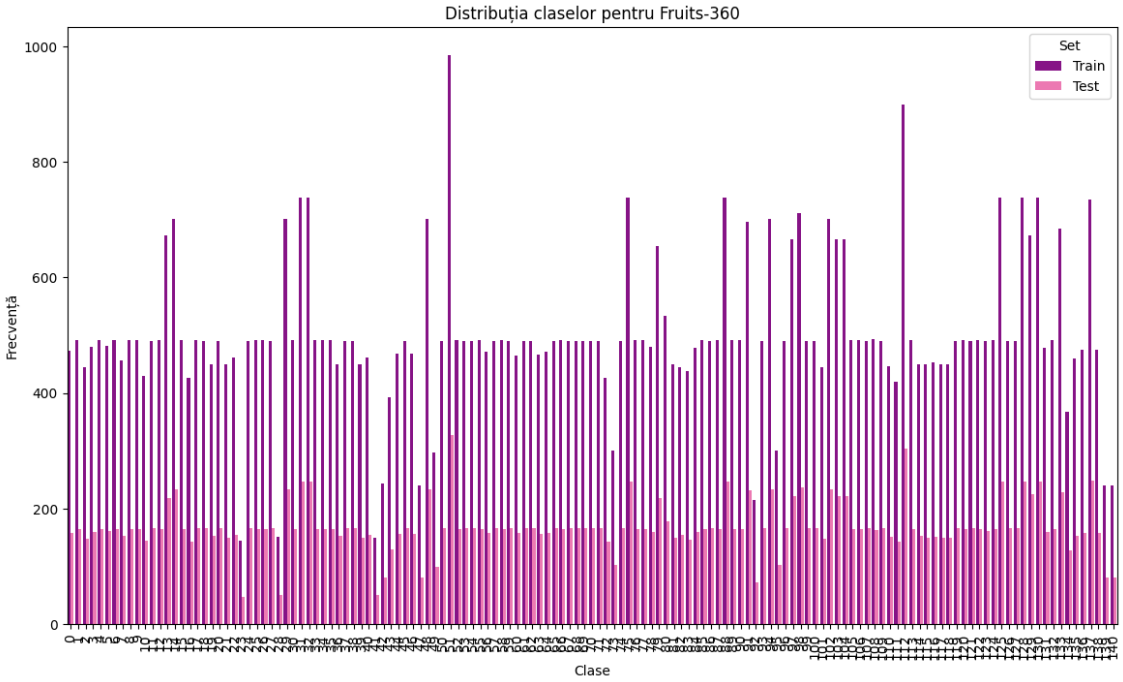


Mai sus se poate observa distribuția claselor în seturile de antrenament și test pentru datasetul Fashion-MNIST. Datasetul Fashion-MNIST este bine echilibrat din punct de vedere al claselor, atât pentru antrenament, cât și pentru testare. Acest lucru sprijină o bună generalizare a modelului pe toate categoriile de obiecte. Setul de antrenament conține aproximativ același număr de exemple pentru fiecare clasă (aproximativ 6.000 de imagini per clasă). Setul de test are, de asemenea aproximativ 1.000 de imagini per clasă. Repartizare egală între clase:

Fiecare clasă este bine reprezentată atât în setul de antrenament, cât și în cel de test, ceea ce reduce riscul de bias între clase în timpul antrenării modelului.

Distribuția echilibrată facilitează învățarea, deoarece modelul va primi un număr egal de exemple pentru fiecare clasă, minimizând posibilitatea ca acesta să favorizeze anumite clase în detrimentul altora. Evaluarea pe setul de test este robustă, deoarece fiecare clasă este reprezentată în mod similar cu setul de antrenament.

Fruits-360



Mai sus se poate observa distribuția claselor în seturile de antrenament și test pentru datasetul Fruits-360. Datasetul Fruits-360 prezintă un dezechilibru moderat între clase, cu variații semnificative în frecvența acestora care poate introduce un risc de bias în clasificare. Deși majoritatea claselor au o distribuție relativ uniformă, unele sunt mult mai reprezentate (aproape 1000 de exemple), în timp ce altele au frecvențe mai reduse.

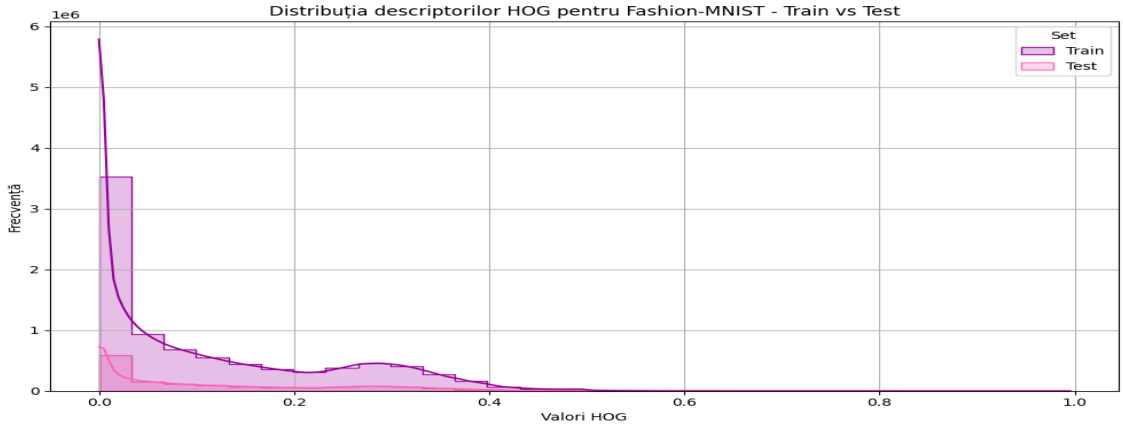
Distribuția între seturile de antrenare și testare este, totuși, proporțională, ceea ce indică o împărțire corespunzătoare a datelor. Deși setul nu este perfect echilibrat, dezechilibrul nu este extrem, iar impactul său poate fi atenuat prin măsuri precum ponderarea claselor sau augmentarea datelor pentru categoriile subreprezentate.

2.2 Vizualizarea cantitativă / calitativă a efectului de extragere a atributelor

Fashion-MNIST

Vizualizare cantitativa

HOG

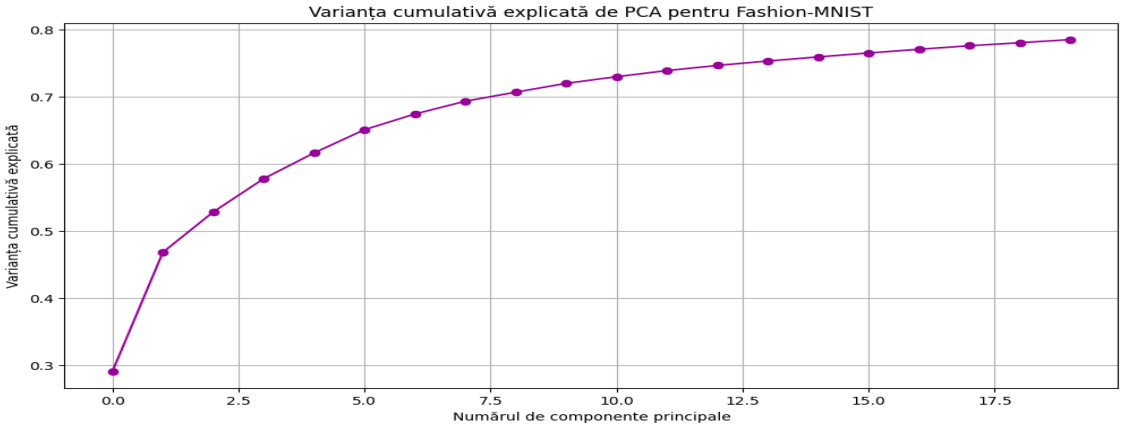


Graficul arată distribuția valorilor descriptorilor HOG pentru seturile de antrenare și testare. Majoritatea valorilor HOG sunt concentrate în intervalul mic (0.0 - 0.2), indicând că cele mai multe caracteristici sunt de intensitate scăzută. Distribuțiile pentru seturile de antrenare și testare sunt foarte similare, ceea ce sugerează o împărțire echilibrată a caracteristicilor între cele două seturi.

Set	count	mean	std	min	25%	50%	75%	max
Test	1440000.0	0.112106	0.123329	0.0	0.005586	0.061235	0.198050	0.888555
Train	8640000.0	0.111845	0.123565	0.0	0.005360	0.060399	0.197438	0.995428

Distribuția aproape identică între seturile de train și test indică faptul că datele sunt bine echilibrate și că extragerea HOG nu introduce variații neașteptate între seturi. Caracteristicile HOG au o concentrație mare a valorilor în jurul valorii zero, ceea ce sugerează că doar o parte mică a imaginii contribuie cu informații semnificative (marginile și contururile).

PCA



Graficul reprezintă varianța cumulativă explicată de componentele principale calculate prin PCA. Primele câteva componente principale explică o proporție semnificativă a variației din date (primele 5 componente explică peste 50% din varianță). După aproximativ 15 componente principale, varianța cumulativă ajunge la o valoare apropiată de 75%, sugerând că un subset redus de componente principale este suficient pentru a captura majoritatea informației din date.

	Număr de componente	Varianță explicată cumulativă (%)
0	5	61.618843
1	10	71.990827
2	15	75.930301
3	20	78.509556

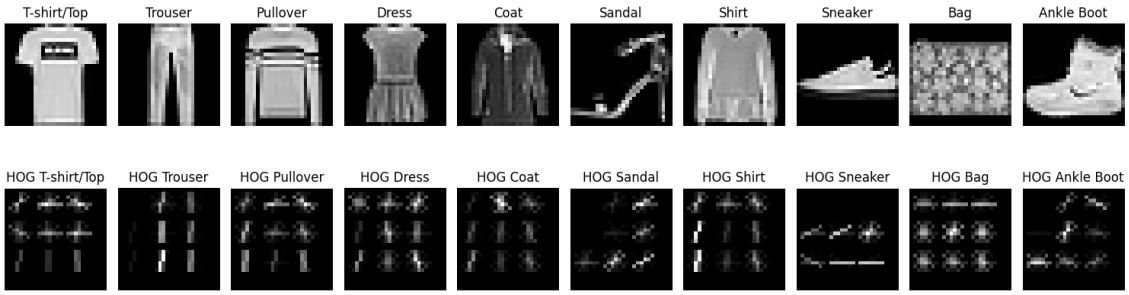
Primele 5 componente principale explică deja 61.61% din varianța totală, ceea ce arată că datele sunt foarte bine comprimate. Cu 10 componente, varianța explicată ajunge la aproape 72%, iar cu 20 componente se atinge 78.5%, ceea ce înseamnă că aceste componente capturează majoritatea informației esențiale din date.

Observăm o scădere a ratei de creștere a variației explicate pe măsură ce adăugăm mai multe componente. De exemplu, trecerea de la 5 la 10 componente adaugă aproximativ 10%, dar de la 15 la 20 componente adaugă doar aproximativ 3%.

Fashion-MNIST

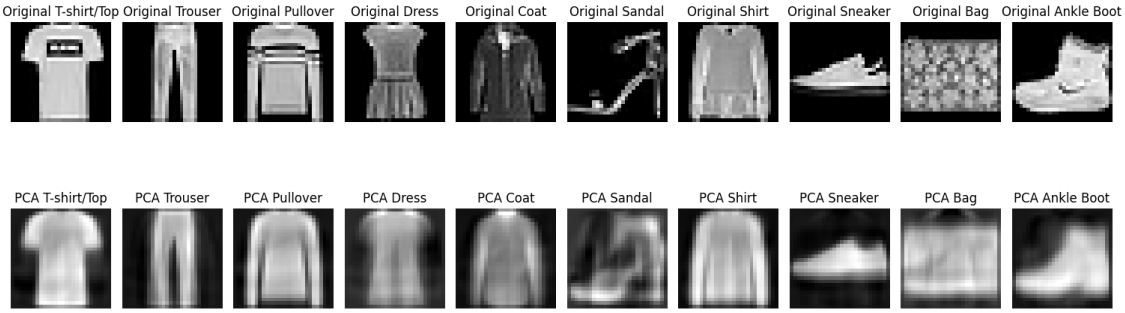
Vizualizare calitativa

HOG



Imaginile rezultate din HOG păstrează contururile principale ale obiectelor (haine, încălțăminte, etc.). Marginile și formele obiectelor sunt bine evidențiate, ceea ce este util pentru clasificare, dar texturile și detaliile fine lipsesc. Este evident că imaginile mai complexe, cum ar fi sandalul sau pantoful sport, au contururi clare, ceea ce subliniază robustețea metodei pentru detectarea marginilor. HOG este mai bun pentru evidențierea contururilor și marginilor, fiind potrivit pentru modele care se bazează pe caracteristici locale.

PCA



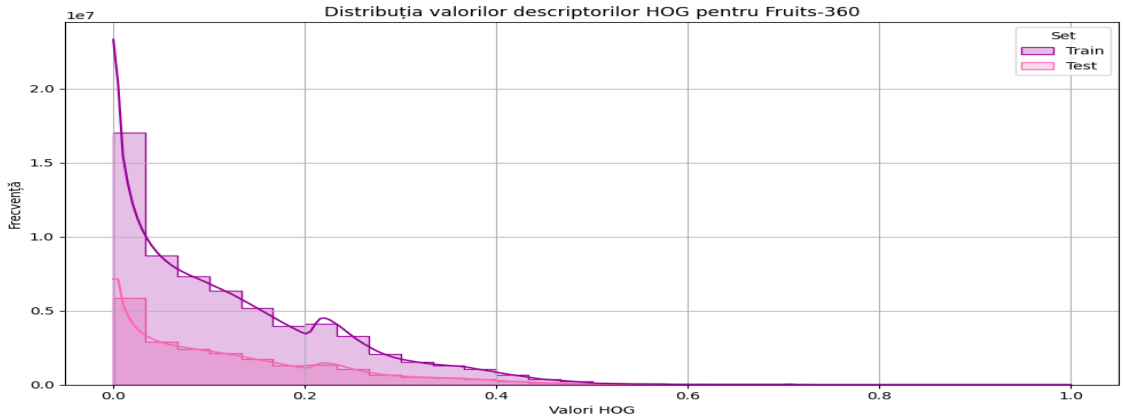
Imaginile reconstruite după aplicarea PCA sunt mai estompate, dar păstrează forma globală și structura obiectelor. Detaliile fine sunt pierdute, însă contururile generale sunt suficient de clare pentru a diferenția clasele. Genții și sandalele sunt mai greu de recunoscut, ceea ce sugerează că PCA nu capturează bine informațiile fine necesare pentru aceste clase. PCA capturează trăsături globale și reduce zgomotul, fiind util pentru reducerea dimensionalității și vizualizarea datelor într-un spațiu comprimat.

Fruits-360

Vizualizare cantitativa

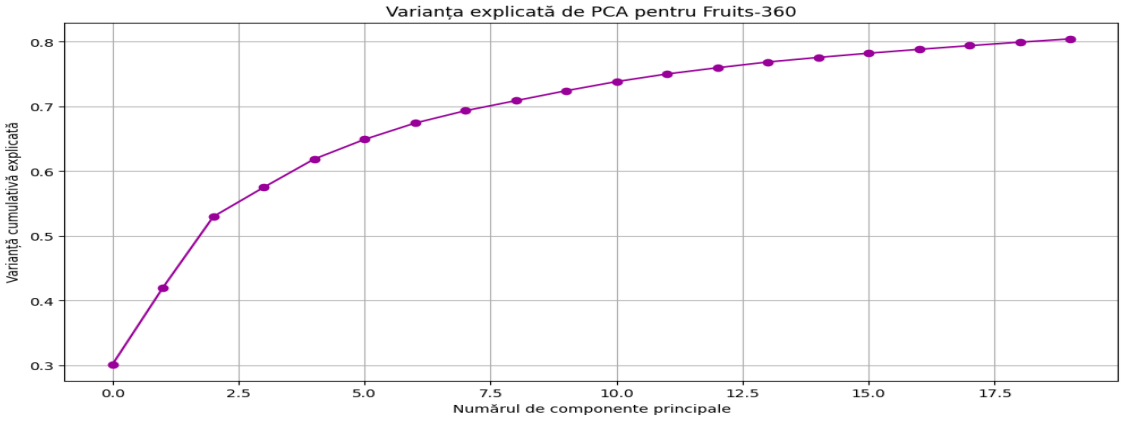
HOG

Set	count	mean	std	min	25%	50%	75%	max
Test	21257100.0	0.122328	0.113159	0.0	0.028224	0.092112	0.189897	1.0
Train	63441900.0	0.122982	0.112421	0.0	0.029592	0.093521	0.190837	1.0



Distribuția similară a valorilor HOG între seturile de Train și Test indică faptul că nu există diferențe majore între cele două seturi, ceea ce contribuie la o bună generalizare a modelului. Majoritatea valorilor HOG sunt mici, ceea ce reflectă faptul că doar o parte a imaginii conține informații semnificative sub formă de contururi sau margini. Zonele cu gradient puternic sunt rare (valoarea maximă este 1, dar 75% din valori sunt sub 0.19), ceea ce subliniază importanța marginilor în procesul de clasificare folosind descriptorii HOG.

PCA



Graficul reprezintă varianța cumulativă explicată de componentele principale calculate prin PCA. Primele câteva componente principale explică o proporție semnificativă a variației din date (primele 5 componente explică peste 50% din varianță). După aproximativ 15 componente principale, varianța cumulativă ajunge la o valoare apropiată de 75%, sugerând că un subset redus de componente principale este suficient pentru a captura majoritatea informației din date.

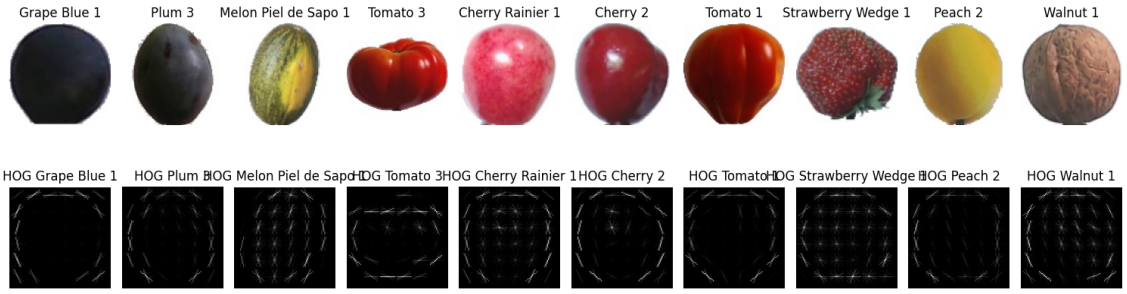
	Număr de componente	Varianță explicată cumulativă (%)
0	5	61.873525
1	10	72.437608
2	15	77.578545
3	20	80.448878

Primele 5 componente principale explică deja 61.87% din variant, a totală, ceea ce arată că datele sunt destul de bine comprimate. Cu 10 componente, varianța explicată ajunge la 72.44%, iar cu 20 componente se atinge 80.45%, ceea ce înseamnă că aceste componente capturează majoritatea informației esențiale din date. Observăm o scădere a ratei de creștere a variației explicate pe măsură ce adăugăm mai multe componente. De exemplu, trecerea de la 5 la 10 componente adaugă aproximativ 10%, dar de la 15 la 20 componente adaugă doar aproximativ 3%.

Fruits-360

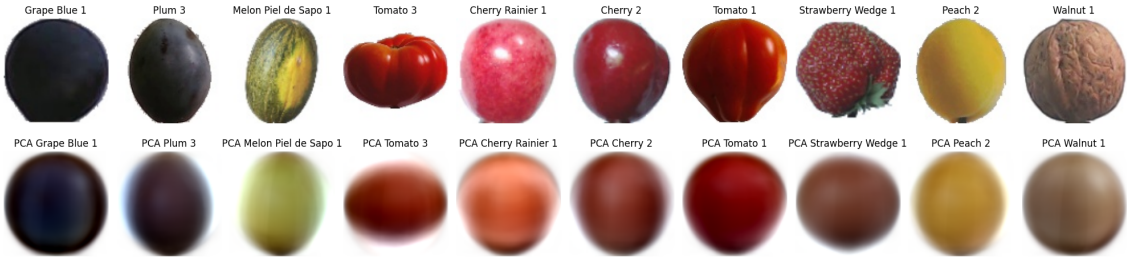
Vizualizare calitativa

HOG



Aceasta evidențiază contururile și structurile orientate ale fructelor, extrăgând caracteristici bazate pe textură și margini. Observațiile: Fructele cu texturi distincte sau contururi puternice, cum ar fi nuca (Walnut 1) și căpșuna (Strawberry Wedge 1), au modele HOG bine definite. Fructele mai netede, cum ar fi strugurii (Grape Blue 1) sau prunele (Plum 3), prezintă un grad mai scăzut de variație în descriptorii HOG.

PCA



Fructele cu culori și forme unice, cum ar fi pepenele galben (Melon Piel de Sapo 1) sau piersica (Peach 2), sunt clar diferențiate chiar și după reducerea dimensionalității. Fructele de culori similare, cum ar fi tomatele (Tomato 1 și Tomato 3), pot părea mai greu de diferențiat doar pe baza PCA, datorită suprapunerii informației de culoare. Metoda se concentrează pe variațiile globale, fiind util pentru clasificarea bazată pe culoare și formă generală.

3 Standardizarea și selecția atributelor

Pentru standardizarea și selecția atributelor în cele două seturi de date, am ales să testez ambele variante pentru a asigura o alegere cât mai clară. Astfel, după combinarea atributelor HOG și PCA folosind funcția np.hstack, le-am scalat cu ajutorul StandardScaler(). În cazul utilizării VarianceThreshold, valoarea pragului a fost setată la 1 pentru ambele seturi de date, iar pentru SelectPercentile, valoarea percentile a fost de 20 pentru setul de date Fruits-360 și de 30 pentru Fashion-MNIST.

Fruits:

- Original feature shape for training: (70491, 920)
- Selected VarianceThreshold shape for training: (70491, 485)
- Select Percentile shape for training: (70491, 97)
- Original feature shape for test: (23619, 920)
- Selected VarianceThreshold shape for test: (23619, 485)

- Select Percentile shape for test: (23619, 97)

Fashion:

- Original feature shape for training: (60000, 164)
- Selected VarianceThreshold shape for training: (60000, 85)
- Select Percentile shape for training: (60000, 49)
- Original feature shape for test: (10000, 164)
- Selected VarianceThreshold shape for test: (10000, 85)
- Select Percentile shape for test: (10000, 49)

În mod evident, am ales cea de-a doua variantă, Select Percentile.

4 Utilizarea algoritmilor de Învățare Automată

În cadrul acestui task, am aplicat pentru fiecare algoritm în parte mai multe seturi de parametriidi l-am ales pe cel ideal cu ajutorul Randomized Search with Cross-Validation.

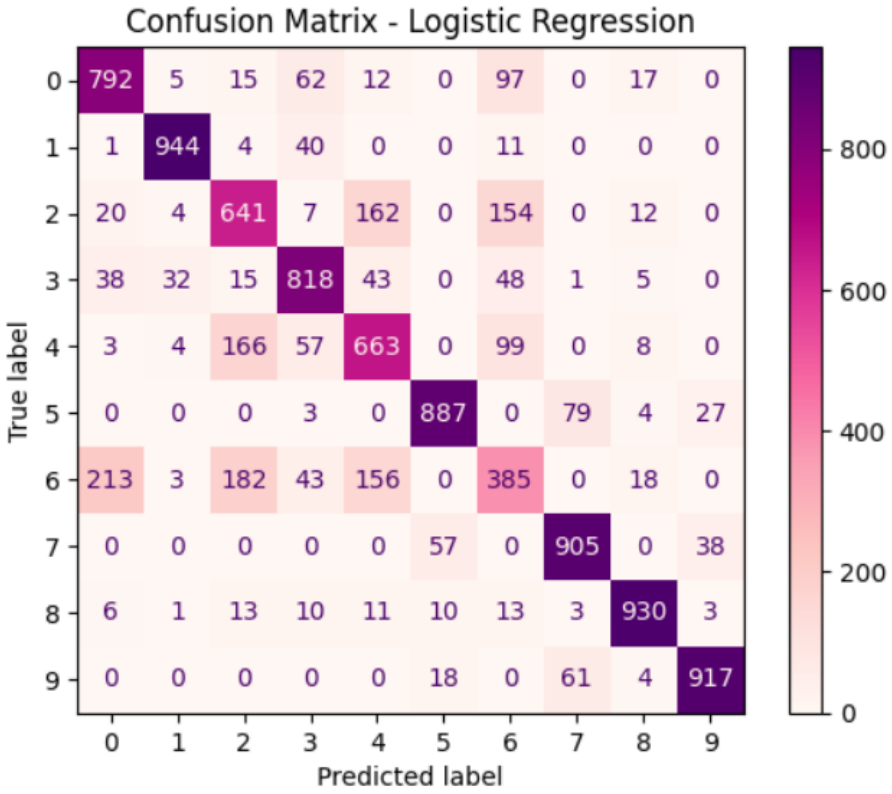
Din considerente de spațiu, nu am inclus în acest raport toate tabelele detaliate cu valorile medii și variațele pentru acuratețe, precizie, recall și F1-score corespunzătoare fiecărui algoritm și configurație de hiper-parametri. Cu toate acestea, toate aceste informații complete pot fi accesate prin următoarele linkuri: Fruits-360 și Fashion-MNIST. Aceste resurse conțin datele integrale, structurate conform cerințelor, incluzând valorile detaliate pentru fiecare clasă și subliniind performanțele maxime.

4.1 Fashion-MNIST

LogisticRegression

Fitting 5 folds for each of 10 candidates, totalling 50 fits.

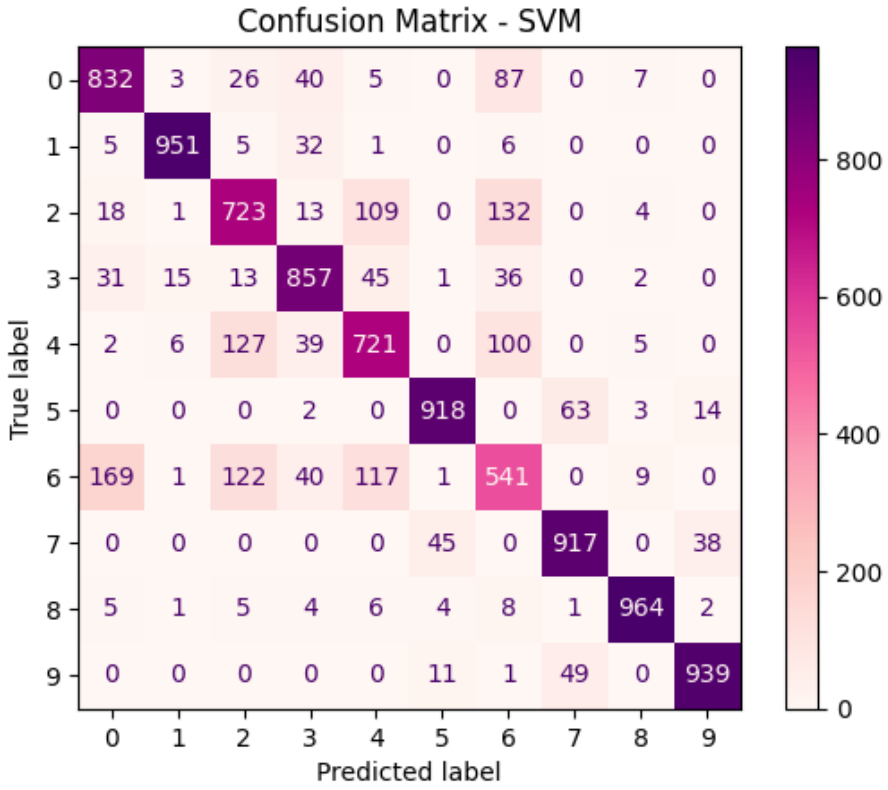
Best Params: 'multi_class': 'multinomial', 'C': 1



	precision	recall	f1-score	support
0	0.74	0.79	0.76	1000
1	0.95	0.94	0.95	1000
2	0.62	0.64	0.63	1000
3	0.79	0.82	0.80	1000
4	0.63	0.66	0.65	1000
5	0.91	0.89	0.90	1000
6	0.48	0.39	0.43	1000
7	0.86	0.90	0.88	1000
8	0.93	0.93	0.93	1000
9	0.93	0.92	0.92	1000
accuracy			0.79	10000
macro avg	0.78	0.79	0.79	10000
weighted avg	0.78	0.79	0.79	10000

SVM

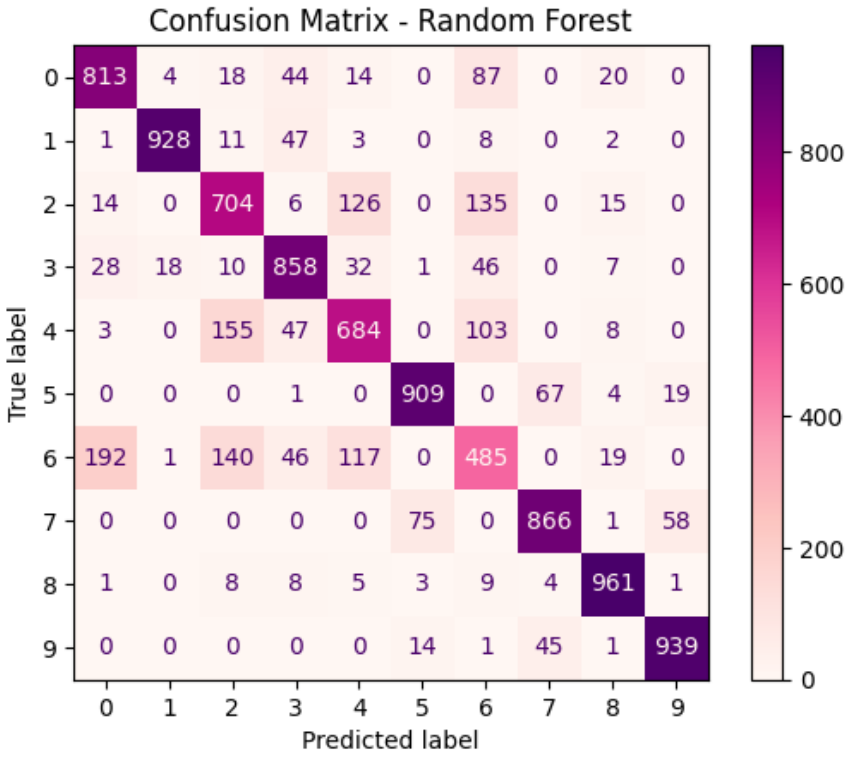
Fitting 5 folds for each of 8 candidates, totalling 40 fits.
Best Params: 'kernel': 'rbf', 'C': 10



	precision	recall	f1-score	support
0	0.78	0.83	0.81	1000
1	0.97	0.95	0.96	1000
2	0.71	0.72	0.72	1000
3	0.83	0.86	0.85	1000
4	0.72	0.72	0.72	1000
5	0.94	0.92	0.93	1000
6	0.59	0.54	0.57	1000
7	0.89	0.92	0.90	1000
8	0.97	0.96	0.97	1000
9	0.95	0.94	0.94	1000
accuracy			0.84	10000
macro avg	0.84	0.84	0.84	10000
weighted avg	0.84	0.84	0.84	10000

RandomForest

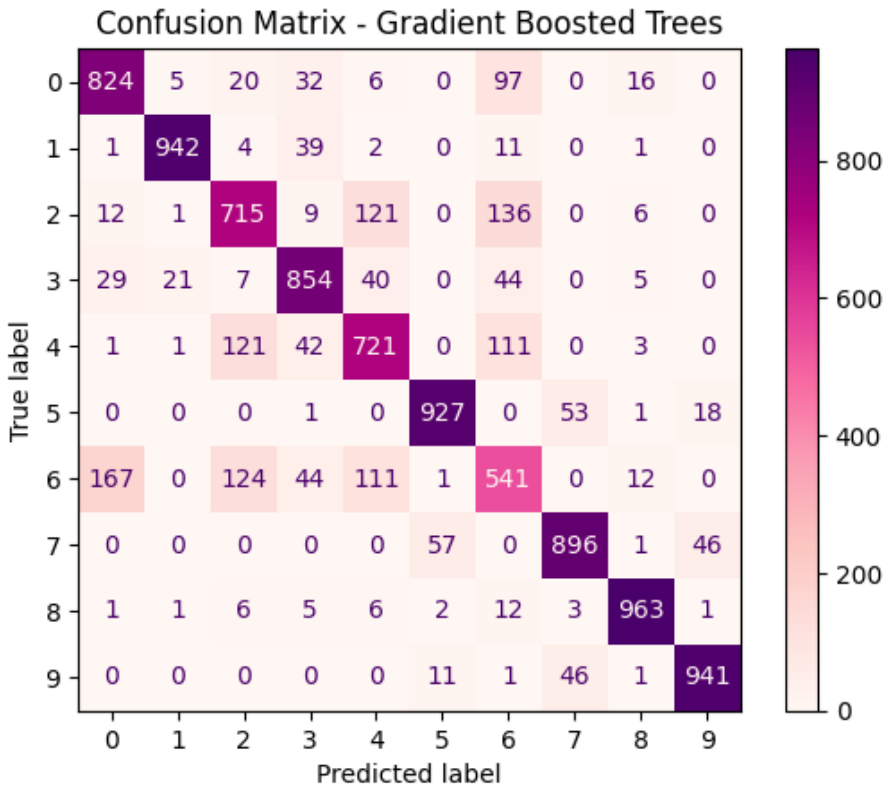
Fitting 5 folds for each of 10 candidates, totalling 50 fits
Best Params: 'n_estimators': 150, 'max_samples': 0.7, 'max_depth': 20



	precision	recall	f1-score	support
0	0.77	0.81	0.79	1000
1	0.98	0.93	0.95	1000
2	0.67	0.70	0.69	1000
3	0.81	0.86	0.83	1000
4	0.70	0.68	0.69	1000
5	0.91	0.91	0.91	1000
6	0.55	0.48	0.52	1000
7	0.88	0.87	0.87	1000
8	0.93	0.96	0.94	1000
9	0.92	0.94	0.93	1000
accuracy			0.81	10000
macro avg	0.81	0.81	0.81	10000
weighted avg	0.81	0.81	0.81	10000

GradientBoosted Trees

Fitting 5 folds for each of 10 candidates, totalling 50 fits.
Best Params: 'n_estimators': 200, 'max_depth': 7, 'learning_rate': 0.1



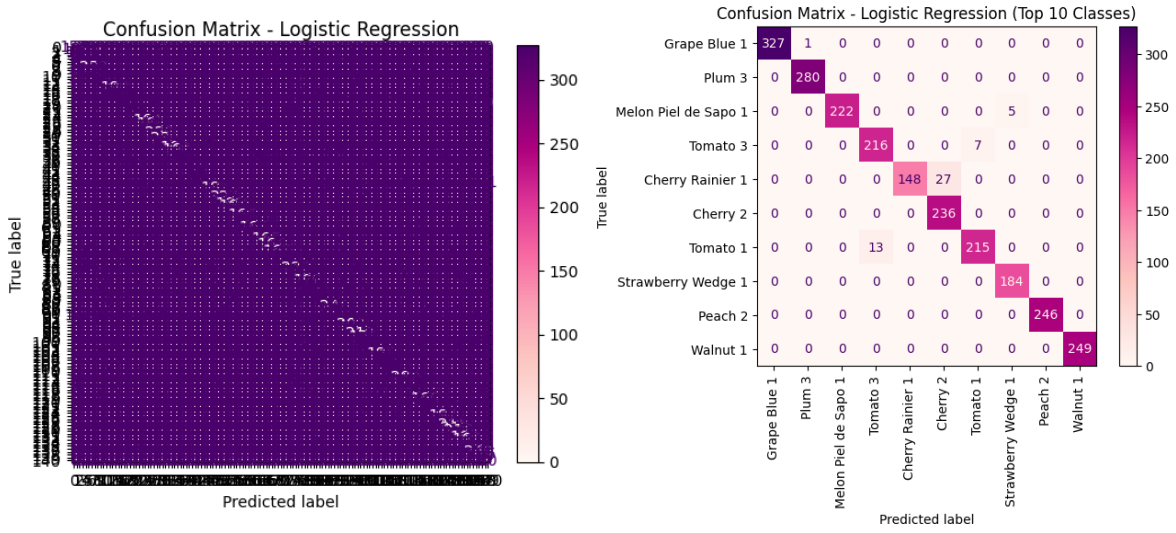
	precision	recall	f1-score	support
0	0.80	0.82	0.81	1000
1	0.97	0.94	0.96	1000
2	0.72	0.71	0.72	1000
3	0.83	0.85	0.84	1000
4	0.72	0.72	0.72	1000
5	0.93	0.93	0.93	1000
6	0.57	0.54	0.55	1000
7	0.90	0.90	0.90	1000
8	0.95	0.96	0.96	1000
9	0.94	0.94	0.94	1000
accuracy			0.83	10000
macro avg	0.83	0.83	0.83	10000
weighted avg	0.83	0.83	0.83	10000

4.2 Fruits-360

De asemenea, din cauza spațiului, am pus in document tabelele cu primelor 20 de clase, dar aici sunt tabelele complete cu toate cele 141 de clase: Fruits-360 Tables

Logistic Regresion

Fitting 5 folds for each of 10 candidates, totalling 50 fits
Best Params: 'multi_class': 'multinomial', 'C': 1



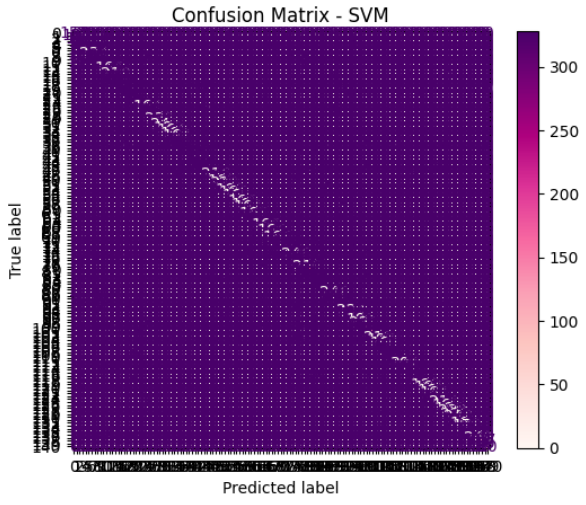
	precision	recall	f1-score	support
0	0.88	1.00	0.94	157
1	0.69	0.77	0.73	164
2	0.71	0.68	0.69	148
3	0.98	0.91	0.94	160
4	0.95	0.77	0.86	164
5	0.66	0.76	0.71	161
6	0.61	0.66	0.63	164
7	0.97	0.99	0.98	234
8	0.65	0.99	0.79	152
9	0.44	0.40	0.42	164
10	0.81	0.79	0.80	164
11	0.55	0.46	0.50	144
12	0.89	0.88	0.88	166
13	0.50	0.50	0.50	164
14	0.95	0.87	0.90	219
15	0.92	0.94	0.93	164
16	0.63	0.71	0.67	143
17	0.82	0.95	0.88	166
18	0.82	0.60	0.69	166
19	0.52	0.69	0.59	152
...				
accuracy			0.82	23619
macro avg	0.82	0.82	0.81	23619
weighted avg	0.82	0.82	0.81	23619

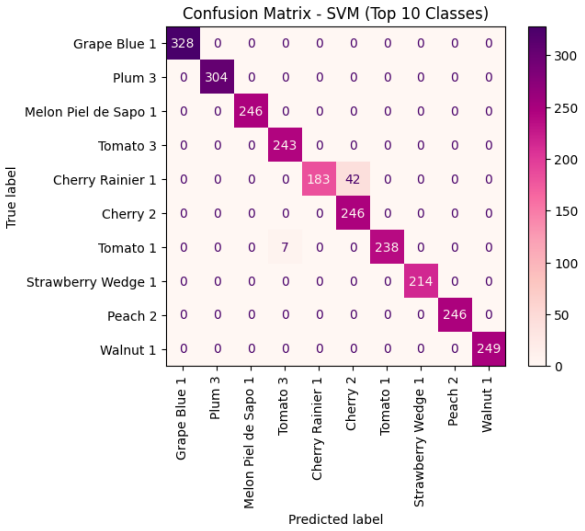
	precision	recall	f1-score	support
Grape Blue 1	1.00	1.00	1.00	328
Plum 3	1.00	0.92	0.96	304
Melon Piel de Sapo 1	1.00	0.90	0.95	246
Tomato 3	0.94	0.88	0.91	246
Cherry Rainier 1	1.00	0.60	0.75	246
Cherry 2	0.90	0.96	0.93	246
Tomato 1	0.97	0.87	0.92	246
Strawberry Wedge 1	0.97	0.75	0.85	246
Peach 2	1.00	1.00	1.00	246
Walnut 1	1.00	1.00	1.00	249
micro avg	0.98	0.89	0.93	2603
macro avg	0.98	0.89	0.93	2603
weighted avg	0.98	0.89	0.93	2603

SVM

Fitting 5 folds for each of 8 candidates, totalling 40 fits

Best Params: 'kernel': 'rbf', 'C': 10





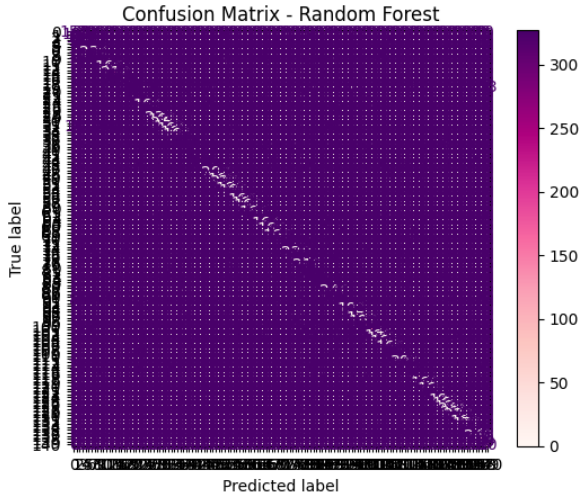
	precision	recall	f1-score	support
0	0.96	1.00	0.98	157
1	0.73	0.76	0.75	164
2	0.91	0.80	0.85	148
3	0.91	0.74	0.82	160
4	0.86	0.97	0.91	164
5	0.66	0.85	0.74	161
6	0.80	0.66	0.72	164
7	0.89	1.00	0.94	234
8	0.90	0.99	0.94	152
9	0.73	0.66	0.69	164
10	0.63	0.85	0.73	164
11	0.66	0.83	0.74	144
12	1.00	1.00	1.00	166
13	0.84	0.96	0.89	164
14	0.95	1.00	0.97	219
15	0.99	0.88	0.93	164
16	0.90	0.92	0.91	143
17	0.98	0.96	0.97	166
18	0.92	0.72	0.81	166
19	0.71	0.75	0.73	152
20	0.78	0.60	0.68	166
...			0.82	23619
accuracy	0.89	23619	0.81	23619
macro avg	0.90	0.89	0.89	23619
weighted avg	0.90	0.89	0.89	23619

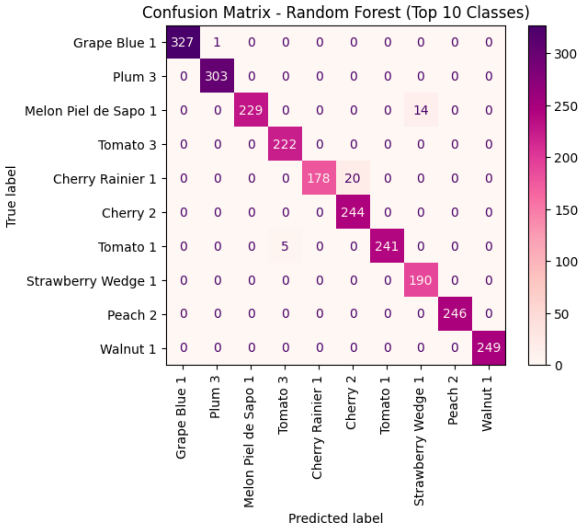
	precision	recall	f1-score	support
Grape Blue 1	1.00	1.00	1.00	328
Plum 3	1.00	1.00	1.00	304
Melon Piel de Sapo 1	1.00	1.00	1.00	246
Tomato 3	0.97	0.99	0.98	246
Cherry Rainier 1	1.00	0.74	0.85	246
Cherry 2	0.85	1.00	0.92	246
Tomato 1	1.00	0.97	0.98	246
Strawberry Wedge 1	1.00	0.87	0.93	246
Peach 2	1.00	1.00	1.00	246
Walnut 1	1.00	1.00	1.00	249
micro avg	0.98	0.96	0.97	2603
macro avg	0.98	0.96	0.97	2603
weighted avg	0.98	0.96	0.97	2603

Random Forest

Fitting 5 folds for each of 10 candidates, totalling 50 fits.

Best Params: 'n_estimators': 100, 'max_samples': 0.9, 'max_depth': None





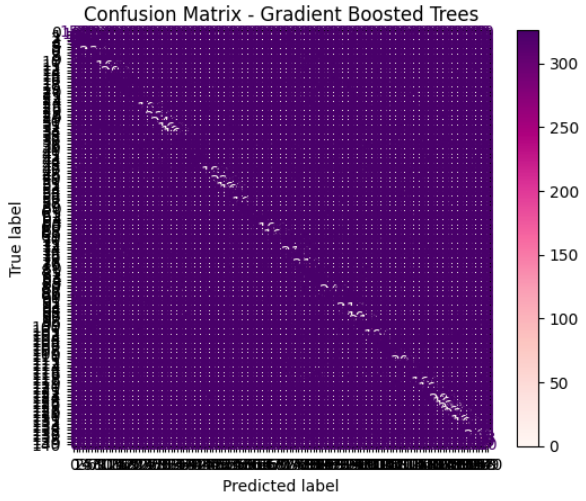
	precision	recall	f1-score	support
0	0.84	1.00	0.91	157
1	0.62	0.66	0.64	164
2	0.77	0.84	0.80	148
3	0.96	0.94	0.95	160
4	0.66	0.95	0.78	164
5	0.63	0.99	0.77	161
6	0.95	0.68	0.79	164
7	0.91	1.00	0.95	234
8	0.77	0.70	0.74	152
9	0.73	0.70	0.71	164
10	0.73	0.76	0.74	164
11	0.79	0.86	0.82	144
12	0.98	1.00	0.99	166
13	0.94	0.77	0.85	164
14	0.85	0.98	0.91	219
15	0.94	0.94	0.94	164
16	0.79	0.97	0.87	143
17	0.91	0.92	0.92	166
18	0.78	0.70	0.74	166
19	0.76	0.65	0.70	152
...	0.78	0.60	0.68	166
accuracy			0.86	23619
macro avg	0.87	0.86	0.86	23619
weighted avg	0.87	0.86	0.86	23619
weighted avg	0.90	0.89	0.89	23619

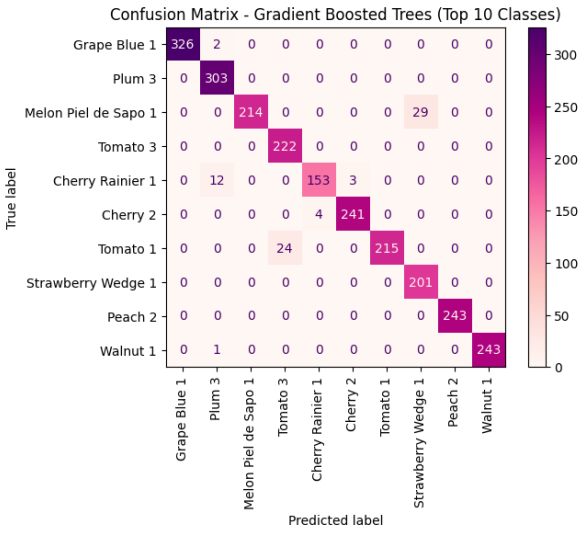
	precision	recall	f1-score	support
Grape Blue 1	1.00	1.00	1.00	328
Plum 3	1.00	1.00	1.00	304
Melon Piel de Sapo 1	1.00	0.93	0.96	246
Tomato 3	0.98	0.90	0.94	246
Cherry Rainier 1	1.00	0.72	0.84	246
Cherry 2	0.92	0.99	0.96	246
Tomato 1	1.00	0.98	0.99	246
Strawberry Wedge 1	0.93	0.77	0.84	246
Peach 2	1.00	1.00	1.00	246
Walnut 1	1.00	1.00	1.00	249
micro avg	0.98	0.93	0.96	2603
macro avg	0.98	0.93	0.95	2603
weighted avg	0.98	0.93	0.96	2603

GradientBoosted Trees

Fitting 5 folds for each of 10 candidates, totalling 50 fits

Best Params: 'n_estimators': 200, 'max_depth': 3, 'learning_rate': 0.1





	precision	recall	f1-score	support
0	0.89	1.00	0.94	157
1	0.64	0.70	0.67	164
2	0.88	0.72	0.79	148
3	0.98	0.78	0.86	160
4	0.79	0.83	0.81	164
5	0.72	1.00	0.83	161
6	0.85	0.71	0.77	164
7	0.97	1.00	0.98	234
8	0.73	0.83	0.78	152
9	0.62	0.65	0.63	164
10	0.87	0.90	0.88	164
11	0.88	0.83	0.85	144
12	0.95	0.99	0.97	166
13	0.58	0.81	0.67	164
14	0.92	0.98	0.95	219
15	0.98	0.70	0.82	164
16	0.82	0.90	0.85	143
17	0.85	0.98	0.91	166
18	0.75	0.65	0.70	166
19	0.74	0.88	0.80	152
...				
accuracy			0.86	23619
macro avg	0.87	0.86	0.86	23619
weighted avg	0.87	0.86	0.86	23619

	precision	recall	f1-score	support
Grape Blue 1	1.00	0.99	0.99	328
Plum 3	0.99	1.00	0.99	304
Melon Piel de Sapo 1	1.00	0.89	0.94	246
Tomato 3	0.98	0.89	0.93	246
Cherry Rainier 1	0.98	0.65	0.78	246
Cherry 2	1.00	0.98	0.99	246
Tomato 1	1.00	0.98	0.99	246
Strawberry Wedge 1	0.90	0.76	0.83	246
Peach 2	1.00	1.00	1.00	246
Walnut 1	1.00	1.00	1.00	249
micro avg	0.99	0.92	0.95	2603
macro avg	0.98	0.91	0.94	2603
weighted avg	0.98	0.92	0.95	2603

Concluzii

Random Forest și Gradient Boosted Trees au obținut cea mai mare acuratețe pe setul Fruits-360 datorită capacității lor de a modela relații complexe între feature-uri, rezistenței la zgomot și outlieri, și concentrării pe cele mai relevante caracteristici. Random Forest oferă robusteză prin diversitatea arborilor săi, iar Gradient Boosted Trees optimizează iterativ performanța, gestionând eficient clase similare. Setul de date Fruits-360 a favorizat performanța acestor metode, care sunt ideale pentru date complexe cu relații non-lineare.

Setul de date Fashion-MNIST a obținut performanțe bune cu SVM și Gradient Boosted Trees datorită separabilității relativ bune între clase și dimensiunii moderate a datelor, care a permis ambelor metode să fie eficiente din punct de vedere computațional. SVM a reușit să separe eficient clasele utilizând kernel-ul adecvat, în timp ce Gradient Boosted Trees a optimizat erorile iterativ, având o performanță robustă, mai ales în fața unor clase vizual similare (de exemplu, tricouri și rochii). Astfel, ambele metode au gestionat bine variabilitatea și complexitatea setului de date.