Tema Învățare Automată

Partea 1

Scopul temei este familiarizarea cu pașii principali dintr-un proiect de machine learning aplicat pe clasificarea imaginilor, și anume:

1. **Analiza exploratorie a datelor (EDA):** înțelegerea și vizualizarea datelor pentru a identifica tipare sau dificultăți.
2. **Extragerea de atribute (Feature Extraction):** prelucrarea imaginilor pentru a le transforma într-un format utilizabil de modele.
3. **Evaluarea modelelor:** antrenarea, compararea și alegerea celui mai bun model pentru clasificarea imaginilor.

Tema folosește două seturi de date (Fashion-MNIST și Fruits-360) pentru a lucra practic la clasificarea corectă a imaginilor, atat in format grayscale, cat si color.

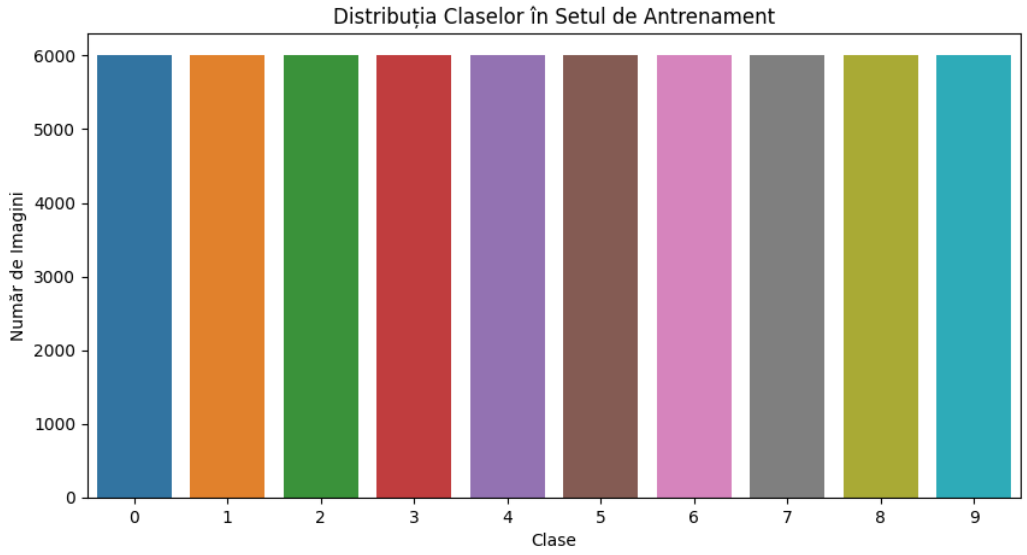
FASHION-MNIST

In urma extragerii datelor am obtinut urmatoarele dimensiuni:

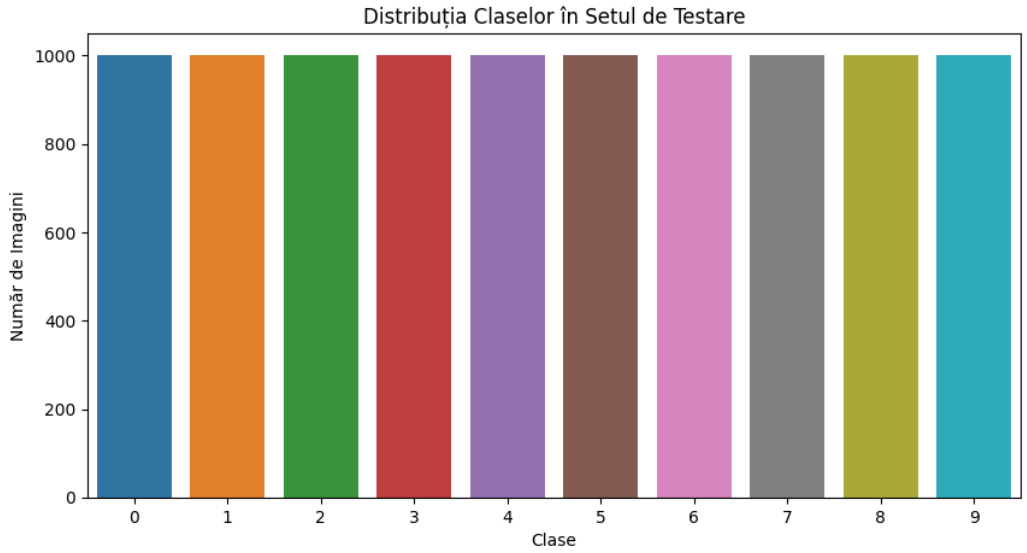
```
Dimensiuni imagini de antrenament: (60000, 784)
Dimensiuni etichete de antrenament: (60000,)
```

Mai jos se poate observa distributia imaginilor pe clase :

- Setul de training



- Setul de test

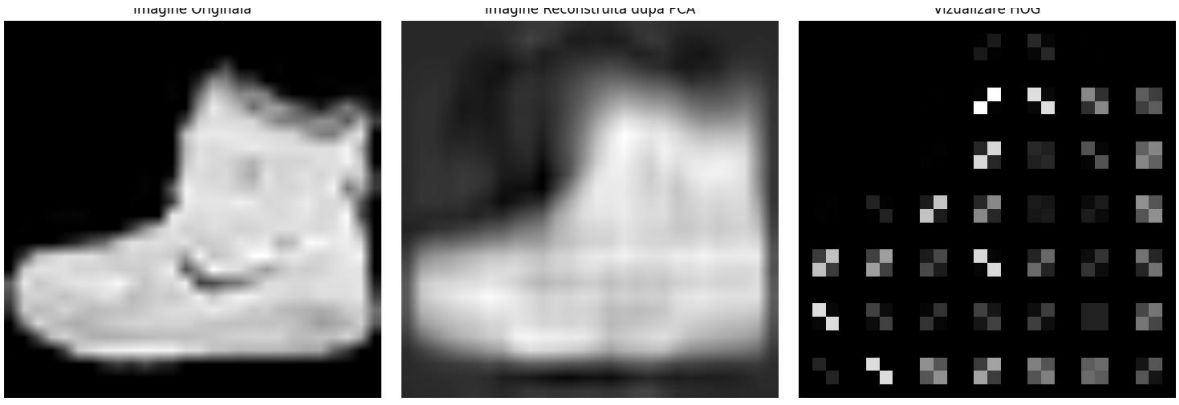


Se poate observa că distribuția imaginilor pe clase este echilibrată, fiecare clasă având aproximativ același număr de exemple (aproximativ 1000 de imagini). Acest echilibru este benefic pentru antrenarea unui model de clasificare, deoarece permite algoritmului să trateze fiecare clasă în mod egal, fără să favorizeze o anumită clasă în detrimentul altora.

Un set de date echilibrat ajută la:

- **Reducerea bias-ului de clasificare:** Modelul nu va învăța să acorde o importanță disproporționată claselor mai numeroase.
- **Îmbunătățirea performanței:** Metricile de evaluare, precum acuratețea, precizia și F1-score-ul, vor reflecta mai bine capacitatea reală a modelului de a clasifica corect toate categoriile.

Extragerea caracteristicilor



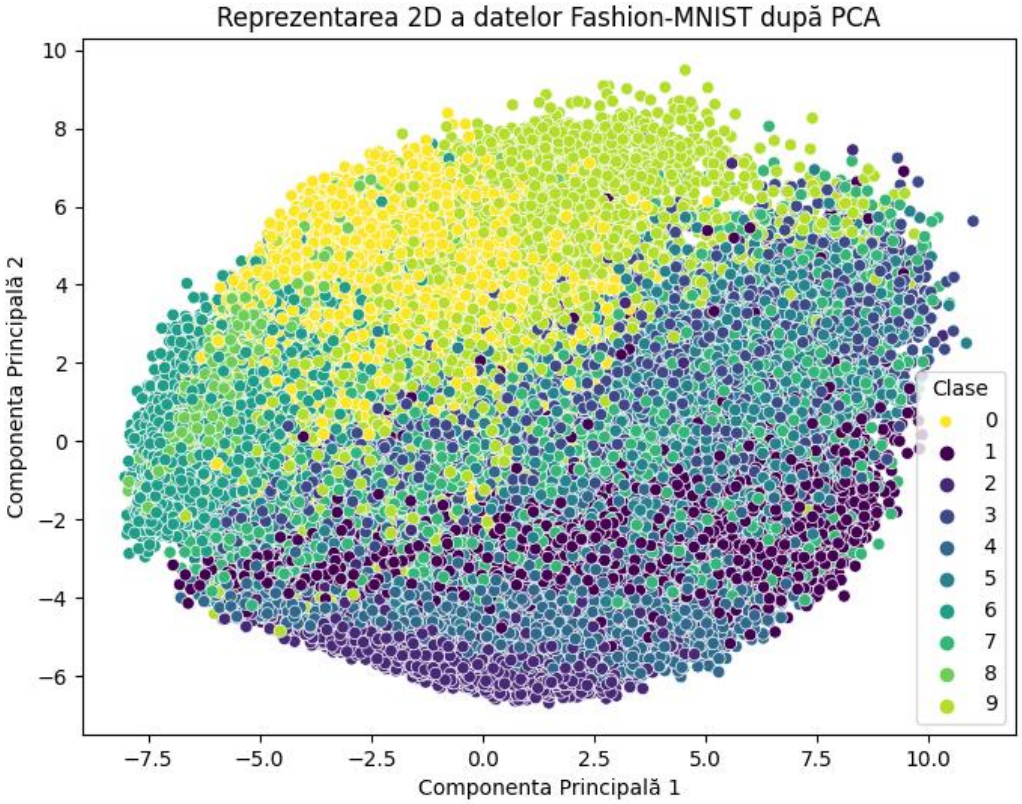
Am reprezentat o comparație între imaginea originală, cea obținută în urma aplicării PCA și cea obținută prin HOG.

Am aplicat aceste metode deoarece:

- **PCA (Principal Component Analysis):**

- Este folosit pentru reducerea dimensionalității datelor, eliminând informațiile redundante și păstrând caracteristicile esențiale ale imaginii. Această tehnică ajută la simplificarea datelor, ceea ce face modelele de machine learning mai eficiente și mai rapide în procesare, reducând complexitatea fără a pierde semnificativ informația vizuală.

- **HOG (Histogram of Oriented Gradients):**
  - Este o tehnică de extragere a caracteristicilor care pune accent pe contururile și marginile obiectelor din imagine. Am utilizat HOG pentru a evidenția structura și forma imaginii, deoarece aceste elemente sunt relevante în problemele de clasificare, oferind o reprezentare mai compactă și specifică a datelor vizuale.



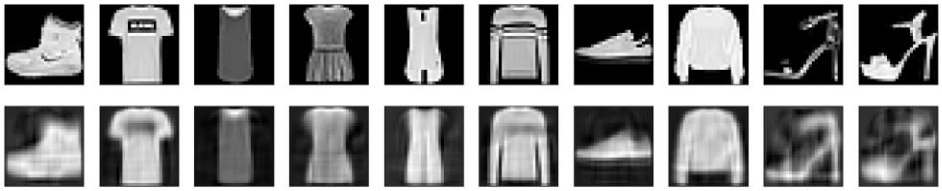
Se poate observa o reprezentare 2D a datelor Fashion-MNIST obținută prin reducerea dimensionalității cu PCA (Principal Component Analysis). Aceasta proiectează datele inițiale într-un spațiu bidimensional utilizând cele mai relevante două componente principale.

Din grafic se poate observa că:

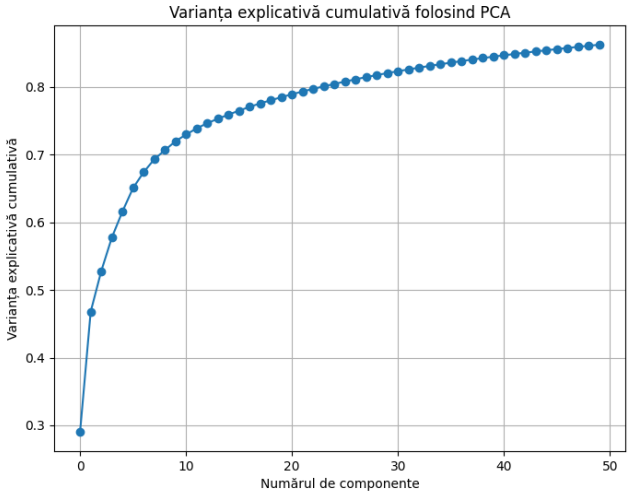
- Fiecare punct reprezintă o imagine din setul de date, iar culorile diferite indică clasele distincte (de exemplu, tipurile de articole vestimentare).
- Anumite clase sunt relativ bine separate în spațiul bidimensional (de exemplu, clasele 0 și 8 par să formeze grupuri distincte), în timp ce alte clase prezintă o suprapunere (de exemplu, clasele 3 și 5).

- Această distribuție sugerează că, deși PCA ajută la vizualizarea structurii datelor și a relațiilor dintre clase, pot exista clase care sunt dificil de separat doar pe baza a două componente principale.

Cateva imagini obtinute in urma aplicarii PCA



Varianta cumulativă



- La început, curba crește rapid, indicând că primele componente principale captează cea mai mare parte a variației din date.
- După un anumit punct (aproximativ 20-30 de componente), creșterea varianței explicative se diminuează, ceea ce sugerează că adăugarea mai multor componente contribuie foarte puțin la explicarea suplimentară a datelor.

```
Varianța explicativă cumulativă:
 [0.29039446 0.46794903 0.52814174 0.5777164  0.6161933  0.65080124
 0.67421836 0.69327265 0.7067712  0.71991396 0.7298428  0.7389754
 0.7466334  0.7532305  0.75930905 0.7652073  0.7707267  0.7759717
 0.7805525  0.7851077  0.78944457 0.79351556 0.79736316 0.8010887
 0.8046998  0.8082064  0.8115388  0.81473243 0.8178131  0.8207458
 0.8235018  0.82615244 0.8287895  0.83134764 0.8338098  0.8361687
 0.8384711  0.84072626 0.8429105  0.8450035  0.84700596 0.84896314
 0.85090965 0.852736   0.8544843  0.8562015  0.8578878  0.8595168
 0.8611127  0.8626583 ]
```

- **Primele 10 componente principale** explică deja o mare parte din varianță (aproximativ 70%).
- Până la **primele 20 de componente principale**, varianța explicativă cumulativă depășește 80%, ceea ce arată că informația relevantă din date poate fi păstrată folosind doar aceste componente, fără pierderi semnificative.
- Pe baza acestui grafic, se poate concluziona că nu este necesar să se utilizeze toate componentele pentru a reprezenta datele

Standardizarea datelor

```
Dimensiuni imagini de antrenament standardizate: (60000, 784)
Dimensiuni imagini de testare standardizate: (10000, 784)
Media datelor de antrenament: -2.2966035e-10
Deviația standard a datelor de antrenament: 1.0000007
```

Media datelor de antrenament:

- Media calculată este foarte aproape de 0 (-2.2966035e-10 este practic zero), ceea ce indică faptul că standardizarea a fost aplicată corect.

Deviația standard a datelor de antrenament:

- Deviația standard este aproape exact 1 (1.000007), confirmând că datele au fost scalate corespunzător.

Codul utilizat:

- StandardScaler este utilizat pentru a ajusta datele, ceea ce este important în special pentru algoritmi de machine learning care sunt sensibili la scala datelor (cum ar fi regresia logistică, SVM sau rețele neuronale).

Selecția percentilă

```
Dimensiuni imagini de antrenament după selecția percentilă: (60000, 79)
Dimensiuni imagini de testare după selecția percentilă: (10000, 79)
Numărul de atribute după selecția percentilă: 79
```

Selecția percentilă este o metodă de reducere a dimensionalității datelor, utilizată pentru a selecta doar atributele cele mai relevante în funcție de un criteriu specific (cum ar fi scorurile calculate pe baza relației dintre atribute și etichetele de clasă).

Observații din rezultatul tău:

1. Dimensiuni inițiale și după selecție:
    - Inițial: Seturile de antrenament și test aveau 784 de atribute (corespunzând pixelilor imaginii).
    - După selecție: Numărul de atribute a fost redus la 79, ceea ce indică faptul că doar aceste 79 de atribute au fost considerate cele mai relevante pentru problemă.

Algoritmi de clasificare

1. Logistic Regression

```
Best parameters for Logistic Regression: {'logisticregression__C': 0.23357214690901212, 'logisticregression__multi_class': 'multinomial'}
Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.70      0.72      1202
           1       0.79      0.85      0.82      1219
           2       0.55      0.61      0.58      1205
           3       0.71      0.67      0.69      1184
           4       0.57      0.61      0.59      1202
           5       0.74      0.76      0.75      1211
           6       0.50      0.41      0.45      1218
           7       0.72      0.68      0.70      1159
           8       0.88      0.88      0.88      1197
           9       0.72      0.76      0.74      1203

    accuracy                           0.69     12000
   macro avg       0.69      0.69      0.69     12000
weighted avg       0.69      0.69      0.69     12000

Accuracy Score: 0.6928333333333333
```

 Performanță pe clase:

- Clasele cu performanță bună:
    - Clasa **8** are cele mai bune valori (precizie, recall, F1-score: 0.88).
    - Clasele **1** și **5** sunt clasificate bine, având valori ridicate pentru precizie și recall.
- Clasele cu performanță slabă:
    - Clasa **3** are un F1-score scăzut (0.59), ceea ce indică dificultăți în identificare.
    - Clasa **4** are cea mai slabă performanță (F1-score 0.45), probabil din cauza suprapunerilor sau lipsei de date reprezentative.

 Metrici globale:

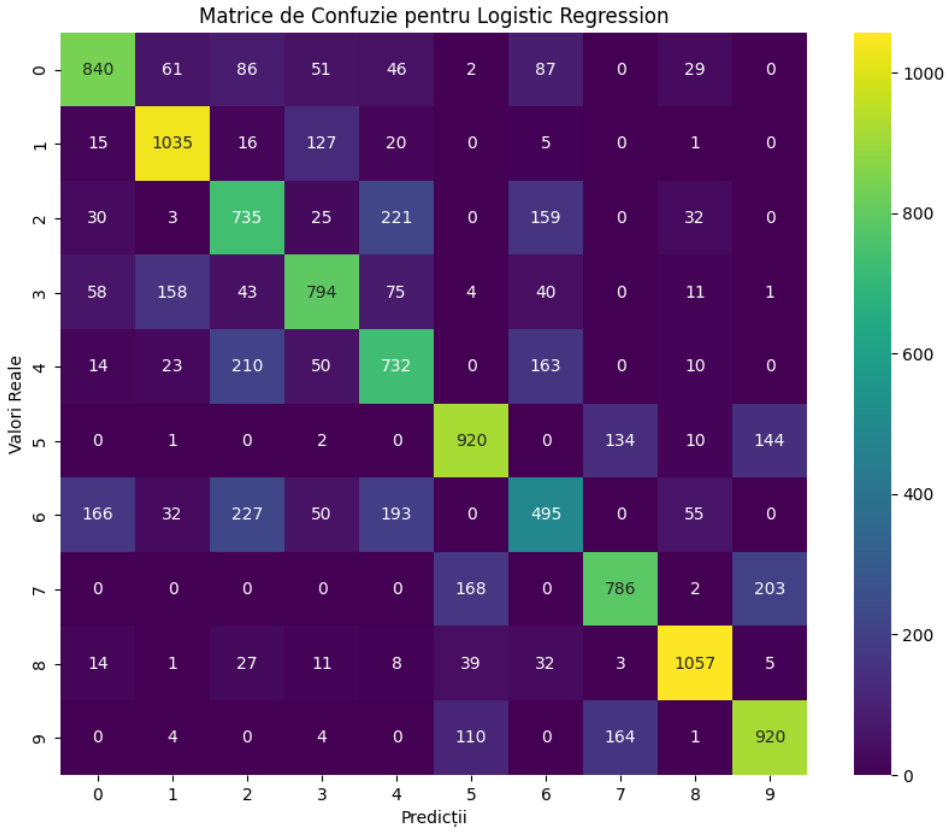
- **Acuratețea:** 69%, adică aproximativ 69% dintre predicții sunt corecte.
- **Macro Avg:** 0.69 (media aritmetică a metricilor, uniform pe toate clasele).
- **Weighted Avg:** 0.69 (media metricilor ponderată în funcție de dimensiunea fiecărei clase).

 Hiperparametrii utilizați:

- **C:** Regularizare de 0.2335, prevenind overfitting-ul.
- **multi_class='multinomial':** Optimizează modelul pentru toate clasele simultan.

Matricea de confuzie

Matricea de confuzie evidențiază punctele forte ale modelului (clasele 1, 5, 8) și punctele slabe (confuziile între clasele 3, 4, 6 și altele). Pentru îmbunătățirea performanței, s-ar putea analiza mai detaliat caracteristicile utilizate sau chiar îmbunătăți setul de date prin adăugarea de exemple mai reprezentative pentru clasele care generează confuzii.



2. SVM

```
Best parameters for SVM: {'svc__C': 100.0, 'svc__kernel': 'rbf'}
Best score for SVM: 0.7729791666666667
Classification Report:
              precision    recall  f1-score   support

           0       0.72      0.76      0.74      1202
           1       0.89      0.91      0.90      1219
           2       0.66      0.74      0.70      1205
           3       0.82      0.79      0.80      1184
           4       0.68      0.71      0.70      1202
           5       0.87      0.83      0.85      1211
           6       0.61      0.52      0.56      1218
           7       0.82      0.75      0.78      1159
           8       0.95      0.92      0.93      1197
           9       0.78      0.87      0.82      1203

    accuracy                           0.78     12000
   macro avg       0.78      0.78      0.78     12000
weighted avg       0.78      0.78      0.78     12000

Accuracy Score: 0.7793333333333333
```

 Performanță pe clase:

- **Clase bine clasificate:** Clasele **8**, **1**, și **5** au cele mai bune valori (F1-score de 0.93, 0.90 și 0.85).
- **Clase slabe:** Clasa **6** are cea mai slabă performanță (F1-score 0.56), iar clasele **2** și **4** au valori medii (F1-score 0.70).

 Metrici globale:

- **Acuratețea generală:** 78%.
- **Macro Avg și Weighted Avg:** Ambele sunt 0.78, indicând o performanță consistentă între clase.

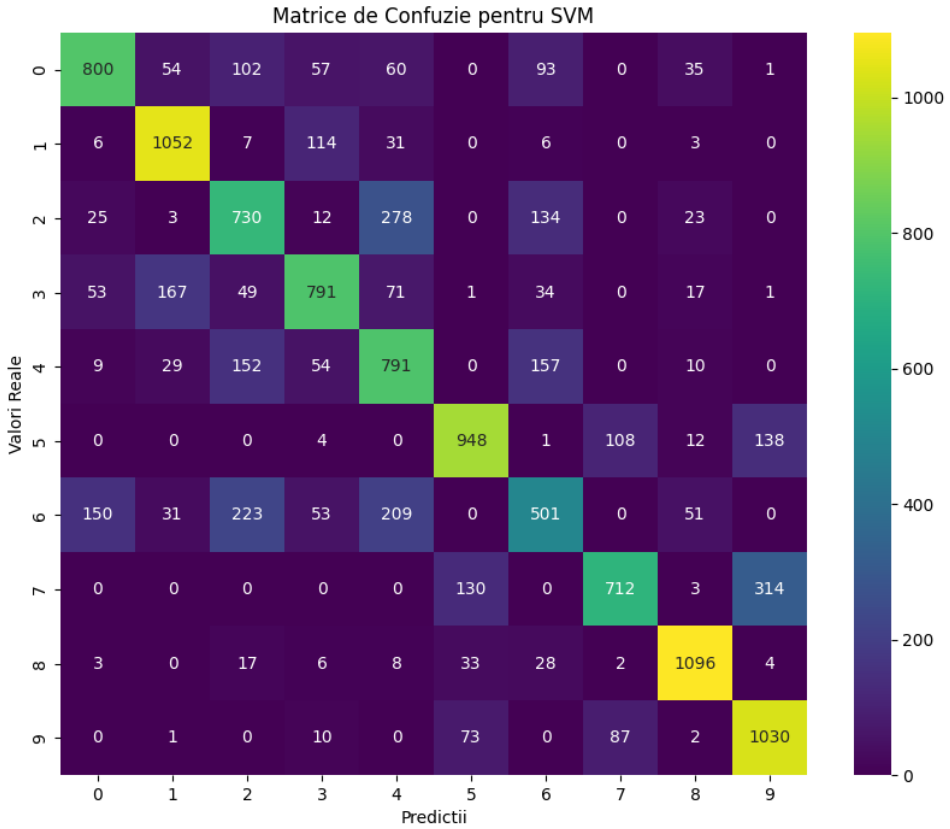
 Parametrii utilizați:

- Kernel RBF și C=100.0C = 100.0C=100.0 (permite clasificare neliniară și o potrivire mai strictă cu datele).

Matricea de confuzie

 Modelul SVM performează bine pentru clase precum 1, 5, și 8, unde erorile sunt minime.

 Clasele 6 și 2 au performanță slabă, necesitând îmbunătățiri în extragerea caracteristicilor sau separarea datelor.



3. Random Forest

```
Best parameters: {'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 4, 'n_estimators': 100}
Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.76      0.76      1202
           1       0.91      0.89      0.90      1219
           2       0.65      0.75      0.70      1205
           3       0.80      0.80      0.80      1184
           4       0.65      0.72      0.68      1202
           5       0.88      0.81      0.84      1211
           6       0.60      0.46      0.52      1218
           7       0.82      0.83      0.82      1159
           8       0.92      0.95      0.94      1197
           9       0.82      0.86      0.84      1203

    accuracy                           0.78     12000
   macro avg       0.78      0.78      0.78     12000
weighted avg       0.78      0.78      0.78     12000

Accuracy Score: 0.78175
```

 Performanță pe clase:

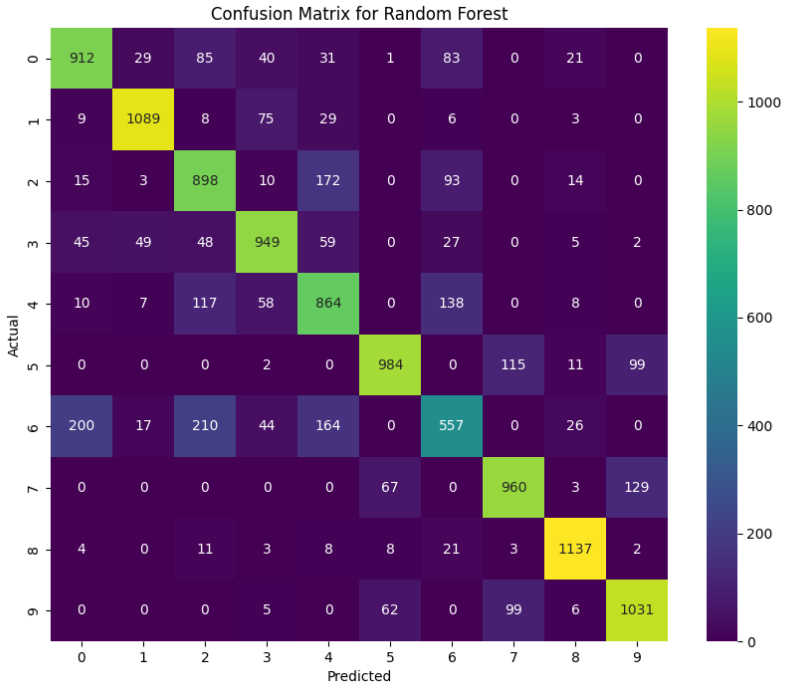
- Clase bine clasificate: **Clasa 1** (F1-score 0.90), **Clasa 8** (F1-score 0.92), și **Clasa 5** (F1-score 0.84).
- Clase problematice: **Clasa 6** (F1-score 0.56) și **Clasa 4** (F1-score 0.66).

 Metrici globale:

- **Acuratețea:** 78.17%.
- **Macro Avg și Weighted Avg:** Ambele 0.78, indicând o performanță echilibrată.

 Parametri: Folosește 100 arbori, fără limită de adâncime, cu un minim de 2 exemple pe frunză și 4 pentru split.

Matricea de confuzie



Random Forest performează bine pentru clase precum **1**, **8**, și **5**, însă clasele **6** și **4** au confuzii majore cu alte clase. Aceste confuzii sugerează necesitatea unei analize suplimentare a datelor sau optimizarea parametrilor modelului.

4. GradientBoosted Trees

```
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.77      0.78      1202
              precision    recall  f1-score   support

           0       0.78      0.77      0.78      1202

           0       0.78      0.77      0.78      1202
           1       0.89      0.91      0.90      1219
           2       0.64      0.73      0.68      1205
           0       0.78      0.77      0.78      1202
           1       0.89      0.91      0.90      1219
           2       0.64      0.73      0.68      1205
           1       0.89      0.91      0.90      1219
           2       0.64      0.73      0.68      1205
           3       0.82      0.81      0.81      1184
           2       0.64      0.73      0.68      1205
           3       0.82      0.81      0.81      1184
           4       0.68      0.71      0.69      1202
           5       0.87      0.83      0.85      1211
           3       0.82      0.81      0.81      1184
           4       0.68      0.71      0.69      1202
           5       0.87      0.83      0.85      1211
```

```
Accuracy Score: 0.7865833333333333
```

 Acuratețe: 78.66%.

 Clase bine clasificate: Clasele **1** (1105 corecte), **8** (1134 corecte), și **5** (1003 corecte).

 Clase problematice: Clasa **6** (592 corecte, mari confuzii cu **0**, **2**, și **4**), clasa **4** (852 corecte, confuzii cu **2** și **6**).

Matricea de confuzie

 Clase bine separate:

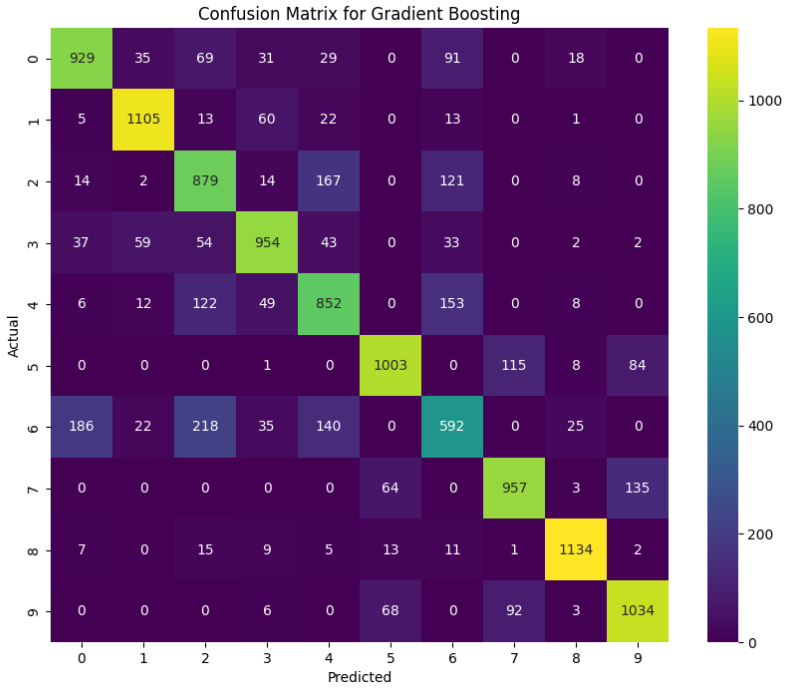
- **Clasa 1:** Are 1105 clasificări corecte și foarte puține erori, ceea ce indică o separare clară.
- **Clasa 8:** Are 1134 clasificări corecte, fiind foarte bine diferențiată.
- **Clasa 5:** 1003 clasificări corecte, cu puține erori.

 Clase problematice:

- **Clasa 6:** Este confundată frecvent cu **0** (186 erori), **2** (218 erori), și **4** (140 erori), indicând dificultăți în separarea caracteristicilor.
- **Clasa 4:** Este confundată cu **2** (167 erori) și **6** (153 erori).

 General:

- Majoritatea claselor sunt bine clasificate, dar clasele **6** și **4** prezintă suprapuneri frecvente cu altele.



FRUITS-360

Extragerea datelor

```
Number of training images: 70491
Number of training labels: 70491
Number of test images: 23619
Number of test labels: 23619
```

Din cauza numărului mare de imagini am ales sa lucrez cu imaginile din 40 de clase selectate in mod aleatoriu.

```
Randomly selected classes:
 ['Papaya 1', 'Pear Kaiser 1', 'Granadilla 1', 'Banana Lady Finger 1', 'Walnut 1', 'Onion Red 1', 'Pineapple 1', 'Pear Forelle 1', 'Apricot 1', 'Watermelon 1',
```

```
'Apple Red 2', 'Apple Red Yellow 1', 'Pear Stone 1', 'Carrot 1', 'Tomato Cherry Red 1', 'Zucchini dark 1', 'Mango 1', 'Plum 2', 'Grapefruit Pink 1'
```

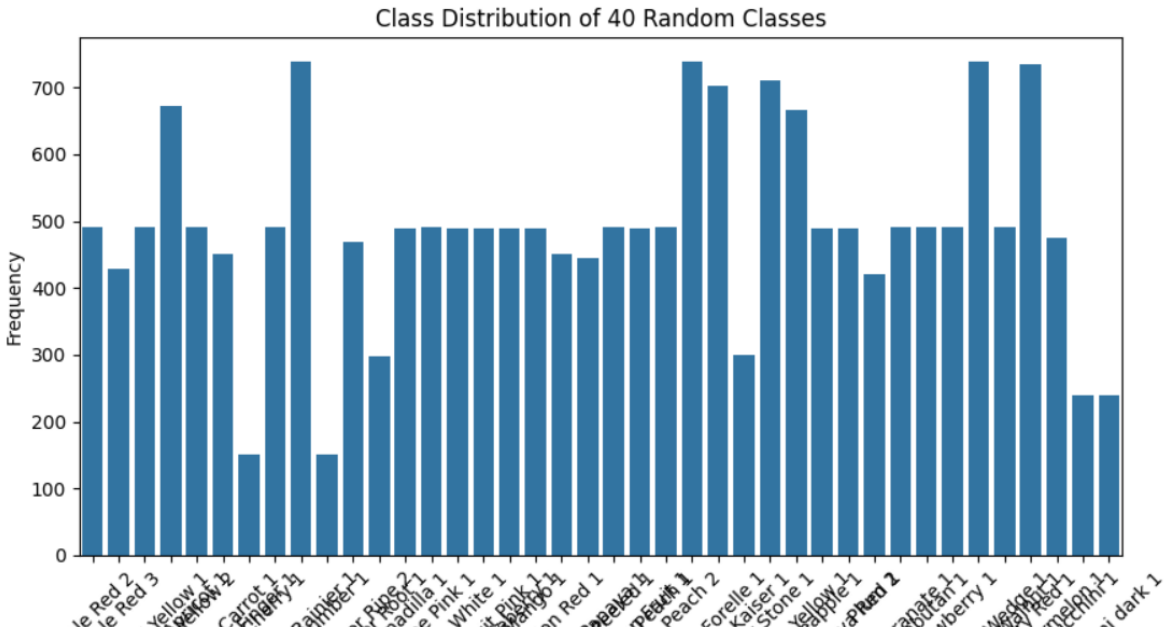
```
'Zucchini 1', 'Cherry 1', 'Cucumber Ripe 2', 'Rambutan 1', 'Grape White 1', 'Pitahaya Red 1', 'Grape Pink 1', 'Passion Fruit 1', 'Apple Red 3', 'Peach 1', 'Ginger Root 1',
```

```
Pepper Yellow 1', 'Peach 2', 'Cucumber 1', 'Strawberry 1', 'Onion Red Peeled 1', 'Strawberry Wedge 1', 'Apple Red Yellow 2', 'Huckleberry 1', 'Cherry Rainier 1', 'Pomegranate 1'
```

In urma selectiei claselor voi lucra cu:

```
Number of training images for selected classes: 19547
Number of test images for selected classes: 6548
Dimensiuni X_train: (19547, 2352)
Dimensiuni X_test: (6548, 2352)
```

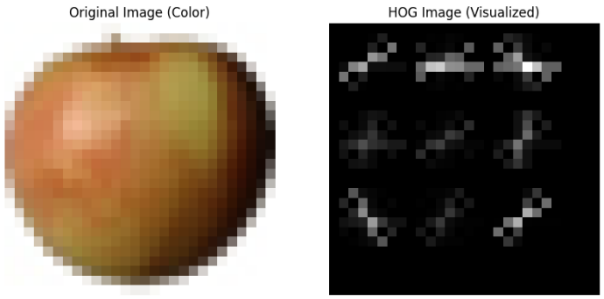
Distributia imaginilor din clasele selectate:



Această distribuție arată un **dezechilibru între clase**, care poate necesita metode de preprocesare, cum ar fi:

- **Reechilibrarea datelor:** Oversampling pentru clasele subreprezentate sau undersampling pentru cele supradimensionate.
- **Ponderarea claselor:** Utilizarea unui model care ajustează importanța claselor în funcție de frecvența lor.

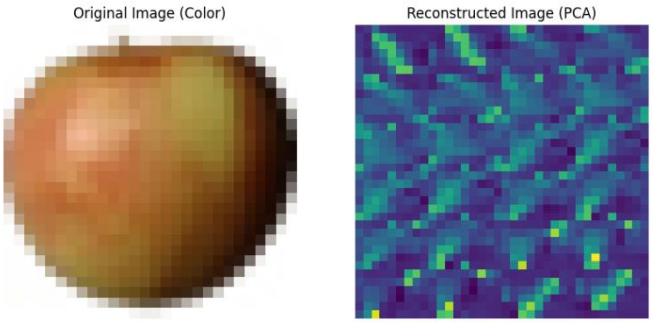
Extragerea caracteristicilor



Am utilizat tot metodele PCA si HOG si pentru acest set de date

Imaginea HOG (Histogram of Oriented Gradients):

- Vizualizarea HOG extrage contururile și gradientele principale din imagine, ignorând informațiile legate de culoare.
- Este o reprezentare utilizată pentru clasificare, punând accent pe structura și marginea obiectelor din imagine.



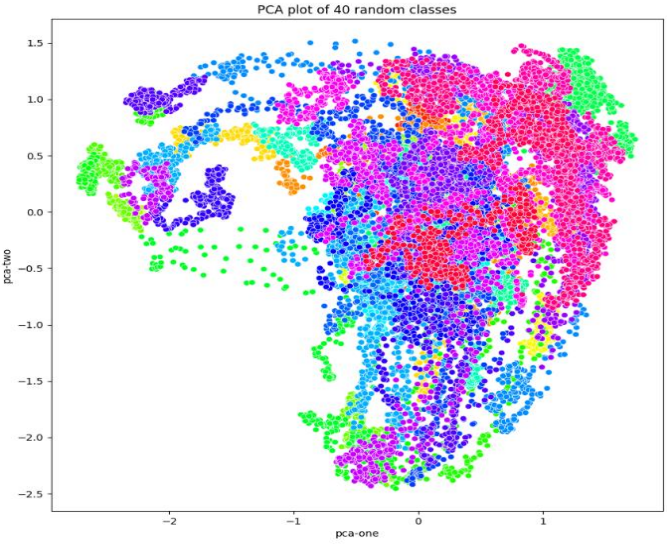
Imaginea reconstruită prin PCA (Principal Component Analysis):

- PCA reduce dimensionalitatea datelor, reținând doar componentele esențiale pentru descrierea imaginii.
- Imaginea reconstruită este mai simplă, detaliile complexe fiind eliminate, dar păstrează informații de bază, cum ar fi forma generală și câteva caracteristici de textură.

Acestea sunt noile dimensiuni cu care voi lucra:

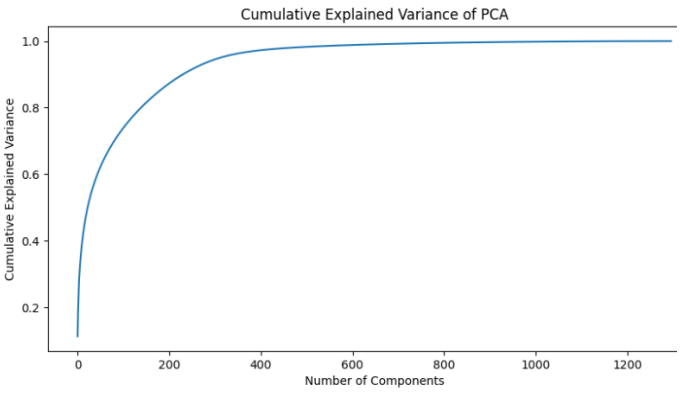
```
Dimensiuni X_train_hog: (19547, 1296)
Dimensiuni X_test_hog: (6548, 1296)
Dimensiuni X_train_pca: (19547, 50)
Dimensiuni X_test_pca: (6548, 50)
```

Reprezentarea 2D a datelor



- Clase bine separate: Unele grupuri de puncte sunt clar delimitate, sugerând că respectivele clase sunt bine separate în spațiul caracteristicilor.
- Clase suprapuse: Există zone în care punctele de culori diferite se suprapun, ceea ce indică dificultăți în separarea acestor clase. Acest lucru poate apărea din cauza similarității între clase sau a unei reprezentări insuficiente a caracteristicilor.

Varianța cumulativă



 Creștere rapidă inițială: Primele 100-200 componente explică o mare parte din varianța totală (aproximativ 80-90%), ceea ce sugerează că aceste componente conțin cele mai relevante informații din date.

 Punctul de saturație: După 400-500 componente, graficul se aplatizează, indicând că adăugarea de componente suplimentare contribuie foarte puțin la explicarea variației.

Standardizarea datelor si selectia percentilă

```
Dimensiunea datelor după standardizare: (14660, 50)
Dimensiunea datelor după standardizare și eliminarea variației: (14660, 5)
```

- După standardizare, dimensiunea datelor este (14660, 50).
- Standardizarea transformă fiecare caracteristică astfel încât să aibă o medie de 0 și o deviație standard de 1, ceea ce:
    - Îmbunătățește performanța algoritmilor sensibili la scala datelor (de exemplu, regresia logistică, SVM).
    - Asigură că toate caracteristicile contribuie în mod egal la procesul de învățare.
- După eliminarea caracteristicilor cu variață scăzută, dimensiunea datelor este redusă la (14660, 5).
- Selectarea percentilei reține doar caracteristicile cele mai relevante, pe baza variației acestora. Caracteristicile cu variață mică sunt mai puțin informative pentru modelele de machine learning.
- Această reducere ajută la:
    - Reducerea dimensionalității setului de date.
    - Îmbunătățirea eficienței algoritmului prin eliminarea redundanței și a informației irelevante.

Algoritmii de clasificare

1. Logistic Regression

```
              accuracy                           1.00      4398
             macro avg       1.00      1.00      1.00      4398
          weighted avg       1.00      1.00      1.00      4398

Accuracy Score: 0.9965893587994543
```

```
Best parameters for Logistic Regression: {'logisticregression__C': np.float64(4.281332398719396), 'logisticregression__multi_class': 'multinomial'}
Classification Report:
                     precision    recall  f1-score   support

        Apple Red 2       1.00      1.00      1.00       106
        Apple Red 3       0.98      0.99      0.98        94
  Apple Red Yellow 1       0.99      1.00      1.00       112
  Apple Red Yellow 2       1.00      1.00      1.00       153
          Apricot 1       1.00      1.00      1.00       112
Banana Lady Finger 1       1.00      1.00      1.00       103
           Carrot 1       1.00      1.00      1.00        38
           Cherry 1       1.00      1.00      1.00       107
    Cherry Rainier 1       1.00      1.00      1.00       167
         Cucumber 1       1.00      1.00      1.00        37
     Cucumber Ripe 2       1.00      1.00      1.00       104
       Ginger Root 1       1.00      1.00      1.00        65
       Granadilla 1       0.99      0.99      0.99       108
        Grape Pink 1       1.00      1.00      1.00       108
       Grape White 1       1.00      1.00      1.00       108
   Grapefruit Pink 1       1.00      1.00      1.00       114
      Huckleberry 1       1.00      1.00      1.00       113
            Mango 1       0.99      1.00      1.00       107
         Onion Red 1       0.95      0.97      0.96       103
  Onion Red Peeled 1       0.98      0.98      0.98        99
```

 Reducerea numărului de clase:

- Modelul a fost antrenat și evaluat pe un subset redus de clase, ceea ce a contribuit la îmbunătățirea performanței. Reducerea claselor a făcut problema mai simplă, eliminând posibile confuzii între clasele similare.

 Performanță globală:

- Accuracy: 99.66%, ceea ce indică un model extrem de performant, cu erori minime.
- Macro avg: 1.00 pentru precizie, recall și F1-score, ceea ce arată că modelul clasifică bine fiecare clasă, indiferent de dimensiunea sa.
- Weighted avg: 1.00, ceea ce confirmă performanța consistentă pe întregul set.

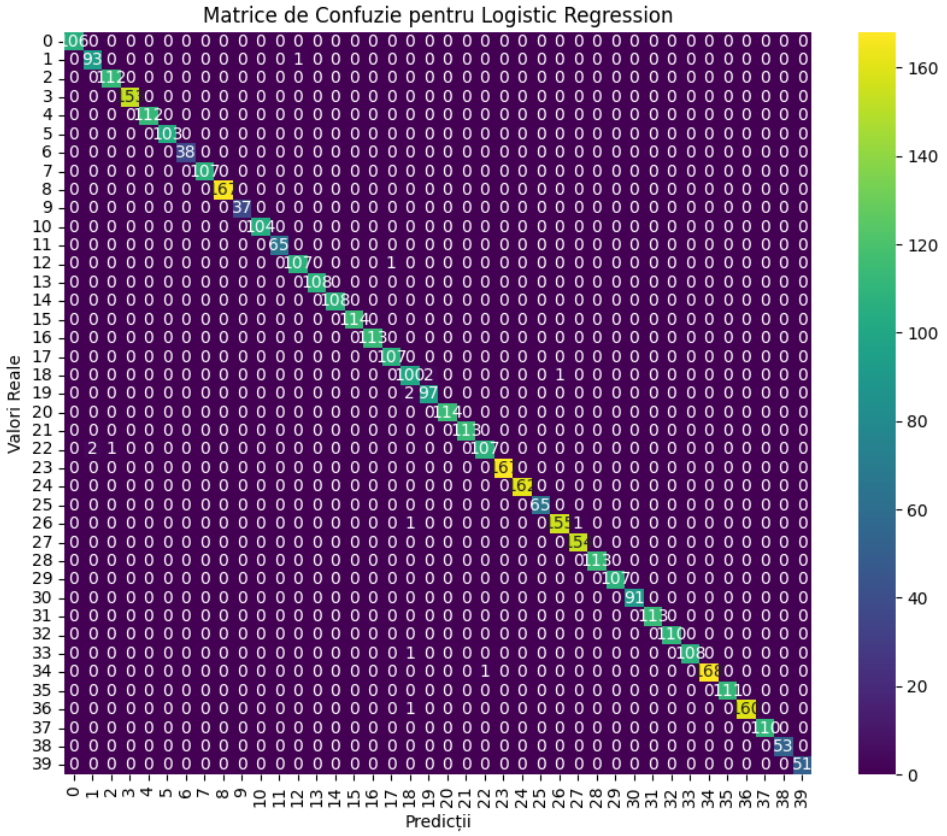
 Performanță pe clase:

- Cele mai multe clase au precizie, recall și F1-score de 1.00, ceea ce sugerează că modelul a clasificat corect toate instanțele acestora.
- Clase cu performanțe ușor mai scăzute (dar excelente):
    - Apple Red 3: F1-score 0.98.
    - Onion Red: F1-score 0.96.
    - Granadilla 1: F1-score 0.99.
- Doar câteva instanțe au fost clasificate greșit în aceste clase.

 Parametrii modelului:

- C = 4.28: Regularizare optimizată pentru a preveni underfitting sau overfitting.
- Multinomial: Configurarea multinomială este adecvată pentru clasificarea multi-clasă.

Matricea de confuzie



Observăm câteva erori izolate, foarte mici, cum ar fi:

- Clasa 19: 2 exemple sunt confundate cu alte clase.
- Clasa 24: O eroare minoră în afara diagonalei.

Aceste erori sunt nesemnificative în raport cu dimensiunea totală a datelor.

2. SVM

```
Best parameters for SVM: {'svc__C': np.float64(10.0), 'svc__kernel': 'rbf'}
Best score for SVM: 0.9991229779603447
Classification Report:
                      precision    recall  f1-score   support

        Apple Red 2       1.00      1.00      1.00       106
        Apple Red 3       1.00      1.00      1.00        94
  Apple Red Yellow 1       1.00      1.00      1.00       112
  Apple Red Yellow 2       1.00      1.00      1.00       153
          Apricot 1       1.00      1.00      1.00       112
Banana Lady Finger 1       1.00      1.00      1.00       103
           Carrot 1       1.00      1.00      1.00        38
           Cherry 1       1.00      1.00      1.00       107
    Cherry Rainier 1       1.00      1.00      1.00       167
         Cucumber 1       1.00      1.00      1.00        37
     Cucumber Ripe 2       1.00      1.00      1.00       104
       Ginger Root 1       1.00      1.00      1.00        65
        Granadilla 1       1.00      1.00      1.00       108
        Grape Pink 1       1.00      1.00      1.00       108
```

Zucchini 1	1.00	1.00	1.00	53
Zucchini dark 1	1.00	1.00	1.00	51
accuracy			1.00	4398
macro avg	1.00	1.00	1.00	4398
weighted avg	1.00	1.00	1.00	4398

Accuracy Score: 1.0

Performanță globală:

- Accuracy: 100%. Toate instanțele din setul de testare au fost clasificate corect, fără erori.
- Macro avg și Weighted avg: 1.00 pentru precizie, recall și F1-score, ceea ce arată că modelul funcționează perfect pe toate clasele, indiferent de dimensiunea acestora.

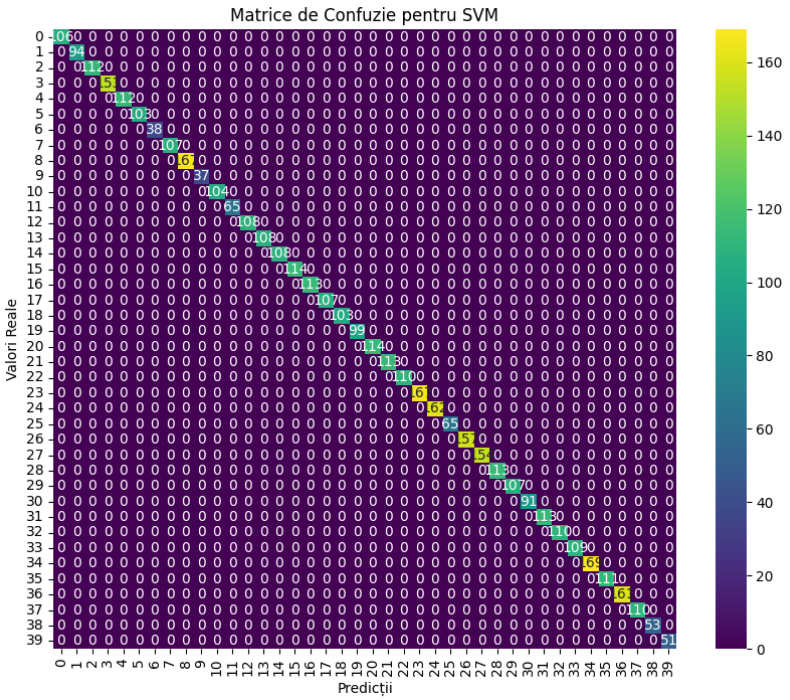
Performanță pe clase:

- Fiecare clasă are precizie, recall și F1-score de 1.00, ceea ce indică o clasificare corectă pentru fiecare instanță.
- Nu există confuzii între clase, ceea ce confirmă o separare clară a datelor.

Parametrii modelului:

- C = 10.0: Valoare optimizată a parametrului de regularizare, care permite modelului să fie suficient de flexibil pentru a captura toate detaliile relevante, dar fără a supraantrena.
- Kernel RBF: Kernelul Radial Basis Function este adecvat pentru seturi de date complexe, deoarece permite modelului să gestioneze relații neliniare între caracteristici.

Matricea de confuzie



Matricea de confuzie confirmă performanța perfectă a SVM. Toate exemplele din setul de testare au fost clasificate corect, fără erori. Modelul este ideal pentru acest subset de date și demonstrează o separare excelentă între aceste clase.

3. Random Forest

```
Best parameters: {'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 4, 'n_estimators': 100}
Classification Report:
                     precision    recall  f1-score   support

        Apple Red 2       1.00      1.00      1.00       106
        Apple Red 3       0.99      1.00      0.99        94
  Apple Red Yellow 1       1.00      1.00      1.00       112
  Apple Red Yellow 2       1.00      1.00      1.00       153
          Apricot 1       0.99      1.00      1.00       112
Banana Lady Finger 1       1.00      0.98      0.99       103
           Carrot 1       1.00      1.00      1.00        38
```

```
          Zucchini 1       1.00      1.00      1.00        53
     Zucchini dark 1       1.00      1.00      1.00        51

           accuracy                           1.00      4398
          macro avg       1.00      1.00      1.00      4398
       weighted avg       1.00      1.00      1.00      4398

Accuracy Score: 0.9956798544793087
```

Performanță globală:

- **Acuratețe:** 99.57%, indicând o performanță aproape perfectă.
- **Macro avg și Weighted avg:** 1.00 pentru precizie, recall și F1-score, ceea ce reflectă o clasificare excelentă pentru toate clasele, indiferent de dimensiunea lor.

Performanță pe clase:

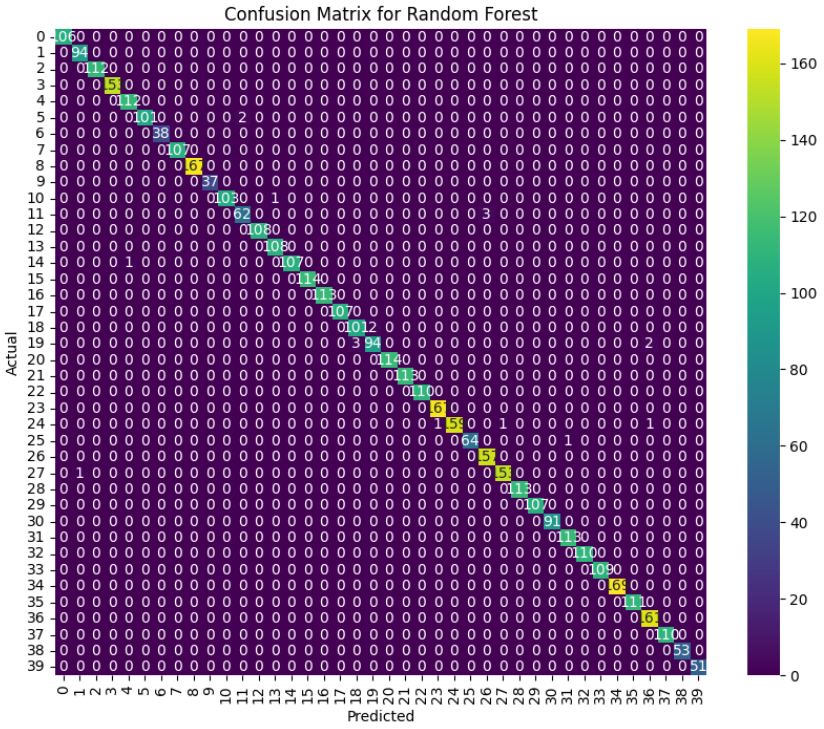
- Majoritatea claselor au **precizie, recall și F1-score de 1.00**, ceea ce arată că modelul a clasificat corect toate instanțele.
- Clasele cu performanțe ușor mai scăzute:
    - **Apple Red 3:** F1-score 0.99 (o eroare minoră).
    - **Banana Lady Finger 1:** F1-score 0.99 (o eroare izolată).

Parametrii modelului:

- **max_depth=None:** Adâncimea arborilor nu este limitată, ceea ce permite modelului să învețe toate detaliile din date.
- **min_samples_leaf=2:** Fiecare frunză trebuie să conțină cel puțin 2 instanțe, prevenind overfitting-ul.
- **min_samples_split=4:** Un nod trebuie să aibă cel puțin 4 instanțe pentru a fi împărțit.

- **n_estimators=100:** Modelul folosește 100 de arbori de decizie pentru a îmbunătăți robustețea și performanța.

Matricea de confuzie



Majoritatea claselor au toate instanțele clasificate corect (valorile pe diagonală corespund dimensiunii reale a claselor).

Clasa 1: 2 instanțe sunt clasificate greșit.

Clasa 23: O instanță este clasificată greșit.

Aceste erori sunt rare și nesemnificative, având un impact minim asupra performanței globale.

4. GradientBoosted Trees

```
Classification Report:
                      precision    recall  f1-score   support

         Apple Red 2       1.00      1.00      1.00       106
         Apple Red 3       0.98      0.99      0.98        94
  Apple Red Yellow 1       0.99      1.00      1.00       112
  Apple Red Yellow 2       1.00      1.00      1.00       153
           Apricot 1       1.00      1.00      1.00       112
Banana Lady Finger 1       1.00      1.00      1.00       103
            Carrot 1       1.00      1.00      1.00        38
            Cherry 1       1.00      1.00      1.00       107

            accuracy                           1.00      4398
           macro avg       1.00      1.00      1.00      4398
        weighted avg       1.00      1.00      1.00      4398

Accuracy Score: 0.9965893587994543
```

- Majoritatea claselor au **precizie, recall și F1-score de 1.00**, ceea ce înseamnă că toate instanțele acestor clase sunt clasificate corect.

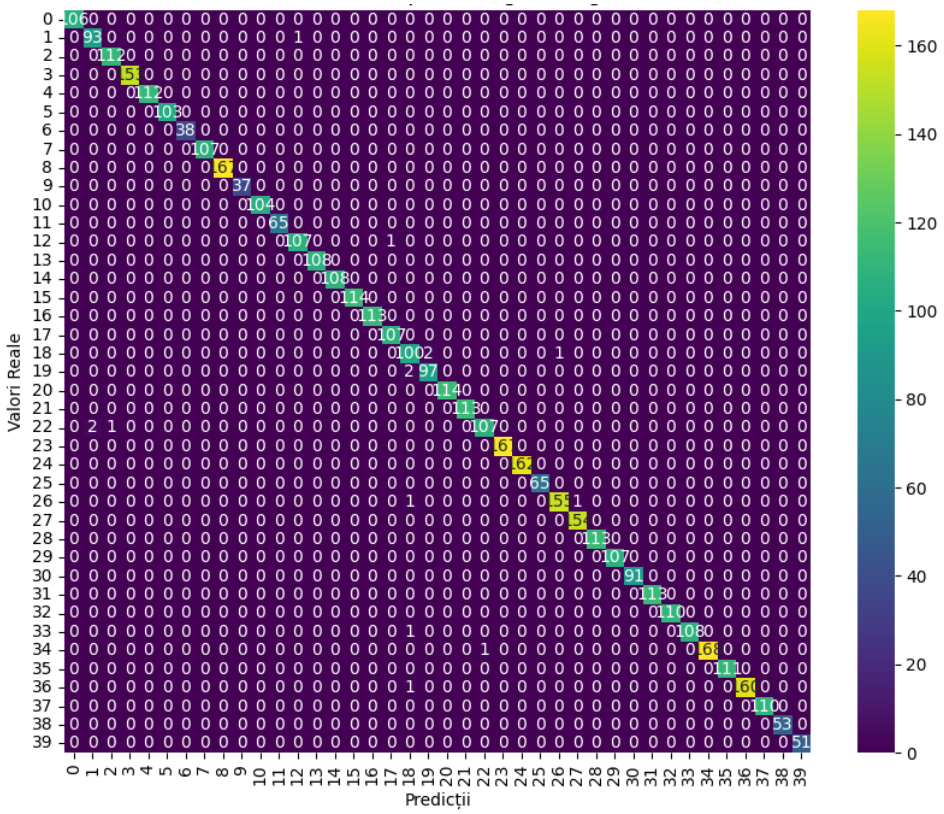
- Clasele cu performanță ușor mai scăzută:
    - **Apple Red 3:** Precizie 0.98, recall 0.99, și F1-score 0.98. Aceasta indică doar câteva instanțe clasificate greșit.
    - **Apple Red Yellow 1:** Precizie 0.99, recall și F1-score 1.00, ceea ce arată o performanță aproape perfectă.

 Observații generale:

- Performanța globală este foarte bună, sugerând că Gradient Boosting gestionează bine separarea claselor și extragerea caracteristicilor esențiale.
- Erorile sunt minore și concentrate doar pe câteva clase, fiind nesemnificative în raport cu dimensiunea setului de date.

Ajustarea parametrilor precum **max_depth=3-5**, **learning_rate=0.1**, și **n_estimators=100-200** poate ajuta la optimizarea suplimentară, menținând un echilibru între acuratețe și eficiență.

Matricea de confuzie



Matricea de confuzie confirmă performanța foarte bună a modelului Gradient Boosting, cu erori rare și clasări corecte pentru majoritatea claselor. Modelul demonstrează o separare clară a datelor, iar micile confuzii observate sunt nesemnificative în raport cu performanța globală.

Concluzii

Performanța modelelor de clasificare

Fashion-MNIST:

- Modelele au avut rezultate bune, dar unele clase, cum ar fi încălțămintea și genții, au fost mai greu de separat din cauza similarităților.
- SVM: A obținut cele mai bune rezultate globale, cu o acuratețe în jur de 78%.
- Random Forest: Performanță bună, dar inferioară SVM pentru clasele mai dificil de separat.
- Gradient Boosting: A oferit o clasificare competitivă, dar cu performanță similară Random Forest.
- Observație: Setul de date Fashion-MNIST are clase suprapuse în spațiul caracteristicilor, ceea ce explică dificultățile întâmpinate de modele.

Fruits-360:

- Performanța modelelor a fost aproape perfectă, datorită separării clare între clase și a unui subset redus și bine reprezentat.
- Logistic Regression: Acuratețe de 99.66%, cu erori nesemnificative.
- SVM: Performanță perfectă (100% acuratețe), confirmând separarea excelentă între clase.
- Random Forest: Acuratețe de 99.57%, cu mici confuzii la câteva clase.
- Gradient Boosting: Aproape perfect, cu erori minime izolate.

Impactul preprocesării și selecției claselor

- Reducerea numărului de clase în Fruits-360 a contribuit semnificativ la obținerea unor rezultate aproape perfecte.
- Standardizarea și selecția caracteristicilor au fost critice pentru a gestiona complexitatea datelor în Fashion-MNIST.

Compararea seturilor de date

- Fruits-360: Separare clară între clase, ceea ce a dus la performanțe remarcabile ale modelelor.
- Fashion-MNIST: Mai multe clase suprapuse și similarități între caracteristici, ceea ce a făcut clasificarea mai dificilă.