```
====================================
> > > > Import des librairies < < < <
====================================


========================================
> > > > Définition des fonctions < < < <
========================================


============================================================
> > > > TRAITEMENT DU JEU DE DONNES D'ENTRAINEMENT < < < <
============================================================


============================================================
Identification des chemins d'accès aux répertoires d'images
============================================================

dataset_path = data/fruits_360_v3b/Training/

image_path =
s3://oc-ds-p8/data/fruits_360_v3b/Training/Corn/,s3://oc-ds-p8/data/fruits_360_v
3b/Training/Orange/,s3://oc-ds-p8/data/fruits_360_v3b/Training/Raspberry/

Nombre de catégories de fruits: 3

2 premières catégories: ['s3://oc-ds-p8/data/fruits_360_v3b/Training/Corn/',
's3://oc-ds-p8/data/fruits_360_v3b/Training/Orange/']
2 dernières catégories: ['s3://oc-ds-p8/data/fruits_360_v3b/Training/Orange/',
's3://oc-ds-p8/data/fruits_360_v3b/Training/Raspberry/']

Durée de l'opération 'Récupération des images': 0.11 s


======================
Calcul des descripteurs
======================


Chargement des images (rdd_images)
==================================

MapPartitionsRDD[21] at coalesce at NativeMethodAccessorImpl.java:0

Nombre de partitions: 18
Dimension: 150

Catégories / Images / Descripteurs (rdd_cat_ima_desc)
====================================================

PythonRDD[23] at RDD at PythonRDD.scala:53
```

```
Catégories / Images / Descripteurs (rdd_cat_ima_desc_f)
=======================================================

PythonRDD[24] at RDD at PythonRDD.scala:53

Catégories (rdd_cat)
====================

PythonRDD[25] at RDD at PythonRDD.scala:53

Identifiants des images (rdd_ima)
=================================

PythonRDD[26] at RDD at PythonRDD.scala:53

Descripteurs (rdd_desc)
=======================

PythonRDD[27] at RDD at PythonRDD.scala:53

Nombre de partitions: 18
Dimension: 11200

Collecte des catégories d'images (list_cat)
===========================================

3 premières occurences: ['Raspberry', 'Raspberry', 'Raspberry']

Collecte des identifiants des images (list_ima)
===============================================

3 premières occurences: ['Raspberry_19_100.jpg', 'Raspberry_19_100.jpg',
'Raspberry_19_100.jpg']
df_ima_cat: (11200, 2)
df_ima_cat (sans dup): (150, 2)

Identifiants des images et des catégories (sdf_ima_cat)
=======================================================

root
 |-- ima: string (nullable = true)
 |-- cat: string (nullable = true)

+--------------------+---------+
|                 ima|      cat|
+--------------------+---------+
|Raspberry_19_100.jpg|Raspberry|
| Raspberry_1_100.jpg|Raspberry|
|Raspberry_20_100.jpg|Raspberry|
+--------------------+---------+

Durée de l'opération 'Extraction des descripteurs des images': 19.57 s
```

```
============================================================
Classification non supervisée des descripteurs avec K-Means
============================================================


Modèle K-Means (km_model)
=========================

<pyspark.mllib.clustering.KMeansModel object at 0x7fbae3ad1c90>

Nombre de clusters: 30

Durée de l'opération 'Clustering K-Means': 10.94 s


========================================
Prédictions des descripteurs avec K-Means
========================================


Prédictions (rdd_km_pred)
=========================

PythonRDD[159] at RDD at PythonRDD.scala:53

Nombre de partitions: 18
Dimension: 11200

Collecte des prédictions (list_km_pred)
=======================================

[21, 18, 24, 9, 11, 9, 10, 9, 7, 24]

Durée de l'opération 'Prédiction K-Means': 1.15 s


========================
Création du bag of words
========================


Encodage des identifiants d'images et concatenation avec les prédictions
(clusters K-Means)
================================================================================
===========

Encodage des identifiants d'images (sdf_ima_label)
--------------------------------------------------

+----------+--------------------+--------+----------+
|Unnamed: 0|                 IMA|image_id|prediction|
+----------+--------------------+--------+----------+
|      1024|Raspberry_24_100.jpg|     116|         4|
|      1025|Raspberry_24_100.jpg|     116|         6|
```

```
|      1026|Raspberry_24_100.jpg|     116|        12|
|      1027|Raspberry_24_100.jpg|     116|        19|
|      1028|Raspberry_24_100.jpg|     116|        19|
|      1029|Raspberry_24_100.jpg|     116|        28|
|      1030|Raspberry_24_100.jpg|     116|         0|
|      1031|Raspberry_24_100.jpg|     116|        19|
|      1032|Raspberry_24_100.jpg|     116|         9|
|      1033|Raspberry_24_100.jpg|     116|        23|
+----------+--------------------+--------+----------+

Prédictions (clusters K-Means) par image (sdf_ima_pred)
-------------------------------------------------------

root
 |-- id: long (nullable = true)
 |-- prediction: long (nullable = true)

+---+----------+
| id|prediction|
+---+----------+
|116|         4|
|116|         6|
|116|        12|
|116|        19|
|116|        19|
|116|        28|
|116|         0|
|116|        19|
|116|         9|
|116|        23|
+---+----------+


Liste des clusters par image (Map + reduceByKey)
================================================

Clusters par image (rdd_words)
------------------------------

PythonRDD[193] at RDD at PythonRDD.scala:53

Nombre de partitions: 8
Dimension: 150

Liste de 'words' par image (sdf_worcds)
---------------------------------------

root
 |-- image_id: long (nullable = true)
 |-- words: array (nullable = true)
 |    |-- element: string (containsNull = true)

+--------+--------------------+
|image_id|               words|
```

```
+--------+--------------------+
|     121|[10, 20, 5, 11, 2...|
|     137|[21, 29, 18, 21, ...|
|       1|[1, 11, 20, 5, 8,...|
|       9|[5, 5, 20, 20, 3,...|
|      17|[21, 21, 8, 3, 5,...|
|      25|[10, 5, 10, 20, 5...|
|      33|[13, 20, 13, 10, ...|
|      41|[20, 11, 20, 20, ...|
|      81|[1, 5, 15, 19, 1,...|
|      89|[5, 1, 15, 5, 1, ...|
+--------+--------------------+


Création du bag of words à partir des listes de 'words' associées aux images
(CountVectorizer)
================================================================================
==============

Bag of words (sdf_bow)
----------------------

root
 |-- image_id: long (nullable = true)
 |-- bag_of_words: vector (nullable = true)

+--------+--------------------+
|image_id|        bag_of_words|
+--------+--------------------+
|     113|(30,[0,1,2,3,4,5,...|
|     129|(30,[0,1,2,3,4,5,...|
|     145|(30,[0,1,2,3,4,5,...|
|      49|(30,[0,1,2,4,5,6,...|
|      57|(30,[0,1,2,4,5,8,...|
|      65|(30,[0,3,4,8,12,1...|
|      73|(30,[0,3,4,8,9,12...|
|     121|(30,[0,1,2,3,4,5,...|
|     137|(30,[0,1,2,3,4,5,...|
|       1|(30,[0,1,2,4,5,6,...|
+--------+--------------------+


==========================
Sauvegarde du bag of words
==========================


Bag of words (df_bow)
=====================

   image_id                                        bag_of_words
0       112  (1.0, 7.0, 12.0, 13.0, 3.0, 11.0, 6.0, 10.0, 7...
1       128  (3.0, 10.0, 5.0, 14.0, 4.0, 18.0, 10.0, 11.0, ...
2       144  (3.0, 16.0, 15.0, 13.0, 5.0, 8.0, 14.0, 11.0, ...
```

```
3        48  (2.0, 2.0, 0.0, 0.0, 12.0, 0.0, 0.0, 0.0, 2.0,...
4        56  (14.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0,...


Bag of words
============

                    ima        cat    0    1     2     3    4    5
0  Raspberry_19_100.jpg  Raspberry  1.0  6.0  12.0  15.0  3.0  8.0
1  Raspberry_19_100.jpg  Raspberry  1.0  6.0  12.0  15.0  3.0  8.0
2  Raspberry_19_100.jpg  Raspberry  1.0  6.0  12.0  15.0  3.0  8.0
3  Raspberry_19_100.jpg  Raspberry  1.0  6.0  12.0  15.0  3.0  8.0
4  Raspberry_19_100.jpg  Raspberry  1.0  6.0  12.0  15.0  3.0  8.0

Dimensions du jeu de données: (11200, 32)

Durée de l'opération 'Création du bag of words': 6.68 s


=========================
Réduction de dimension PCA
=========================


Résultats de la PCA (sdf_features)
==================================

root
 |-- features: vector (nullable = true)

+--------------------+
|            features|
+--------------------+
|[27.7780969539286...|
|[41.7540870225420...|
|[4.57301250203883...|
|[4.13710522990711...|
|[0.68545033385841...|
|[2.19846535106123...|
|[3.44101296354694...|
|[1.8486051622748,...|
|[0.15371664637558...|
|[-1.0405542269321...|
+--------------------+


Jointure entre les ids des images et les features (sdf_ima_features)
====================================================================

root
 |-- image_id: long (nullable = true)
 |-- bag_of_words: vector (nullable = true)
 |-- features: vector (nullable = true)
```

```
+--------+--------------------+--------------------+
|image_id|        bag_of_words|            features|
+--------+--------------------+--------------------+
|     104|(30,[0,1,2,3,4,5,...|[32.0138285416642...|
|      96|(30,[0,6,8,10,12,...|[-0.4808992597576...|
|      64|(30,[0,3,8,12,13,...|[-0.5642262867535...|
|      72|(30,[0,2,3,6,11,1...|[0.27871775964171...|
|       0|(30,[0,1,2,3,4,8,...|[5.31670707242512...|
|     128|(30,[0,1,2,3,4,5,...|[34.3690822290198...|
|       8|(30,[0,1,2,4,8,9,...|[1.59824745168281...|
|      48|(30,[0,1,4,8,10,1...|[2.08119494153099...|
|      24|(30,[0,1,2,4,6,8,...|[2.80325872564979...|
|      16|(30,[0,1,4,8,10,1...|[1.72208502977390...|
+--------+--------------------+--------------------+
```

Jointure entre les catégories et les features (sdf_cat_features)
==========

```
root
 |-- IMA: string (nullable = true)
 |-- image_id: long (nullable = true)
 |-- Unnamed: 0: long (nullable = true)
 |-- features: vector (nullable = true)
 |-- cat: string (nullable = true)
```

```
+--------------------+--------+----------+--------------------+---------+
|                 IMA|image_id|Unnamed: 0|            features|      cat|
+--------------------+--------+----------+--------------------+---------+
|     Corn_45_100.jpg|      24|      7404|[2.80325872564979...|     Corn|
|   Orange_11_100.jpg|      72|      9041|[0.27871775964171...|   Orange|
|  Orange_112_100.jpg|      64|      8809|[-0.5642262867535...|   Orange|
|  Orange_112_100.jpg|      64|      8823|[-0.5642262867535...|   Orange|
|  Orange_112_100.jpg|      64|      8819|[-0.5642262867535...|   Orange|
|     Corn_45_100.jpg|      24|      7383|[2.80325872564979...|     Corn|
|    Orange_1_100.jpg|      88|      9427|[0.33037642947585...|   Orange|
|    Orange_1_100.jpg|      88|      9452|[0.33037642947585...|   Orange|
|    Orange_1_100.jpg|      88|      9446|[0.33037642947585...|   Orange|
|Raspberry_42_100.jpg|     136|      4110|[33.6321024449893...|Raspberry|
+--------------------+--------+----------+--------------------+---------+
```

Encodage de la variable catégories (sdf_lab_features)
==========

```
root
 |-- label: double (nullable = false)
 |-- features: vector (nullable = true)
```

```
+-----+--------------------+
|label|            features|
+-----+--------------------+
|  0.0|[1.8486051622748,...|
|  1.0|[34.5145081686740...|
```

```
|  2.0|[4.13710522990711...|
|  0.0|[1.8486051622748,...|
|  0.0|[1.8486051622748,...|
|  2.0|[-0.1318028715449...|
|  0.0|[1.8486051622748,...|
|  2.0|[4.13710522990711...|
|  1.0|[0.15371664637558...|
|  1.0|[34.5145081686740...|
+-----+--------------------+
```

Bag of words après réduction de dimension (df_lab_features)
==========

	label	0	1	2	3	4	5
0	2.0	0.330376	-4.698919	-2.556379	-1.753178	-1.902630	1.298467
1	0.0	0.153717	-3.672065	0.463361	-1.445074	-0.231556	0.819848
2	2.0	3.441013	2.309796	-1.730758	-3.248047	-4.133691	-0.994865
3	1.0	34.514508	0.356139	7.945553	-5.846689	-5.059411	2.386031
4	0.0	4.573013	7.235424	0.762018	-5.468033	-1.932407	2.238392
5	2.0	-0.516230	-3.932645	-2.017393	-2.830665	-3.392697	0.640334
6	1.0	4.137105	0.086613	-2.316639	-5.521217	-4.984313	-0.406035
7	1.0	31.670925	-0.758153	-6.523613	-0.356006	-0.143185	0.033581
8	0.0	27.778097	-0.274003	-2.831368	2.885279	-8.110952	3.122179

Dimensions du nouveau jeu de données avec les étiquettes (df_lab_features):
(150, 22)

Durée de l'opération 'Réduction de dimension': 14.7 s

==========
Classification RF
==========

Durée de l'opération 'Classification': 98.27 s

==========
> > > > TRAITEMENT DU JEU DE DONNEES DE TEST < < < <
==========

==========
Identification des chemins d'accès aux répertoires d'images
==========

dataset_path = data/fruits_360_v3b/Test/

image_path =
s3://oc-ds-p8/data/fruits_360_v3b/Test/Corn/,s3://oc-ds-p8/data/fruits_360_v3b/Test/Orange/,s3://oc-ds-p8/data/fruits_360_v3b/Test/Raspberry/

Nombre de catégories de fruits: 3

```
2 premières catégories: ['s3://oc-ds-p8/data/fruits_360_v3b/Test/Corn/',
's3://oc-ds-p8/data/fruits_360_v3b/Test/Orange/']
2 dernières catégories: ['s3://oc-ds-p8/data/fruits_360_v3b/Test/Orange/',
's3://oc-ds-p8/data/fruits_360_v3b/Test/Raspberry/']

Durée de l'opération 'Récupération des images - Test': 0.13 s


======================
Calcul des descripteurs
======================


Chargement des images (rdd_images)
==================================

MapPartitionsRDD[508] at coalesce at NativeMethodAccessorImpl.java:0

Nombre de partitions: 18
Dimension: 75

Catégories / Images / Descripteurs (rdd_cat_ima_desc)
=====================================================

PythonRDD[510] at RDD at PythonRDD.scala:53

Catégories / Images / Descripteurs (rdd_cat_ima_desc_f)
=======================================================

PythonRDD[511] at RDD at PythonRDD.scala:53

Catégories (rdd_cat)
====================

PythonRDD[512] at RDD at PythonRDD.scala:53

Identifiants des images (rdd_ima)
=================================

PythonRDD[513] at RDD at PythonRDD.scala:53

Descripteurs (rdd_desc)
=======================

PythonRDD[514] at RDD at PythonRDD.scala:53

Nombre de partitions: 18
Dimension: 5352

Collecte des catégories d'images (list_cat)
===========================================

3 premières occurences: ['Corn', 'Corn', 'Corn']
```

```
Collecte des identifiants des images (list_ima)
================================================

3 premières occurences: ['Corn_19_100.jpg', 'Corn_19_100.jpg',
'Corn_19_100.jpg']
df_ima_cat: (5352, 2)
df_ima_cat (sans dup): (75, 2)

Identifiants des images et des catégories (sdf_ima_cat)
=======================================================

root
 |-- ima: string (nullable = true)
 |-- cat: string (nullable = true)

+---------------+----+
|            ima| cat|
+---------------+----+
|Corn_19_100.jpg|Corn|
|Corn_20_100.jpg|Corn|
|Corn_21_100.jpg|Corn|
+---------------+----+

Durée de l'opération 'Extraction des descripteurs des images - Test': 7.82 s


========================================
Prédictions des descripteurs avec K-Means
========================================


Prédictions (rdd_km_pred)
=========================

PythonRDD[524] at RDD at PythonRDD.scala:53

Nombre de partitions: 18
Dimension: 5352

Collecte des prédictions (list_km_pred)
=======================================

[5, 5, 8, 3, 5, 13, 9, 16, 12, 2]

Durée de l'opération 'Prédiction K-Means - Test': 1.06 s


=======================
Création du bag of words
=======================


Encodage des identifiants d'images et concatenation avec les prédictions
```

(clusters K-Means)
==========
==========

Encodage des identifiants d'images (sdf_ima_label)

+----------+--------------+--------+----------+
|Unnamed: 0| IMA|image_id|prediction|
+----------+--------------+--------+----------+
669	Corn_51_100.jpg	17	2
670	Corn_51_100.jpg	17	18
671	Corn_51_100.jpg	17	15
672	Corn_51_100.jpg	17	4
673	Corn_51_100.jpg	17	3
674	Corn_51_100.jpg	17	4
675	Corn_51_100.jpg	17	3
676	Corn_51_100.jpg	17	24
677	Corn_51_100.jpg	17	5
678	Corn_51_100.jpg	17	20
+----------+--------------+--------+----------+

Prédictions (clusters K-Means) par image (sdf_ima_pred)
---

root
 |-- id: long (nullable = true)
 |-- prediction: long (nullable = true)

+---+----------+
| id|prediction|
+---+----------+
0	5
0	5
0	8
0	3
0	5
0	13
0	9
0	16
0	12
0	2
+---+----------+

Liste des clusters par image (Map + reduceByKey)
==========

Clusters par image (rdd_words)

PythonRDD[558] at RDD at PythonRDD.scala:53

Nombre de partitions: 8

```
Dimension: 75

Liste de 'words' par image (sdf_worcds)
----------------------------------------

root
 |-- image_id: long (nullable = true)
 |-- words: array (nullable = true)
 |    |-- element: string (containsNull = true)

+--------+--------------------+
|image_id|               words|
+--------+--------------------+
|      24|[3, 13, 11, 8, 20...|
|      32|[5, 10, 5, 5, 10,...|
|      56|[29, 5, 21, 11, 1...|
|      64|[10, 20, 13, 21, ...|
|      72|[1, 1, 21, 29, 21...|
|       0|[5, 5, 8, 3, 5, 1...|
|       8|[13, 8, 3, 5, 20,...|
|      16|[10, 8, 3, 20, 5,...|
|      40|[13, 13, 5, 5, 5,...|
|      48|[5, 15, 1, 15, 1,...|
+--------+--------------------+


Création du bag of words à partir des listes de 'words' associées aux images
(CountVectorizer)
================================================================================
==============

Bag of words (sdf_bow)
----------------------

root
 |-- image_id: long (nullable = true)
 |-- bag_of_words: vector (nullable = true)

+--------+--------------------+
|image_id|        bag_of_words|
+--------+--------------------+
|       1|(30,[0,1,2,3,4,6,...|
|       9|(30,[0,1,2,3,4,5,...|
|      17|(30,[0,1,2,6,8,9,...|
|      41|(30,[0,3,8,9,10,1...|
|      49|(30,[0,1,3,8,9,12...|
|      57|(30,[0,1,2,3,4,5,...|
|      65|(30,[0,1,2,3,4,5,...|
|      25|(30,[0,8,9,12,14,...|
|      33|(30,[0,1,7,8,9,12...|
|      73|(30,[0,1,2,3,4,5,...|
+--------+--------------------+
```

```
==========================
Sauvegarde du bag of words
==========================


Bag of words (df_bow)
=====================

   image_id                                        bag_of_words
0        24  (7.0, 2.0, 1.0, 2.0, 0.0, 1.0, 6.0, 0.0, 1.0, ...
1        32  (6.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 10.0,...
2        56  (4.0, 11.0, 10.0, 8.0, 5.0, 7.0, 2.0, 10.0, 3....
3        64  (5.0, 9.0, 6.0, 8.0, 12.0, 11.0, 5.0, 6.0, 2.0...
4        72  (1.0, 6.0, 6.0, 4.0, 10.0, 13.0, 4.0, 5.0, 6.0...


Bag of words
============

               ima   cat    0    1    2    3    4    5
0  Corn_19_100.jpg  Corn  6.0  1.0  2.0  5.0  2.0  1.0
1  Corn_19_100.jpg  Corn  6.0  1.0  2.0  5.0  2.0  1.0
2  Corn_19_100.jpg  Corn  6.0  1.0  2.0  5.0  2.0  1.0
3  Corn_19_100.jpg  Corn  6.0  1.0  2.0  5.0  2.0  1.0
4  Corn_19_100.jpg  Corn  6.0  1.0  2.0  5.0  2.0  1.0

Dimensions du jeu de données: (5352, 32)

Durée de l'opération 'Création du bag of words - Test': 2.24 s

==========================
Réduction de dimension PCA
==========================


Résultats de la PCA (sdf_features)
==================================

root
 |-- features: vector (nullable = true)

+--------------------+
|            features|
+--------------------+
|[2.50340126532030...|
|[3.71680866938770...|
|[1.10049207047198...|
|[0.02883850545651...|
|[0.55296496252741...|
|[28.9645332268840...|
|[2.26995971440568...|
|[-0.2474420526789...|
|[26.8194851947211...|
|[26.4787808133587...|
```

```
+--------------------+


Jointure entre les ids des images et les features (sdf_ima_features)
======================================================================

root
 |-- image_id: long (nullable = true)
 |-- bag_of_words: vector (nullable = true)
 |-- features: vector (nullable = true)

+--------+--------------------+--------------------+
|image_id|        bag_of_words|            features|
+--------+--------------------+--------------------+
|       0|(30,[0,1,2,3,4,5,...|[28.9645332268840...|
|      32|(30,[0,1,7,8,9,12...|[3.71680866938770...|
|      64|(30,[0,1,2,3,4,5,...|[0.02883850545651...|
|      56|(30,[0,1,2,3,4,5,...|[1.10049207047198...|
|      24|(30,[0,1,2,3,5,6,...|[2.50340126532030...|
|      16|(30,[0,2,3,4,6,8,...|[-0.2474420526789...|
|      40|(30,[0,1,3,4,8,9,...|[26.8194851947211...|
|      72|(30,[0,1,2,3,4,5,...|[0.55296496252741...|
|      33|(30,[0,1,7,8,9,12...|[0.47186315328416...|
|      25|(30,[0,8,9,12,14,...|[-0.1397734003102...|
+--------+--------------------+--------------------+


Jointure entre les catégories et les features (sdf_cat_features)
================================================================

root
 |-- IMA: string (nullable = true)
 |-- image_id: long (nullable = true)
 |-- Unnamed: 0: long (nullable = true)
 |-- features: vector (nullable = true)
 |-- cat: string (nullable = true)

+-----------------+--------+----------+--------------------+------+
|              IMA|image_id|Unnamed: 0|            features|   cat|
+-----------------+--------+----------+--------------------+------+
|Orange_30_100.jpg|      25|       968|[-0.1397734003102...|Orange|
|Orange_30_100.jpg|      25|       969|[-0.1397734003102...|Orange|
|Orange_30_100.jpg|      25|       970|[-0.1397734003102...|Orange|
|Orange_30_100.jpg|      25|       971|[-0.1397734003102...|Orange|
|Orange_30_100.jpg|      25|       972|[-0.1397734003102...|Orange|
|Orange_30_100.jpg|      25|       973|[-0.1397734003102...|Orange|
|Orange_30_100.jpg|      25|       974|[-0.1397734003102...|Orange|
|Orange_30_100.jpg|      25|       975|[-0.1397734003102...|Orange|
|Orange_30_100.jpg|      25|       976|[-0.1397734003102...|Orange|
|Orange_30_100.jpg|      25|       977|[-0.1397734003102...|Orange|
+-----------------+--------+----------+--------------------+------+


Encodage de la variable catégories (sdf_lab_features)
```

```
==================================================

root
 |-- label: double (nullable = false)
 |-- features: vector (nullable = true)

+-----+--------------------+
|label|            features|
+-----+--------------------+
|  0.0|[28.9645332268840...|
|  0.0|[28.9645332268840...|
|  0.0|[28.9645332268840...|
|  0.0|[28.9645332268840...|
|  0.0|[28.9645332268840...|
|  0.0|[26.4787808133587...|
|  0.0|[26.4787808133587...|
|  0.0|[26.4787808133587...|
|  1.0|[2.26995971440568...|
|  1.0|[2.26995971440568...|
+-----+--------------------+


Bag of words après réduction de dimension (df_lab_features)
============================================================

   label          0         1         2         3         4         5
0    0.0  26.478781 -2.585510  8.934875  0.965440 -2.861286  4.534787
1    0.0  26.819485 -0.661571 -0.499717  4.247502 -4.049971  4.023794
2    0.0  28.964533 -3.924289 -3.520938 -3.054572 -3.331665  4.996848
3    1.0   1.100492  3.950441 -0.925458  1.192264 -4.986692  5.336048
4    2.0  -0.247442 -7.608611 -0.092081  1.064961 -2.722220  4.731834
5    1.0   2.269960  1.872251 -0.709663 -0.015734 -4.595592  6.048048
6    0.0  27.759339 -0.008410  5.628048 -1.152293 -2.320900  3.628365
7    0.0  25.178903 -0.874379  4.402495 -3.225218  1.629329  1.524403
8    1.0   2.280804  2.677976 -1.912545  0.774386 -3.600531  5.166144

Dimensions du nouveau jeu de données avec les étiquettes (df_lab_features): (75,
22)

Durée de l'opération 'Réduction de dimension - Test': 7.0 s


==============
Prédictions RF
==============


Prédictions RF (test_lab_pred)
==============================

<class 'pyspark.sql.dataframe.DataFrame'>
DataFrame[features: vector, prediction: double]

Prédictions (predictionAndLabels)
```

```
================================

<class 'pyspark.sql.dataframe.DataFrame'>
DataFrame[prediction: double, label: double]
+----------+-----+
|prediction|label|
+----------+-----+
|       1.0|  1.0|
|       2.0|  2.0|
|       1.0|  1.0|
|       1.0|  1.0|
|       2.0|  2.0|
|       0.0|  0.0|
|       1.0|  1.0|
|       0.0|  0.0|
|       0.0|  0.0|
|       0.0|  0.0|
+----------+-----+

+----------+-----+--------------------+
|prediction|label|            features|
+----------+-----+--------------------+
|       0.0|  0.0|[28.9645332268840...|
|       0.0|  0.0|[28.9645332268840...|
|       0.0|  0.0|[28.9645332268840...|
|       0.0|  0.0|[28.9645332268840...|
|       0.0|  0.0|[28.9645332268840...|
|       0.0|  0.0|[28.9645332268840...|
|       0.0|  0.0|[26.8194851947211...|
|       0.0|  0.0|[26.8194851947211...|
|       1.0|  1.0|[1.10049207047198...|
|       0.0|  0.0|[26.8194851947211...|
|       1.0|  1.0|[1.10049207047198...|
+----------+-----+--------------------+

Test set accuracy (RF) = 0.72

Durée de l'opération 'Prédiction - Test': 15.37 s


==========
Evaluation
==========


Jointure entre les identifiants des images et les features (sdf_ima_features)
================================================================================

root
 |-- image_id: long (nullable = true)
 |-- bag_of_words: vector (nullable = true)
 |-- features: vector (nullable = true)
 |-- prediction: double (nullable = false)

+--------+--------------------+--------------------+----------+
```

```
|image_id|        bag_of_words|            features|prediction|
+--------+--------------------+--------------------+----------+
|       0|(30,[0,1,2,3,4,5,...|[2.50340126532030...|       1.0|
|      32|(30,[0,1,7,8,9,12...|[-0.2474420526789...|       2.0|
|      64|(30,[0,1,2,3,4,5,...|[26.8194851947211...|       0.0|
|      56|(30,[0,1,2,3,4,5,...|[28.9645332268840...|       0.0|
|      24|(30,[0,1,2,3,5,6,...|[2.26995971440568...|       1.0|
|      16|(30,[0,2,3,4,6,8,...|[1.10049207047198...|       1.0|
|      40|(30,[0,1,3,4,8,9,...|[0.02883850545651...|       2.0|
|      72|(30,[0,1,2,3,4,5,...|[26.4787808133587...|       0.0|
|      73|(30,[0,1,2,3,4,5,...|[27.7593390809938...|       0.0|
|      10|(30,[0,2,3,5,6,7,...|[2.28080426931557...|       1.0|
+--------+--------------------+--------------------+----------+


Jointure entre les catégories et les features (sdf_cat_features)
================================================================

root
 |-- IMA: string (nullable = true)
 |-- image_id: long (nullable = true)
 |-- Unnamed: 0: long (nullable = true)
 |-- features: vector (nullable = true)
 |-- prediction: double (nullable = false)
 |-- cat: string (nullable = true)

+--------------------+--------+----------+--------------------+----------+------
---+
|                 IMA|image_id|Unnamed: 0|            features|prediction|
cat|
+--------------------+--------+----------+--------------------+----------+------
---+
|     Corn_19_100.jpg|       0|         6|[28.9645332268840...|       0.0|
Corn|
|     Corn_19_100.jpg|       0|        26|[28.9645332268840...|       0.0|
Corn|
|     Corn_19_100.jpg|       0|        28|[28.9645332268840...|       0.0|
Corn|
|     Corn_29_100.jpg|       8|       296|[2.26995971440568...|       1.0|
Corn|
|    Orange_8_100.jpg|      48|      1676|[26.4787808133587...|       0.0|
Orange|
|Raspberry_97_100.jpg|      72|      5023|[0.55296496252741...|
2.0|Raspberry|
|Raspberry_97_100.jpg|      72|      4972|[0.55296496252741...|
2.0|Raspberry|
|Raspberry_81_100.jpg|      56|      2633|[1.10049207047198...|
1.0|Raspberry|
|Raspberry_89_100.jpg|      64|      3805|[0.02883850545651...|
2.0|Raspberry|
|Raspberry_81_100.jpg|      56|      2701|[1.10049207047198...|
1.0|Raspberry|
+--------------------+--------+----------+--------------------+----------+------
---+
```

```
Encodage de la variable catégories (sdf_lab_features)
=====================================================

root
 |-- ima: string (nullable = true)
 |-- cat: string (nullable = true)
 |-- label: double (nullable = false)
 |-- prediction: double (nullable = false)
 |-- features: vector (nullable = true)

+--------------------+---------+-----+----------+--------------------+
|                 ima|      cat|label|prediction|            features|
+--------------------+---------+-----+----------+--------------------+
|Raspberry_97_100.jpg|Raspberry|  0.0|       0.0|[26.4787808133587...|
|      Corn_9_100.jpg|     Corn|  1.0|       1.0|[2.26995971440568...|
|      Corn_9_100.jpg|     Corn|  1.0|       1.0|[2.26995971440568...|
|Raspberry_97_100.jpg|Raspberry|  0.0|       0.0|[26.4787808133587...|
|Raspberry_97_100.jpg|Raspberry|  0.0|       0.0|[26.4787808133587...|
|      Corn_9_100.jpg|     Corn|  1.0|       1.0|[2.26995971440568...|
|      Corn_9_100.jpg|     Corn|  1.0|       1.0|[2.26995971440568...|
|Raspberry_98_100.jpg|Raspberry|  0.0|       0.0|[33.3821764457442...|
|Raspberry_98_100.jpg|Raspberry|  0.0|       0.0|[33.3821764457442...|
|Raspberry_98_100.jpg|Raspberry|  0.0|       0.0|[33.3821764457442...|
+--------------------+---------+-----+----------+--------------------+


Catégories réelles (label) vs Prédictions (prediction)
======================================================

prediction   0.0   1.0   2.0
label
0.0         51.0  20.0  30.0
1.0         17.0  58.0  25.0
2.0         35.0  16.0  49.0

Durée de l'opération 'Evaluation - Test': 25.24 s

Durée de l'opération 'Fin des traitements': 0.0 s
Durée totale de traitement: 00 h 03 m 30 s

Durée des opérations
--------------------
                                       Opération  Durée Estimation
0                         Récupération des images   0.11
1          Extraction des descripteurs des images  19.57
2                              Clustering K-Means  10.94
3                              Prédiction K-Means   1.15
4                         Création du bag of words   6.68
5                           Réduction de dimension  14.70
6                                   Classification  98.27
7                  Récupération des images - Test   0.13
8   Extraction des descripteurs des images - Test   7.82
```

```
9                       Prédiction K-Means - Test   1.06
10                 Création du bag of words - Test   2.24
11                   Réduction de dimension - Test   7.00
12                               Prédiction - Test  15.37
13                               Evaluation - Test  25.24
14                            Fin des traitements   0.00

====================================
> > > > Traitements finalisés < < < <
====================================
```