



CLASSIFICATION DES ESPÈCES DE PLANTES À PARTIR DE L'IMAGE DE LA FEUILLE

**Présenté par Adama
SAMAKE**

Plan de la présentation

01 Introduction

Contexte du projet, importance de la classification des espèces, problématique

02 Objectifs

Objectifs spécifiques, bénéfices potentiels, proposition de solution

03 Méthodologie

Collecte des données, pré-traitement, extraction des caractéristiques

04 Traitement d'images

Techniques de traitement, visuels du process

05 Modèle de classification

Choix algorithme de machine learning, entraînement du modèle

06 Résultats

Évaluation de la performance du modèle, test avec des exemples d'images de feuilles

07 Analyse et interprétation

Domaines d'application, défis rencontrés

08 Conclusion

Recapitulatif, perspectives futures

Introduction

- La biodiversité végétale constitue un élément crucial de notre écosystème, jouant un rôle fondamental dans l'équilibre de la vie sur Terre
- Le projet vise l'automatisation du processus de classification des espèces de plantes, processus qui est ardu manuellement



Problématique

- Espèces végétales en voix de disparition
- Dépendance de l'expertise humaine
- Temps de la classification manuelle



Objectifs

- L'objectif principal est d'utiliser des techniques de computer vision couplé à celle du machine learning
- Comme solution à cette problématique:
Création d'un modèle de machine learning, qui pourra être utilisé pour créer des outils destinés aux chercheurs en botanique.



Objectifs spécifiques du projet



Précision de la Classification



Qualité du pre-traitement des images



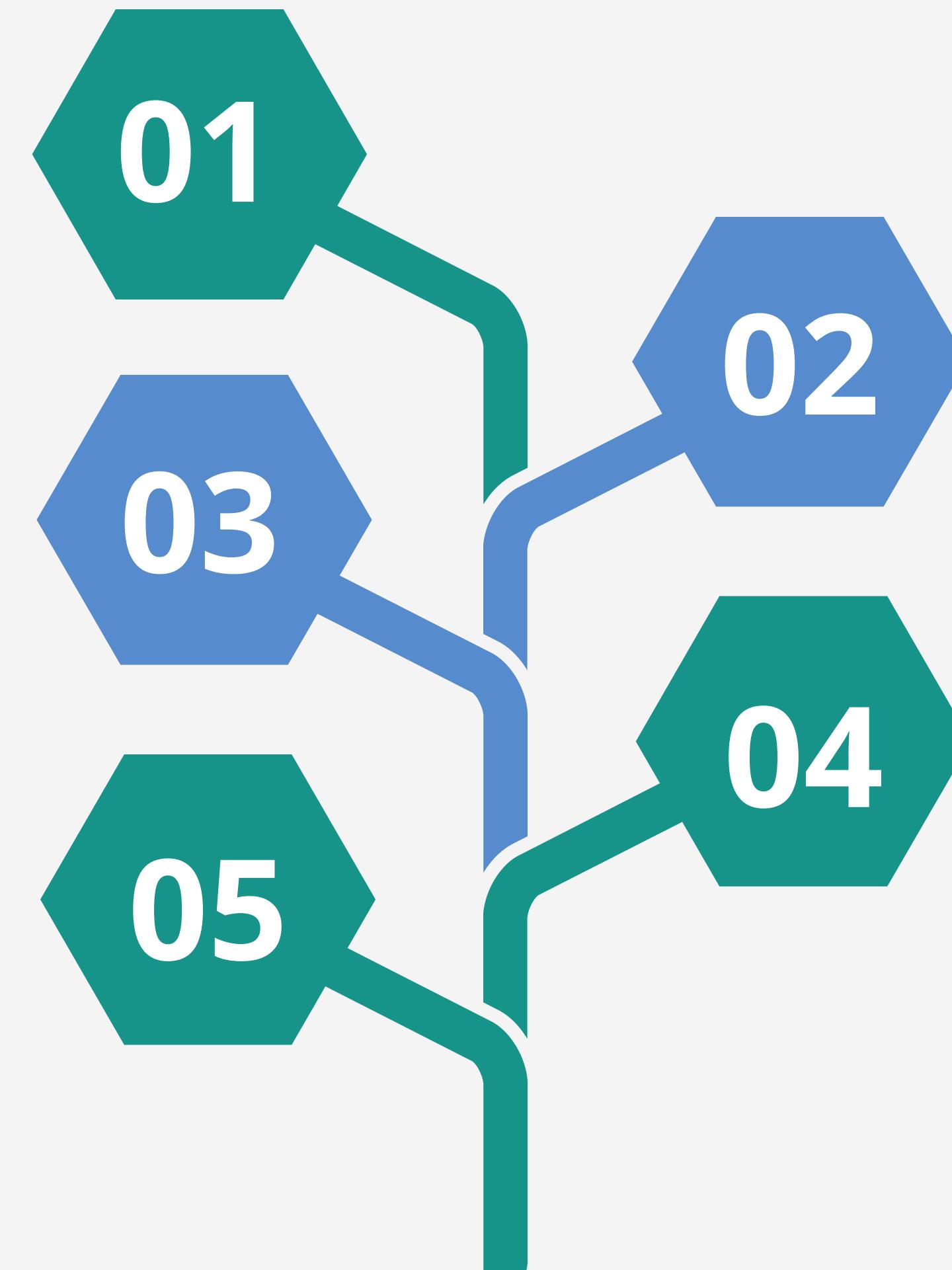
Évaluation des Performances



Sensibilisation à l'Importance de la Classification Automatisée

Methodologie

- 01 Collecte des données
- 02 Pré-traitement des données
- 03 Extraction des caractéristiques
- 04 Choix de l'algorithme et entraînement du modèle
- 05 Validation et tests



Collecte des données

Plusieurs sources de données d'images existent. On distingue entre autres:

- **Flavia dataset:** <https://flavia.sourceforge.net/>
- **ImageCLEF dataset:**
<https://www.imageclef.org/datasets>
- Potentiellement du web scraping

La source de données utilisée tout au long du projet est disponible sur kaggle:

<https://www.kaggle.com/datasets/meetnagadia/collection-of-different-category-of-leaf-images>



Collecte des données

La source de données a été constituée en recueillant les images à l'Université Shri Mata Vaishno Devi, à Katra en Inde.

La constitution a été faite en 2019.



Collecte des données

La source de données contient des images de feuilles de 12 espèces de plantes différentes qui sont:

- Mango (Mangue)
- Arjun (*Terminalia arjuna*)
- Alstonia Scholaris
- Guava (Goyave)
- Bael, Jamun, Jatropha
- Pongamia Pinnata
- Basil (Basilic)
- Pomegranate, Chinar
- Lemon (Citron)



Collecte des données

Nous retiendrons quatre espèces de plantes. Voici un aperçu des images.



#1 Gauva



#2 Basil



#3 Lemon



#4 Mango



Traitement d'images



Segmentation par thresholding

Convertir une image noir-blanc en une image binaire, où les pixels sont classés comme faisant partie du premier plan ou de l'arrière-plan.

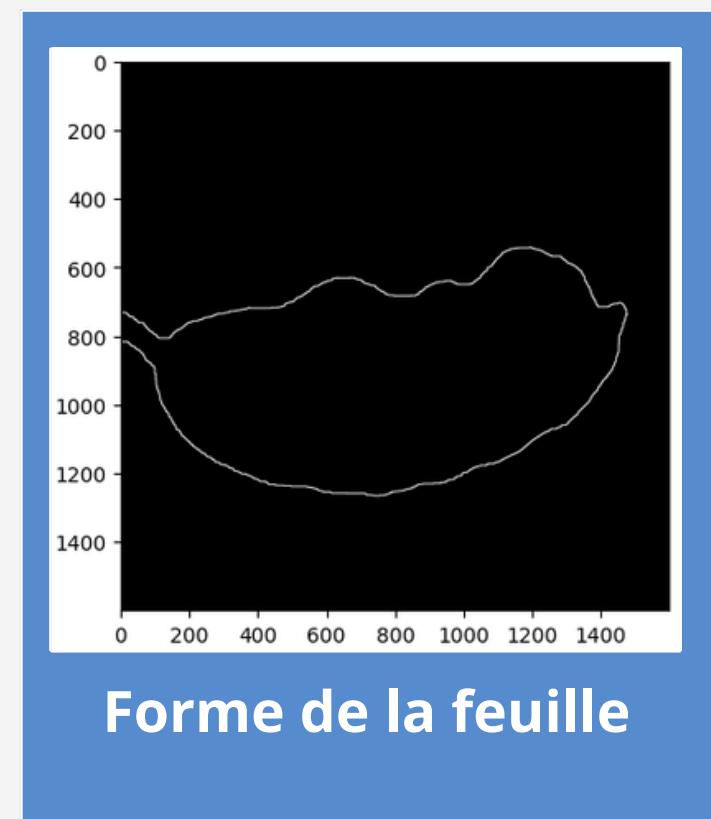
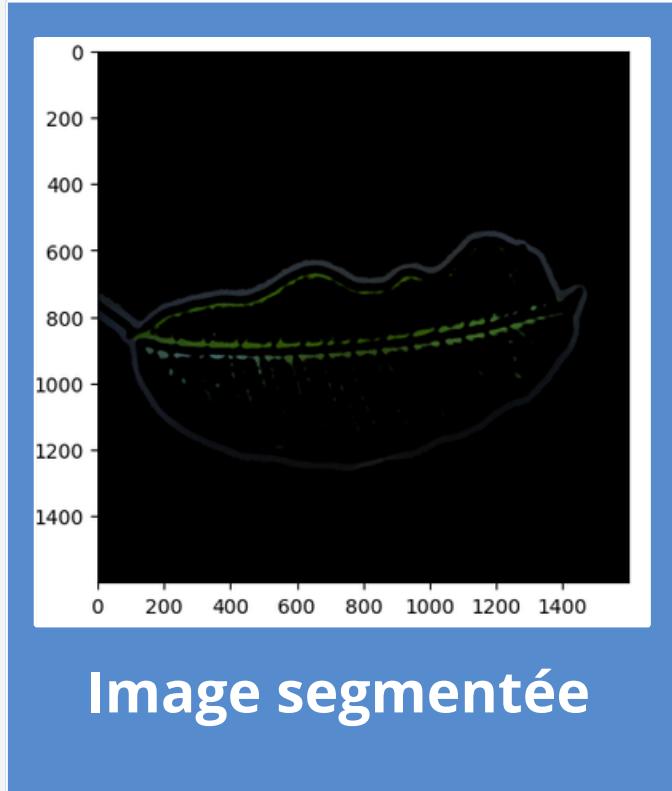
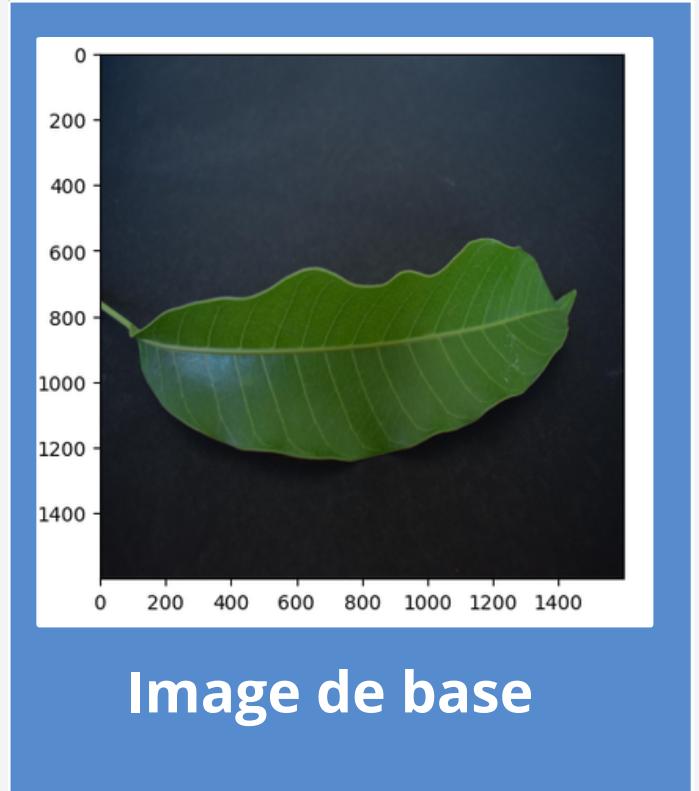
Segmentation par Kmeans clustering

Partitionner l'image en clusters, où chaque pixel appartient au cluster avec la couleur moyenne la plus proche.

Traitement d'images



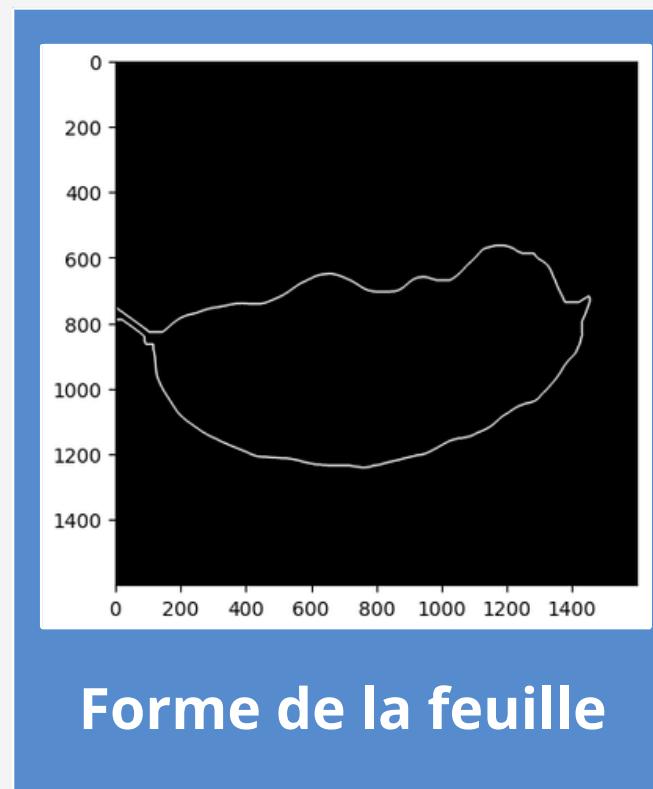
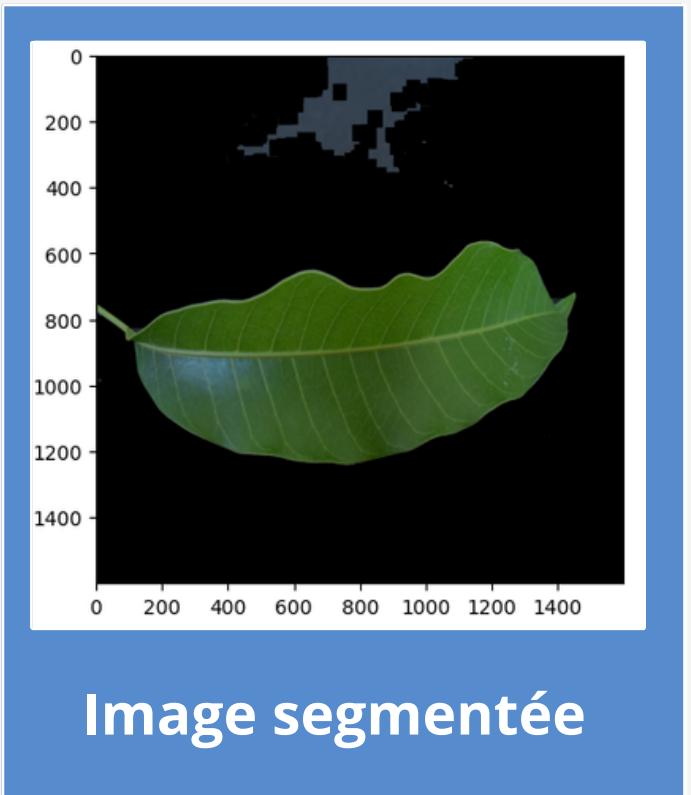
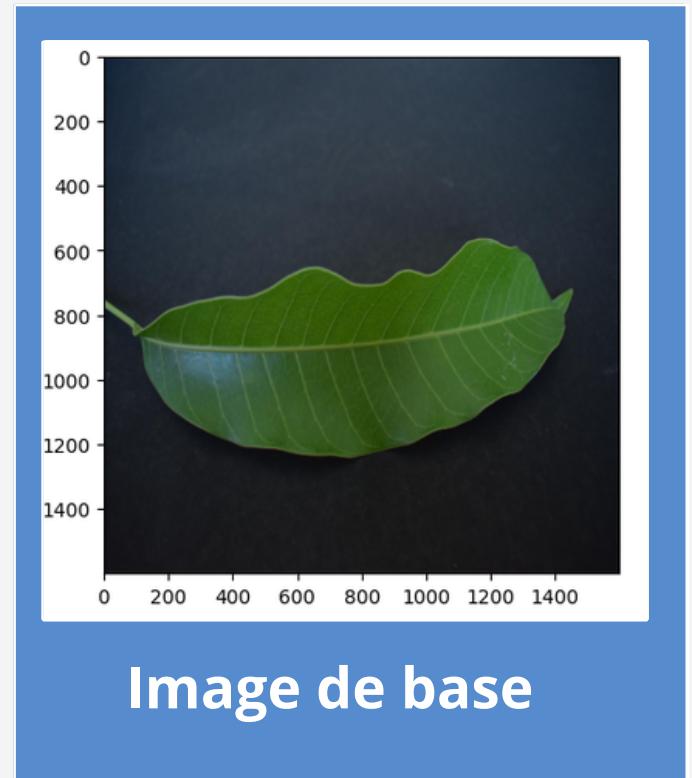
Segmentation par thresholding



Traitement d'images



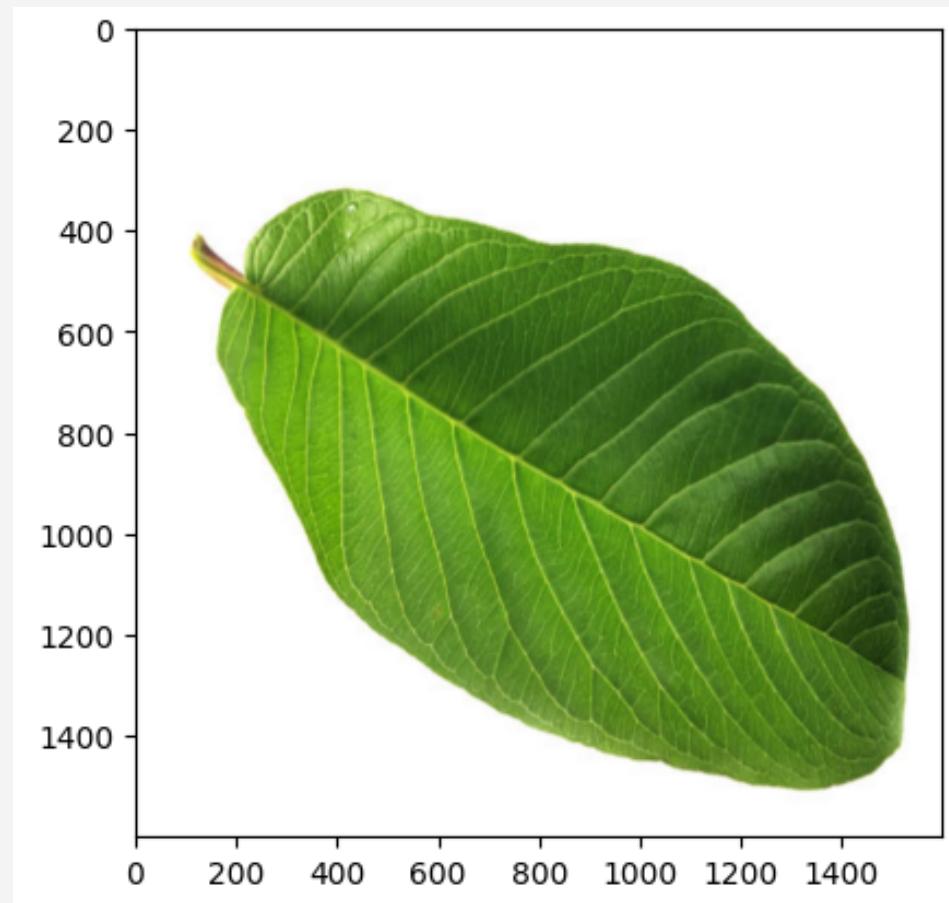
Segmentation par Kmeans clustering



Extraction des caractéristiques

Couleur	Forme	Aspect Morphologique	Texture
Moyenne des couleurs rouges, vertes et bleues	Perimètre	Rectangularité	Contraste
Écart-type des couleurs	Aire	Aspect ratio	Entropie et Moment de différence inverse
	Diamètre	Circularité	Correlation
	Longueur et largeur		

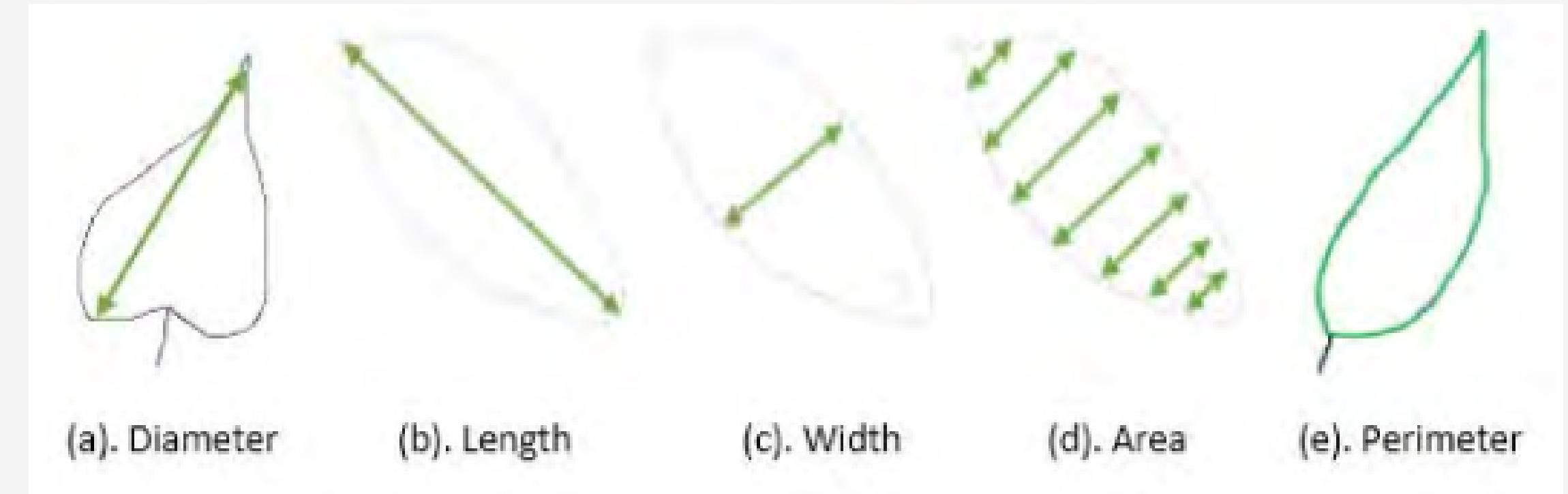
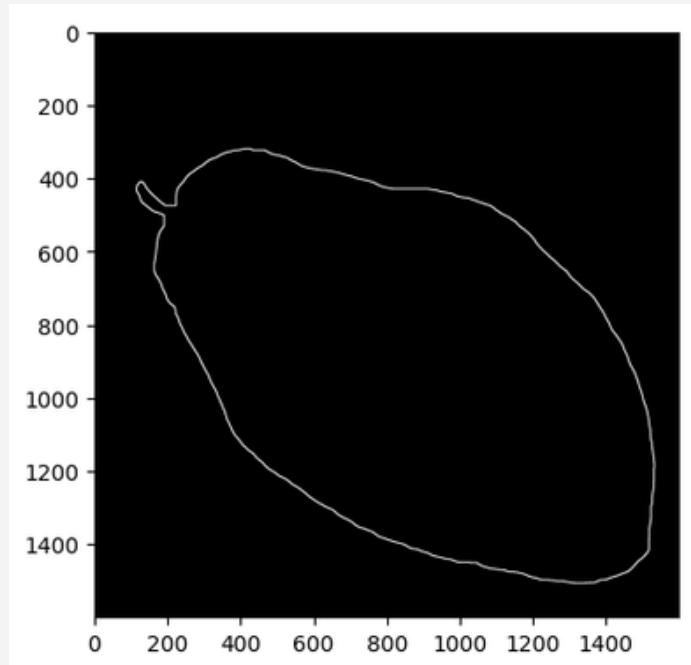
Extraction des caractéristiques



			row	0	1	2					
			column	0	.392	.482	.576				
			0	.478	.63	.169	.263	.376			
			1	.580	.79	.263	.44	.306	.376	.451	
			2	.373	.60	.376	.478	.561	.443	.569	.674
			0								
			1								
			2								

Caractéristiques de couleurs

Extraction des caractéristiques



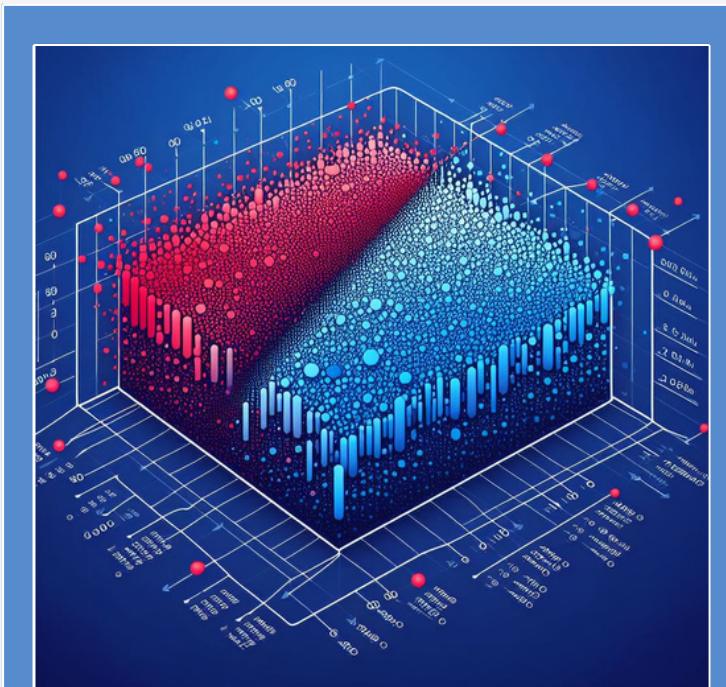
Caractéristiques de forme

Extraction des caractéristiques

Après l'extraction des caractéristiques, les données sont exportées dans un fichier au format csv pour l'entraînement des modèles de machine learning

	area	perimeter	physiological_length	physiological_width	aspect_ratio	rectangularity	circularity	diameter	mean_red	mean_green	mean_blue	stddev_red	stddev_green	stddev_blue	contrast	correlation	inverse_difference_moments	entropy	species	
0	628280.0	3801.149612		1111	1037	1.071360	1.833748	22.997292	894.399766	1.397607	1.667350	0.977093	8.513442	10.994798	6.016809	138.493835	0.773885	0.997870	0.058886	Lemon
1	450661.0	2956.792186		916	791	1.158028	1.607763	19.399549	757.495483	1.021457	1.170891	0.737389	7.486983	9.178166	5.533433	107.741151	0.775270	0.998343	0.047660	Lemon
2	529974.5	3133.863253		1039	789	1.316857	1.546812	18.531267	821.452671	1.248682	1.440288	0.931602	8.727181	10.828482	6.321649	114.809864	0.773587	0.998234	0.050078	Lemon
3	278406.5	2432.961667		796	585	1.360684	1.672590	21.261366	595.380689	0.870884	1.034409	0.704435	6.821553	8.540862	5.674191	88.811867	0.774945	0.998634	0.040276	Lemon
4	452268.5	2878.074365		950	692	1.372832	1.453561	18.315032	758.845267	0.824042	0.899913	0.661069	6.154941	7.254806	5.029955	105.746045	0.771518	0.998374	0.046360	Lemon

Modèle de classification



Architecture du SVM

Support Vector Machine (SVM) est un algorithme d'apprentissage supervisé efficace pour la classification et la régression.

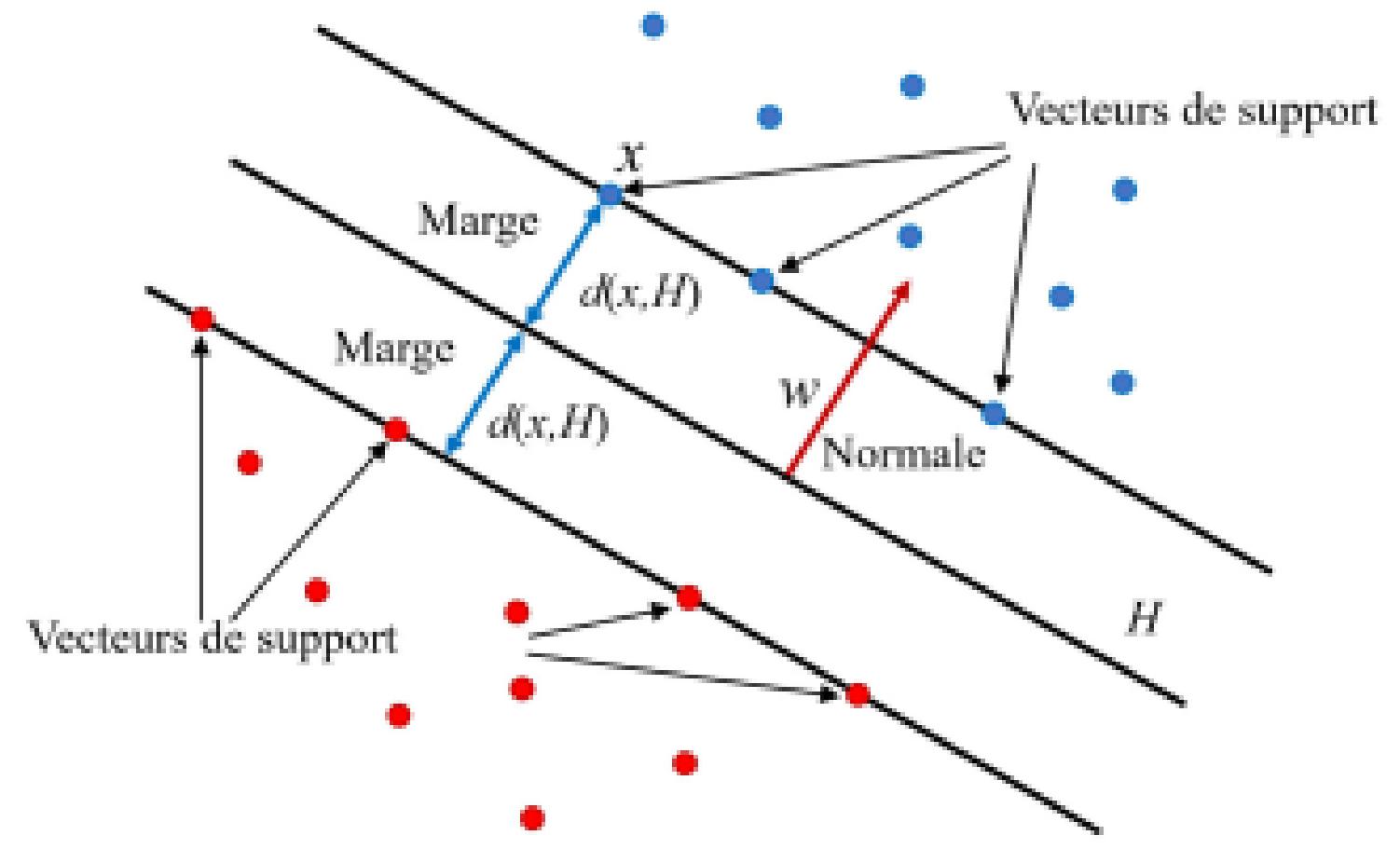
L'objectif est de trouver un hyperplan optimal qui sépare les données en classe distinct dans un espace multidimensionnel.

La fonction coût maximise la marge entre les classes tout en minimisant les erreurs de classification.

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$$

$$J(W, b) = \frac{1}{2} \|W\|^2 + C \sum_{i=1}^m \max(0, 1 - y^{(i)}(W^T X^{(i)} + b))$$

C est le paramètre de régularisation. Plus **C** est grand, moins la marge est large, mais les erreurs de classification sont pénalisées davantage.



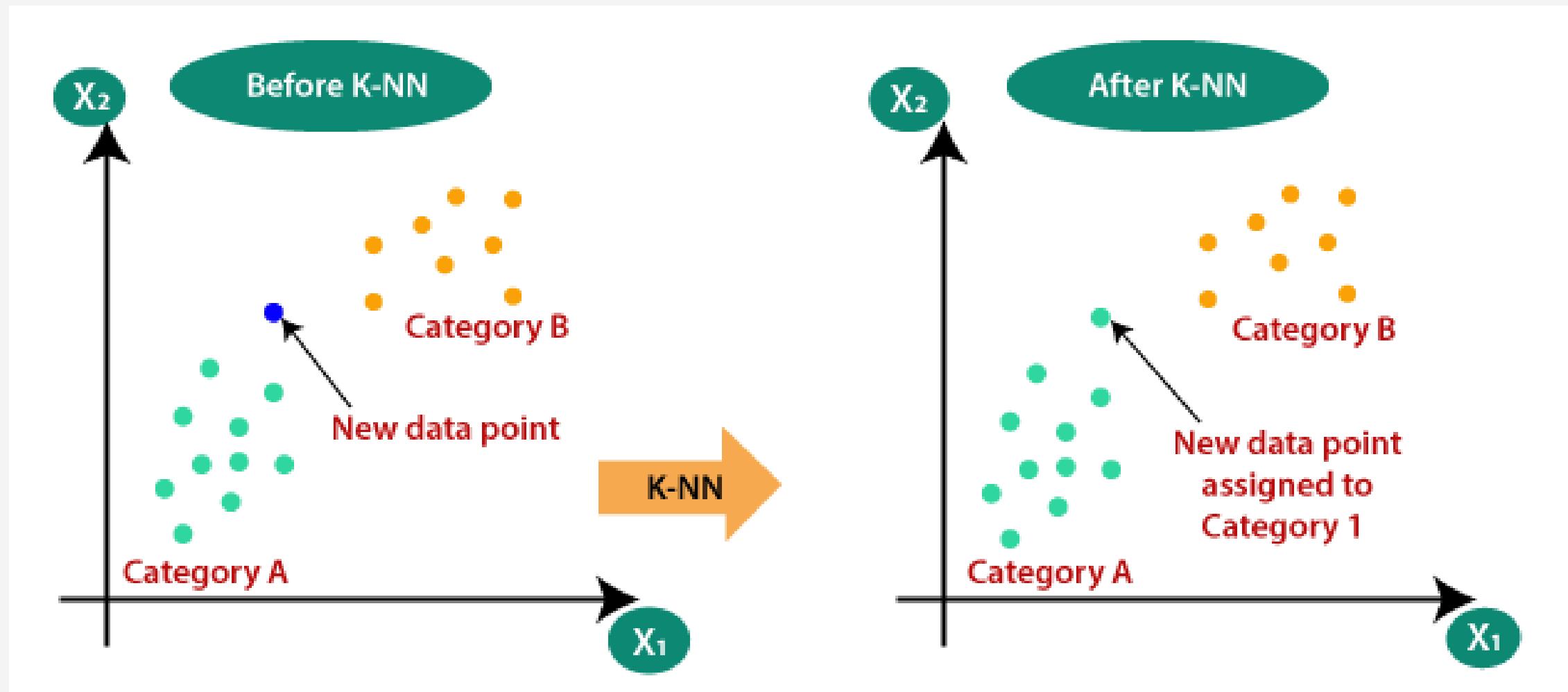
Architecture du KNN

L'algorithme de KNN est un algorithme d'apprentissage supervisé dont l'objectif est de donner un label à des données en se basant sur le voisinage.

Le paramètre clé est le **K** qui détermine le nombre de voisins à considérer lors de la classification

Le K choisi, On calcule les distances d entre la nouvelle donnée et ses voisins déjà classés.

Les métriques de distance peut être ajustés. On distingue la distance Euclidienne ou encore la distance de Minkowski

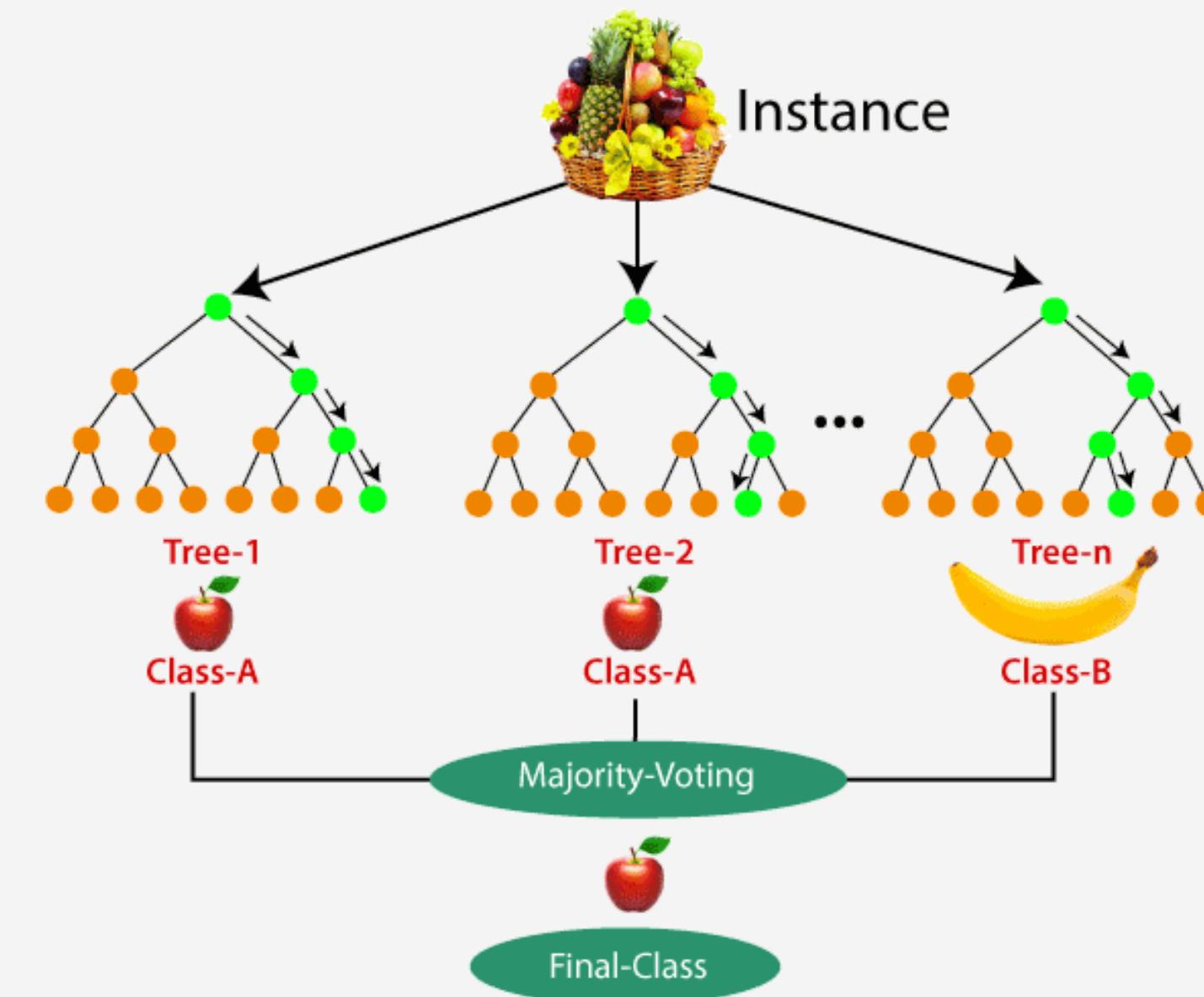


$$d(P, Q) = \left(\sum_{i=1}^n |p_i - q_i|^p \right)^{1/p}$$

Architecture du Random Forest

Random Forest est un modèle d'ensemble (rassembler les prédictions de plusieurs algorithmes de Machine Learning pour obtenir un résultat optimal) basé sur des arbres de décision.

Les hyperparamètres les plus importants sont le **nombres d'arbres** de décision et la **profondeur maximale** de l'arbre



Résultats

Modèles / Indicateurs	SVM	KNN	Random Forest
Précision	92%	88%	88%
Accuracy	92%	86%	87%
Recall	92%	91%	88%
F1-score	92%	86%	87%

Analyse et interprétations

SVM

La plus performante des modèles après application du GridSearchCV

KNN

Simple à comprendre et à mettre en place sur des datasets moyens.

Random forest

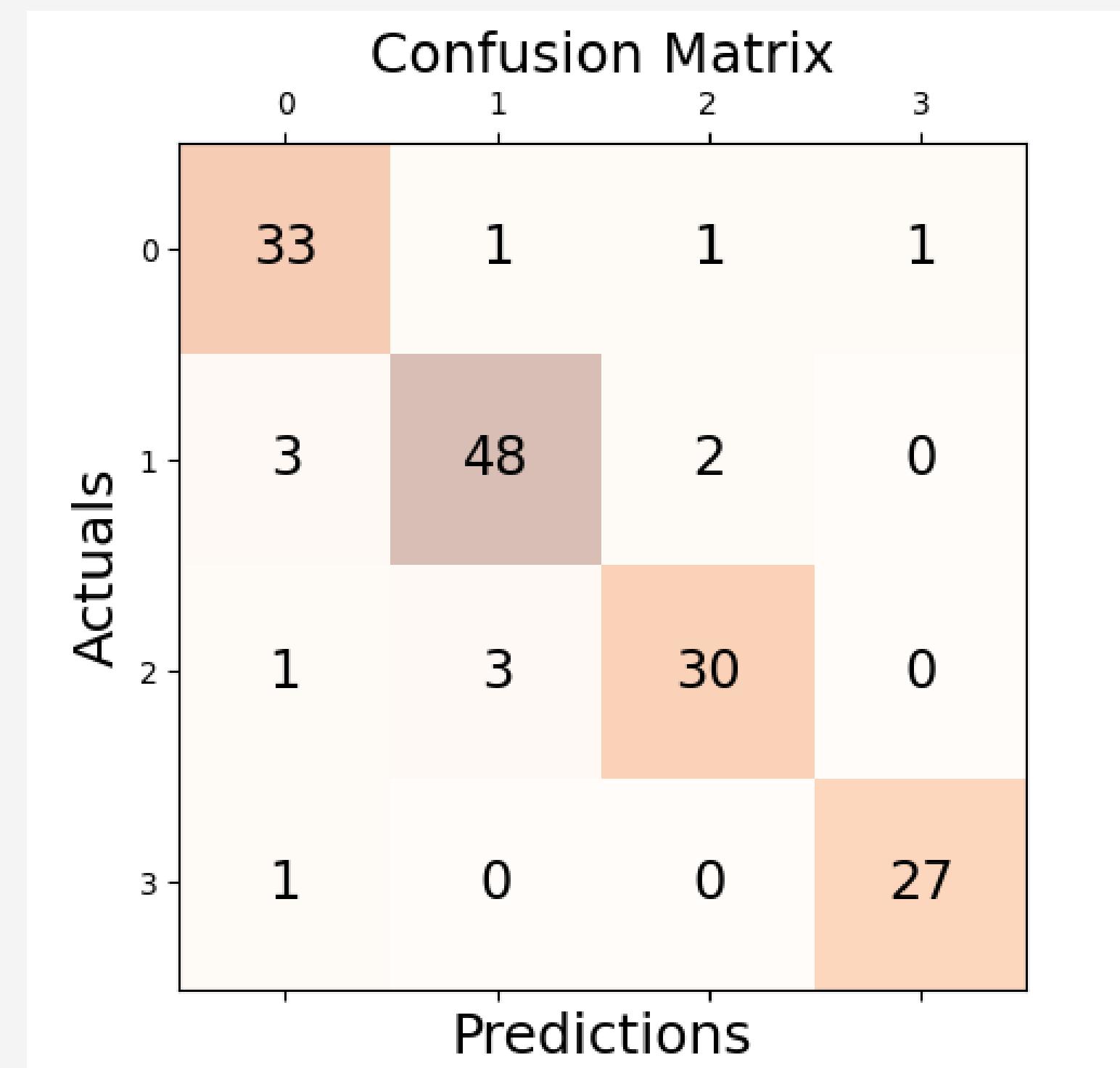
Efficace également puisqu'il généralise mieux que le KNN et le SVM dans certains cas

Analyse et interprétations

#1. SVM

Il a tendance à confondre:

- les espèces de citron (Lemon) avec la goyave (Gauva) et inversement

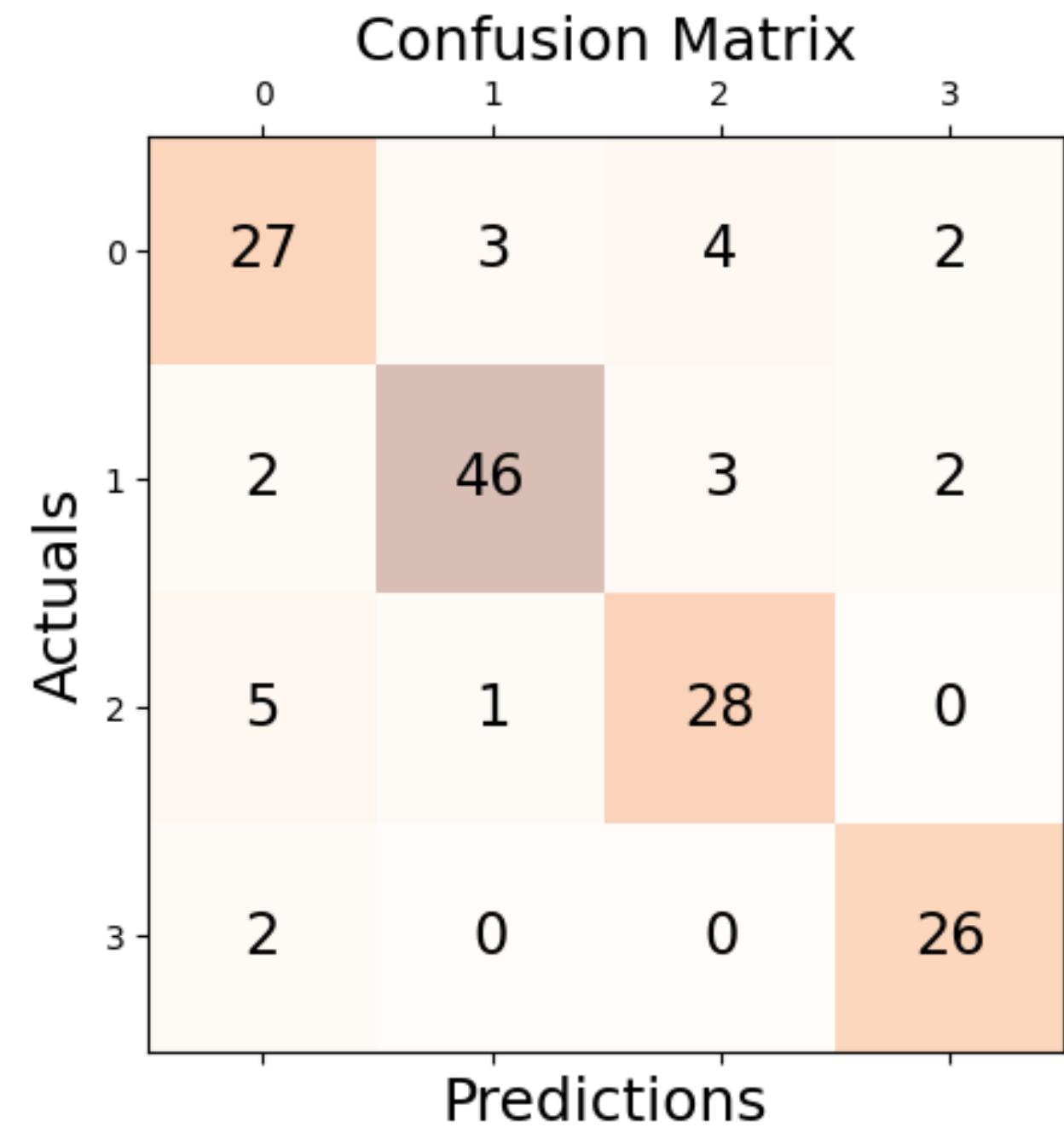


Analyse et interprétations

#2. KNN

Il a tendance à confondre:

- les espèces de basilic (Basil) avec la goyave (Gauva) et le citron (Lemon)
- les espèces de citron (Lemon) avec celle du basilic (Basil)
- Et la goyave (Gauva) avec le citron (Lemon)

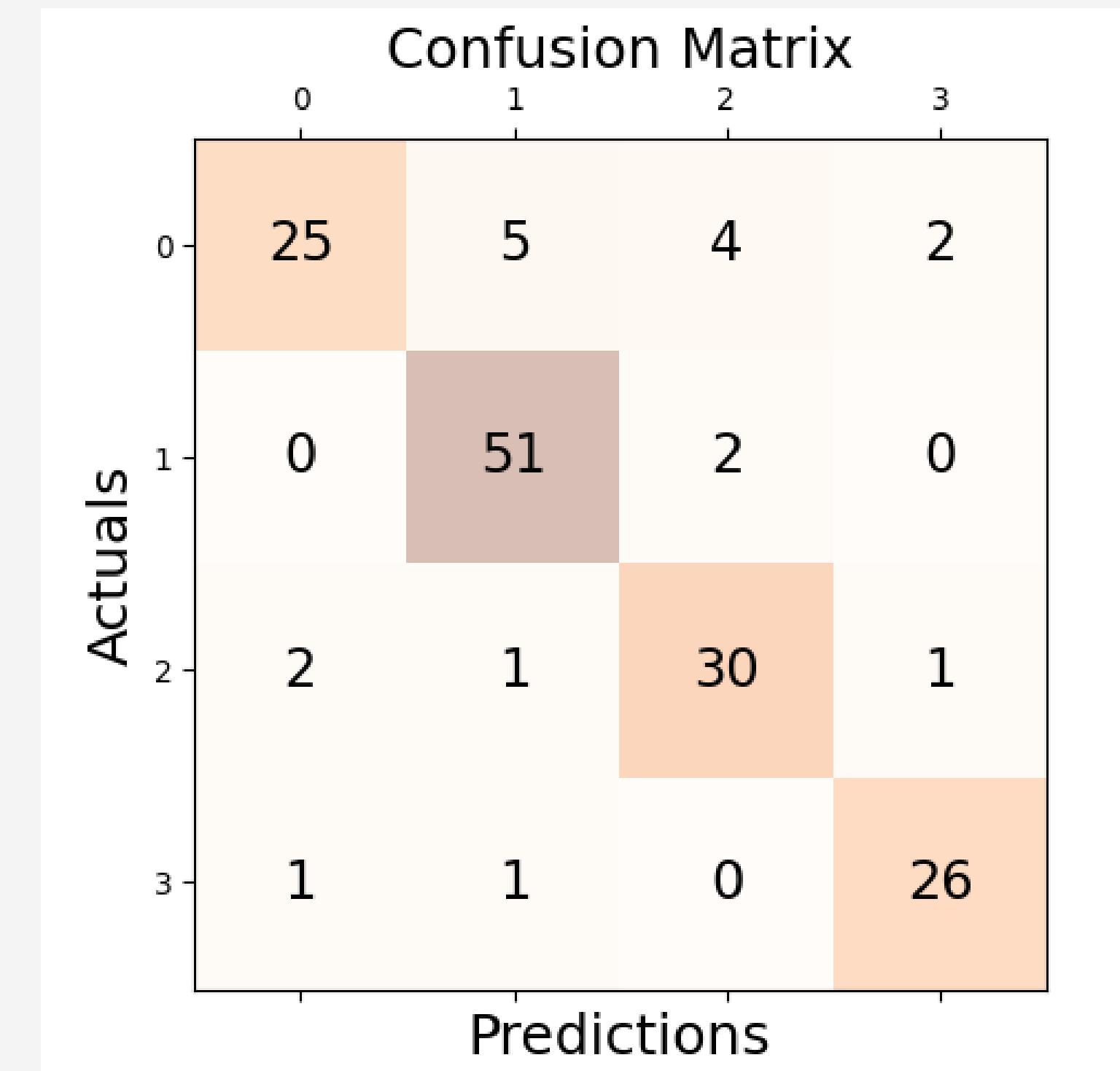


Analyse et interprétations

#3. Random Forest

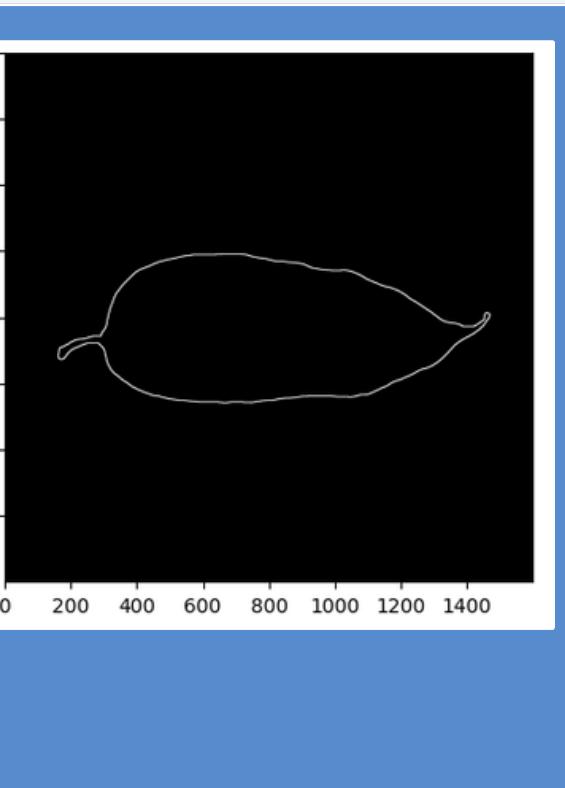
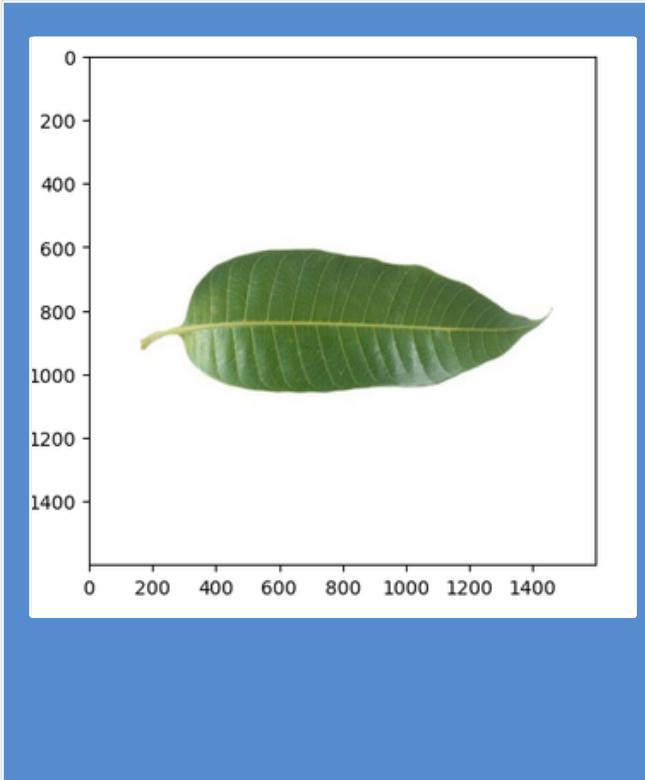
Il a tendance à confondre:

- les espèces de basilic (Basil) avec la goyave (Gauva) et le citron (Lemon)
- les espèces de citron (Lemon) avec celle du basilic (Basil)
- Et la goyave (Gauva) avec le citron (Lemon)



Test des modèles

#1 Mango



SVM Predictions:

Sample 1: Predicted: Mango

RF Predictions:

Sample 1: Predicted: Mango

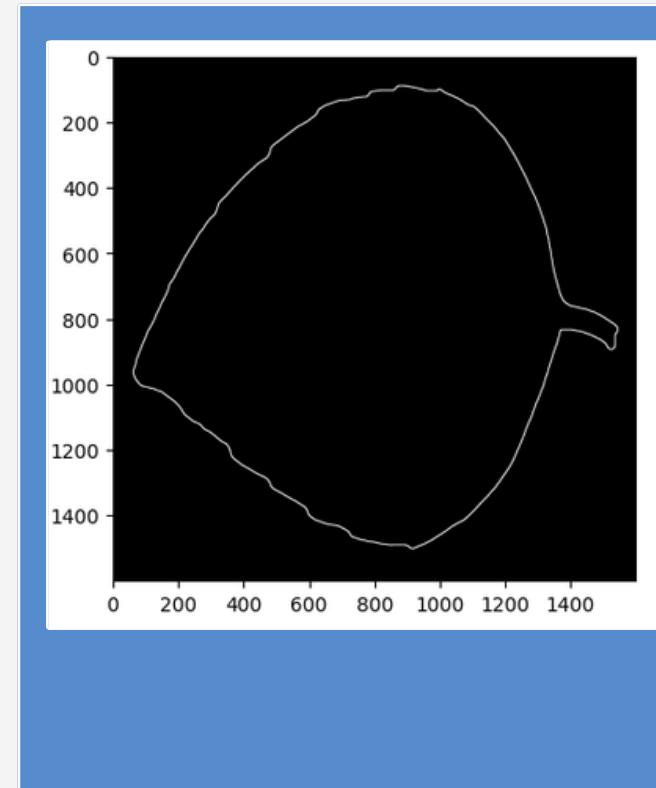
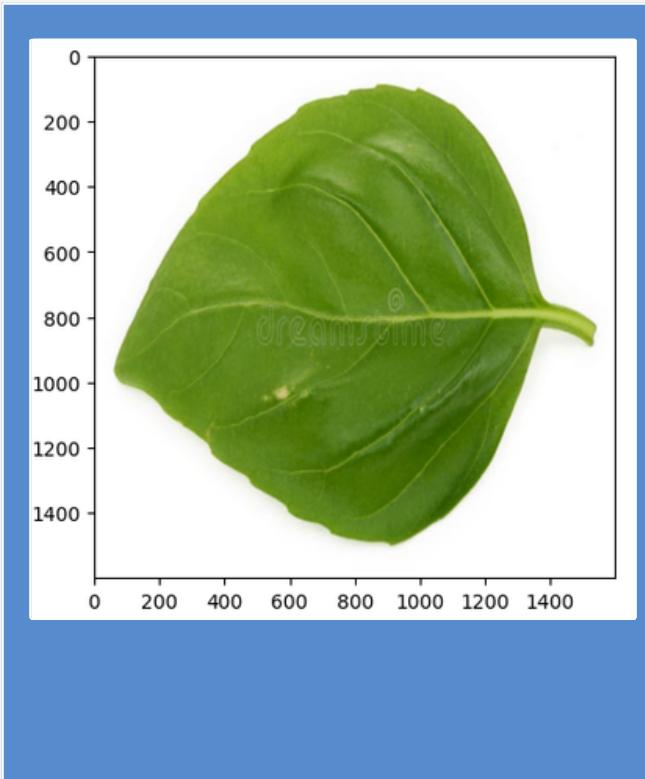
KNN Predictions:

Sample 1: Predicted: Basil

Le test est effectué avec une image issu d'internet. Seuls les modèles de SVM et de Random Forest ont réussi à prédire correctement l'espèce de cette feuille.

Test des modèles

#2 Basil



SVM Predictions:

Sample 1: Predicted: Mango

RF Predictions:

Sample 1: Predicted: Guava

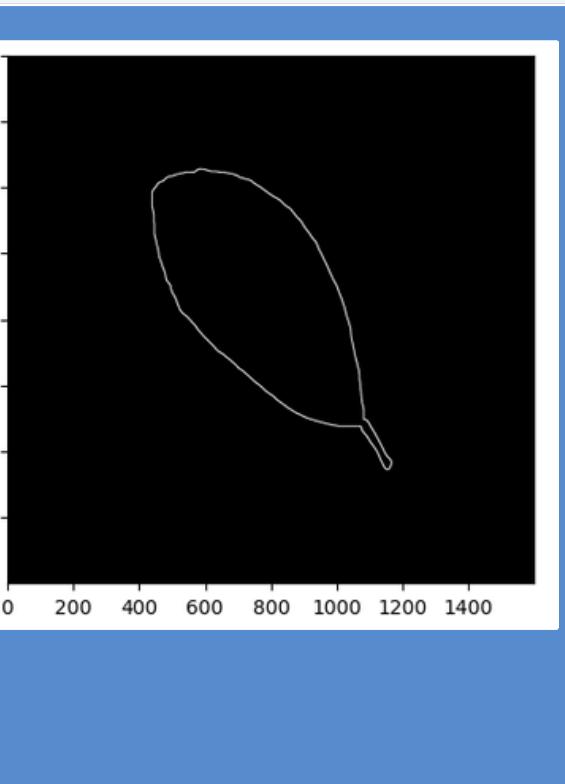
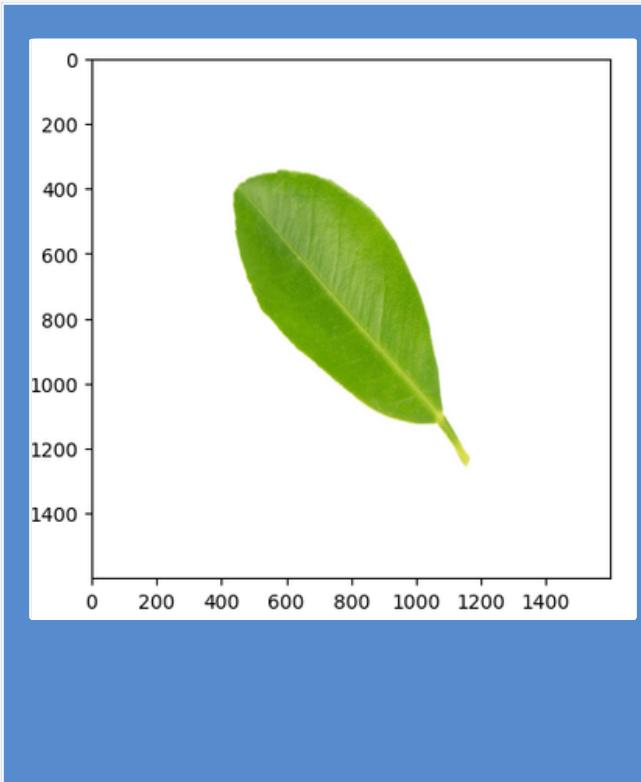
KNN Predictions:

Sample 1: Predicted: Basil

Le test est effectué avec une image issu d'internet. Seuls le modèle de KNN a réussi à prédire correctement l'espèce de cette feuille.

Test des modèles

#3 Lemon



SVM Predictions:

Sample 1: Predicted: Mango

RF Predictions:

Sample 1: Predicted: Lemon

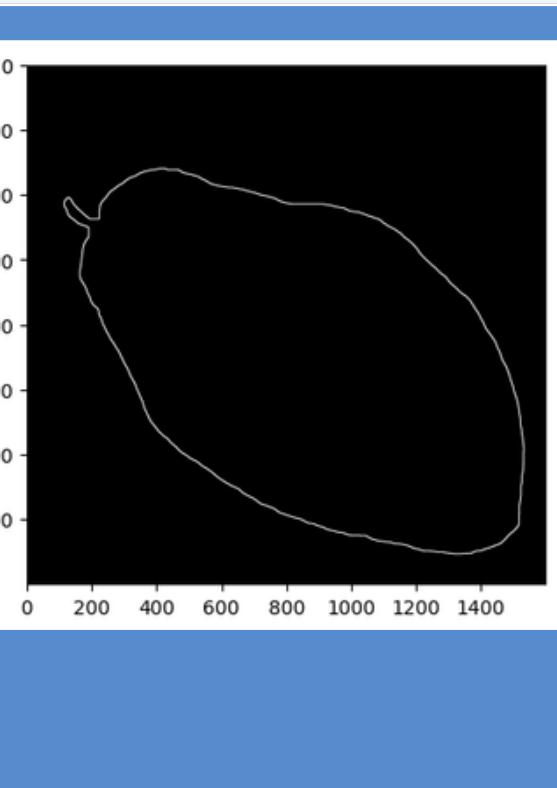
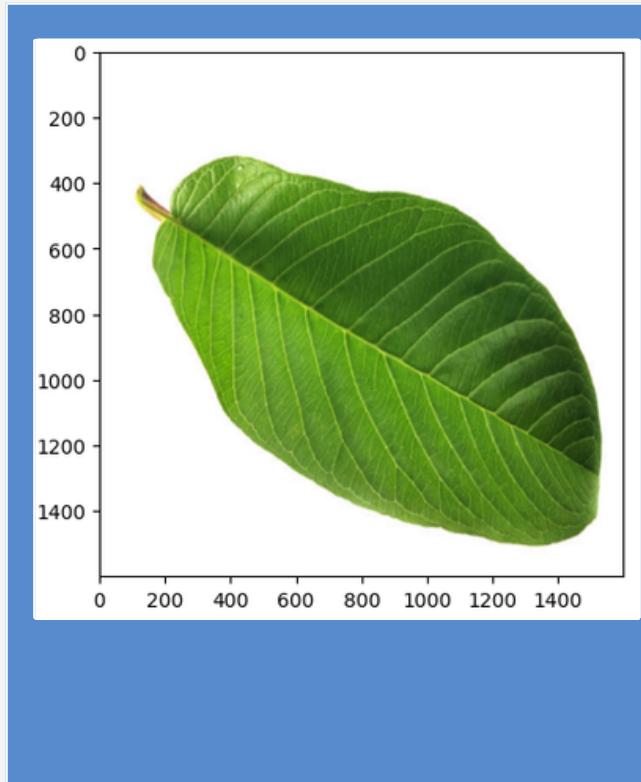
KNN Predictions:

Sample 1: Predicted: Basil

Le test est effectué avec une image issu d'internet. Seuls le modèle de Random Forest a réussi à prédire correctement l'espèce de cette feuille.

Test des modèles

#4 Gauva



SVM Predictions:

Sample 1: Predicted: Mango

RF Predictions:

Sample 1: Predicted: Guava

KNN Predictions:

Sample 1: Predicted: Basil

Le test est effectué avec une image issu d'internet. Seuls le modèle de Random Forest a réussi à prédire correctement l'espèce de cette feuille.

Défis rencontrés

- Banque d'images pas assez grande (À peu près 500 images)
- Ressource computationnelle limitée
- Technique de segmentation



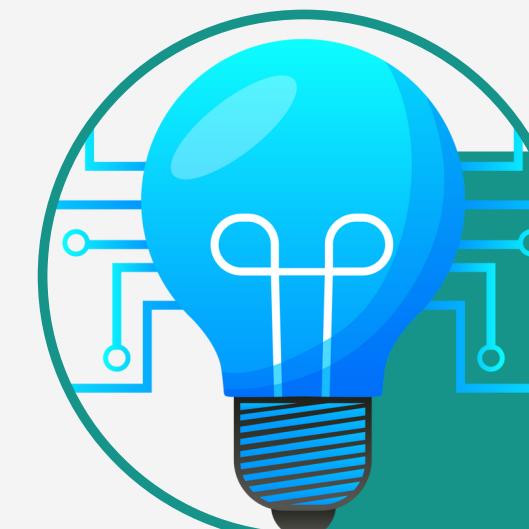
Conclusions

- Méthodologie de travail favorable à une bonne compréhension et utile pour un apprentissage de la computer vision
- Le SVM et le Random Forest sont à privilégier compte tenu des performances
- Le pré-traitement, étape la plus critique et essentielle du projet



Domaine d'application

Développement d'un outil utilisable par des chercheurs en botanique



Perspectives futures

- Mettre à disposition les données extraites
- Optimiser les modèles en variant les hyperparamètres
- Améliorer le pré-traitement pour une meilleure extraction des caractéristiques des images
- Faire de la data augmentation

MERCI POUR VOTRE ATTENTION

Contacter moi



Adama SAMAKE

- 92 02 90 08
- adama.samake.work@gmail.com

