# Reconnaissance de plantes

DataScientest - Promotion Février 2024

## Équipe

Felipe Souto
Nicolas Papegaey

# Introduction au Projet

## Contexte

### Contexte d'insertion du projet dans votre métier

Ce projet s'inscrit dans le cadre de notre formation en Data Science chez DataScientest, visant à appliquer des techniques avancées de traitement de données et de machine learning pour résoudre des problèmes réels. L'objectif principal est de diagnostiquer des maladies des plantes à partir d'images, ce qui peut avoir des implications importantes dans l'agriculture, permettant d'améliorer les rendements et de réduire les pertes.

### Du point de vue technique

Le projet utilise le dataset "PlantVillage" disponible sur Kaggle. Ce dataset contient des images de feuilles de plantes présentant différentes maladies, ainsi que des feuilles saines. Nous utilisons des techniques d'exploration de données et de data visualisation pour comprendre et préparer les données avant d'appliquer des modèles de deep learning.

### Du point de vue économique

L'agriculture est un secteur crucial de l'économie mondiale. Les maladies des plantes peuvent entraîner des pertes significatives de récoltes, impactant les revenus des agriculteurs et la sécurité alimentaire. Un système efficace de diagnostic des maladies peut réduire ces pertes, améliorer les rendements et augmenter la rentabilité agricole.

### Du point de vue scientifique

Scientifiquement, ce projet explore l'application de la vision par ordinateur (computer vision) dans l'agriculture, un domaine en pleine expansion. Il s'agit de développer des modèles capables de reconnaître les symptômes des maladies des plantes à partir d'images, contribuant ainsi à la recherche dans le domaine de l'agritech.

## Objectifs

Les principaux objectifs sont :

- Explorer et comprendre le dataset PlantVillage.
- Pré-processer les images pour les rendre utilisables par des modèles de machine learning.
- Visualiser les données pour identifier les outliers et éventuellement les supprimer.

C'est la première fois que nous travaillons sur un projet de vision par ordinateur. Ce projet a pour objectif de nous familiariser avec les concepts de base et les techniques couramment utilisées, représentant ainsi une opportunité d'approfondir nos connaissances dans ce domaine.

Nous ne sommes pas en contact avec des experts métiers pour affiner la problématique et les modèles sous-jacents.

# Compréhension et Manipulation des Données

## Cadre

Le dataset utilisé est le "PlantVillage Dataset" disponible sur Kaggle. Les données sont librement accessibles et disponibles à l'adresse suivante : PlantVillage Dataset.

Ce dataset contient plus de 50 000 images de feuilles de plantes.

Le dataset contient 3 répertoires racines représentant un traitement différent appliqués aux images:

- *color* : photographies d'origine
- *greyscaled* : photographies converties en noir et blanc
- *segmented* : photographies segmentées pour isoler les zones d'intérêt

Au sein des répertoires, les photographies sont classées en 38 catégories de maladies ou de feuilles saines.

Les répertoires color, greyscaled et segmented contenant les mêmes catégories et photos (seul le traitement de l'image diffère), nous focaliserons l'exploration des données sur le répertoire *color*.

## Pertinence

Les variables pertinentes dans ce contexte sont les images elles-mêmes classées par répertoire, représentant les différentes classes de maladies. La variable cible est la catégorie de la maladie ou l'état sain de la feuille.

### Particularité

Les particularités seront couvertes en détail dans l'analyse exploratoire, néanmoins nous pouvons déjà relever certaines caractéristiques :

- Les images ont toutes les mêmes dimensions : 256*256
- Les images sont toutes au format JPEG et au modèle de couleur RGB
- Les images sont dépourvues de métadonnées (ce qui aurait pu être intéressant à exploiter)
- Le dataset est assez déséquilibré en termes de classes

### Limitations des données

Certaines images peuvent être de mauvaise qualité ou en doublon, ce qui peut affecter la performance des modèles.

## Pre-processing et Feature Engineering

### Nettoyage et Traitement des Données

Compte tenu du type de données (images), nous n'avons pas eu à effectuer un pré-processing ou un feature engineering avancé. En revanche, nous avons construit un dataframe regroupant les caractéristiques des différentes images pour ensuite faire une analyse exploratoire.

**Démarche Utilisée**

La création du dataframe a pour but de structurer et organiser les données d'images de plantes afin de faciliter l'analyse. Les données sont organisées de la manière suivante :

- **Chemin** : Enregistrer le chemin complet de chaque image pour pouvoir y accéder facilement.
- **Classe** : Identifier la combinaison de la plante et de la maladie/état de santé pour chaque image.
- **Nom_image** : Conserver le nom du fichier pour des références futures et des vérifications.
- **Dimensions de l'image** : Stocker la largeur et la hauteur des images pour des analyses basées sur les dimensions.
- **Hash MD5** : Calculer et stocker le hash MD5 de chaque image pour détecter et gérer les doublons.
- **Format** : Format de l'image (JPEG, PNG, etc.).
- **Mode** : Retourner le mode de l'image (RGB, RGBA, L, etc.).

Deux colonnes supplémentaires ont été ajoutées en découpant le champ "Classe" :

- **Plante** : Enregistrer le nom de la plante.
- **Maladie/Statut** : Enregistrer le statut de la plante (en bonne santé ou avec virus/maladie).

Nous avons ensuite fait une recherche de doublon.

La constitution de ce premier dataframe nous a permis de récupérer un premier niveau d'information sur les images présentes dans notre dataset, confirmant l'uniformité de certaines caractéristiques (dimensions, format, mode). Il est à noter que ce processus initial de traitement a également permis de vérifier qu' aucune image n'était corrompue.

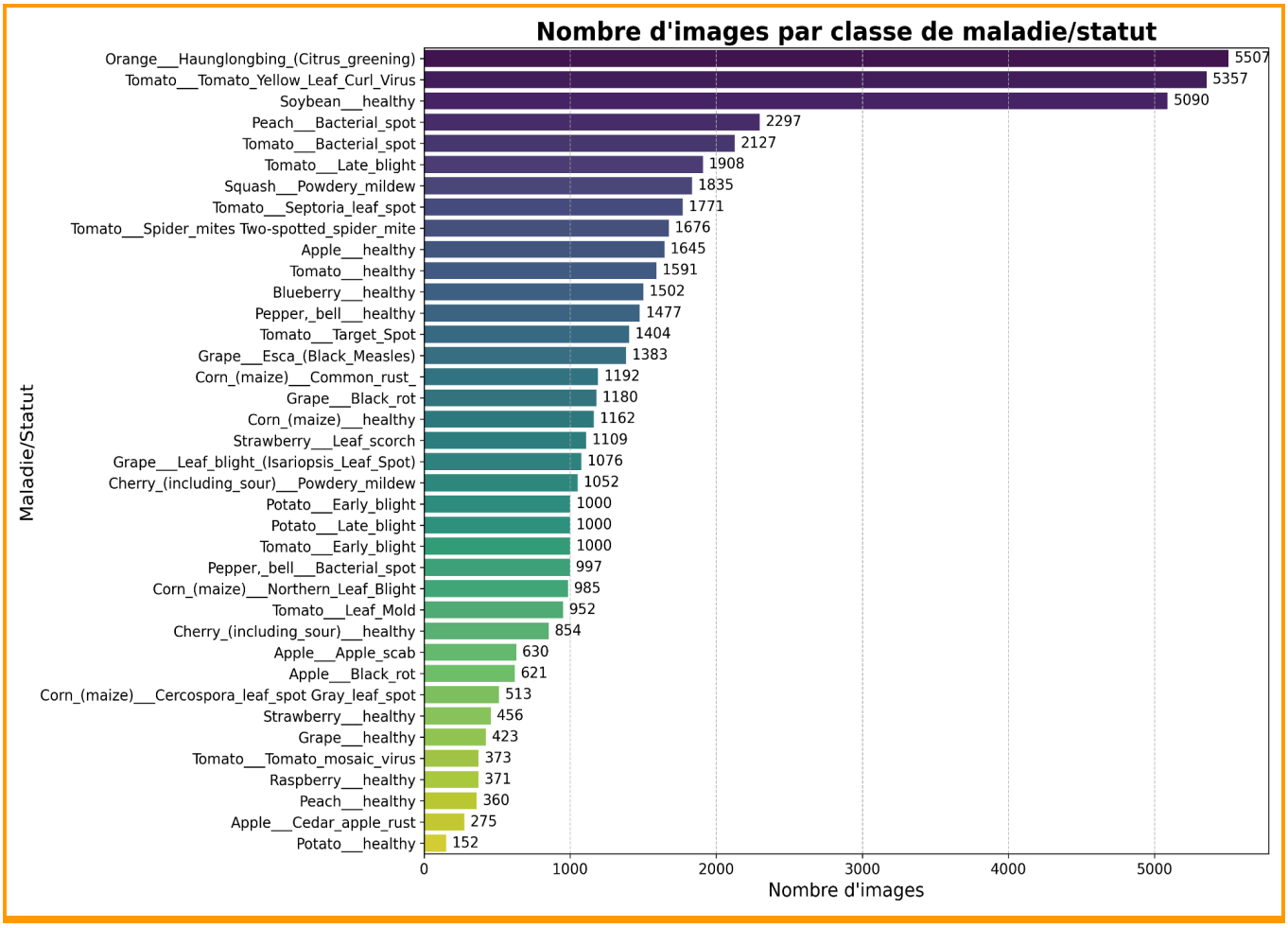
# Visualisations et Statistiques

## Analyse exploratoire

### Distribution des classes

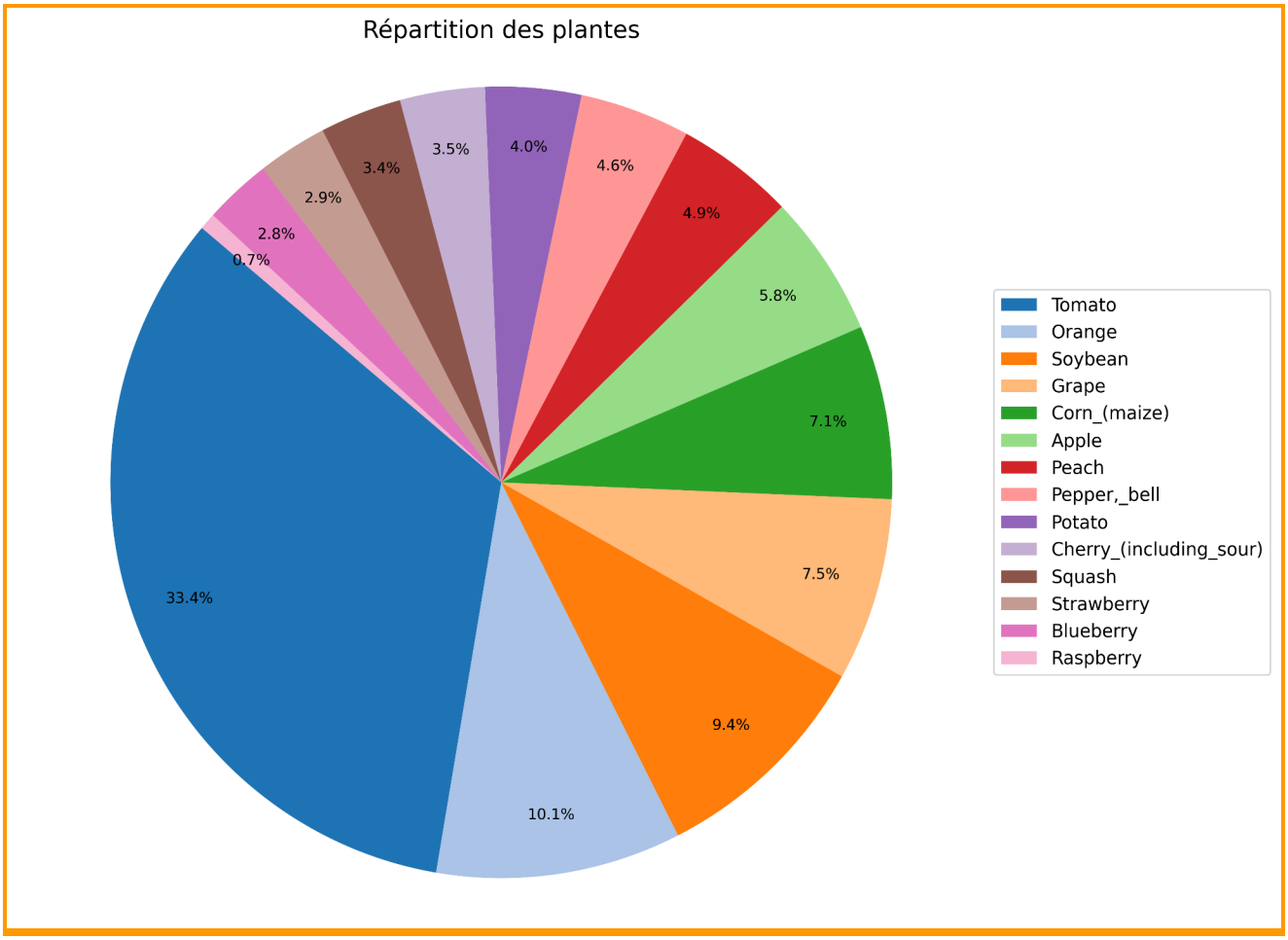
Nous pouvons constater une distribution inégale du nombre d'images par classe avec certaines classes, comme Orange___Haunglongbing (Citrus greening) et Tomato___Tomato Yellow Leaf Curl Virus, étant fortement représentées, tandis que d'autres, comme Apple___Cedar apple rust et Potato___healthy, sont beaucoup moins fréquentes.

Le déséquilibre dans la représentation des classes peut conduire à un biais du modèle vers les classes majoritaires, il faudra envisager une augmentation des données pour les classes minoritaires.



Répartition des espèces

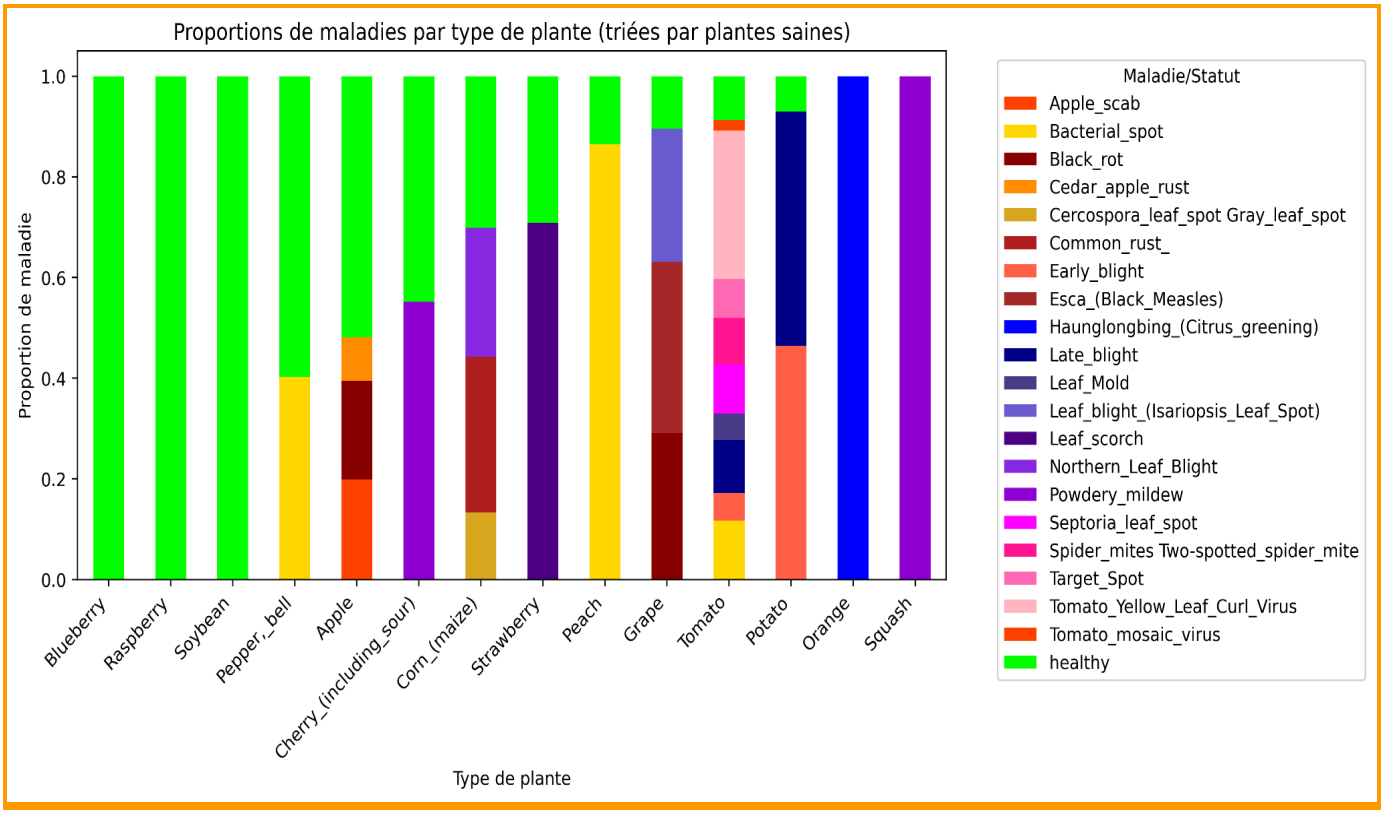
La répartition des espèces de plantes dans le dataset révèle quelles espèces sont sur-représentées et lesquelles sont sous-représentées. Par exemple, les tomates représentent 33,4% des images, tandis que d'autres plantes comme la courge, la fraise, la myrtille et la framboise sont sous-représentées.



Répartition des maladies par espèce de plante

Cette analyse montre la proportion de chaque maladie/statut pour chaque type de plante. Les plantes saines sont triées en premier, ce qui permet de voir immédiatement quelles plantes sont principalement en bonne santé et quelles plantes sont principalement affectées par des maladies spécifiques. Cela aide à identifier les priorités pour la modélisation et la détection des maladies.

On note une grande variabilité dans la prévalence des différentes maladies selon le type de plante. Certaines plantes, comme la myrtille, la framboise et le soja sont entièrement classées comme "saines", tandis que d'autres, comme les oranges et les courges, sont entièrement affectées par une maladie spécifique. Les pommes, le maïs ou bien les tomates présentent une grande diversité de maladies.

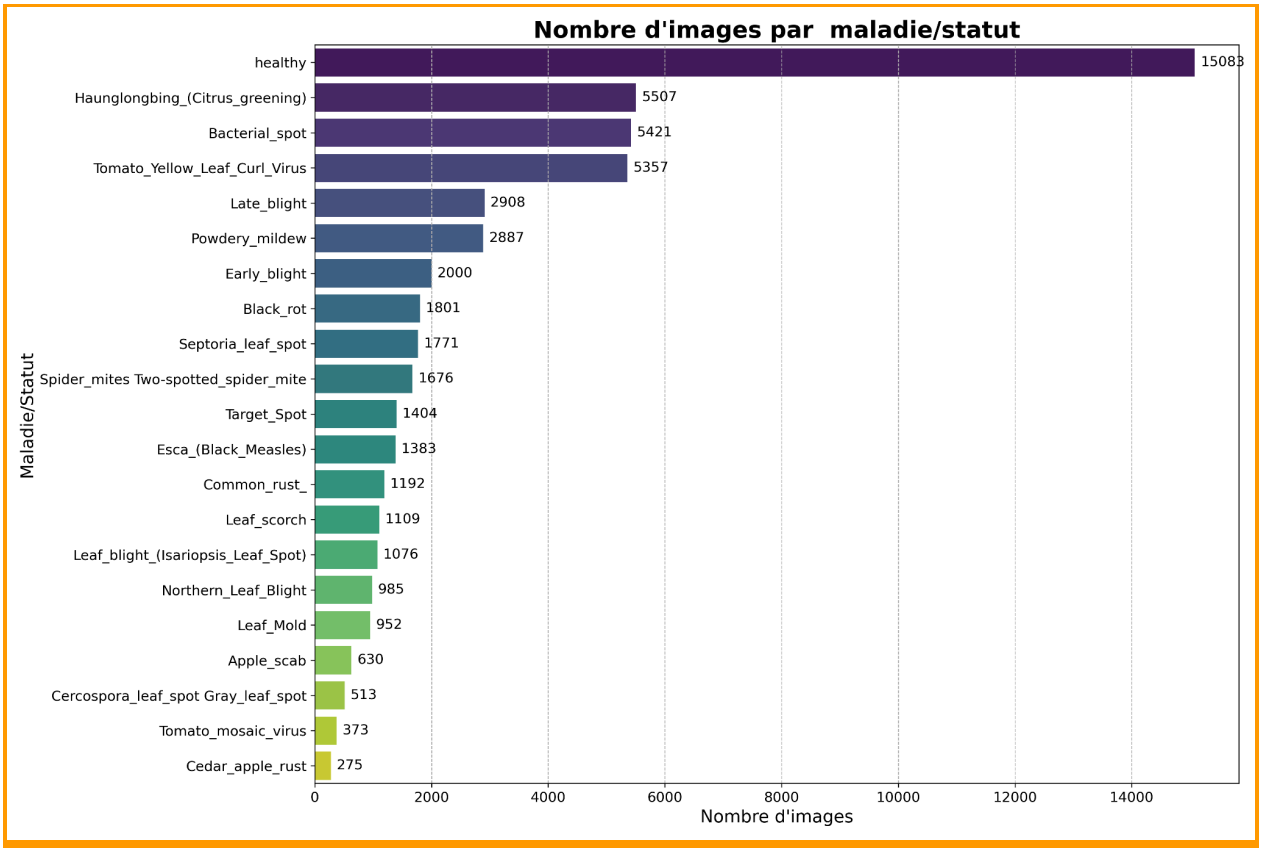


## Répartition des maladies

L'analyse de la distribution des maladies montre que la majorité des images des plantes sont en bonne santé (27.78% classée comme "healthy"). Cependant, plusieurs maladies comme Haunglongbing (Citrus greening), Bacterial spot, et Tomato Yellow Leaf Curl Virus sont relativement fréquentes, affectant environ 10% des plantes chacune.

Il est à noter que certains virus peuvent se retrouver au niveau de plusieures espèces :

- Bacterial_spot (pêche ,tomate ou poivron)
- Late_blight ou Early_blight (tomate ,pomme de terre)




## Identification des doublons

La colonne "hash MD5" du dataframe a permis de détecter 42 possibles doublons d'images dans l'ensemble des données. La vérification visuelle des images a confirmé que ces 42 cas sont effectivement des duplicatas.

Dans le cadre de la phase de pré-processing, on ne conservera qu'une version de ces images.

## Analyse des Outliers

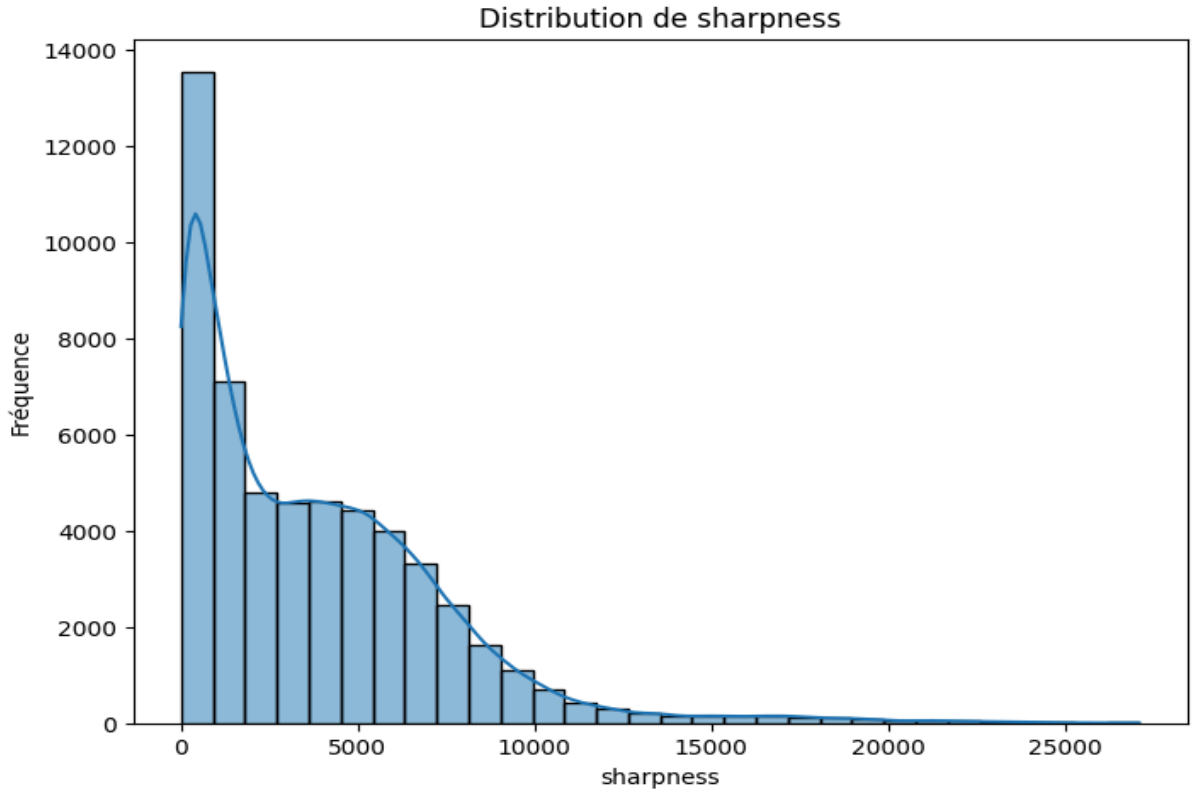
Afin de mieux connaître la qualité des images présentes dans le jeu de données, un certain nombre de métriques ont été calculées. Ces métriques fournissent des informations sur différents aspects des images, notamment leur netteté, leur luminosité, leur bruit, leur contraste, leur complexité , leur saturation, etc...

L'étude du résultat de ces métriques s'est ensuite réalisée en trois étapes :

- une analyse de la distribution globale du jeu de données
- la visualisation des principaux outliers min/max pour chaque métrique
- une comparaison qualitative des caractéristiques des outliers par rapport à la distribution générale des données non-outliers de même espèce/maladie

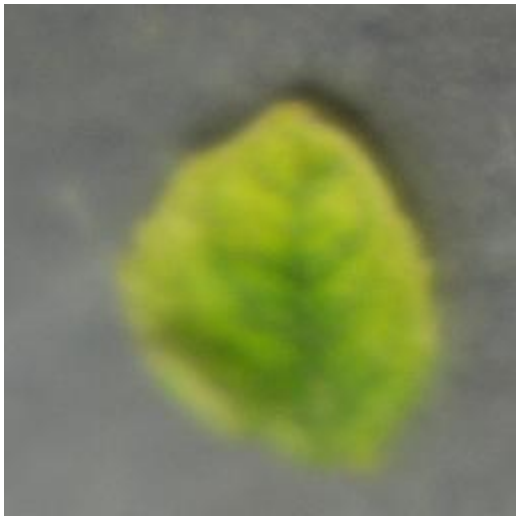
### Sharpness (Netteté)

On observe une distribution avec une forte asymétrie à gauche, ce qui indique un proportion importante d'images pouvant présenter moins de détails , voire être floues.




Les images avec une netteté minimale apparaissent floues, indiquant des problèmes lors de la prise de vue.

Pour les outliers min, on peut distinguer deux types de cas illustrés ci-dessous :



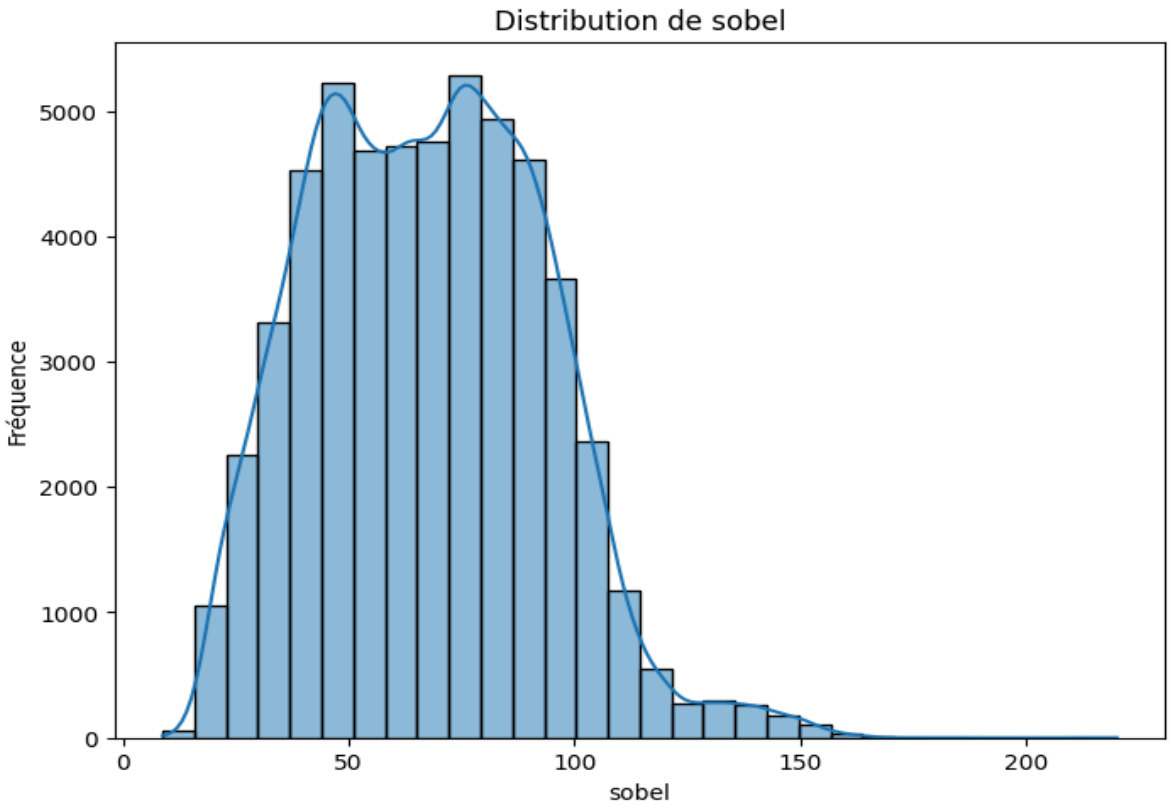

Dans le premier cas, la faible résolution rend difficile la distinction des détails fins. Dans le second cas, on ne perçoit pas les contours à cause du zoom, mais on perçoit quelques détails des rainures.

Les outliers max présentent des détails fins, notamment au niveau des nervures et de la forme.


Sobel

La métrique Sobel évalue la netteté en détectant les variations d'intensité des pixels.

On observe une distribution présentant une courbe en cloche avec deux pics.




Les images avec des valeurs minimales de Sobel sont souvent floues et peuvent avoir subi un traitement de segmentation.

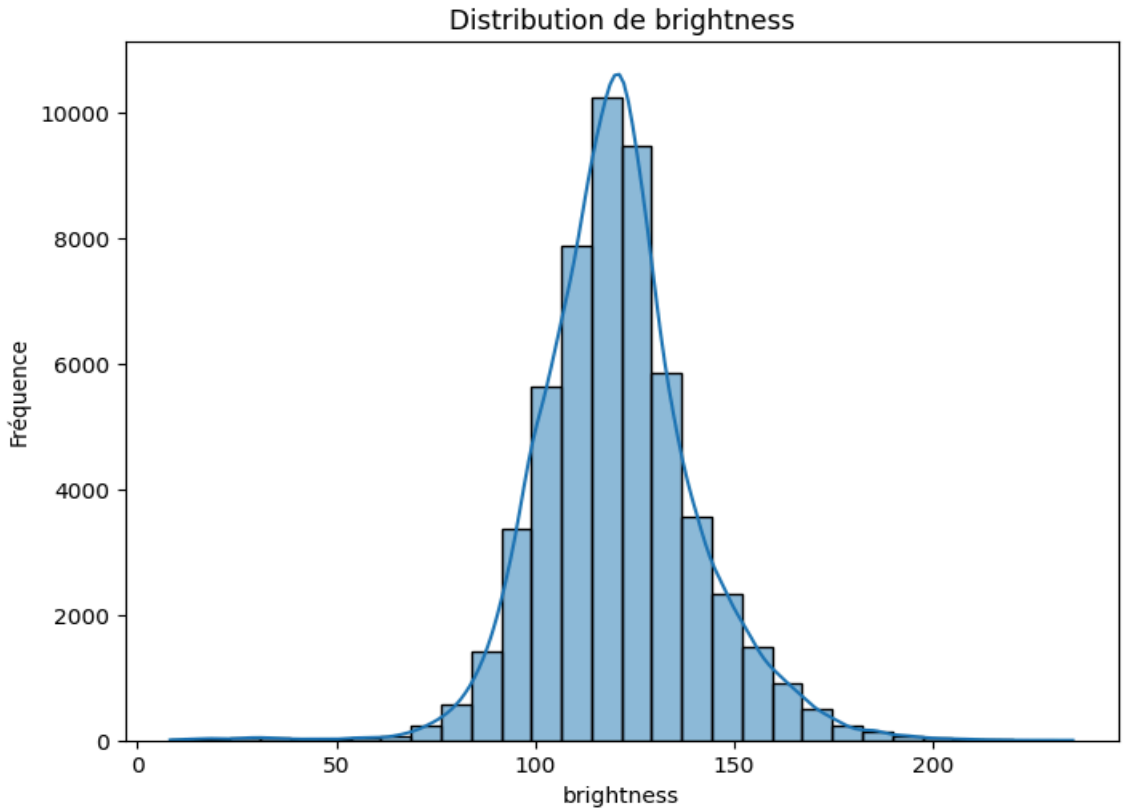
On remarque dans les outliers min des images ayant été déjà remontées par la métrique sharpness.

Des outliers ont également été identifiés, caractérisés principalement par des feuilles sur un fond noir uni, ce qui met néanmoins en évidence la texture et la nervure de la feuille.


Un outlier maximal montre une perspective différente avec des feuilles capturées en pleine nature. La présence de divers types de plantes peut générer du bruit dans la mise en place d'un modèle.


Brightness (Luminosité)

La distribution de la luminosité montre une courbe en forme de cloche ce qui indique que la plupart des images ont une luminosité moyenne autour de la valeur centrale.

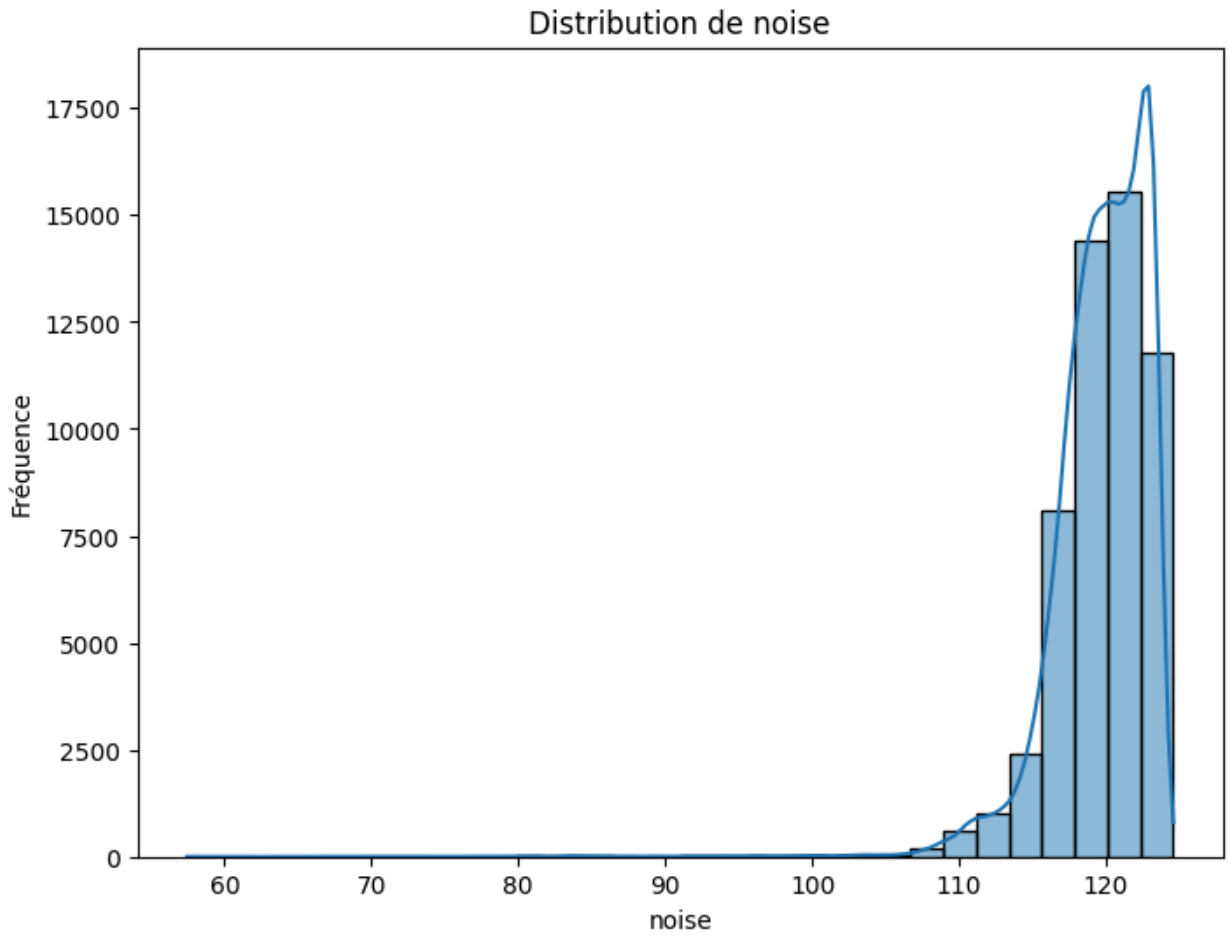



Les images avec une luminosité minimale ont souvent un fond noir dû à la segmentation comme les images suivantes :

Les outliers maximaux montrent des images très lumineuses, suggérant une surexposition.

Noise (Bruit)

La distribution du bruit semble être asymétrique et décalée vers la droite avec une forte concentration autour du pic autour de 120 unités.

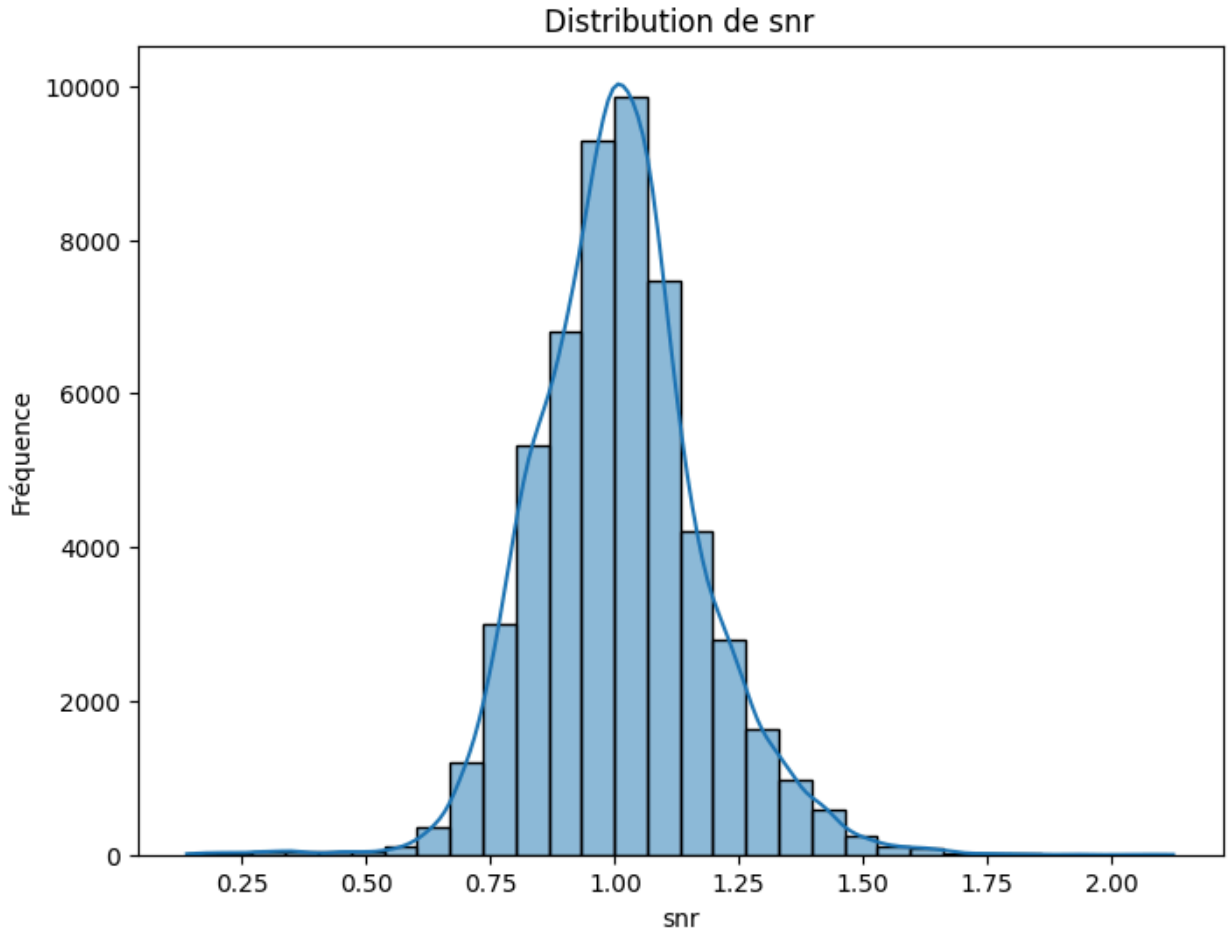


Les images avec un bruit minimal ont souvent un fond noir déjà recontré lors de l'analyse d'autres métriques. Les feuilles semblent montrer des dommages liés à la maladie : décoloration ou flétrissement.


Les outliers maximaux présentent un niveau de zoom élevé, révélant des détails fins.


SNR (Rapport Signal sur Bruit)

La distribution du SNR montre une courbe en forme de cloche ce qui indique que la plupart des images ont un SNR moyen autour de la valeur centrale 1.

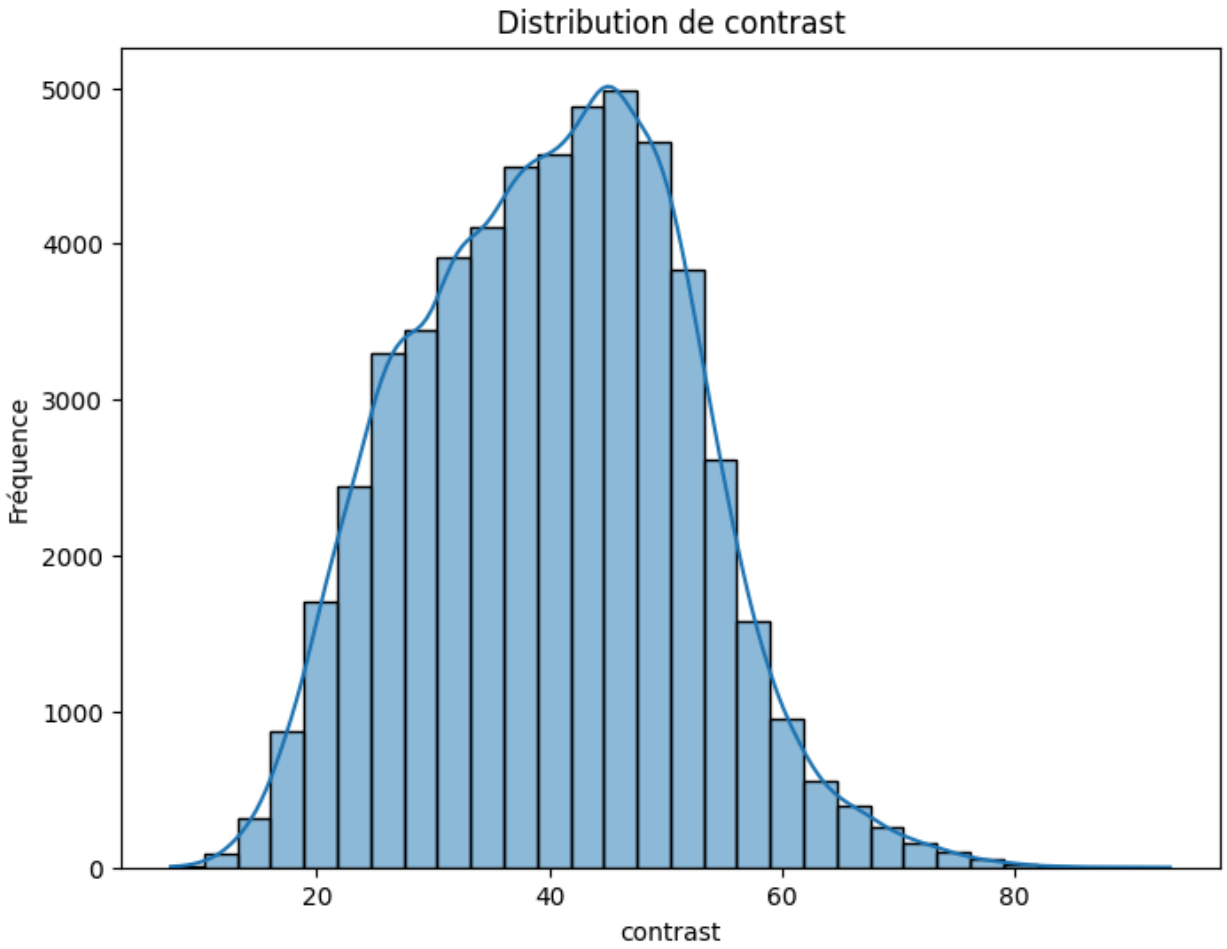


Les outliers minimaux de SNR correspondent aux mêmes images que celles identifiées avec la métrique de bruit (fond noir uni)

Les outliers maximaux de SNR correspondent à des images avec une forte luminosité où certains détails sont bien perçus.

Contrast (Contraste)

La distribution du contraste montre une courbe en forme de cloche avec un écart important montrant la diversité des conditions de prises de vue des images de plantes.

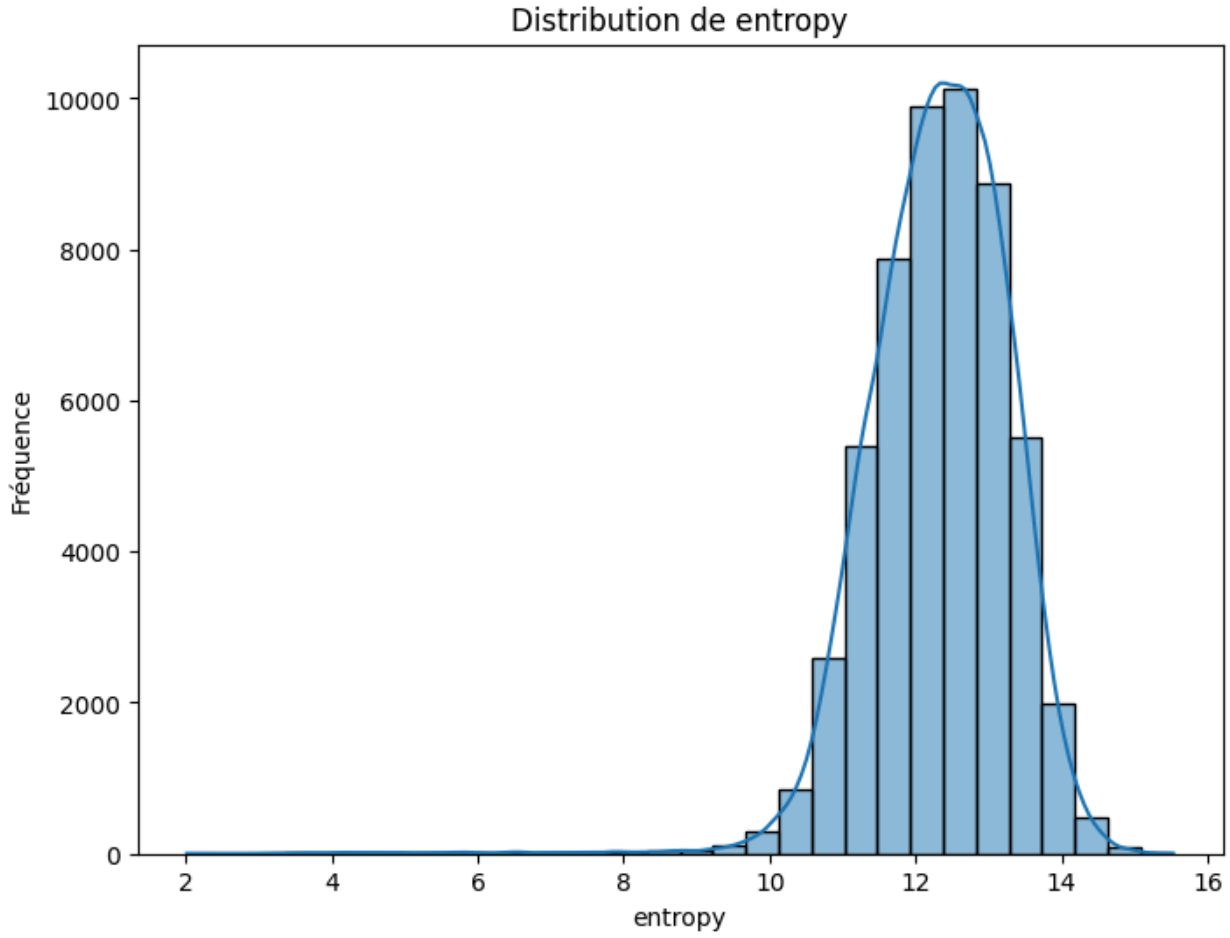


Les images avec un contraste minimal ont un fond gris et une couleur uniforme.


Les outliers maximaux montrent un contraste élevé avec des tons forts et des signes de maladie (tâches brun-rougeâtre par exemple).


Entropy (Entropie)

On observe une distribution de l'entropie en forme de cloche avec une grande partie des images dont la valeur est comprise entre 10 et 14. Cela montre un dataset dont la complexité est globalement uniforme.

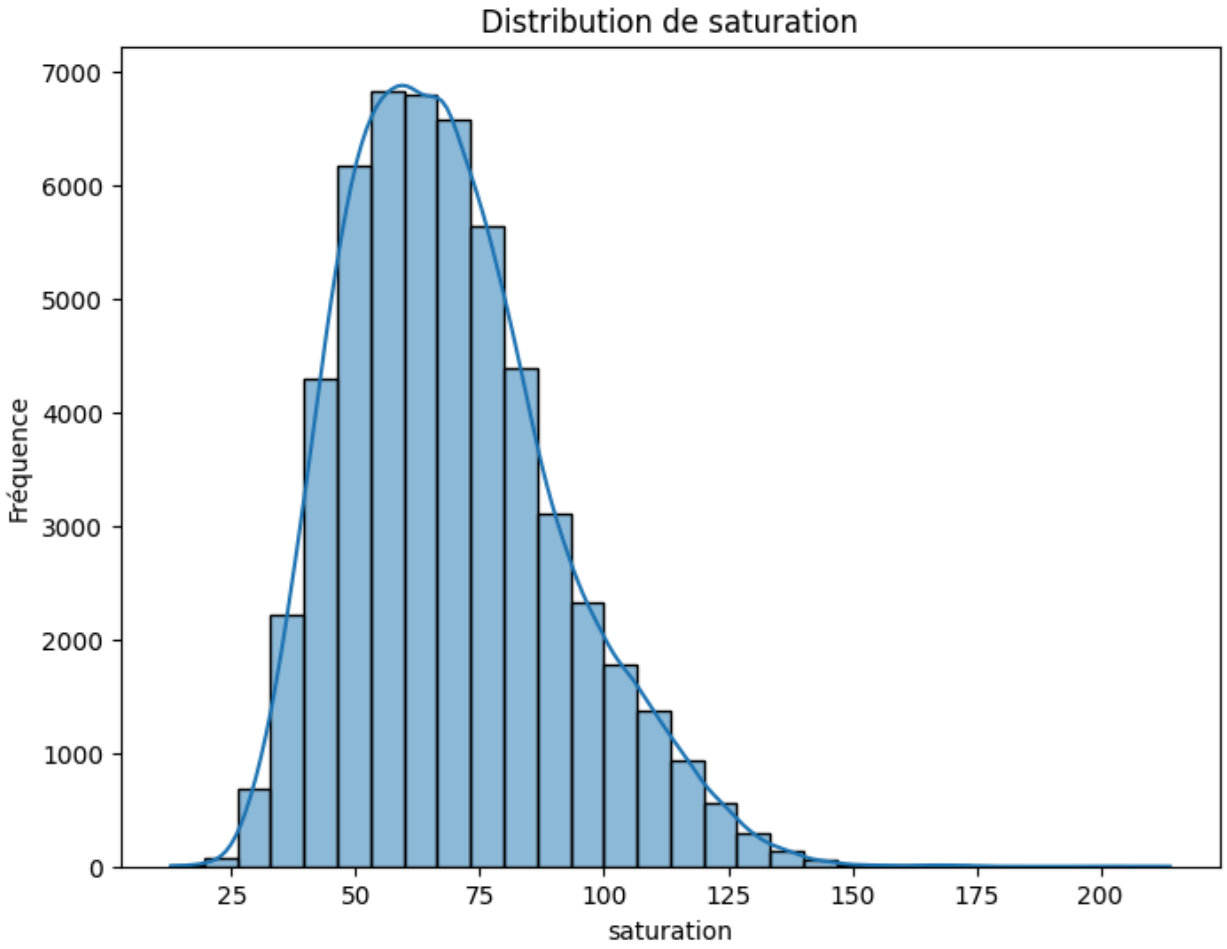


Les outliers minimaux montrent des images avec un fond noir uni déjà rencontrées lors de l'évaluation d'autres métriques.

Les outliers maximaux montrent des images avec une grande diversité de motifs et des signes de maladie très visibles.

Saturation

La saturation mesure l'intensité des couleurs dans une image. La distribution ressemble à une courbe en forme de cloche avec un décalage vers la gauche.



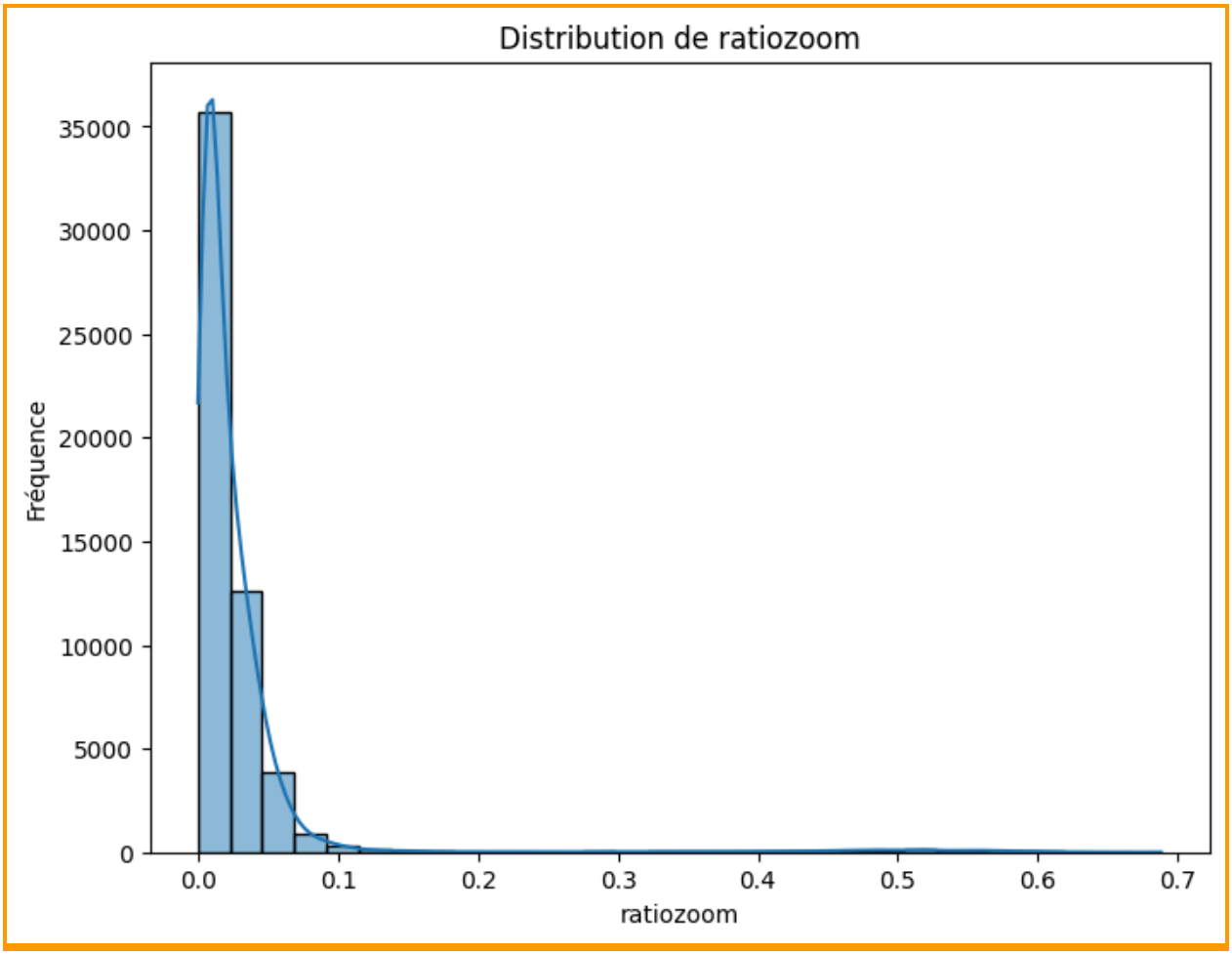
Les images avec une saturation minimale ont des couleurs très atténuées, parfois sans feuille visible.


Les outliers maximaux montrent une saturation élevée avec des feuilles et des fruits très colorés voire des fleurs.


Ratiozoom

Cette métrique mesure combien de la surface de l'image est occupée par les contours détectés.

La distribution montre une forte asymétrie à gauche laissant à penser qu'un partie importante du dataset est constituée d'images dont la surface délimitée par les contours est relativement faible par rapport à la surface de l'image.



Les outliers minimaux montrent des images très zoomées qui montrent les détails des rainures ,de la texture et des tâches caractéristiques de maladie.

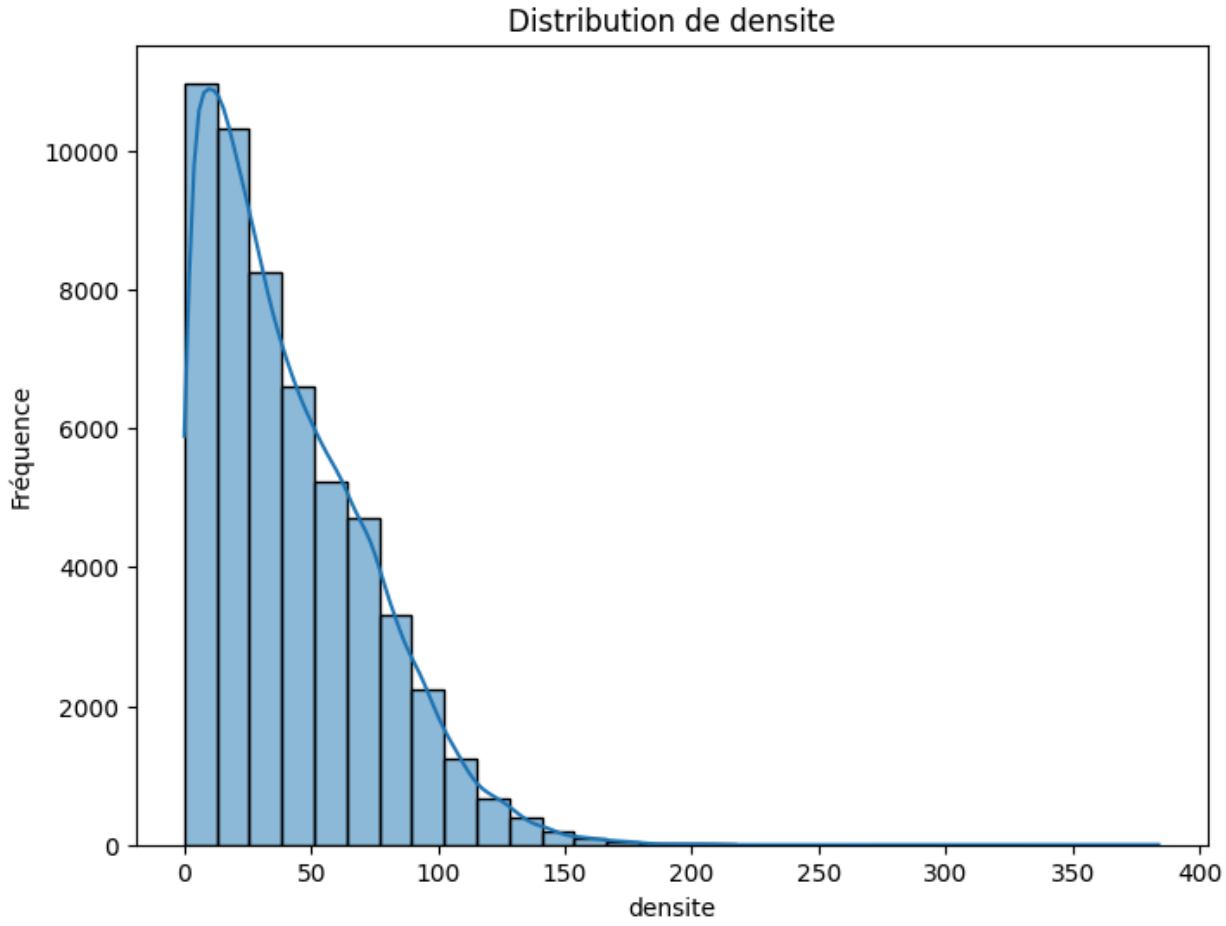
Il est à noter que ce niveau de zoom est caractéristique des feuilles de maïs. Le jeu de données présente 4 catégories d'états pour les feuilles de maïs et dans les 4 cas la distribution du ratiozoom est très similaire.

Les outliers maximaux mettent en évidence les contours et les détails de la feuille. Il s'agit d'une prise de vue optimale pour prendre en compte le plus de caractéristiques possibles.

Densité

Cette métrique tente de mesurer le nombre d'objets présents dans l'image, en se basant sur le nombre de contours.

La distribution présente une forte asymétrie à gauche suggérant qu'une majorité d'images présente peu de contours.



Les outliers minimaux ont déjà été identifiés par d'autres métriques et correspondent soit à des images très zoomées, soit à des images avec des feuilles de forme très régulière.

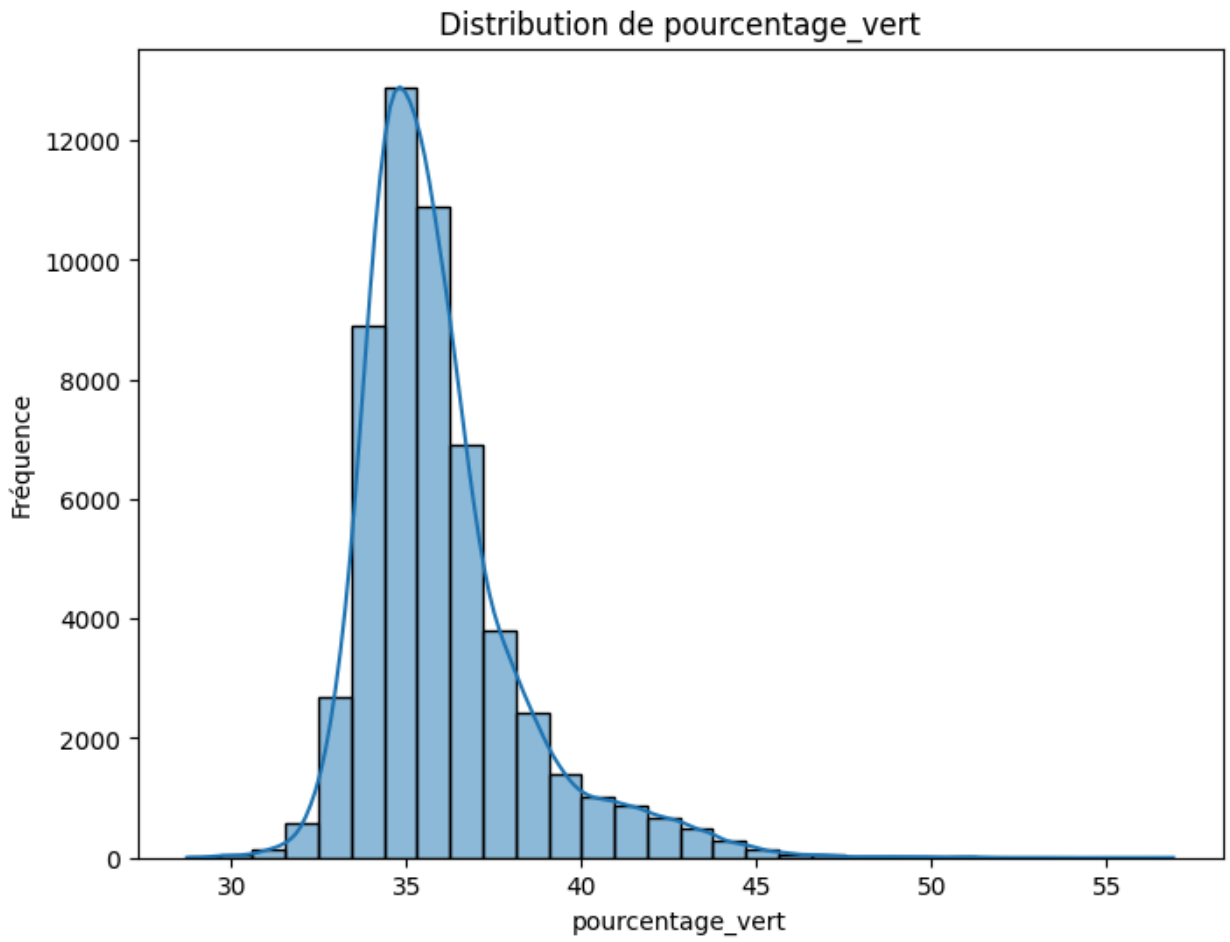
Les outliers mettent en valeur une complexité plus importante que l'information attendue avec la présence d'objets externes à la catégorie des feuilles de plantes ou bien la présence de plusieurs feuilles d'une même espèce (avec des statuts différents) ou de différentes espèces.




Pourcentage de vert

Cette métrique mesure le pourcentage de vert présent dans l'image.

La distribution suit une courbe en cloche avec une valeur moyenne autour de 35%.



Les outliers minimaux présentent des images de plantes avec un fond violet et une couleur caractéristique de maladie (ici le blanc ou le rouge).


Les outliers maximaux correspondent à des images déjà remontées par d'autres métriques qui montrent des détails de la plante et les variations de couleur avec une prédominance de vert (zoom).


## Conclusions Préliminaires

Dans l'élaboration d'un modèle de détection de maladies dans les plantes, il est essentiel de focaliser l'attention sur les feuilles des plantes pour permettre au modèle de se concentrer sur les détails susceptibles de révéler la présence de maladies.

Dans cette optique, nous recommandons de restreindre notre jeu de données en excluant uniquement les images qui ne répondent pas aux critères suivants :

- Cohérence des plantes : Exclure les images présentant une diversité trop large de plantes afin de garantir une cohérence dans les caractéristiques visuelles traitées par le modèle. Les images doivent se concentrer sur des types de plantes spécifiques pour une meilleure précision.
- Éléments perturbateurs : Éliminer les images contenant des éléments externes trop visibles, tels que des morceaux de bois, de carton, ou autres objets non pertinents, qui pourraient perturber l'analyse visuelle du modèle.
- Caractéristiques non ciblées : Exclure les images mettant en avant des caractéristiques telles que des fruits ou des fleurs. L'objectif est de maintenir la concentration sur les feuilles, qui sont les principaux indicateurs de la santé des plantes dans ce contexte.
- Absence de feuilles : Retirer les images où les feuilles de plantes sont absentes, car elles ne fournissent pas les informations nécessaires pour l'analyse.
- Doublons : Éliminer les images en doublon pour éviter les redondances et garantir que chaque image apporte une valeur unique au dataset.