Telecom Saint-Étienne
FISE 2

MALLOWMETER
– Données –

Étude en Filière Ingénieur sous Statut Étudiant
28.02.2025

Maxime JOURNOUD Lucas LESCURE Aubin SIONVILLE
Ruben VERCHERE

Table des Matières

I. Rappel du projet 3

II. État de l'art 3

III. Définition de la base de données 3

IV. Procédures d'acquisition 4

V. Tri de la base de données 4

VI. Critères d'évaluation 4

VII. Métriques d'évaluation 5

VIII. Annexes 6

I. Rappel du projet

L'objectif est de classifier des photos de marshmallows blancs en fonction de leur niveau de cuisson. L'application devra fournir, à partir d'une photo de marshmallow, une étiquette correspondant à sa classe parmi les **4 degrés de cuisson** possibles:

1Pas cuit 2Peu cuit 3Bien cuit 4 Trop cuit

II. État de l'art

N'ayant pas trouvé de base d'images de marshmallows répondant à nos besoins (différents degrés de cuisson, nombre important d'images, diversité des cuissons), nous avons décidé de la constituer nous-mêmes, principalement en réunissant des clichés trouvés sur le web. En parallèle, la recherche d'images sur Internet sera complétée par une acquisition manuelle afin de garantir un nombre conséquent d'échantillons et une diversité dans les cuissons.

Nous disposons de **62 images de marshmallows acquises depuis le web et 121 par acquisition manuelle**, réparties ainsi :

Degré de cuisson	Images Web	Images Manuelles	Total
Pas cuit	15	27	42
Peu cuit	9	32	41
Bien cuit	24	33	57
Trop cuit	13	29	42

Table 1. Répartition des images selon le degré de cuisson

Afin d'équilibrer notre base, nous prévoyons une option afin de limiter le nombre d'échantillons utilisés à un même nombre de chaque classe (40 par défaut). On laissera tout de même la possibilité d'utiliser toutes les données pour réaliser une étude sur l'impact du déséquilibre.

On aura éventuellement recours à l'enrichissement dans le cas où les données seraient insuffisantes, notamment dans un second temps pour la classification CNN.

III. Définition de la base de données

Notre base de données sera composée de **4 classes**, avec **40 images minimum par classe**, soit un total d'au moins **160 images**. Dans chaque image, le marshmallow sera extrait du fond.

Les images seront toutes placées dans un dossier commun et respecteront les critères suivants :

- **Format** : PNG 512×512 pixels
- **Fond** : Neutre, avec le marshmallow au premier plan
- **Type de marshmallow** : Principalement de couleur blanche

Les fichiers seront nommés selon la structure suivante, où `[X]` est le numéro de l'échantillon et `[0-1-2-3]` la classe déterminée manuellement :

- `img_[0-1-2-3]_[X].png`

IV. Procédures d'acquisition

Les images collectées sur internet seront choisies arbitrairement en prenant en compte le détachement des marshmallows par rapport à leur fond, et par rapport à la cuisson. On évitera de prendre des images trop pixelisées et on fera attention à vérifier les droits d'images afin d'être sûr qu'elles soient libres d'utilisation.

D'autre part, les images qui seront prises à la main seront prises en faisant attention à ce que le fond puisse également bien se détacher.

Nous effectuerons ensuite une **suppression de l'arrière-plan** de façon à n'avoir que le marshmallow comme objet d'intérêt. Ceci sera réalisé à l'aide d'outils comme RemoveBG ou RemovalAI. Les images seront ensuite **rognées**, de façon à obtenir un ratio objet/image d'environ 30 à 75% sur toutes les images. Enfin, on réalisera de façon adéquate une **mise à l'échelle** pour s'assurer que la taille des images soit toutes au format **512x512** pixels.

V. Tri de la base de données

Nous réalisons des **boîtes à moustaches** sur l'ensemble des descripteurs envisagés afin de déterminer les plus pertinents *(voir Annexes)*. Ensuite, à partir de ces caractéristiques, nous **éliminons les images aberrantes** pour conserver une cohérence dans chaque classe.

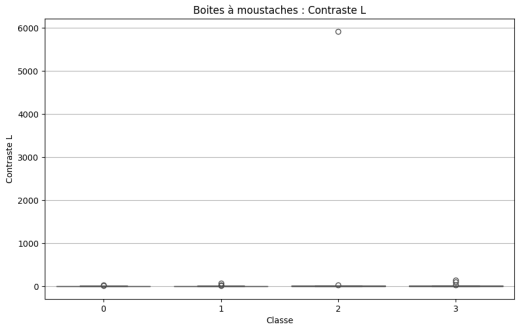


Figure 1. Présence d'images aberrantes dans la classe 2

VI. Critères d'évaluation

Nous allons tout d'abord effectuer une **analyse descriptive des données** afin d'évaluer la distribution des caractéristiques visuelles des images et détecter d'éventuels déséquilibres dans notre base. Cette analyse portera sur les aspects suivants :

- **Moyenne et écart-type des composantes L et b dans l'espace Lab pour chaque classe** : L'espace colorimétrique Lab permet une meilleure représentation des variations de couleur et de luminosité. La composante **L** correspond à la luminosité (du noir au blanc), tandis que la composante **b** représente la balance entre les tons jaunes et bleus, particulièrement pertinente pour identifier les degrés de cuisson des marshmallows.
  Nous calculerons **la moyenne et l'écart-type** de ces valeurs pour chaque classe afin de vérifier si les classes sont bien différenciées en termes de couleur et si certaines d'entre elles présentent une trop grande **variabilité intra-classe**.

Ces analyses nous permettront de détecter d'éventuels déséquilibres et de prendre les mesures nécessaires pour améliorer notre base de données. Il est également essentiel d'évaluer la qualité de notre base de données afin de garantir des performances correctes du modèle de classification. Deux critères sont particulièrement importants :

- **La représentativité** : notre base doit être variée, afin de couvrir un **éventail assez large de cas possibles** et d'**éviter un modèle spécifique à un contexte particulier**. Une mauvaise représentativité risquerait de réduire la capacité de généralisation du modèle et d'entraîner une mauvaise classification sur des données inconnues.
- **Équité** : chaque classe doit être **représentée de manière équilibrée** afin d'éviter des biais dans l'apprentissage. Une classe sur-représentée pourrait amener le modèle à privilégier certaines prédictions au détriment des autres, ce qui réduirait sa précision globale.

VII. Métriques d'évaluation

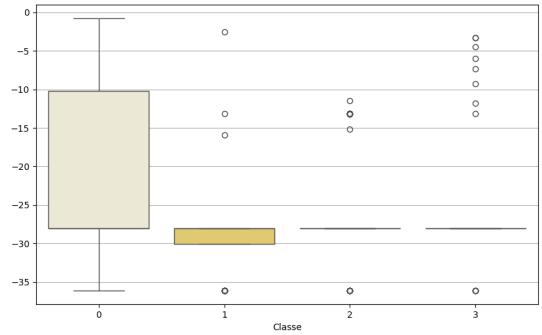
- **Évaluation de l'équilibre des classes** : Notre base de données est composée de $k = 4$ classes, notées $i \in \{0, 1, 2, 3\}$, contenant chacune N_i échantillons, pour une proportion $p_i = \frac{N_i}{\sum_{j=0}^{k-1} N_j}$ Notre mesure de l'équilibre entre les classes sera **l'entropie de Shannon normalisée**, définie par :

  $$H_n = -\frac{1}{\log_2(k)} \sum_{i=0}^{k-1} p_i \log_2(p_i),$$

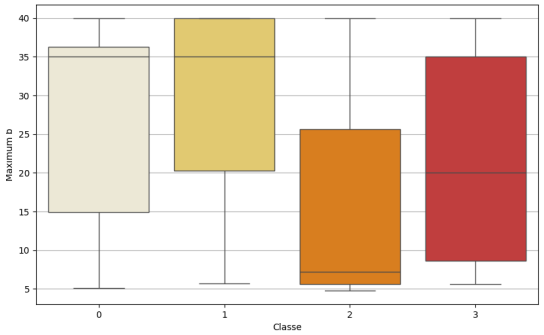
  où H_n varie entre 0 (base très déséquilibrée) et 1 (répartition parfaitement uniforme). Nous considérons que la base est suffisamment équilibrée si $H_n > 0.75$.

  Après traitements, on obtient une entropie $H_n = 0.993$
- **Vérification de la diversité intra-classe** : Au-delà de la répartition globale des échantillons, il faut s'assurer que **chaque classe couvre une variété suffisante de cas** pour éviter un **sur-apprentissage** sur des caractéristiques trop spécifiques. Pour cela, nous analysons la variation des conditions de prise de vue: les images doivent inclure **différentes directions d'éclairage et angles de vue** afin de rendre le modèle robuste aux variations du contexte. Cette partie sera vérifiée empiriquement au vu de la difficulté de l'établissement d'un critère de mesure objectif.
  L'objectif est d'éviter que le modèle ne se base sur des **artefacts visuels non pertinents** (ex: un fond particulier présent majoritairement dans une classe) et qu'il puisse **généraliser** efficacement à des images inconnues.

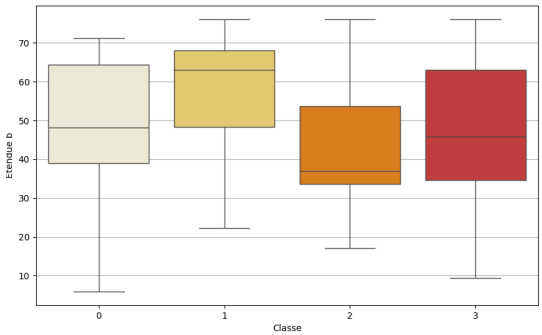
VIII. Annexes



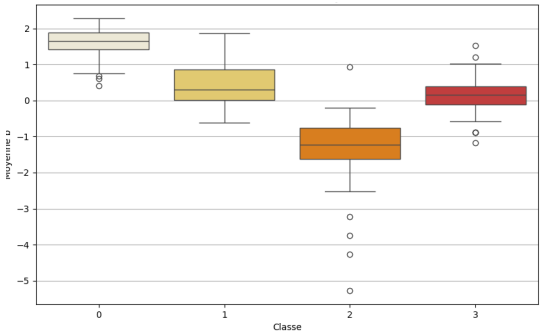
(a) Min b



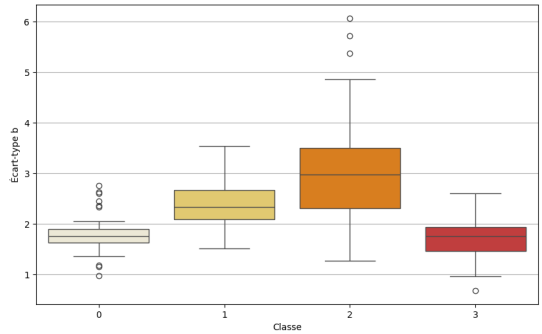
(b) Max b



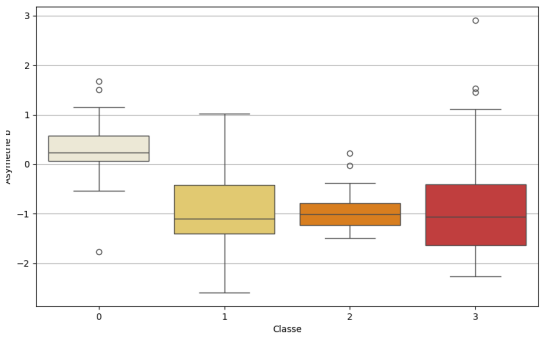
(c) Étendue b



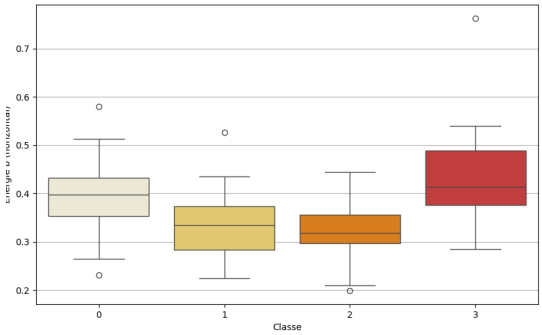
(d) Moyenne b



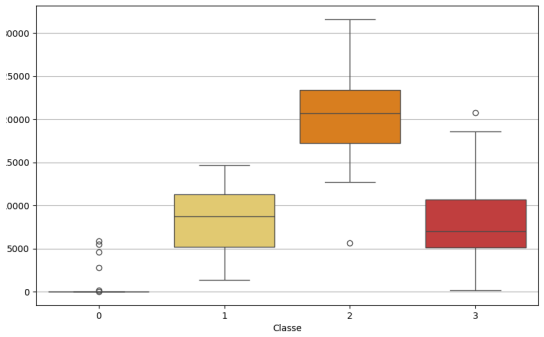
(e) Écart-type b



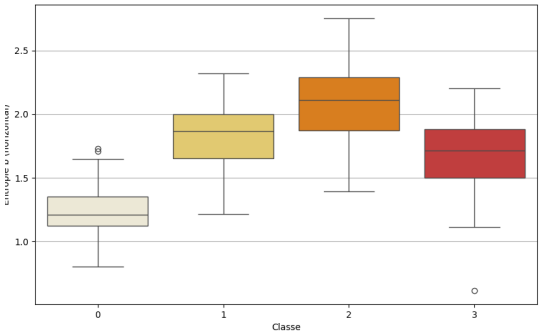
(f) Asymétrie b



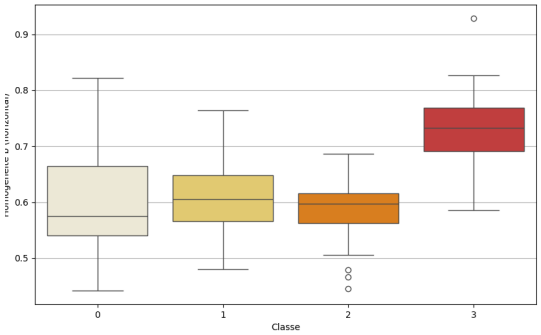
(g) Énergie b



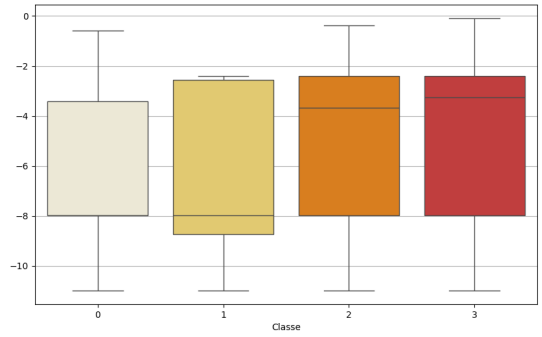
(h) Contraste b



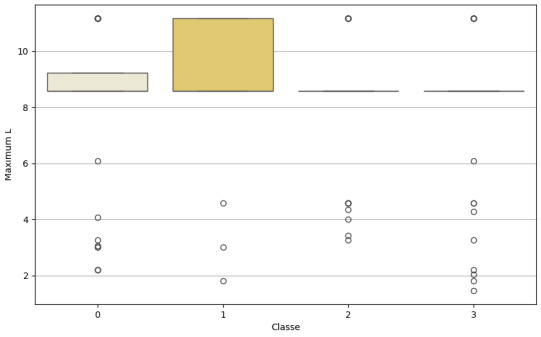
(a) Entropie b



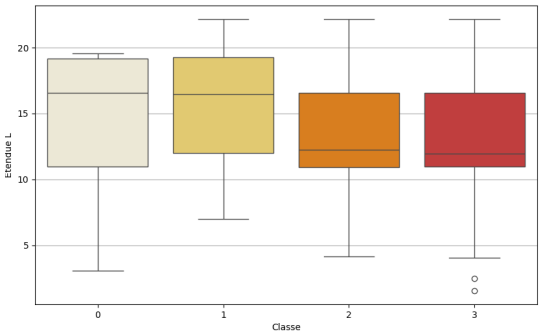
(b) Homogénéité b



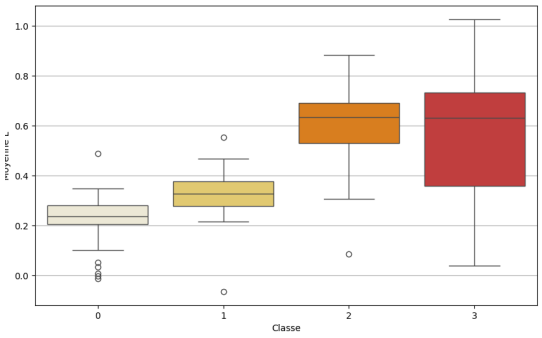
(c) Min L



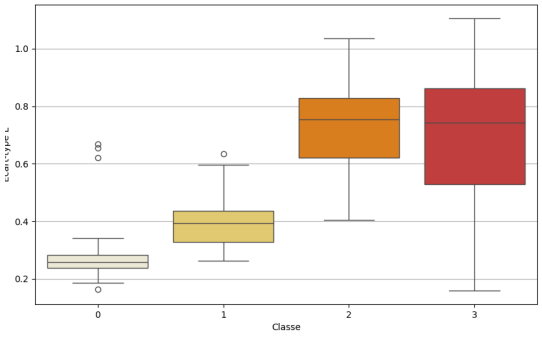
(d) Max L



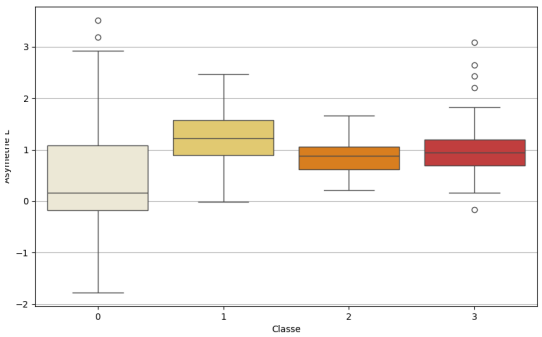
(e) Étendue L



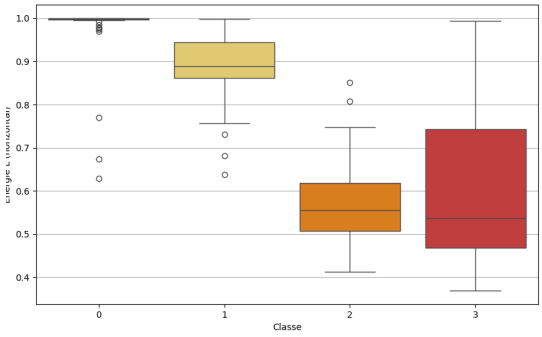
(f) Moyenne L



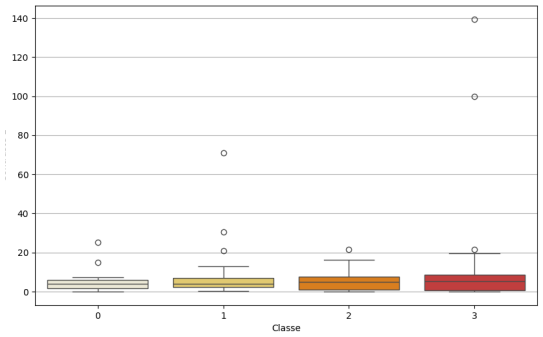
(g) Écart-type L



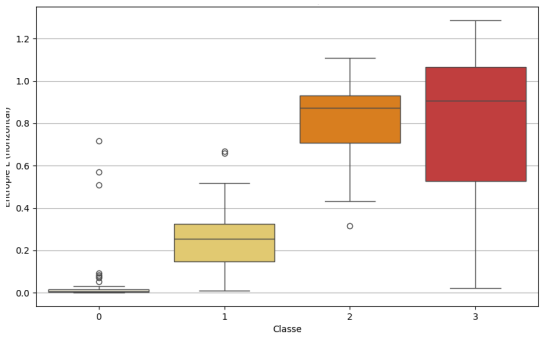
(h) Asymétrie L



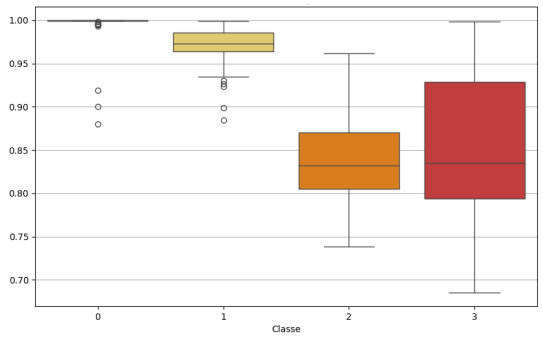
(a) Énergie L



(b) Contraste L



(c) Entropie L



(d) Homogénéité L