République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique

Université Kasdi Merbah - Ouargla

Faculté des Nouvelles Technologies de l'information et la Communication
Département de l'informatique et Technologies de l'information

Mémoire présente en vue de l'obtention du
diplôme Master

Domaine : Informatique et Technologie de l'Information
Filière : Informatique
Spécialité : Administration et sécurité des réseaux

Présenté par :

Baba Saci Hadjer
Bengana Ikram

Titre :

Evaluation de qualité d'image sous Spark

Le 08/ 06/ 2024

Membres du jury :

Encadrant	Dr. Naima MERZOUGUI	Univ Ouargla
Président	Dr. Bachir MAHDJOUB	Univ Ouargla
Examinateur	Dr. Charafeddine MECHALIKH	Univ Ouargla

Année Universitaire : 2023-2024

Dedicase

À ma mère et mon père

Piliers de ma vie et sources inépuisables d'amour et de soutien.

À mes deux petits frères

Hani et Djed, compagnons de jeux et de joie, qui ont toujours été là pour me faire sourire.

À ma grande sœur

Nesrine, modèle de force et d'inspiration, qui m'a montré le chemin avec bienveillance.

À ma grand-mère

Gardienne de sagesse et de tendresse, dont les conseils éclairés ont guidé mes pas.

À mes copines

Sara, Amira, Marwa, Melissa complices de fous rires et de confidences, dont le soutien inconditionnel illumine mes journées.

À ma compagne de travail

Ikram, ensemble, nous avons surmonté les obstacles et atteint de nouveaux sommets.

À mes collègues

compagnons de route dans cette aventure académique, dont l'entraide et la camaraderie ont rendu chaque défi surmontable.

À vous tous, je dédie ce projet avec gratitude et affection. Votre soutien indéfectible a été ma plus grande force et je suis honorée de vous avoir à mes côtés dans ce voyage.

Avec tout mon amour

Baba Saci Hadjer

Dedicase

À la mémoire de mon père

Dont les conseils, l'amour, la confiance et les encouragements me manquent chaque jour.

À ma mère

Dont le sourire et la bonté illuminent mes journées, et dont l'amour et le soutien inconditionnels m'accompagnent toujours.

À ma seconde mère, Maman Aïcha

Un exemple vivant de force, de compassion et d'amour maternel infini.

À mes adorables frère et sœurs

Redouane, Sara et Fella, témoins de mes joies, gardiens de mes secrets, et soutiens indéfectibles.

À mes copines

Houria, Salma, Samah et Kouka, qui sont plus que des amies, mais aussi des sœurs de cœur sur qui je peux toujours compter.

À ma compagne de travail

Hadjer, merci d'avoir été un partenaire de confiance et une amie fidèle tout au long de ce parcours.

À mes collègues exceptionnelles Marwa et Besma

À toute ma famille et mes amis

Bengana Ikram

Remerciement

Nous commençons par louer Dieu, dont la grâce nous a permis de persévérer dans les moments difficiles et de réaliser notre projet.

Nous tenons à exprimer notre profonde reconnaissance à notre promoteur, Dr. Merzougui Naima, pour son encadrement précieux et ses conseils avisés.

Nous sommes reconnaissants pour sa confiance en nous et pour les connaissances qu'elle nous a transmises.

Nous remercions également les membres du jury pour avoir pris le temps d'évaluer notre projet.

Nous tenons à remercier toutes les personnes qui ont contribué d'une manière ou d'une autre à la réalisation de ce projet.

Ce projet de fin d'études a été une étape importante dans nos vies, et nous sommes fiers de l'avoir mené à bien.

Nous sommes reconnaissants envers tous ceux qui nous ont soutenus et encouragés tout au long de ce parcours.

Baba Saci Hadjer.
Bengana Ikram.

Résumé

Ce travail aborde le défi de l'évaluation objective de la qualité des images dans le contexte de l'évaluation sans référence. Alors que la technologie continue de progresser et que la demande en contenu visuel de haute qualité augmente, des méthodologies d'évaluation de la qualité des images précises et efficaces sont essentielles. Malgré l'abondance des métriques d'évaluation objectives existantes, la complexité des scénarios de dégradation d'images du monde réel présente des défis persistants.

Dans ce mémoire, nous proposons une approche qui intègre des techniques d'apprentissage par ensemble, plus particulièrement la régression en forêt, avec une validation croisée k-fold pour une évaluation robuste et efficace de la qualité des images, et pour améliorer les performances prédictives. Les résultats expérimentaux sur nos ensembles de données démontrent l'efficacité de l'approche proposée dans l'évaluation précise de la qualité des images. En exploitant la puissance de l'apprentissage par ensemble et de la validation croisée, notre méthodologie atteint des performances supérieures par rapport aux approches traditionnelles. De plus, la scalabilité du cadre proposé, mis en œuvre à l'aide de PySpark, permet un traitement efficace des ensembles de données d'images.

Mots-clés : évaluation de la qualité des images, big-data, régression, Machine-Learning, Apache Spark, PySpark.

Abstract

This work addresses the challenge of objective image quality assessment in the context of no-reference evaluation. As technology continues to advance and demand for high-quality visual content increases, accurate and effective image quality assessment methodologies are essential. Despite the abundance of existing objective evaluation metrics, the complexity of real-world image degradation scenarios presents persistent challenges.

In this paper, we propose an approach that integrates ensemble learning techniques, specifically forest regression, with k-fold cross-validation for robust and efficient image quality assessment, and to improve predictive performance. Experimental results on our datasets demonstrate the effectiveness of the proposed approach in accurately image quality assesment. By leveraging the strength of ensemble learning and cross-validation, our methodology achieves superior performance compared to traditional metrics. Moreover, the scalability of the proposed framework, implemented using PySpark enables efficient processing of image datasets.

Keywords: image quality assessment, big data, regression, Machine-Learning, Apache Spark, PySpark.

ملخص

يتناول هذا العمل التحدي المتمثل في التقييم الموضـوعي لجودة الصـورة في سـياق التقييم غير المرجعي. مع اسـتمرار تقدم التكنولوجيا وزيادة الطلب على المحتوى المرئي عالي الجودة، أصـــبحت منهجيات تقييم جودة الصـــورة الدقيقة والفعالة ضـــرورية. على الرغم من وفرة مقاييس التقييم الموضـــوعي الحالية، فإن تعقيد ســـيناريوهات تدهور الصـــورة في العالم الحقيقي يمثل تحديات مستمرة.

في موضوع تخرجنا هذا، نقترح نهجًا يدمج تقنيات التعلم الجماعي، وتحديدًا "Forest regression "مع "k-fold cross-validation " لتقييم جودة الصورة بشكل قوي وفعال، ولتحسين الأداء التنبئي. توضح النتائج التجريبية على مجموعات البيانات لدينا فعالية النهج المقترح في تقييم جودة الصورة بدقة. ومن خلال الاستفادة من قوة التعلم الجماعي والتحقق المتبادل، تحقق منهجيتنا أداءً فائقًا مقارنة بالمناهج التقليدية. علاوة على ذلك، فإن قابلية تطوير الإطار المقترح، والتي تم تنفيذها باستخدام PySpark ، تتيح المعالجة الفعالة لمجموعات بيانات الصور.

الكلمات المفتاحية: تقييم جودة الصورة، البيانات الضخمة، الانحدار، التعلم الآلي، Apache Spark، PySpark

Table de Matière

Liste des Figures

Liste des tableaux

Introduction générale

L'évaluation de la qualité de l'image (IQA : Image Quality Assessement) est un domaine crucial dans le traitement d'image numérique, englobant l'analyse et la mesure de la perception humaine de la qualité visuelle d'une image. Elle joue un rôle essentiel dans divers domaines tels que la photographie, la vidéo, l'imagerie médicale et la radiographie, permettant aux développeurs de systèmes d'imagerie d'optimiser leurs algorithmes et de garantir une expérience utilisateur optimale.

Traditionnellement, IQA s'est appuyée sur des approches subjectives impliquant des évaluations par des observateurs humains. Ces méthodes, bien que fournissant une référence ultime en matière de qualité d'image perçue, présentent des inconvénients majeurs, notamment leur coût élevé, leur temps d'évaluation important et leur variabilité interindividuelle. Ces limitations ont motivé le développement de techniques d'évaluation de la qualité d'image à l'aide de la machine qui visent à automatiser le processus d'évaluation de la qualité d'image sans nécessiter de référence de qualité subjective.

Les méthodes d'évaluation objectives offrent de nombreux avantages par rapport aux approches subjectives traditionnelles. Elles permettent de réduire considérablement les coûts d'évaluation, d'améliorer l'objectivité et la reproductibilité des mesures, et d'évaluer de grands volumes d'images de manière efficace. De plus, ces techniques peuvent être intégrées aux systèmes d'imagerie en temps réel, permettant une évaluation et une adaptation continues de la qualité d'image.

L'exploration de ces méthodes efficaces et précises est un domaine de recherche actif, avec des contributions majeures provenant de divers domaines. Les recherches récentes se sont concentrées sur le développement d'algorithmes basés sur l'apprentissage automatique qui exploitent des ensembles de données d'images annotées avec des scores de qualité d'image subjectifs pour apprendre à prédire la qualité d'image à partir de caractéristiques d'image extraites.

Cette thèse s'inscrit dans le cadre de la recherche sur l'évaluation de la qualité d'image basée sur l'apprentissage automatique. Nous explorons l'utilisation du modèle Random Forest, un algorithme d'apprentissage par ensemble robuste et performant, pour prédire la qualité d'image sans référence (NR-IQA : No Referece). Nous employons l'ensemble de données KADID-10K, largement utilisé dans la communauté NR-IQA, pour développer et évaluer notre système.

Dans notre étude, l'outil et la technologie sont destinés à tirer parti de la capacité de traitement distribué d'Apache Spark, un framework open source pour le traitement rapide des données à grande échelle. La proximité de Spark par rapport à notre implémentation est dans la manipulation du traitement des données. Ici, avant l'entrée du modèle d'apprentissage automatique, nous l'utilisons pour précharger et traiter les données d'image pour assurer une efficacité et une praticité du traitement du volume de données et préparer au modèle Random Forest.

Notre mémoire se compose de quatre chapitres principaux. Le premier présente une introduction générale au domaine de la qualité d'image, en soulignant l'importance de l'évaluation NR-IQA et les différentes approches existantes. Le deuxième est concentré sur les frameworks Hadoop et Apache Spark, leurs objectifs et leurs fonctionnalités. Concernant le troisième chapitre est basé sur l'apprentissage automatique, nous avons décrit la méthodologie employée dans notre étude. Nous avons présenté aussi en détail le modèle Random Forest. Et le dernier chapitre sera consacré à expliquer notre nouvelle méthode ainsi que nous exposerons les différents résultats obtenus, accompagnés par des comparaisons et des commentaires.

Chapitre 1 : Évaluation de qualité d'image

1.1 Introduction

Dans le monde d'aujourd'hui, où la technologie progresse rapidement, la demande d'images de haute qualité ne cesse d'augmenter. Que ce soit dans un contexte professionnel, artistique ou personnel, la recherche de la meilleure qualité d'image possible est un objectif constant non seulement agréable à l'œil [1]. La qualité d'une image joue un rôle crucial dans sa capacité à communiquer de manière efficace des informations, à provoquer des émotions et à offrir une expérience utilisateur optimale [2]. L'objectif de ce chapitre est d'analyser les différentes facettes de la qualité d'image, les techniques d'évaluation et les difficultés liées à la production d'images de grande qualité.

1.2 Définition de l'image

Une image est une représentation visuelle ou à une reproduction de quelque chose ou de quelqu'un. En particulier, dans le domaine de la technologie numérique et de la photographie, une image est une représentation en deux dimensions de la lumière capturée par un appareil photo ou générée par un logiciel informatique.

Dans le domaine de la vision, l'image correspond à ce que nous voyons à travers nos sens visuels. Il peut s'agir d'une représentation physique, telle qu'une peinture ou une photographie, ou d'une représentation mentale, telle que celle que nous créons dans notre esprit lorsqu'on a une image.
Dans le domaine de l'informatique, une image numérique correspond à une représentation numérique d'une scène ou d'un objet, qui est stockée en binaires. Il est possible d'obtenir ces images en utilisant des appareils de capture d'images, comme des caméras numériques, ou en utilisant des logiciels spécifiques à des fins artistiques, scientifiques ou techniques [3].

1.3 Définition La qualité d'une image

La qualité d'une image est un facteur crucial qui détermine son utilité et son efficacité. Elle fait référence à la quantité de dégradation présente dans une image, qui peut affecter sa clarté, sa couleur et son attrait visuel général. En tant que telle, une image de haute qualité est toujours souhaitable, car elle peut transmettre des informations plus efficacement et laisser une impression durable sur le spectateur. Grâce aux images numériques, de nombreuses personnes à la possibilité d'évaluer et de classer des images en fonction des bits de données ou des pixels. Par exemple, si deux images numériques sont présentes, l'une avec un méga de pixels et l'autre avec 4 mégas de pixels d'information, la majorité des individus affirmeraient que l'image avec plus de pixels est de meilleure qualité. Et cela serait vrai dans de nombreuses situations. Mais si on prenait une image de 4 méga pixels dans l'obscurité totale et une image de 1 méga pixels dans la lumière régulière, la plupart des gens seraient

d'accord pour dire qu'une image ayant plus de méga pixels n'a pas forcément une meilleure qualité d'image [4].

L'un des éléments les plus importants de la qualité de l'image :

- **Résolution :** La résolution de l'image fait référence au nombre de pixels qu'elle contient. La haute résolution permet une description détaillée de l'image, ce qui est important pour les applications nécessitant une haute précision, comme la radiologie médicale ou la photographie professionnelle [5].
- **Netteté :** La netteté d'une image fait référence à la clarté de ses parties et détails. De nombreux facteurs l'affectent, tels que la mise au point lors de la prise de vue et les techniques de post-traitement pour la rendre nette [6].
- **Couleur et contraste :** La qualité des couleurs et le contraste entre les différentes parties de l'image contribuent également à sa qualité globale. Une reproduction précise des couleurs est essentielle dans de nombreux domaines, tels que la photographie, le graphisme et la médecine [7].
- **Réduction du bruit** : Le bruit dans une image peut affecter sa clarté et sa lisibilité, rendant difficile l'interprétation des détails. Des techniques de réduction du bruit peuvent être appliquées pour améliorer la qualité de l'image [8].
- **Incendie :** L'éclairage joue un rôle important dans la qualité de l'image. Un éclairage approprié peut mettre en évidence les détails importants de l'image et améliorer sa qualité globale [9].
- **Compression :** Bien que la compression puisse être utile pour réduire la taille d'un fichier image, elle peut également l'aggraver. Il est important de trouver un équilibre entre compression et qualité d'image pour répondre aux besoins spécifiques de chaque application [10].

Cependant, il n'est pas toujours facile d'obtenir des images de haute qualité, car divers facteurs peuvent affecter la qualité d'une image.
Par exemple, des distorsions peuvent se produire pendant l'acquisition, la compression, le stockage et la décompression de l'image, ce qui peut encore dégrader la qualité de l'image. En outre, la qualité des images prises par un appareil photo peut varier en fonction de plusieurs facteurs [11].

Nous aborderons brièvement certaines des malformations les plus courantes :

- **Bruit** : Le bruit est l'une des causes les plus courantes de distorsion de l'image. Il peut être causé par divers facteurs tels que les interférences électriques, le bruit du capteur ou même les conditions atmosphériques. La présence de bruit dans une image peut entraîner une perte de détails et de clarté, ce qui rend difficile l'interprétation et l'analyse précises de l'image [12].

Figure 1. Exemple de bruit [13]

- **Compression** : Un autre facteur qui peut contribuer à la distorsion de l'image est l'utilisation de techniques de compression. Si la compression peut être utile pour réduire la taille des fichiers et améliorer l'efficacité du stockage, elle peut aussi entraîner une perte de qualité de l'image. Cette perte de qualité peut être particulièrement visible sur les images très détaillées ou présentant des textures fines [14].

Figure 2. Exemple de compression [13]

- **flou de mouvement** : peut survenir lorsque des objets en mouvement rapide sont capturés pendant une exposition prolongée. Ce manque de netteté et de clarté des détails dans l'image peut diminuer, en particulier si l'objet en mouvement est l'objet principal de la scène [12].

Figure 3. Exemple de flou de mouvement [14]

- **L'éclairage :** Les conditions d'éclairage peuvent également avoir un impact significatif sur la qualité d'une image. Un mauvais éclairage peut entraîner une sous-exposition ou une surexposition des images, ce qui se traduit par une perte de détails dans les ombres ou les

hautes lumières. De même, des réglages incorrects de l'appareil photo, tels que la vitesse d'obturation, l'ouverture et la sensibilité ISO, peuvent également entraîner une distorsion de l'image et une mauvaise qualité [9].

- **Filtrage :** Le flou peut être causé par certains traitements visant à réduire quelques distorsions, comme le filtrage du bruit [15]. Un exemple de flou causé par l'utilisation d'un filtrage passe bas est présenté par

Figure 4. Exemple de filtrage [16]

En garantissant des images de haute qualité, nous pouvons transmettre efficacement des informations, laisser des impressions durables et améliorer notre communication visuelle.

1.4 Évaluation de la qualité d'image

l'évaluation de la qualité d'une image (IQA : Image Quality Assessement) est un processus à multiples facettes qui prend en compte plusieurs facteurs. Il est important de s'assurer que les images sont de haute qualité et qu'elles répondent à l'objectif visé. En prêtant attention à la résolution, à la précision des couleurs, à la netteté et à la réduction du bruit, les créateurs peuvent s'assurer que leurs images sont de la meilleure qualité possible.[17]

L'évaluation de la qualité d'une image est essentielle pour s'assurer qu'elle est adaptée à l'usage prévu et qu'elle transmet efficacement le message souhaité. Voici quelques raisons pour lesquelles il est important d'évaluer la qualité d'une image [18] :

- Communication efficace : Une image de haute qualité communique mieux avec le public cible, qu'il s'agisse d'informer, d'éduquer ou de divertir.
- Image de marque : Pour les entreprises et les marques, la qualité des images utilisées dans la publicité, le marketing ou les supports de communication reflète leur image de marque. Les images de haute qualité renforcent la crédibilité et le professionnalisme d'une marque.
- Expérience utilisateur : Dans le domaine de la conception et des applications web, la qualité des images joue un rôle crucial dans l'expérience utilisateur. Des images claires et nettes améliorent l'interaction et la satisfaction des utilisateurs.

- Précision des informations : Dans des domaines comme la médecine, l'astronomie ou la recherche scientifique, la qualité de l'image est cruciale pour une interprétation précise et fiable des données visuelles.
- Normes et réglementations : Certaines industries ont des normes strictes en matière de qualité d'image, notamment dans les domaines de l'imprimerie, de la photographie professionnelle et de la production de films.
- Faire des ajustements ou des améliorations peut aider à optimiser les ressources et les efforts de production.

Pour s'assurer que les produits et les services répondent aux normes de qualité requises, deux méthodes principales peuvent être utilisées : l'évaluation subjective et l'évaluation objective.

1.5 Méthodes existantes pour l'évaluation de la qualité d'image

La meilleure évaluation de la qualité visuelle des images est l'évaluation par les êtres humains (évaluation subjective). L'évaluation subjective est basée sur des opinions et des perceptions personnelles. Cette méthode est souvent utilisée dans les domaines où les critères quantifiables sont difficiles à mesurer, comme la satisfaction des clients ou l'expérience globale de l'utilisation d'un produit. Ce processus peut être prendre beaucoup de temps, et malheureusement, les applications en temps réel exigent un temps de réponse rapide, et ce processus ne peut tout simplement pas suivre le rythme. Par conséquent, en utilise l'évaluation objective.

L'évaluation objective fait appel à des critères quantifiables et mesurables, tels que le nombre de défauts d'un produit ou le temps nécessaire à la réalisation d'un service. Cette méthode est souvent préférée car elle est plus précise et plus scientifique, ce qui permet de comparer plus facilement différents produits ou services [17].

1.5.1 L'évaluation subjective de la qualité

En réalité, l'évaluation la plus fiable de la qualité de l'image est l'évaluation subjective effectuée par des observateurs humains. Lors de tests subjectifs, un groupe de personnes est invité à donner son avis sur la qualité de chaque image.

Historiquement, le groupe d'observateurs est composé d'observateurs "experts" et "non experts". Un observateur novice peut se concentrer davantage sur la situation globale, tandis qu'un observateur expérimenté peut accorder une plus grande attention aux détails.
A l'issu de ces tests, une note subjective appelée MOS (Mean Opinion Score) est obtenue.

Le MOS est donné par la formule suivante :

$$MOS(i) = \frac{1}{N}\sum_{j=1}^{N} Note_i(j)$$

Où N est le nombre total de participants et Notei(j) la note affectée `à l'image i par l'observateur j.

successivement, le score final peut également être interprété comme un score d'opinion moyen différentiel (DMOS), qui représente la différence de MOS entre l'image déformée et sa référence correspondante.
Pendant l'exécution de ces tests, il est crucial de prêter une attention particulière à certains facteurs qui peuvent affecter les jugements des participants comme la distance d'observation, Conditions de visualisation, Ecran, Le choix des images, Facteurs psychologiques, Les observateurs [19].

1.5.2 Évaluation objective de la qualité d'image

L'évaluation objective de la qualité des images est un aspect crucial du traitement et de l'analyse des images. Elle implique l'utilisation de techniques automatisées pour évaluer la qualité d'une image sans se fier à la perception humaine subjective. L'objectif de l'IQA (image quality assessment) objective est de concevoir des modèles mathématiques capables de prédire la qualité d'une image avec précision et automatiquement. Une méthode IQA objective idéale devrait être capable d'imiter les prédictions de qualité d'un observateur humain moyen. Les méthodes objectives d'IQA ont une grande variété d'applications. Cette approche permet de comparer différentes images ou de quantifier l'impact des modifications apportées à une image.

La qualité d'une image peut avoir un impact sur son utilité dans divers domaines, et les méthodes objectives d'IQA ont une grande variété de applications :

- Ils sont adaptés à la surveillance de la qualité des images dans les systèmes de contrôle de la qualité. Par exemple, les systèmes d'acquisition d'images peuvent se surveiller et s'ajuster automatiquement pour obtenir des données d'image de qualité supérieure en utilisant une métrique IQA objective.
- Ils peuvent servir à comparer les différents algorithmes de traitement d'image. Par exemple, si plusieurs algorithmes d'amélioration d'image sont disponibles, il est possible d'utiliser une métrique IQA objective afin de sélectionner celui qui offre des images de meilleure qualité.
- Ils peuvent servir à améliorer l'efficacité des systèmes de traitement et de transmission d'images. Par exemple, dans le cadre d'un réseau de communication visuelle, il est possible d'utiliser une métrique IQA objective afin d'améliorer les algorithmes de pré-filtrage et d'attribution de bits au niveau du codeur, ainsi que les algorithmes de post-filtrage et de reconstruction.

Il existe différentes catégories de méthodes qui peuvent être utilisées pour évaluer la qualité d'une image, et l'utilisation de ces méthodes offre une approche complète de l'évaluation de la qualité des images et permet de s'assurer qu'elles répondent aux normes requises.
En se basant sur la présence d'une image de référence sans distorsion et de qualité parfaite, on peut classer les méthodes objectives d'IQA en trois catégories. L'évaluation de la qualité de l'image sans référence (NRIQA : No Reference) est la première catégorie, où ni l'image de référence ni ses caractéristiques ne sont accessibles pour une évaluation de la qualité. L'évaluation de la qualité de l'image à référence réduite (RR-IQA : Reduced Reference) est une autre catégorie où seules des informations partielles sur l'image de référence sont accessibles. L'évaluation de la qualité de l'image de référence complète (FR-IQA : Full Reference) est la troisième catégorie, où l'image de référence est entièrement accessible [19].

Dans les sous-sections suivantes, les caractéristiques des trois principales catégories de l'IQA objectif sont décrites :

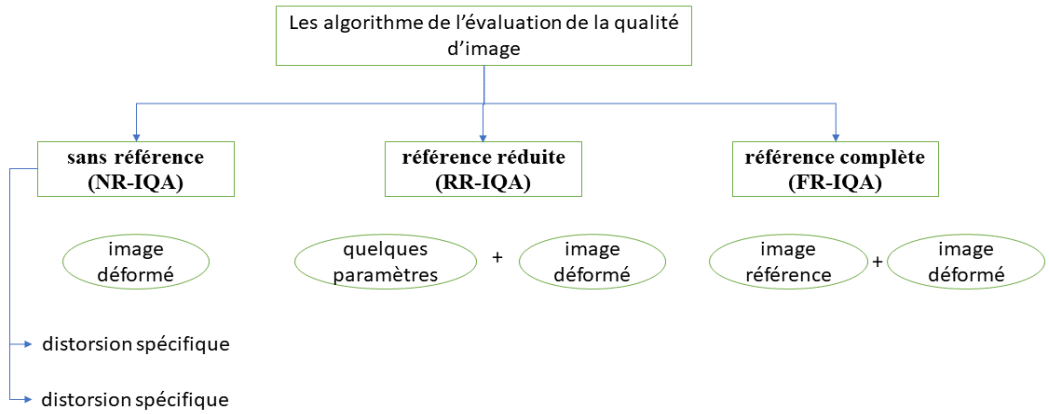

Figure 5. Classification objective des algorithmes d'évaluation de la qualité de l'image

- **Les méthodes sans référence (NR-IQA)** sont utilisées lorsqu'il n'y a pas d'image de référence disponible pour comparer la qualité de l'image évaluée. L'image de référence n'est pas disponible dans de nombreuses applications du monde réel, comme les systèmes de communication d'images, et l'évaluation de la qualité ne repose que sur l'image de test. Dans l'état actuel de notre projet, nous avons opté pour cette méthode étant donné l'absence d'images de référence. NR-IQA présente une difficulté supérieure à celle des méthodes RR-IQA et FRIQA. Toutefois, les individus ont généralement la capacité d'évaluer de manière efficace la qualité d'une image de test sans avoir besoin d'une image de référence. Il est probable que cela soit dû au fait que notre cerveau possède de nombreuses informations sur la manière dont une image devrait ou ne devrait pas être dans la réalité [20].

- **Les méthodes à référence réduite (RR-IQA)** sont utilisées lorsqu'il existe une image de référence limitée. L'image de référence n'est pas entièrement accessible dans RR-IQA. Plutôt que cela, plusieurs caractéristiques sont tirées de l'image de référence. Le processus d'évaluation de la qualité utilise ces caractéristiques comme données secondaires afin d'évaluer la qualité de l'image de test. Les techniques RR-IQA peuvent être employées dans divers domaines. Par exemple, il est possible d'utiliser ces outils pour surveiller en temps réel le niveau de détérioration de la qualité visuelle des données d'image et vidéo transmises via les réseaux de communication visuelle [21].

- **Les méthodes de référence complète (FR-IQA)** sont utilisées lorsqu'une image de référence complète est disponible pour comparer la qualité de l'image évaluée. Lorsque l'algorithme a accès à une version parfaite de l'image, il peut comparer la version dégradée [19].

1.6 Défis et Problématiques

L'évaluation d'images présente de nombreux défis et problèmes qui nécessitent un examen attentif pour garantir des résultats précis et fiables. Ces défis peuvent être liés aux différents facteurs affectant la qualité de l'image et aux méthodes de mesure utilisées pour mesurer cette qualité. Dans cette section, nous aborderons certains des principaux défis rencontrés lors de l'évaluation de la qualité de l'image [22].

1.6.1 La complexité des facteurs d'influence

La qualité d'une image peut être influencée par divers facteurs, tels que l'éclairage, la résolution, le bruit et la compression. Chacun de ces facteurs peut avoir un impact considérable sur la qualité globale de l'image, ce qui rend difficile l'identification et la quantification de leur contribution individuelle à la qualité de l'image [22].

1.6.2 Subjectivité vs Objectivité

comme nous avons vu l'évaluation subjective de la qualité de l'image, bien qu'elle constitue une méthode fiable, peut prendre du temps et être coûteuse en termes de temps et de ressources humaines. D'un autre côté, la méthode objective, bien qu'elle soit efficace et rapide, ne reflète pas toujours l'opinion d'une personne sur la qualité de l'image, ce qui peut conduire à des résultats négatifs [19].

1.6.3 Normalisation et standardisation

Les différences dans les normes et formats d'image utilisés dans différentes régions et applications peuvent poser des problèmes lors de la configuration et de l'ajustement de la taille de l'image. Cela peut rendre difficile la comparaison et l'évaluation de la qualité d'image entre différentes applications et systèmes [22].

1.6.4 Flexibilité et généralisation

Les techniques et les mesures employées pour évaluer la qualité des images doivent être suffisamment adaptables pour convenir à différents types d'images et scénarios d'utilisation. En outre, elles doivent également être généralisables pour garantir leur applicabilité dans différents contextes et environnements [22].

1.7 Conclusion

L'évaluation de la qualité d'image tente de quantifier une qualité visuelle, une quantité de distorsion dans une image donnée. Par conséquent, il existe une demande pour une technique informatisée permettant de concevoir le plus fidèlement possible la qualité visuelle perçue par l'homme. Le principal inconvénient des méthodes d'évaluation de qualité d'images est le long temps de traitement dû à la grande complexité de leurs algorithmes. Cela s'aggrave lorsque ces algorithmes doivent être traités séquentiellement avec de grands ensembles de données.

Le chapitre suivant présente des solutions possibles pour le cas de traitement de grands ensembles de données.

Chapitre 2 : Apache Hadoop et Apache Spark

1.1 Introduction

L'avènement du Big Data a bouleversé le monde des entreprises, les obligeant à gérer un flux constant de données énormes et diverses. Dans le passé, les données étaient stockées et traitées sur des ordinateurs personnels, cette approche traditionnelle s'avère efficace pour des quantités de données modestes et un nombre limité de tâches de traitement, mais pas pour le traitement de données volumineuses.
Dans ce chapitre nous allons présenter les frameworks Hadoop et Apache Spark, leurs objectifs, et leurs fonctionnalités. Ces deux outils sont parfois considérés comme des concurrents. Il est souvent admis qu'ils fonctionnent encore mieux quand ils sont ensemble. On va faire un aperçu de leurs caractéristiques et de leurs différences.

2.2 Définition de Big Data

Le Big Data, ou "mégadonnées" en français, désigne de jeux de données complexes et volumineuses. C'est ce que l'on appelle les cinq « V ». Il fait référence à la collecte, au traitement et à l'analyse de vastes ensembles de données qui ne peuvent être traités par les outils traditionnels de gestion de données. Mais ces énormes volumes de données peuvent être utilisés pour résoudre des problèmes que vous n'auriez jamais pu résoudre auparavant.
Ces données proviennent de diverses sources telles que les médias sociaux, les transactions en ligne, les capteurs IoT, etc. [23]

2.2.1 Les 5 « V » du Big Data

Les cinq V du Big Data (Volume, Vélocité, Variété, Véracité, Valeur) sont les caractéristiques les plus importantes du Big Data [24]. Voici une description concise de chaque "V" :

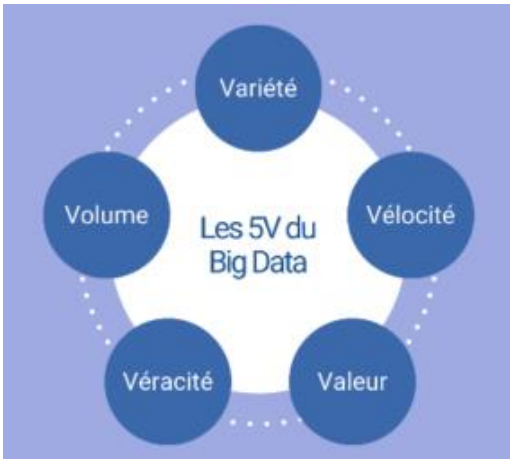

Figure 6. Les 5 V de Big Data [25]

1. **Volume :** Avec le Big Data, vous devrez traiter de gros volumes de données qui peuvent être structurées, semi-structurées ou non structurées, souvent mesurées en téraoctets, pétaoctets ou même exaoctets.

2. **Vélocité :** Les données sont reçues et éventuellement traitées à grande vitesse, en temps réel ou quasi-réel. Normalement, les données haute vitesse sont transmises directement à la mémoire, plutôt que d'être écrites sur le disque.

3. **Variété :** fait allusion aux nombreux types de données disponibles, y compris textuelles, numériques, graphiques, sonores, vidéos, etc. et peuvent être structurées, semi-structurées ou non structurées.

4. **La véracité :** fait référence à l'incertitude des données en raison de leur incohérence et de leur caractère incomplet. Lorsqu'elles traitent le Big Data, les organisations doivent tenir compte de l'incertitude des données.

5. **Valeur :** Les données sans informations n'ont aucun sens. Le Big Data est inutile tant que nous ne le transformons pas en valeur. Le simple fait de collecter des données volumineuses et de les stocker ne sert à rien tant que les données ne sont pas analysées et qu'un résultat utile n'est pas généré.

2.2.2 Les avantages de Big Data

Le Big Data permet aux entreprises et autres organisations (publiques et privées) de prendre des décisions plus efficaces et plus intelligentes. De plus, l'utilisation du Big Data présente les avantages suivants [26] :

- Permet aux entreprises de tirer des informations précieuses des vastes volumes de données pour prendre des décisions éclairées.
- Permet d'obtenir des réponses plus complètes, signifient plus de confiance dans les données, car le volume d'informations est plus important.
- Identifier les besoins des clients et développer de nouveaux produits et services qui répondent à ces besoins.
- Réduire le coût de la publicité : toute partie peut envoyer aux clients des publicités personnalisées en fonction de leurs intérêts, directement via le système appelé (Microtargeting)

2.2.3 Les défis du Big Data

Lorsqu'il s'agit de Big Data, les défis sont nombreux, en voici quelques-uns [26] :

1. **Coût élevé de stockage et de maintenance :** Gérer de grandes quantités de données peut être très coûteux en termes d'infrastructure et de ressources.
2. **Risques pour la vie privée :** L'analyse des données peut compromettre la confidentialité des individus.

3. **Croissance des données :** L'augmentation massive des données rend leur analyse et leur utilisation de plus en plus difficiles. Il est crucial de trouver des méthodes efficaces pour extraire les informations importantes d'une quantité de données toujours croissante.

2.2.4 Cas d'utilisation du Big Data

Le Big Data est utilisé dans différents domaines, comme la réalisation de diverses activités commerciales, de l'expérience client à l'analytique. Voici quelques exemples concrets :

- Les entreprises comme Netflix de développement de produits utilisent le Big Data pour créer des modèles prédictifs pour anticiper la demande des clients, classer les caractéristiques clés des produits ou services passés et actuels et modéliser la relation entre ces caractéristiques et le succès commercial de leurs offres.
- Les soins de santé, les mégadonnées peuvent être utilisés pour aider les hôpitaux à travailler plus efficacement, en utilisant une analyse des mégadonnées pour comprendre comment distribuer au mieux les services, les soins et les procédures dans un hôpital. D'un autre côté, les mégadonnées peuvent également être utilisées dans la recherche pour découvrir et guérir les maladies. [27]
- Différentes techniques utilisées pour le traitement d'image par exemple traitement d'images médicales, traitement d'images satellites ; techniques pour la segmentation d'images, compression d'images. [28]

2.2.6 Pile de technologies Big Data

Il existe de nombreux outils et technologies de mégadonnées pour traiter ces quantités massives de données : [29]

- **Apache Hadoop :** Il s'agit d'un framework de traitement distribué open source. Il est préférable pour le traitement par lots.
- **Apache Spark :** Il s'agit d'un framework de traitement en temps réel open-source. Il dispose d'une capacité de calcul en mémoire.
- **Apache Hive :** Il s'agit d'un outil d'entrepôt de données open source permettant d'interroger une énorme quantité de données stockées dans Hadoop HDFS.

2.3 Apache Hadoop

2.3.1 Historique

Hadoop a été lancé en 2002 par Doug Cutting et Mike Cafarella, initialement comme un projet appelé Apache Nutch pour créer un moteur de recherche capable d'indexer un milliard de pages. Cependant, les défis d'évolutivité et les coûts élevés liés au stockage et au traitement de grandes quantités de données ont conduit à la séparation de l'informatique distribuée de Nutch et à la création de Hadoop en 2006.

Doug Cutting, inspiré par les travaux de Google sur GFS et MapReduce, a rejoint Yahoo et a pris Hadoop avec lui. Hadoop a été nommé d'après l'éléphant jaune en peluche de son fils. Il est rapidement devenu une solution pour gérer efficacement de vastes ensembles de données à moindre coût.
En 2009, Hadoop a prouvé son efficacité dans le traitement de données à grande échelle, ce qui a incité Cutting à rejoindre Cloudera pour promouvoir davantage l'adoption de Hadoop. Le projet a été transféré à l'Apache Software Foundation en 2008 et a atteint des jalons significatifs avec les versions 1.0 en décembre 2011, 2.0.6 en août 2013 et 3.0 en décembre 2017, marquant son évolution en tant que technologie fondamentale à grande échelle. [30]

2.3.2 Présentation de Hadoop

Hadoop est un Framework open source d'Apache écrit en Java et est utilisé pour stocker des processus et analyser des données de très gros volume. Il est largement utilisé dans le domaine du Big Data pour répondre aux défis liés à la gestion de données à grande échelle.
Hadoop est basé sur les composants principaux : Hadoop Distributed File System (HDFS), un système de planification des traitements (YARN), un Framework de traitement (MapReduce). [30]

2.3.3 Caractéristique d'Hadoop

Hadoop caractérisé par [31] :

- **Robuste :** Hadoop est conçu pour être robuste en cas de panne matérielle. Si un nœud de calcul tombe en panne, ses tâches sont automatiquement distribuées aux autres nœuds disponibles dans le cluster. De plus, les blocs de données sont répliqués sur plusieurs nœuds pour garantir la disponibilité des données même en cas de panne du nœud.
- **Coût :** Hadoop permet une meilleure utilisation des ressources en distribuant le traitement des données sur plusieurs nœuds au sein du cluster. Cela permet de réduire les coûts liés à l'acquisition et à la gestion du matériel informatique très performant, en utilisant plutôt un grand nombre de serveurs moins coûteux.
- **Compatibilité avec plusieurs langages :** Hadoop prend en charge plusieurs langages pour le traitement et le stockage des données, ce qui renforce sa polyvalence dans les tâches d'analyse des données.
- **Haute disponibilité :** Les données sont hautement disponibles et accessibles malgré une panne matérielle due à plusieurs copies de données. Si la machine ou le matériel tombe en panne, les données seront accessibles depuis un autre chemin.

2.3.4 Architecture

Le Framework Hadoop principal comprend trois modules qui fonctionnent ensemble pour former l'écosystème Hadoop [32] :

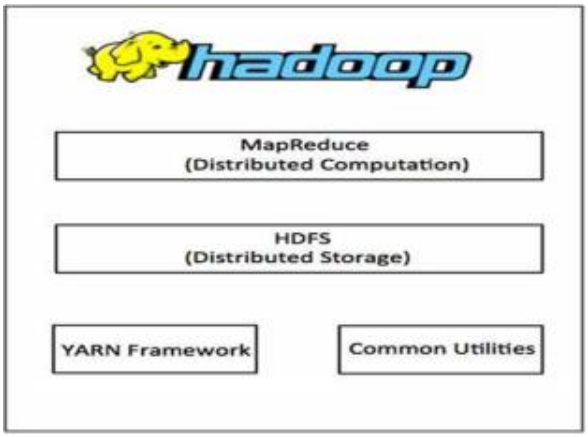

Figure 7. Architecture Hadoop [33]

2.3.4.1 Hadoop Distributed File System (HDFS)

Le système de fichiers distribué Hadoop (HDFS) est un système conçu pour stocker de manière fiable de très grands ensembles de données (données structurées, non structurées et semi-structurées), Il permet aussi d'accéder aux données sur les clusters Hadoop.
HDFS gère, prend en charge et analyse de très gros volumes de données. Contrairement au système de fichiers local HDFS est un progiciel open source. Il peut donc être utilisé sans frais de licence et de support. Et en termes de vitesse, les clusters Hadoop peuvent lire et écrire plus d'un téraoctet de données en une seconde.
HDFS copie les données plusieurs fois et distribue les copies aux nœuds individuels, ce qui garantit la fiabilité. Les nœuds sont un groupe d'ordinateurs faisant partie d'un cluster connectés entre eux via un réseau à haut débit pour effectuer des tâches de traitement de données. Chaque nœud d'un cluster Hadoop est un ordinateur distinct doté de son propre processeur.

2.3.4.1.1 Caractéristiques de HDFS

- **Stockage et traitement distribués** : HDFS est conçu pour stocker de grands ensembles de données de manière distribuée sur un cluster de serveurs, ce qui permet un stockage et un traitement parallèles de données.
- **Interface de commande :** Hadoop fournit une interface de commande qui permet aux utilisateurs d'interagir avec HDFS pour effectuer des opérations telles que la création de répertoires, le chargement de fichiers, la suppression de fichiers, etc.
- **Serveurs namenode et datanode :** HDFS utilise une architecture maître-esclave avec un namenode principal qui gère le système de fichiers et des datanodes qui stockent effectivement les données. Cette architecture permet aux utilisateurs de vérifier facilement l'état du cluster et de comprendre la distribution des données.
- **Accès continu aux données :** HDFS offre un accès continu aux données du système de fichiers, permettant aux applications de lire et d'écrire des données de manière transparente, quel que soit l'emplacement physique des données sur le cluster.

- **Autorisations de fichiers et authentification :** HDFS fournit un système de contrôle d'accès basé sur les autorisations de fichiers pour sécuriser les données stockées. Il prend également en charge l'authentification des utilisateurs pour contrôler l'accès aux données.

2.3.4.2 MapReduce

Le modèle de programmation MapReduce est une approche puissante pour traiter de grands ensembles de données. En utilisant des algorithmes informatiques parallèles et distribués, MapReduce aide à décomposer des tâches complexes en étapes plus simples, facilitant ainsi le traitement efficace de grandes quantités de données. L'algorithme MapReduce se déroule en deux tâches principales : la phase de mapping (Map) et la phase de réduction (Reduce).

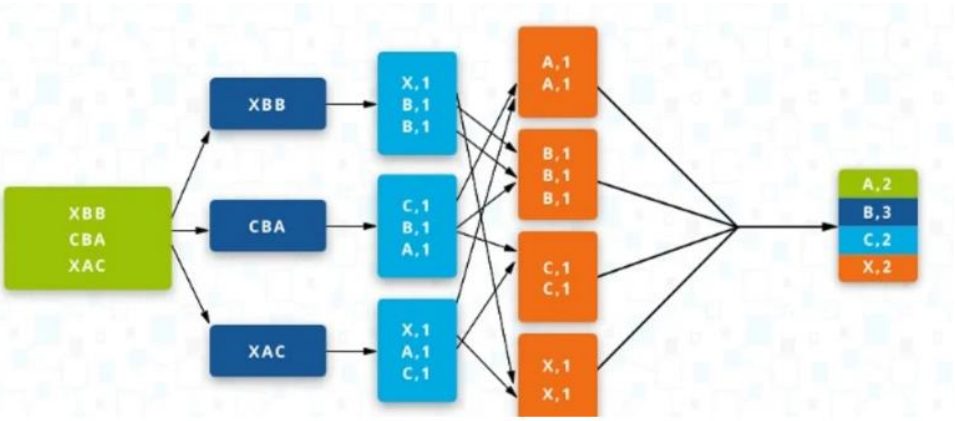

Figure 8. MapReduce fonctionnement [34]

L'algorithme MapReduce contient deux tâches importantes, Map et Reduce. Map prend un ensemble de données et le convertit en un autre ensemble de données, où les éléments individuels sont décomposés en tuples (paires clé / valeur). Reduce réduit la tâche qui prend la sortie d'une carte comme entrée et combine ces tuples de données en un ensemble plus petit de tuples. Comme la séquence du nom MapReduce l'indique, la tâche de réduction est toujours effectuée après la tâche de map. À un niveau élevé, MapReduce divise les données d'entrée en fragments et les distribue sur différentes machines. Les fragments d'entrée sont constitués de paires clé-valeur. Les tâches de mappage parallèle traitent les données fragmentées sur les machines d'un cluster. La sortie de mappage sert ensuite d'entrée pour l'étape de réduction. La tâche de réduction combine le résultat dans une sortie de pair clé-valeur particulière et écrit les données dans HDFS.

2.3.4.3 YARN

YARN signifie « Yet Another Resource Negotiator » est la couche de gestion des ressources de Hadoop. Dans les versions précédentes d'Hadoop, MapReduce était responsable du traitement des données et de l'allocation des ressources. Et avec le temps, la nécessité de séparer ces deux fonctions a conduit au développement de YARN. YARN se situe entre la couche de stockage, représentée par HDFS, et le moteur de traitement MapReduce. De plus, YARN son rôle principal est d'assurer l'allocation efficace des ressources du cluster. Il fournit

une interface générique qui permet l'implémentation de nouveaux moteurs de traitement qui s'adaptent à différents types de données.

2.3.4.4 Hadoop commun

Hadoop Common représente le cœur même du framework Apache Hadoop (Hadoop Core). Il s'agit d'un élément essentiel du framework Apache Hadoop, aux côtés de Hadoop Distributed File System (HDFS), Hadoop YARN et Hadoop MapReduce. Hadoop Common fait référence à l'ensemble d'utilitaires et de bibliothèques fondamentaux nécessaires au fonctionnement des autres modules de Hadoop.
Le package Hadoop Common constitue la base/le cœur du framework, fournissant des services essentiels et des processus de base tels que l'abstraction du système d'exploitation sous-jacent et de son système de fichiers. Il contient également les fichiers JAR et les scripts nécessaires au démarrage de Hadoop. Le package Hadoop Common fournit également du code source et de la documentation, ainsi qu'une section de contribution qui comprend différents projets de la communauté Hadoop.

2.4 Apache Spark

2.4.1. Historique

Apache Spark a débuté en tant que projet de recherche au laboratoire AMP de l'Université de Californie à Berkeley en 2009 et est devenu open source début 2010. De nombreuses idées fondamentales ont été présentées dans divers documents de recherche au fil des ans. Initialement axé sur les algorithmes d'apprentissage automatique distribués, Spark a été conçu pour offrir des performances élevées dans les applications itératives. Après sa sortie, Spark est rapidement devenu une vaste communauté de développeurs et a rejoint l'Apache Software Foundation en 2013. Il est désormais développé en collaboration par des centaines de développeurs de nombreuses organisations. [35]

2.4.2 Définition

Apache Spark, un framework de traitement parallèle open source, prend en charge le traitement en mémoire et conçu pour améliorer les performances des applications dédiées au Big Data. Grâce à Spark, il est possible de traiter d'énormes quantités de données en mémoire, ce qui est beaucoup plus rapide que les méthodes de traitement reposant sur l'accès aux données sur disque. Spark fournit un écosystème riche et diversifié. Il intègre des bibliothèques avec des API composables pour l'apprentissage automatique (MLlib), SQL pour les requêtes interactives (Spark SQL), le traitement de flux (Structured Streaming) pour interagir avec les données en temps réel et le traitement de graphiques (Graph X) [36]. Il prend également en charge des API telles que Java, Python, R et Scala.

Discutons en détail de l'utilisation de ces langages :

1. **La Scala**
   Bien que Spark soit construit sur la programmation Scala, il permet en conséquence d'accéder à certaines de ses fonctionnalités intéressantes. Ces fonctionnalités peuvent ne pas être disponibles dans d'autres langues qui le prennent en charge.
2. **Python**
   Ce langage fournit une vaste gamme de bibliothèques dédiées à l'analyse de données, telles que NumPy, Pandas, Matplotlib et Scikit-learn, qui offrent des fonctionnalités avancées pour le traitement, la manipulation et la visualisation des données, ainsi que pour l'apprentissage automatique.
3. **Langage R**
   Ce langage fournit une plate-forme riche pour l'apprentissage automatique et l'analyse statistique. Améliore également la productivité des développeurs. Pour gérer le traitement sur une seule machine, nous pouvons utiliser le langage R avec Spark via Spark R.
4. **Java**
   Java est définitivement un bon choix pour les développeurs venant d'un background Java + Hadoop.

2.4.3 Caractéristiques d'Apache Spark

Apache Spark possède de nombreuses fonctionnalités qui en font un excellent choix en tant que moteur de traitement de Big Data. Voici quelques-unes des principales caractéristiques qui le distinguent de ses concurrents [37] :

- **Rapidité :** Spark est jusqu'à 100 fois plus rapide que MapReduce pour traiter de grandes quantités de données.
- **Temps réel :** grâce à son traitement en mémoire, il offre un calcul en temps réel et une faible latence.
- **Multi-Language Support :** Spark fournit des API intégrées en Java, Scala ou Python. Par conséquent, on peut rédiger des applications dans différentes langues.
- **Analyse avancée :** il prend en charge les requêtes SQL, le streaming de données, machine learning (ML) et les algorithmes graphiques.
- **Calcul en mémoire :** Spark effectue des calculs en mémoire, ce qui accélère les vitesses de traitement par rapport aux systèmes traditionnels basés sur disque comme Hadoop MapReduce.
- **Intégré à Hadoop :** Apache Spark s'intègre bien au système de fichiers Hadoop HDFS. Il prend en charge plusieurs formats de fichiers tels que parquet, json, csv, etc. Hadoop peut être facilement exploité en utilisant Spark comme source ou destination des données d'entrée.

Voici quelques exemples d'entreprises qui utilisent Spark :

- Netflix : Spark est utilisé par Netflix pour suggérer des films et des séries à ses utilisateurs.
- Amazon : Spark est utilisé par Amazon pour proposer des produits personnalisés à ses clients.
- Google : Spark est employé par Google pour améliorer ses services de recherche et de publicité.
- Facebook : Spark est utilisé par Facebook pour cibler les publicités et améliorer l'expérience des utilisateurs.

2.3.4 Architecture d'Apache Spark

Il s'agit d'une architecture maître-esclave composée d'un pilote, qui s'exécute en tant que nœud maître, et de nombreux exécuteurs qui s'exécutent en tant que nœuds de travail dans le cluster [38], voir la figure 4.

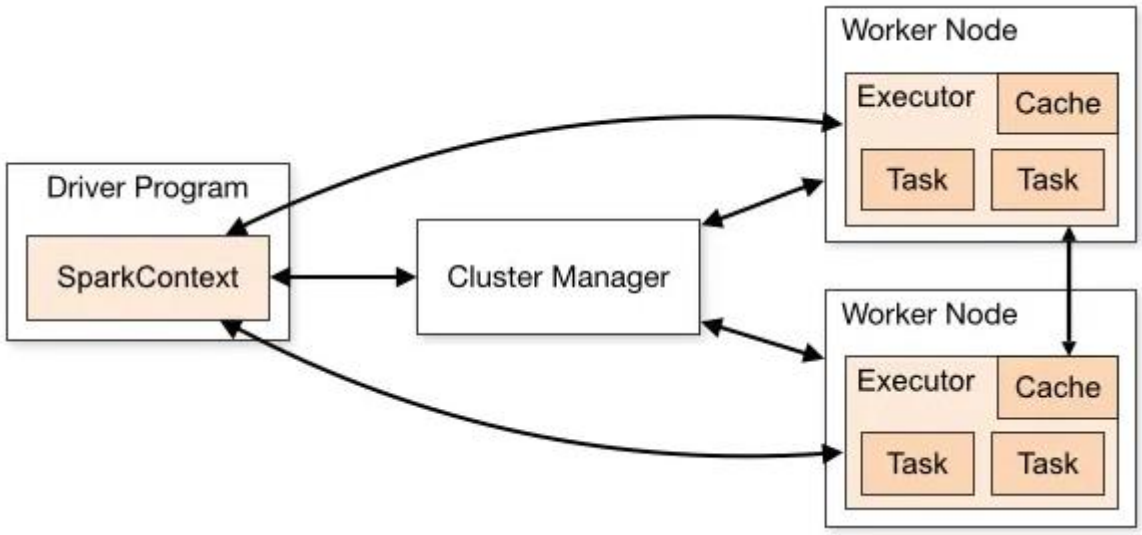

Figure 9. Architecture d'Apache Spark [39]

Le Driver

Le driver ou le pilote est le programme ou le processus chargé de coordonner l'exécution de l'application Spark. Il exécute la fonction principale et crée le SparkContext, qui se connecte au gestionnaire de cluster.

- **Les exécuteurs**
  Les exécuteurs sont des processus de travail responsables de l'exécution des tâches dans les applications Spark. Ils sont lancés sur les nœuds Worker et communiquent avec le programme pilote et le gestionnaire de cluster. Les exécuteurs exécutent des tâches simultanément et stockent les données en mémoire ou sur disque pour la mise en cache et le stockage intermédiaire.
- **The cluster manager**
  Le gestionnaire de cluster est responsable de l'allocation des ressources et de la gestion du cluster sur lequel l'application Spark s'exécute. Spark prend en charge divers

gestionnaires de clusters comme Apache Mesos, Hadoop YARN et standalone cluster manager.

- **SparkContext**
  SparkContext est le point d'entrée de toute fonctionnalité Spark. Il représente la connexion à un cluster Spark et peut être utilisé pour créer des RDD (Resilient Distributed Datasets), des accumulateurs et des variables de diffusion. SparkContext coordonne également l'exécution des tâches.
- **Task**
  C'est une tâche. Elle est la plus petite unité de travail dans Spark, représentant une unité de calcul qui peut être effectuée sur une seule partition de données. Le programme pilote divise la tâche Spark en tâches et les attribue aux nœuds d'exécution pour exécution.

2.4.5 Composants d'Apache Spark (EcoSystem)

Maintenant que nous avons une idée générale de Spark, plongeons plus profondément et comprenons ses composants. Apache Spark est composé du moteur Spark Core, de Spark SQL, de Spark Streaming, de MLlib, de GraphX et de Spark R. Vous pouvez utiliser le moteur Spark Core avec l'un des cinq autres composants mentionnés précédemment. Il n'est pas nécessaire d'utiliser tous les composants Spark ensemble. Selon le cas d'utilisation et l'application, un ou plusieurs de ces composants peuvent être utilisés avec le Spark Core.

Plongeons dans les détails de chaque composant [23] :

Figure 10. Composants de Spark [40]

Spark Core : est le cœur du framework Apache Spark. Il fournit le moteur d'exécution pour la plateforme Spark, nécessaire et utilisé par d'autres composants construits au-dessus de Spark Core selon les besoins. Le Spark Core offre la capacité de calcul en mémoire intégrée et la référence des ensembles de données stockés dans des systèmes de stockage externes. Il incombe au Spark Core d'effectuer toutes les fonctions de base d'E/S, de planification, de surveillance, etc. De plus, la récupération en cas de panne et une gestion efficace de la mémoire sont d'autres fonctions importantes du Spark Core.

Spark Streaming : est un module du framework Apache Spark qui permet le traitement des flux de données en temps réel. Il offre aux développeurs la possibilité de travailler avec des flux de données continus, ce qui ouvre la voie à une gamme d'applications en temps réel telles que le monitoring, la détection d'anomalies, l'analyse en temps réel, et bien plus encore.

Spark SQL : Spark SQL permet d'exécuter des requêtes SQL et d'effectuer des manipulations de données structurées à l'aide de Spark. Cela permet aux utilisateurs d'écrire des requêtes SQL ou d'utiliser des API en langage naturel pour interagir avec les données, ce qui facilite l'intégration de Spark dans les applications SQL existantes.

MLlib : MLlib, la bibliothèque d'apprentissage automatique de Spark, offre une solution robuste pour répondre aux besoins croissants des entreprises axées sur les données centrées sur le client. Ces entreprises recherchent des produits et services de données qui utilisent l'apprentissage automatique pour générer des informations prédictives, des recommandations personnalisées et des résultats adaptés aux besoins spécifiques des clients. Cette bibliothèque prend en charge toutes les API telles que Java, Scala et Python dans le cadre des applications Spark.

La bibliothèque MLlib a des implémentations pour plusieurs algorithmes d'apprentissage automatique courants. Comme :

- Clustering
- Classification
- Régression

Apache Spark GraphX : est un moteur de calcul de graphes construit sur Spark. Il permet aux utilisateurs de créer, transformer et raisonner sur des données à grande échelle sous forme de graphes. GraphX est déjà disponible avec une bibliothèque d'algorithmes courants, ce qui facilite les manipulations graphiques.

Spark R : Le langage de programmation R est largement utilisé par les Data scientists en raison de sa simplicité et de sa capacité à exécuter des algorithmes complexes. Mais R souffre d'un problème : sa capacité de traitement des données est limitée à un seul nœud. Cela rend inutilisable lors du traitement d'une énorme quantité de données. Le problème est résolu par SparkR qui est un package R dans Apache Spark.

Chaque composant de Spark est conçu pour répondre à des besoins spécifiques dans le traitement et l'analyse des données, offrant ainsi une flexibilité et une extensibilité significatives dans le développement d'applications de traitement de données distribuées.

2.4 comparaisons entre Hadoop et Spark Apache

La popularité croissante d'Apache Spark est un élément clé dans la bataille actuelle entre Spark et Hadoop. Dans le monde du Big Data, Spark et Hadoop sont des projets Apache populaires. Nous pouvons dire qu'Apache Spark est une amélioration par rapport au composant Hadoop MapReduce d'origine [32].
Donc, nous allons maintenant explorer pourquoi Spark est souvent le choix préféré pour les tâches de régression d'apprentissage automatique.

1. **Calcul en mémoire et traitement sur disque :** Spark peut conserver les données en mémoire pendant le traitement, ce qui accélère considérablement les performances des

algorithmes itératifs courants dans les tâches d'apprentissage automatique telles que Random Forest Régression, où les mêmes données sont traitées plusieurs fois.

2. **Traitement de données volumineuses :** Spark est conçu pour le traitement de grandes quantités de données en mémoire.
3. **Chargement de données :** Spark charge de grands ensembles de données à l'aide des capacités de chargement de données de Spark à partir de diverses sources telles que HDFS, S3 ou des fichiers locaux.
4. **Intégration et support :** Spark supporte Python et dispose de nombreuses références et ressources communautaires, fournissant un large éventail de documentation et de support pour les développeurs.

En conséquence, Spark présente d'excellentes performances et est très rentable. Il prend également en charge le traitement de données en mémoire. Cependant, il est compatible avec toutes les sources de données et formats de fichiers Hadoop. De plus, il dispose d'API faciles à utiliser et est disponible en plusieurs langues.
Hadoop fournit des fonctionnalités que Spark ne possède pas, comme un système de fichiers distribué et Spark fournit en temps réel, le traitement en mémoire pour les ensembles de données dont on a besoin. Pour les tâches de régression en apprentissage automatique telles que la régression par forêt aléatoire, Spark offre des avantages significatifs par rapport à Hadoop, notamment en termes de performance, de facilité d'utilisation, de capacités de traitement itératif. Même sur un seul nœud, le traitement en mémoire de Spark et la gestion efficace des algorithmes itératifs offrent des avantages substantiels, ce qui en fait un meilleur choix pour le développement et le déploiement de modèles d'apprentissage automatique.

2.6 Conclusion

Dans ce chapitre, nous avons parlé du big data. Nous l'avons défini et parlé de ses caractéristiques et de ses avantages, puis nous avons parlé sur les deux piliers fondamentaux de l'écosystème Big Data : Apache Spark et Apache Hadoop. Nous avons examiné en profondeur leurs caractéristiques, leurs fonctionnalités et leurs distinctions, ainsi que leur rôle crucial dans le traitement et l'analyse des données à grande échelle.

Chapitre 3 : Apprentissage automatique

3.1 Introduction

L'intelligence artificielle (IA) est considérée comme l'une des avancées technologiques les plus prometteuses de notre époque. L'utilisation de l'apprentissage automatique, une sous-catégorie de l'intelligence artificielle, est courante dans divers domaines professionnels. Les méthodes et applications de ces techniques sont expliquées dans le chapitre présent dans votre possession.

3.2 L'intelligence artificiel

L'intelligence artificielle (IA) peut être définie comme l'imitation de l'intelligence humaine par des systèmes informatiques, lui faisant effectuer des tâches similaires à celles des humains et répétant la pensée et le comportement humains. Il permet également à l'intelligence artificielle de simuler une intelligence capable de reconnaître des choses, de résoudre des problèmes et d'effectuer d'autres actions qu'un humain peut effectuer. L'intelligence artificielle comprend à la fois l'apprentissage automatique et l'apprentissage profond [41][42].

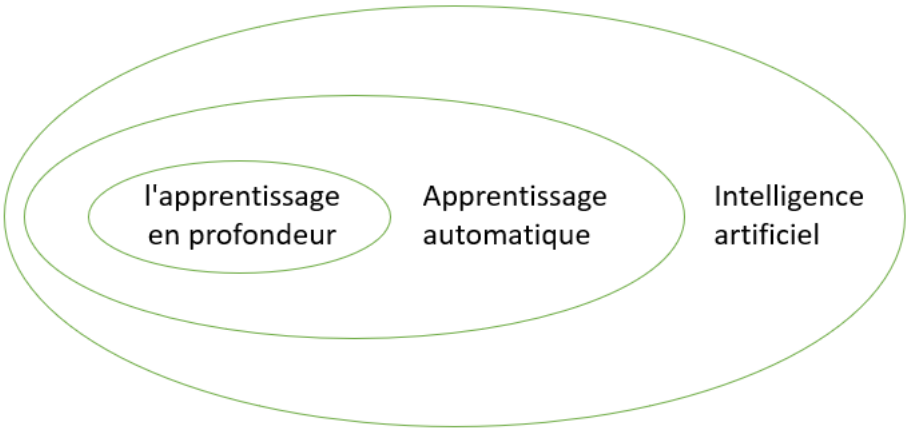

Figure 11. L'intelligence artificiel inclus l'apprentissage automatique et l'apprentissage en profondeur

3.3 définitions de l'apprentissage automatique

L'apprentissage automatique est un outil permettant de convertir des informations en connaissances, en fournissant des données aux ordinateurs et aux machines et en leur faisant utiliser des algorithmes informatiques pour les convertir en modèles utilisables **[43]**.

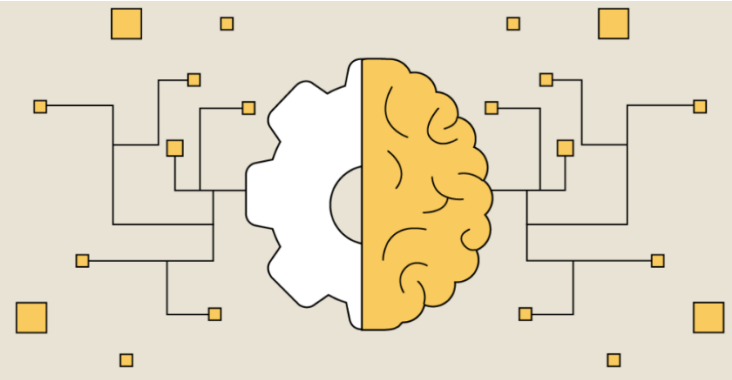

Figure 12. Apprentissage automatique [44]

3.4 L'importance de l'apprentissage automatique

l'apprentissage automatique revêt une grande importance pour diverses raisons, telles que [45]:

- Donne aux entreprises la possibilité de voir les tendances du comportement des clients et des modèles de fonctionnement de l'entreprise, tout en soutenant le développement de nouveaux produits.
- L'intelligence artificielle joue un rôle crucial dans les opérations de nombreuses grandes entreprises contemporaines comme Facebook, Google et Uber.
- L'utilisation de l'apprentissage automatique est devenue un élément compétitif essentiel pour de nombreuses entreprises.

3.5 Types d'apprentissage automatique

Les classificateurs d'apprentissage automatique se répartissent en trois catégories principales [46]:

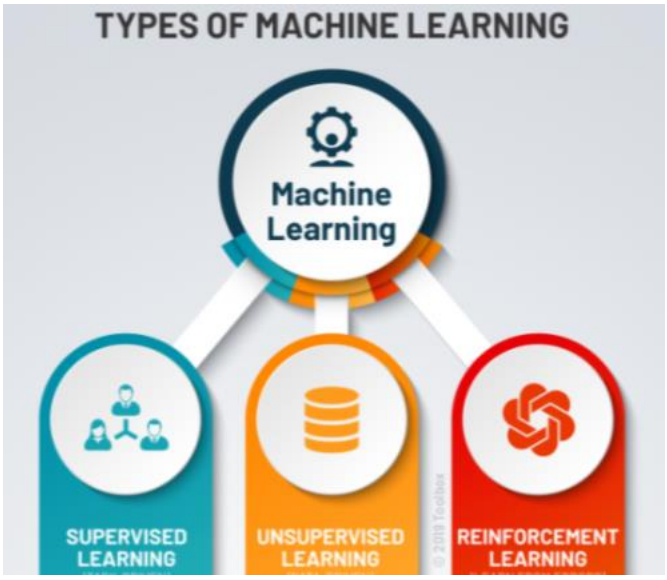

Figure 13. Types d'apprentissage automatique [46]

1. Apprentissage supervisé
2. Apprentissage non supervisé
3. Apprentissage par renforcement

Afin de superviser et guider notre programme, nous utilisons l'apprentissage supervisé dans notre projet.

3.5.1 Apprentissage supervisé

L'apprentissage supervisé signifie superviser ou diriger une certaine activité et s'assurer qu'elle est effectuée correctement. Dans ce type d'apprentissage, la machine apprend sous-direction.
La méthode d'apprentissage supervisé consiste à former un modèle sur un ensemble de données étiquetées, ce qui permet de prédire ou de classer de nouvelles données non étiquetées, ce qui signifie que la sortie vous est déjà connue, le modèle a juste besoin de mapper les entrées sur la sortie. Dans le cadre de l'évaluation de la qualité d'une image, l'utilisation de l'apprentissage supervisé permet de créer des modèles qui peuvent prédire la qualité d'une image en se basant sur différents attributs ou caractéristiques de celle-ci [45][46].

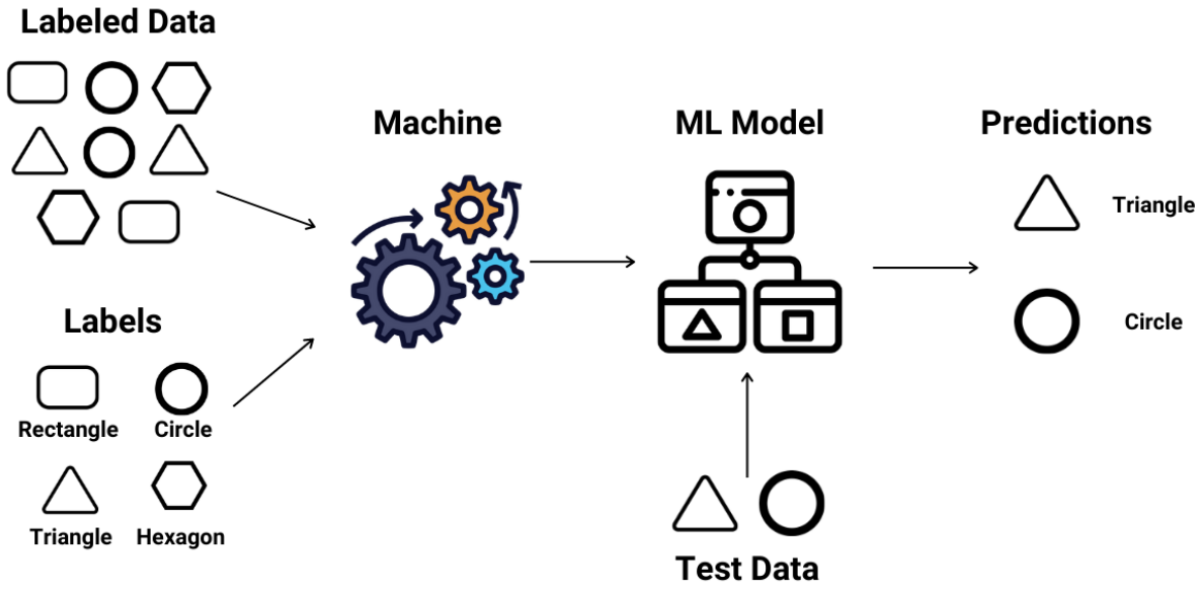

Figure 14. Apprentissage supervisé [47]

Malgré les avantages de l'apprentissage supervisé pour évaluer la qualité d'image, il pose également des défis particuliers [48] :

- **Collecte de Données Étiquetées :** L'une des principales conditions de l'apprentissage supervisé est la présence d'un ensemble de données étiquetées de grande qualité. La collecte et l'analyse de ces informations peuvent être onéreuses et demander du temps.
- **Interprétabilité du Modèle :** La compréhension des facteurs influençant la qualité d'image peut être limitée par la transparence et l'interprétabilité des modèles d'apprentissage

supervisé, notamment les modèles complexes tels que les forêts aléatoires ou les réseaux neuronaux.

- **Généralisation et Adaptabilité :** Même si les modèles supervisés ont la capacité de prédire de manière précise l'ensemble des données sur lesquelles ils ont été élaborés, leur capacité à généraliser et à s'adapter à de nouveaux types d'images ou à des scénarios d'utilisation différents peut être restreinte.

L'apprentissage supervisé peut être divisé en deux types : la classification et la régression.

1. **Classification :** Il s'agit d'un processus par lequel un ensemble particulier de données, qu'elles soient structurées ou non, est classé dans une catégorie **[**49**]**.
2. **Régression :** est une méthode de modélisation prédictive qui permet de déterminer la corrélation entre deux variables ou plus. L'utilisation principale de la régression est la prédiction et l'inférence causale. Les algorithmes de régression les plus couramment utilisés sont la régression linéaire et la régression logistique **[**50**]**.

Dans cette situation, notre programme utilise la méthode de regression.

L'apprentissage supervisé de la régression vise à prédire des valeurs continue, à la différence de la classification qui cherche à prédire des classes discrètes. Son objectif est de élaborer un modèle mathématique qui prévoit la valeur d'une variable cible en fonction d'une ou plusieurs variables prédictives.

Types de régression : il y a deux types :

Régression linéaire:[51]

- Les variables prédictives et la variable cible sont linéaires.
- Modèle facile à comprendre.

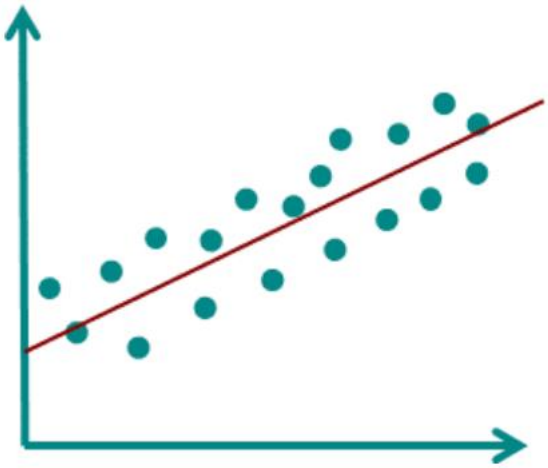

Figure 15. Exemple de graphe de régression linear [52]

Régression non linéaire[53]:

- Relation non linéaire entre la variable cible et les variables prédictives.
- Modèles plus sophistiqués et flexibles.

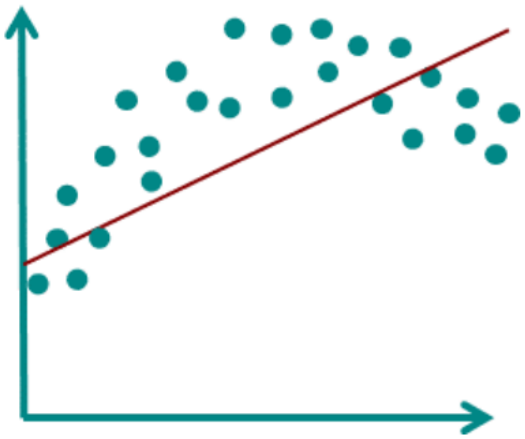

Figure 16. Exemple de graphe de regression non-linear [52]

Les techniques d'apprentissage automatique supervisé impliquent de nombreux algorithmes et méthodes de calcul. dans notre cas, nous utilisons principalement les techniques suivantes :

3.5.1.1 Régression par Forêts Aléatoires (Random Forest Regression)

Dans le programme que nous avons développé, nous avons utilisé la régression pour résoudre un problème spécifique, pour prédire une variable continue (dans ce cas, la qualité de l'image). Plus précisément, nous avons utilisé la Régression par Forêts aléatoires, une technique de régression non linear. Son mécanisme consiste à générer une variété d'arbres de décision pendant l'entraînement et à extraire la moyenne des prédictions individuelles de ces arbres afin de fournir une estimation finale. Une explication approfondie de son fonctionnement dans le cadre de la régression est fournie [54]:

1. Mise en place d'arbres de décision : Prélèvement par bootstrapping : Par échantillonnage avec remplacement, de nombreux sous-ensembles de données sont générés à partir de l'ensemble de données d'entraînement. Chaque composant est employé afin de former un arbre de décision. Sélection de caractéristiques aléatoires : Un sous-ensemble aléatoire des caractéristiques est choisi pour chaque nœud de chaque arbre. Ceci offre la possibilité de varier les arbres, car divers arbres seront édifiés à partir de différents ensembles de caractéristiques.

2. Stimulation des arbres : Tous les arbres sont conçus de façon à réduire au minimum l'erreur sur propre échantillon en utilisant le bootstrapping, qui garantit de ne pas utiliser les mêmes données pour chaque arbre. La sélection aléatoire des caractéristiques permet de réduire la corrélation entre les arbres.

Si chaque caractéristique est utilisée, la plupart des arbres auront les mêmes nœuds de décision et agiront de la même manière.

3. Prévision : Prédiction personnalisée : Afin de prédire une observation future, chaque arbre de décision dans la forêt émet une prédiction individuelle pour une nouvelle observation. Ensemble des prédictions : Par la suite, les prédictions des divers arbres sont regroupées, et une agrégation est réalisée en prenant la moyenne des prévisions de tous les arbres.

4. Adaptation et vérification : Les résultats de la forêt aléatoire sont mesurés à l'aide d'ensembles de validation croisée ou d'un ensemble de test particulier. Il est possible de modifier plusieurs paramètres afin d'améliorer les performances du modèle, comme le nombre

d'arbres, la profondeur maximale des arbres, le nombre minimum d'échantillons par feuille, etc.

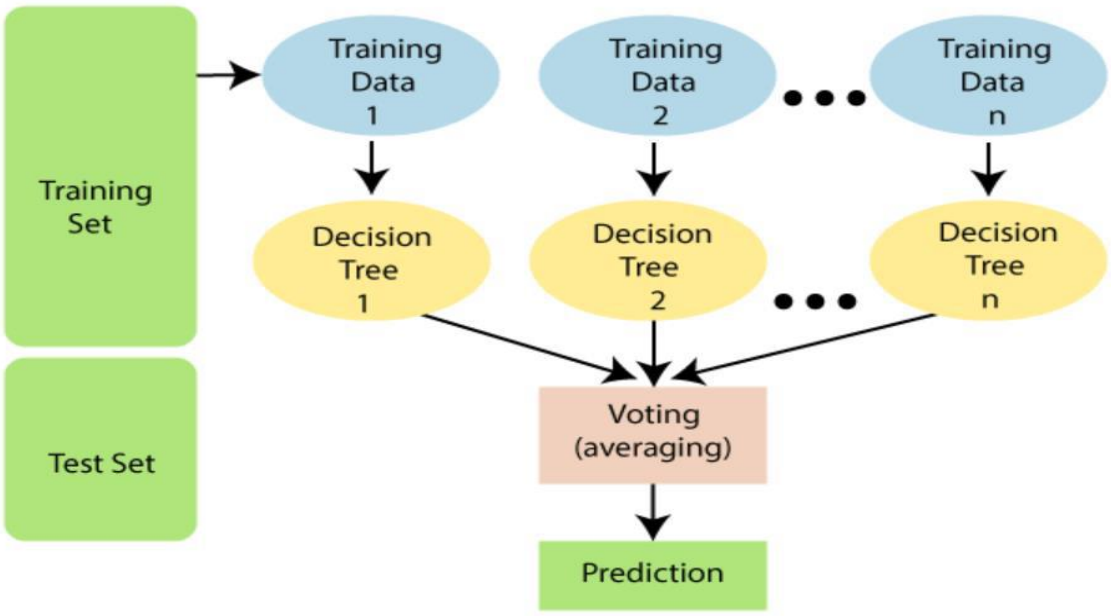

Figure 17. diagramme d'une forêt aléatoire pour la régression

Explication du diagramme :

- Haut du diagramme :
  "Données d'apprentissage" : Représente l'ensemble de données original utilisé pour entraîner la forêt aléatoire.
  "1", "2", ..., "n" : Représentent les différents sous-ensembles de données créés par échantillonnage avec remise.
- Centre du diagramme :
  "Arbre de décision 1", "Arbre de décision 2", ..., "Arbre de décision n" : Représentent les différents arbres de décision construits à partir des sous-ensembles de données.
  "Vote (moyenne)" : Indique que la prédiction finale pour une nouvelle observation est obtenue en faisant la moyenne des prédictions de tous les arbres de la forêt.
- Bas du diagramme :
  "Ensemble de test" : Représente un ensemble de données distinct utilisé pour évaluer la performance de la forêt aléatoire.
  "Prédiction" : Indique que la forêt aléatoire est utilisée pour faire des prédictions pour les données du test.
  Dans le programme précédent, la Régression par Forêts Aléatoires est utilisée pour prédire la variable cible en fonction d'un grand nombre de variables prédictives. Le modèle est entraîné sur un ensemble d'entraînement, puis évalué sur un ensemble de test pour mesurer ses performances à l'aide de différentes métriques.

3.5.1.2 Validation Croisée en k-fold

Depuis 40 ans, cette méthode est largement utilisée par les praticiens pour évaluer la performance d'une méthode. La validation croisée est employée pour résoudre de nombreux problèmes tels que la sélection de modèle, l'adaptabilité et l'identification. On peut

considérer cela comme une application spécifique des méthodes appelées rechantillonnage. Dans cette situation, on utilise la validation croisée en k-fold afin d'évaluer de manière plus fiable et robuste la performance du modèle de régression (Random Forest) et réduire le surapprentissage(overfitting) [55].

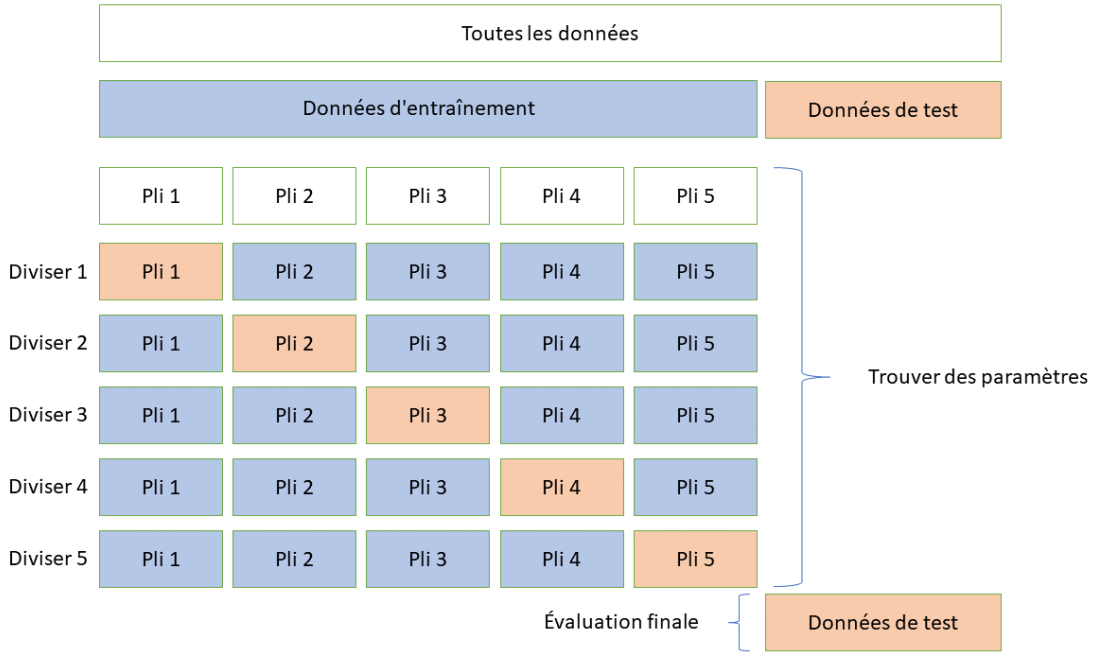

Figure 18. Le processus de validation croisée [56]

Voici pourquoi elle est utilisée et comment elle est liée au surapprentissage [57]:

1. **Évaluation de la performance du modèle:**
   - k-fold est une méthode de formation et de validation utilisée pour évaluer les performances du modèle à l'aide de différents sous-ensembles de données. Au lieu de diviser en un ensemble d'entraînement et un ensemble de test, cela fournit une estimation plus précise des performances du modèle.
2. **Surapprentissage (Overfitting) :**
   - Lorsque le modèle apprend avec trop de précision les caractéristiques spécifiques des données d'entraînement, un surapprentissage se produit, ce qui correspond à une perte de capacité et à une incapacité à bien généraliser à des données nouvelles, invisibles et inédites.
   - Grâce à la méthode de validation croisée en k-fold, il devient possible d'évaluer les compétences du modèle sur divers ensembles de données. Cette approche simplifie la détection de l'état de surapprentissage du modèle.
   - Si le modèle parvient à obtenir de bons résultats lors de l'entraînement mais affiche de mauvais résultats lors de la validation, cela peut indiquer un surapprentissage.
3. **Validation croisée pour eviter le surapprentissage :**
   - Les données sont divisées en k plis lors de la validation croisée.
   - Pour chaque pli :
     - Les données sont réparties en un ensemble d'entraînement et un ensemble de test pour chaque pli.

- L'entraînement du modèle se déroule sur l'ensemble d'entraînement.
- Les résultats du modèle sont mesurés à travers l'ensemble des tests.

On évalue les performances du modèle en moyennant les scores de performance sur tous les plis afin d'obtenir une évaluation globale.

La validation croisée k-fold est un outil crucial pour évaluer de manière objective les performances d'un modèle et combattre le surapprentissage est la validation croisée k-fold, qui assure que le modèle peut être généralisé à de nouvelles données et fournir des prédictions précises.

3.6 Métriques d'évaluer les performances des modèles

Il y a des métriques spécifiques sont nécessaires pour quantifier les performances d'un modèle ou d'une méthode. Ces métriques jouent un rôle crucial dans la comparaison des performances, l'optimisation des algorithmes et la validation des résultats. Voici les métriques les plus couramment utilisées pour évaluer la précision des prédictions dans le cadre de la régression [54][58]:

3.6.1 Erreur quadratique moyenne (MSE)

L'erreur quadratique moyenne (MSE) est une mesure courante pour les tâches de régression qui mesure l'erreur quadratique moyenne entre les valeurs prédites et les valeurs réelles.

Elle est définie par la formule suivante :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

où y_i représente la valeur réelle, $\hat{y}_i$ représente la valeur prédite, et n est le nombre total d'échantillons.

3.6.2 Erreur quadratique moyenne (RMSE)

L'erreur quadratique moyenne Racine (RMSE) est simplement la racine carrée de l'erreur quadratique moyenne (MSE). Elle est utile pour exprimer l'erreur dans les mêmes unités que la variable de sortie, ce qui facilite l'interprétation.

$$RMSE = \sqrt{MSE}$$

3.6.3 Corrélation de Spearman

La corrélation de Spearman mesure la corrélation monotone entre deux variables. Elle est utilisée pour évaluer la corrélation entre les valeurs prédites et les valeurs réelles, et est particulièrement utile lorsque les relations entre les variables ne sont pas linéaires.

$$SROCC(A,B) = \frac{\sum_{i=1}^{n}(A_i - \hat{A})(B_i - \hat{B})}{\sqrt{\sum_{i=1}^{n}(A_i - \hat{A})^2}\sqrt{\sum_{i=1}^{n}(B_i - \hat{B})^2}},$$

3.6.4 Corrélation de Pearson

La corrélation de Pearson mesure la corrélation linéaire entre deux variables. Elle est également utilisée pour évaluer la corrélation entre les valeurs prédites et les valeurs réelles, mais elle est plus sensible aux relations linéaires.

$$PLCC(A,B) = \frac{\sum_{i=1}^{n}(A_i - \overline{A})(B_i - \overline{B})}{\sqrt{\sum_{i=1}^{n}(A_i - \overline{A})^2}\sqrt{\sum_{i=1}^{n}(B_i - \overline{B})^2}},$$

Interprétation :

- La force et la direction d'une relation monotone entre deux variables sont évaluées par des mesures statistiques telles que le coefficient de corrélation de Spearman et le coefficient de corrélation de Pearson.
- Le coefficient de corrélation a une valeur allant de -1 à 1.
  - Une valeur qui se rapproche de 1 témoigne d'une corrélation positive importante, ce qui indique que les variables ont tendance à augmenter ou à diminuer au même moment.
  - La présence d'une valeur proche de -1 témoigne d'une corrélation négative importante, ce qui indique que les variables ont tendance à évoluer différemment.
  - Une valeur proche de 0 témoigne d'une corrélation faible ou absente.
- Le coefficient de Spearman est une mesure non paramétrique, ce qui signifie qu'il ne fait aucune hypothèse sur la distribution des données.
- Le coefficient de Pearson est une mesure paramétrique, ce qui signifie qu'il suppose une distribution linéaire entre les variables.

En utilisant ces mesures, nous pouvons acquérir une compréhension approfondie des performances de notre modèle de régression RandomForestRegressor et évaluer sa capacité à prédire efficacement la qualité de l'image. Ces mesures nous permettent également d'identifier les domaines dans lesquels des améliorations peuvent être apportées pour optimiser les performances du modèle.

3.7 Conclusion

Ce chapitre offre un aperçu complet des techniques utilisées pour modéliser et prédire des valeurs continues, en définissant l'intelligence artificielle et l'apprentissage automatique, en explorant les différents types d'apprentissage supervisé et en se concentrant sur la régression. Dans le chapitre suivant, en présentera notre nouvelle méthode qui se base sur une méthode de régression pour la prédiction de qualité d'image.

Chapitre 4 : Implémentation et résultats

4.1 Introduction

Notre objectif est de proposer une nouvelle méthode pour l'évaluation de la qualité des images sans référence sous un apache Spark. Cette méthode est le résultat d'une régression par Random Forest des caractéristiques des images de la base.
Dans ce chapitre, nous présenterons en détails notre méthodes tout en commençant par les outils utilisés, une description de la base d'image utilisés et l'approche proposé. Nous expliquerons enfin les résultats obtenus par différentes méthodes d'évaluation.

4.2 Environnement de travail

4.2.1 Matériel

Fabricant : HP
Processeur : Intel(R) Core (TM) i7-6600U CPU @ 2.60GHz 2.81 GHz.
RAM: 8,00 Go.
Type de système : système d'exploitation 64 bits.

4.2.2 Langage de programmation python

Python est le langage informatique le plus répandu et le plus employé, en particulier dans les domaines de la Science des Données et de Machine Learning. En outre, Python offre une compatibilité avec différents systèmes d'exploitation, tels que Windows, MacOs et Linux, ce qui en fait une option parfaite pour les développeurs qui travaillent dans divers environnements [59].

4.2.3 Pourquoi Python

Nous avons choisi Python plutôt que d'autres langages pour les raisons suivantes [60]:

1. **Simplicité syntaxique :** les autres langages ont tendance à être plus verbeux que Python en termes de syntaxe, ce qui peut rendre le code plus long et plus complexe à écrire et à comprendre. Cela peut ralentir le développement initial du projet.

2. **La flexibilité :** Python est souvent considéré comme plus flexible que les autres langages en raison de son typage dynamique, de sa syntaxe concise et de sa gestion plus souple de la mémoire.
3. **Écosystème des bibliothèques :** Bien que d'autres langages disposent d'un écosystème de bibliothèques assez vaste, il pourrait être moins développé ou moins spécialisé dans certains domaines, notamment la machine learning et la manipulation de données par rapport à Python. Ce qui facilite d'autant plus son utilisation pour effectuer des actions complexes.
4. **Performance relative :** Bien que d'autres machines virtuelles dans d'autres langages fonctionnent très bien, elles peuvent ne pas être aussi efficaces que Python pour certaines tâches, en particulier en ce qui concerne le développement de prototypes rapides. Cela pourrait entraîner des temps de développement plus longs ou une moindre réactivité lors du développement et du test de nouvelles fonctionnalités.

4.2.4 Bibliothèques Python spécifiques

Dans notre programme nous avons utilisés les bibliothèques Python ci-dessous pour effectuer l'objectif de notre recherche :

- **PySpark et SparkSession [61] :** Sont essentiels pour traiter efficacement les gros volumes de données de manière distribuée.
  - **PySpark :** Interface Python pour Apache Spark, un framework de traitement de données distribué.
  - **SparkSession :** Permet de créer une session Spark pour charger et traiter de gros volumes de données, comme l'ensemble de données que nous avons utilisé.
- **scikit-learn [61] :** Offre des algorithmes d'apprentissage automatique performants et des outils d'évaluation pour la régression prédictive et l'analyse des résultats.
  - **RandomForestRegressor :** Modèle d'apprentissage par ensemble utilisé pour la régression prédictive de la qualité d'image.
  - **KFold :** Fonction pour effectuer une validation croisée K-fold afin d'évaluer la performance du modèle de manière impartiale.
  - **spearmanr et pearsonr :** Fonctions pour calculer les coefficients de corrélation de Spearman et de Pearson, respectivement, pour mesurer la relation entre les valeurs prédites et les valeurs réelles de la qualité d'image.
- **Pandas et NumPy [62] :** facilitent le prétraitement des données tabulaires, les manipulations mathématiques et la préparation des données d'entrée pour le modèle.
  - **Pandas :** Permet de charger, manipuler et nettoyer les données tabulaires, comme le fichier CSV contenant notre donnée.
  - **NumPy :** Offre des fonctions pour les opérations mathématiques et les manipulations de tableaux, notamment pour le prétraitement des données et la préparation des données d'entrée pour le modèle de régression.

- **matplotlib.pyplot [63] :** est un module de la bibliothèque Matplotlib, qui est une bibliothèque de visualisation de données en Python. Ce module fournit une interface de style MATLAB pour créer des graphiques. Avec matplotlib.pyplot, on peut créer des graphiques 2D, des diagrammes à barres, des nuages de points, des histogrammes et bien d'autres types de visualisations pour explorer et présenter les données de manière efficace.

4.3 Approche Proposée et Résultats

4.3.1 L'ensemble de données et le système

4.3.1.1 Présentation de l'ensemble de données KADID-10K

Dans cette étude, nous avons mené une analyse de régression pour prédire la qualité d'image sans référence (NR-IQA) en utilisant l'ensemble de données KADID-10K. Cet ensemble de données largement utilisé dans la communauté NR-IQA, se compose de 10 125 images déformées dérivées de 81 images de référence vierges (sans distorsion) utilisant 25 types de distorsion différents à 5 niveaux de distorsion différents.

De plus, chaque image est associée à une valeur subjective MOS différentielle.
Les images de l'ensemble de données KADID-10K ont été prétraitées en redimensionnant et en normalisant leurs pixels. Ensuite, les caractéristiques statistiques globales ont été extraites de chaque image, telles que la moyenne, la variance, l'entropie et la dimension fractale. En total, 132 caractéristiques pour chaque image ont été utilisées comme variables prédictives dans un modèle de régression par forêt aléatoire pour prédire les scores de qualité d'image [64].

Figure 19. Les 81 images vierges de KADID-10k [65]

4.3.1.2 Le système

Le système utilisé dans ce programme est une combinaison de PySpark (une bibliothèque de traitement des données distribuée), et scikit-learn (une bibliothèque d'apprentissage automatique pour Python).
L'ensemble de données utilisé est stocké dans un fichier CSV. Ce fichier contient des données utilisées pour effectuer une analyse de régression, où certaines caractéristiques sont utilisées pour prédire une variable cible représentée par la valeur subjective de la qualité de l'image.

4.3.2 Méthodologie

Dans cette section, nous décrivons en détail la méthodologie employée pour l'analyse de régression visant à prédire la qualité d'image sans référence (NR-IQA) à l'aide de l'ensemble de données KADID-10K et du modèle Random Forest.

4.3.2.1 Chargement et préparation des données

La première étape consiste à charger le fichier CSV contenant les données de l'ensemble de données KADID-10K. Ce fichier est supposé se trouver dans un chemin spécifié. Les données sont stockées dans un DataFrame Pandas nommé regres_data. La dernière colonne est définie comme la variable cible à prédire (qualité d'image). Tandis que les autres colonnes représentent les variables prédictives (caractéristiques statistiques des images) et qui sont regroupées dans la variable X.

```
regres_data = pd.read_csv('C:/regres.csv')
X = regres_data.drop(columns=["ec"]).values
y = regres_data["ec"].values
```

Figure 20. Chargement des données

4.3.2.2 Initialisation et configuration du modèle RandomForestRegressor

```
model = RandomForestRegressor(random_state=42, n_jobs=-1)
```

Figure 21. Initialiser le modèle RandomForestRegressor

Le modèle RandomForestRegressor est initialisé avec les paramètres spécifiés instancié à l'aide de la bibliothèque scikit-learn.

Paramètres :

- **random_state :** Ce paramètre définit la graine aléatoire « Random seed » pour la reproductibilité. Le paramètre « random_state » est fixé à 42 pour garantir la reproductibilité des résultats

- **n_jobs** : Nombre de tâches à exécuter en parallèle pour l'ajustement et la prévision (fit &predict). Et « n_jobs= -1 » indique que le modèle doit exploiter tous les cœurs de processeur disponibles pour le calcul parallèle.

4.3.2.3 Initialisation de la validation croisée en k-fold

```
kf = KFold(n_splits=5, shuffle=True, random_state=42)
```

Figure 22. Initialiser la validation croisée

Une stratégie de validation croisée en 5 plis est mise en place à l'aide de la fonction KFold() de scikit-learn.

Paramètres :

- **n_splits=5 : ce** qui signifie que les données sont divisées en 5 parties.
- **shuffle=True :** est également utilisée, ce qui signifie que les données seront mélangées avant d'être divisées en 5 parties. Cela garantit que les données sont réparties aléatoirement entre les plis de la validation croisée.

4.3.2.4 Validation croisée en K-fold pour une évaluation impartiale

```
rmse_scores = np.sqrt(-cross_val_score(model, X, y, cv=kf, scoring="neg_mean_squared_error"))
```

Figure 23. Calculer les scores RMSE de la validation croisée

Le processus d'entraînement et d'évaluation est répété 5 fois, en utilisant chaque sous-ensemble comme ensemble de test à tour de rôle. Les scores RMSE (Root Mean Squared Error) sont calculés pour chaque pli de la validation croisée et stockés dans le tableau rmse_scores. Le meilleur score RMSE est identifié comme le score RMSE minimal parmi tous les plis, indiquant la performance prédictive attendue la plus élevée.

Paramètres :

cross_val_score() : gère automatiquement l'entraînement du modèle sur chaque pli de la validation croisée

- **model :** C'est le modèle de machine learning que nous avons initialisé précédemment (dans ce cas, un RandomForestRegressor).

- **X :** Ce sont les données d'entrée, c'est-à-dire les caractéristiques.

- **y :** Ce sont les valeurs cibles.

- **cv :** Détermine la stratégie de division de validation croisée et la nôtre était une stratégie de validation kfold.

- **Scoring :** dans la fonction de validation renverra les valeurs MSE

```
best_rmse = np.min(rmse_scores)
```

Figure 24. Sélectionner le meilleur score RMSE

- **Sélectionement du meilleur score RMSE :** Le meilleur score RMSE parmi les plis de la validation croisée est sélectionné.

4.3.2.5 Entraînement du modèle sur l'ensemble des données

```
model.fit(X, y)
```

Figure 25. L'entraînement le modèle

model.fit() : Une fois la validation croisée terminée et le meilleur score RMSE sélectionné, le modèle est entraîné sur l'ensemble des données (y compris les données d'entraînement et de test utilisées dans la validation croisée). Cela permet au modèle d'apprendre les relations entre les variables prédictives et la variable cible.

4.3.2.6 Prédictions avec le modèle entraîné

```
predictions = model.predict(X)
```

Figure 26. Prédiction

La fonction predict() du modèle entraîné est utilisée pour générer des prédictions de qualité d'image pour toutes les données. Les prédictions sont stockées dans le tableau predictions.

4.3.3 Évaluation des performances

Les coefficients de corrélation de Spearman (spearman_corr) et de Pearson (pearson_corr) sont calculés entre les valeurs réelles de qualité d'image (y) et les valeurs prédites (predictions). Ces mesures indiquent la force et la direction de la relation linéaire entre les variables. Ces mesures sont utiles pour évaluer la performance et la validité du modèle par rapport aux données observées. À la fin, le meilleur score RMSE, les corrélations de Spearman et de Pearson sont affichées à la console et un graphique est créé pour visualiser la relation entre les valeurs prédites et les valeurs réelles de la qualité d'image.

4.3.4 Évaluation des résultats

4.3.4.1 Le coefficient RMSE

- Nous avons utilisé un tableau contenant toutes les valeurs prédites. Ce tableau permet d'analyser la distribution des prédictions et d'identifier d'éventuels écarts importants entre les valeurs prédites et les valeurs réelles de la qualité d'image.
- Nous avons obtenu la valeur : RMSE=0.5569014861018617, cette valeur de RMSE qui présente l'erreur moyenne quadratique entre les valeurs subjectives et les valeurs prédites est très petite. Ce qui nous indique qu'il est capable d'adapter au mieux l'ensemble de données.

4.3.4.2 Corrélations de Spearman et de Pearson

- Selon la définition de ces coefficients, plus que leurs valeurs sons grandes plus que les résultats sont bons. Le coefficient de corrélation de Spearman est de 0.9865570308305986, tandis que le coefficient de corrélation de Pearson est de 0.9864240394795287. Ces valeurs élevées indiquent une forte corrélation positive entre les valeurs prédites et les valeurs réelles de la qualité d'image. Cela signifie que le modèle prédit la qualité d'image de manière très cohérente avec les évaluations subjectives humaines.

4.3.4.3 Diagramme de dispersion

- Un graphique est créé pour visualiser la relation entre les valeurs prédites et les valeurs réelles de la qualité d'image. Ce graphique permet d'observer la distribution des points et d'évaluer la précision des prédictions du modèle. On remarque que les points ne sont pas dispersés (sont très proches) ce qui signifie une cohérence des résultats par rapport aux valeurs subjectives.

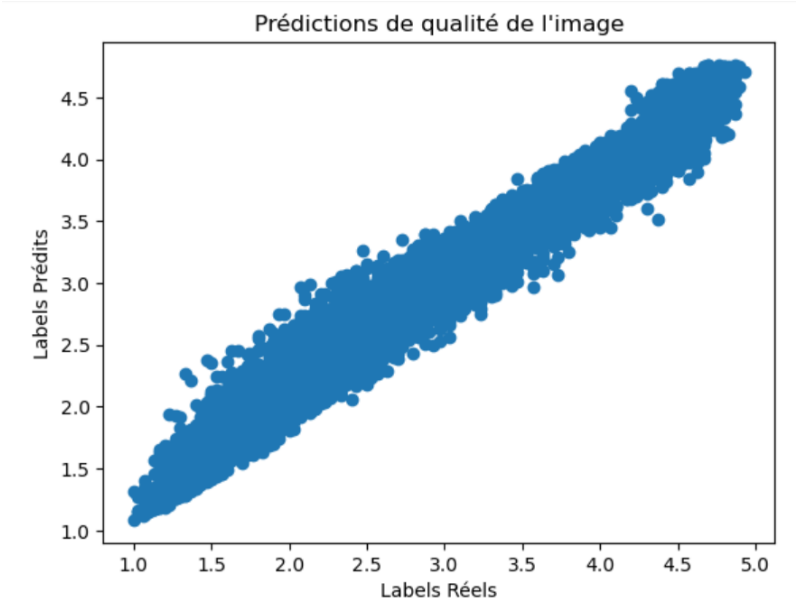

Figure 27. Diagramme de dispersion

4.4 Comparaison avec les études antérieures

Pour comparer les résultats de notre méthode avec les études antérieures appliquées à l'ensemble de données kadid10-k, nous nous référons aux trois méthodes suivantes : ENIQA [66], NBIQA [67] et GSF-IQA [68]. Les résultats sont résumés dans le tableau ci-dessous :

	KADID-10K	
Méthode	**PLCC**	**SROCC**
ENIQA	0.634	0.636
NBIQA	0.635	0.626
GSF-IQA	0.737	0.725
RFR-IQA	**0.986**	**0.987**

Tableau 1.*Comparaison des résultats de RFR-IQA méthode avec les études antérieures*

Comparé aux méthodes antérieures pour l'ensemble de données KADID-10K, le modèle RFR surpasse les performances de ENIQA, NB-IQA, et GSF-IQA. Les coefficients de corrélation obtenus avec RFR sont significativement plus élevés. Plus précisément, le coefficient de corrélation de Spearman est de 0.9865570308305986 et le coefficient de corrélation de Pearson est de 0.9864240394795287, ce qui témoigne de la capacité du modèle à prédire la qualité d'image de manière très précise et cohérente avec les évaluations subjectives humaines. Cela représente une avancée majeure par rapport aux méthodes précédentes, mettant en évidence l'efficacité et la fiabilité du modèle RFR pour l'évaluation de la qualité des images.

4.5 Conclusion

Le dernier chapitre de cette étude contient les outils que nous avons utilisés pour effectuer une régression afin de prédire la qualité de l'image. Il contient également les résultats que nous avons obtenu grâce à une analyse de régression qui a démontré la capacité du modèle Random Forest à prédire efficacement la qualité d'image sans référence à l'aide de l'ensemble de données KADID-10K.

Notre analyse de régression a permis d'obtenir des scores de qualité d'image prédits précis pour les images KADID-10K, avec une corrélation élevée entre les valeurs prédites et les scores réels. Ces résultats indiquent que notre modèle est capable de capturer efficacement les caractéristiques statistiques globales qui influencent la qualité d'image.

Conclusion générale

Ce projet a abordé l'utilisation de l'algorithme de régression sous un apache Spark pour évaluer la qualité des images où son évaluation est un défi qui implique à la fois des aspects techniques et perceptuels.

Des mesures objectives visent à fournir une évaluation automatisée et numérique de la qualité de l'image, sans s'appuyer sur des jugements humains subjectifs. Ces méthodes analysent divers attributs de l'image tels que le bruit, le contraste, la couleur, etc. pour générer un score numérique en corrélation avec la qualité de l'image perçue.

Dans notre projet au premier chapitre, nous avons discuté deux environnements : Apache Hadoop et Apache Spark, et leurs étonnantes capacités de traitement des données. Nous avons montré ses composants et caractéristiques, puis nous les avons comparés et avons conclu lequel était le plus adapté à notre projet.

Dans le deuxième chapitre, nous avons défini la qualité d'image, examiné ses composants et présenté les méthodes d'évaluation subjective et objective, en mettant en évidence les défis liés à cette évaluation. Le troisième chapitre explore les méthodes appliquées, mettant en lumière l'intelligence artificielle et l'apprentissage automatique. Il examine également l'utilisation de techniques telles que la régression par forêts aléatoires pour prédire la qualité des images, ainsi que la validation croisée en k-fold pour évaluer la performance du modèle.

Finalement, Nous avons implémenté et effectué une régression pour prédire la qualité d'image sans référence (NR-IQA) à l'aide de l'ensemble de données KADID-10K. Nous avons obtenu un modèle prédit puis nous avons évalué la performance du modèle et sa capacité à prédire avec précision la qualité d'image perçue par les humains par trois indices de performance (SRCC, PCC, RMSE).

Perspectives

Notre futur objectif vise à améliorer continuellement notre travail. Nous reconnaissons bien que les mesures objectives puissent apporter efficacité et cohérence, elles ne reflètent pas toujours pleinement la perception humaine. Ainsi, notre démarche vise à affiner et valider constamment nos méthodes par rapport aux évaluations subjectives, afin d'améliorer la précision et l'applicabilité de nos résultats.

References

[1]. Vanessa Hojda. How images impact conversion rates on marketplaces [en ligne]. November 28, 2023 disponible sur l'url: https://www.photoroom.com/blog/images-conversion-rates

[2]. Hassan Khan. The Psychology of Images and Infographics in Content Marketing [en ligne]. Aug 10, 2015 disponible sur l'url : https://visme.co/blog/the-psychology-of-images-and-infographics/

[3]. GONZALES, R. C. et WOODS, R. E. Digital image processing 4th edition. 2018.

[4]. SEGHIR, Zianou Ahmed. *Evaluation de la qualité d'image*. 2012. Thèse de doctorat. Université de Mentouri–Constantine.

[5]. FREEMAN, Michael. *The complete guide to digital photography*. Sterling Publishing Company, Inc., 2008.

[6]. VINSONNEAU, Emile, DOMENGER, Jean-Philippe, et CHERIF, Anne. Mesure de la netteté sur une image seule dans des documents anciens. In : *CORIA-CIFED*. 2014. p. 139-152.

[7].22 Adobe. Color Quality and Image Quality: Understanding the Relationship [en ligne]. May 24, 2023.disponible sur l'url : https://helpx.adobe.com/photoshop/using/understanding-color-management.html

[8]. Bentler R, Chiou L-K. Digital Noise Reduction: An Overview[en ligne].*Trends in Amplification*. 2006;10(2):67-82. disponible sur l'url :10.1177/1084713806289514

[9]. JÄHNE, Bernd. *Digital image processing*. Springer Science & Business Media, 2005.

[10]. JAYARAMAN, Subramania, ESAKKIRAJAN, S., et VEERAKUMAR, T. *Digital image processing*. New Delhi : Tata McGraw Hill Education, 2009.

[11]. CHIU, Tai-Yin, ZHAO, Yinan, et GURARI, Danna. Assessing image quality issues for real-world problems. In : *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020. p. 3646-3656.

[12]. SEGHIR, Zianou Ahmed. *Evaluation de la qualité d'image*. 2012. Thèse de doctorat. Université de Mentouri–Constantine.

[13]. L. Zhang, X. Mou and D. Zhang. "FSIM: A Feature Similarity Index for Image Quality Assessment,". (Aug. 2011) [jpg]. in *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp.

[14]. Merzougui, Naima. Métaheuristiques pour l'évaluation objective de la qualité d'images et de vidéos. [en ligne] These de doctorat. Université de mohamed kheider biskra, (2022). Foramat pdf disponible sur : < http://thesis.univ-biskra.dz/id/eprint/5764>

[15]. A. Kazi, S. D. Sawarkar and D. J. Pete, "Image Restoration using Blind Deconvolution," *2019 IEEE Pune Section International Conference (PuneCon)*, Pune, India, 2019, pp. 1-4, doi: 10.1109/PuneCon46936.2019.9105910.

[16]. Harri,Stojka. Filtre passe-bas. (30.08.2008). [jpg] In : Wikipedia. Disponible sur : https://fr.wikipedia.org/wiki/Filtre_passe-bas#/media/Fichier:Lowpass_picture.jpg (18.04.2024).

[17]. K. -H. Thung and P. Raveendran, "A survey of image quality measures," *2009 International Conference for Technical Postgraduates (TECHPOS)*, Kuala Lumpur, Malaysia, 2009, pp. 1-4, doi: 10.1109/TECHPOS.2009.5412098.

[18]. L. Zhang, L. Zhang, X. Mou and D. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," *2012 19th IEEE International Conference on Image Processing*, Orlando, FL, USA, 2012, pp. 1477-1480, doi: 10.1109/ICIP.2012.6467150.

[19]. MOHAMMADI, Pedram, EBRAHIMI-MOGHADAM, Abbas, et SHIRANI, Shahram. Subjective and objective quality assessment of image: A survey. arXiv preprint arXiv:1406.7799, 2014.

[20]. D. Chen, Y. Wang and W. Gao, "No-Reference Image Quality Assessment: An Attention Driven Approach," in *IEEE Transactions on Image Processing*, vol. 29, pp. 6496-6506, 2020, doi: 10.1109/TIP.2020.2990342.

[21]. Z. Wang and A. C. Bovik, "Reduced- and No-Reference Image Quality Assessment," in *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 29-40, Nov. 2011, doi: 10.1109/MSP.2011.942471

[22]. Damon M. Chandler, "Seven Challenges in Image Quality Assessment: Past, Present, and Future Research", International Scholarly Research Notices, vol. 2013, Article ID 905685, 53 pages, 2013. https://doi.org/10.1155/2013/905685

[23]. VIKTOR, Mayer-Schönberger, et KENNETH, Cukier. "Big Data: A Revolution That Will Transform How We Live, Work, and Think". Boston : Houghton Mifflin Harcourt, 2013.

[24]. ROBERT, Slane. Big Data Essentials. New York: Data Insights Press, 2018.

[25]. Business intelligence et big data (22 DÉCEMBRE 2022) [JPG] IN: tim free. Disponible sur: https://timfree.fr/business-intelligence-et-big-data/ (17 avril 2024).

[26]. Shahid Husain, M., Zunnun Khan, M., et Siddiqui, T. "Big Data Concepts, Technologies, and Applications (1st ed.) ". Auerbach Publications. 2023.

https://doi.org/10.1201/9781003441595

[27]. SAYAN, G., AMIT, K. D., SOURABH, M. "Big Data Simplified". Pearson Education India.2019

[28]. Dumka, A., Ashok, A., Verma, P., et Verma, P. Advanced Digital Image Processing and Its Applications in Big Data (1st ed.). CRC Press. 2020 https://doi.org/10.1201/9780429351310

[29]. HURWITZ, J., NUGENT, A., HALPER, F. et KAUFMAN, M. "Big Data for Dummies". John Wiley & Sons, Hoboken. 2013.

[30]. TOM, White. "Hadoop: The Definitive Guide, 4th Edition". O'Reilly Media, Inc. 2015.

[31]. SULTANA, Afreen. "Using Hadoop to Support Big Data Analysis: Design and Performance Characteristics". Culminating Projects in Information Assurance. 27. 2015.

[32]. JEYARAJ, R., PUGALENDHI, G., et PAUL, A. "BIG DATA WITH HADOOP MAPREDUCE: A classroom approach". Apple Academic Press. 2020.

https://doi.org/10.1201/9780429321733

[33]. PANKAJ, Dadhich. Introduction to Apache Hadoop in IoT (09 December 2022) [JPG] IN: DR. PANKAJ DADHICH. Disponible sur : https://www.drpankajdadhich.com/2022/12/introduction-to-apache-hadoop-in-iot.html (17 Mai 2024)

[34]. Map Reduce 101 (11 Février 2020) [JPG] IN: Vipanchi Reddy Disponible sur : https://vipanchikatthula.github.io/post/mapper-reducer-implementation/ (17 Mai 2024)

[35]. RAJDEEP, D., MANPREET, S., et NICK, P. "Machine Learning with Spark - Second Edition". Packt Publishing Ltd. 2017.

[36]. AMJI, J. S., WENIG, B., DAS, T., et LEE, D. "Learning Spark: Lightning-Fast Data Analytics". Sebastopol: O'Reilly Media 2020.

Disponible sur : https://www.oreilly.com/library/view/learning-spark/9781492050032/

[37]. MEZZOUDJ, Saliha. *"Approches de classification des images à grande échelle sur des architectures massivement parallèles."*. Doctorat thésis, Université de Batna 2. 2020. Disponible sur : http://eprints.univ-batna2.dz/1898/

[38]. HOLDEN, K., ANDY, K., PATRICK, W., et MATEI, Z. "Learning Spark". O'Reilly Media, Inc. 2015.

[39]. HEERA, Swati et BAI, Anita. "A Novel Map Reduced Based Parallel Feature Selection and Extreme Learning for Micro Array Cancer Data Classification". Wireless Personal Communications.2022.

[40]. Apache Spark [JPG] IN :nextdicision Disponible sur : https://www.next-decision.fr/editeurs-bi/etl/apache-spark (19 Avril 2024)

[41]. CAMPESATO, Oswald. *Artificial intelligence, machine learning, and deep learning*. Mercury Learning and Information, 2020.

[42]. Boden, Margaret A., *Artificial Intelligence: A Very Short Introduction*, Very Short Introductions (Oxford, 2018; online edn, Oxford Academic, 23 Aug. 2018), https://doi.org/10.1093/actrade/9780199602919.001.0001

[43]. Gavin Edwards. *Machine Learning | An Introduction*. 18 november 2018. URL: https : / / towardsdatascience . com / machine - learning - an - introduction-23b84d51e6d0#d3ea.

[44]. Jérémy Robert. Machine Learning : Définition, fonctionnement, utilisations.(18 Nov 2020).[JPG] IN : data scientest. Disponible sur : https://datascientest.com/machine-learning-tout-savoir (01 mai 2024).

[45]. DE MATTEIS, Ludovic, JANNY, S., NATHAN, S., *et al.* Introduction à l'apprentissage automatique. 2022.

[46]. ANALYTICS, Potentia. What is machine learning: Definition, types, applications and examples. *Potentia Analytics*, 2019, vol. 19.

[47]. Par Équipe Blent. L'apprentissage supervisé. (12 avr. 2022). [JPG] In : blent. Disponible sur : https://blent.ai/blog/a/apprentissage-supervise-definition (03 mai 2024)

[48]. NAGORNY, Pierre, PILLET, Maurice, et PAIREL, Eric. Contrôle Qualité 2.0: Apprentissage supervisé de la notion de Qualité, application à l'injection plastique. In : *CIGI-QUALITA 2019*. 2019.

[49]. Sotiris B Kotsiantis, Ioannis Zaharakis, P Pintelas, et al. "Supervised machine learning: A review of classification techniques". In: *Emerging artificial intelligence applications in computer engineering* 160.1 (2007), pp. 3–24.

[50]. Gavin Edwards. *Machine Learning | An Introduction*. 18 november 2018. URL: https : / / towardsdatascience . com / machine - learning - an - introduction-23b84d51e6d0#d3ea.

[51]. Xiaogang Su, Xin Yan, and Chih-Ling Tsai. "Linear regression". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.3 (2012), pp. 275–294.

[52]. DATAtab Team (2024). DATAtab: Online Statistics Calculator. DATAtab e.U. Graz, Austria. URL https://datatab.net

[53]. ARCHONTOULIS, Sotirios V. et MIGUEZ, Fernando E. Nonlinear regression models and applications in agricultural research. *Agronomy Journal*, 2015, vol. 107, no 2, p. 786-798.

[54]. Yanli Liu, Yourong Wang, and Jian Zhang. "New Machine Learning Algorithm: Random Forest". In: Information Computing and Applications. Ed. by Baoxiang Liu, Maode Ma, and Jincai Chang. Springer Berlin Heidelberg, 2012.

[55]. CORNEC, Matthieu et BERTAIL, P. Validation Croisée et Modèles Statistiques Appliquées. *Nanterre, Paris (France)*, 2009.

[56]. Balaji Nalawade. The Essential Guide to K-Fold Cross-Validation in Machine Learning. (Mar 19, 2024). [PNG] In: medium. Disponible sur : https://medium.com/@bididudy/the-essential-guide-to-k-fold-cross-validation-in-machine-learning-2bcb58c50578 (03 mai 2024)

[57]. ARLOT, Sylvain. Validation croisée. *Apprentissage statistique et donn\'ees massives*, 2018.

[58]. DE WINTER, Joost CF, GOSLING, Samuel D., et POTTER, Jeff. Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological methods*, 2016, vol. 21, no 3, p. 273.

[59]. DERFOUFI, Younes. Programmation en langage Python. 2019.

[60]. MCKINNEY, Wes. *Python for data analysis*. " O'Reilly Media, Inc.", 2022.

[61]. TESTAS, Abdelaziz. Decision Tree Regression with Pandas, Scikit-Learn, and PySpark. In: *Distributed Machine Learning with PySpark: Migrating Effortlessly from Pan`das and Scikit-Learn*. Berkeley, CA: Apress, 2023. p. 75-113.

[62]. MCKINNEY, Wes. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.

[63]. MORUZZI, Giovanni et MORUZZI, Giovanni. Plotting with matplotlib. *Essential Python for the Physicist*, 2020, p. 53-69.

[64]. VARGA, Domonkos. No-reference image quality assessment with global statistical features. *Journal of Imaging*, 2021, vol. 7, no 2, p. 29.

[65]. An Improved SPSIM Index for Image Quality Assessment - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Konstanz-Artificially-Distorted-Image-quality-Database-KADID-10k-reference-images_fig4_350301845 [accessed 19 May, 2024]

[66]. Chen,X.; Zhang, Q.; Lin, M.; Yang, G.; He, C. No-reference color image quality assessment: From entropy to perceptual quality. EURASIP J. Image Video Process. 2019, 2019, 77.

[67]. Ou, F.Z.; Wang, Y.G.; Zhu, G. A novel blind image quality assessment method based on refined natural scene statistics. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1004–1008.

[68]. Lin, H., Hosu, V., & Saupe, D. (2019). KADID-10k: A large-scale artificially distorted IQA database. Dans Proceedings of the 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX) (pp. 1-3). Berlin, Germany: IEEE: Piscataway.