Rapport de Mini-Projet : Classification d'Images avec l'Algorithme des K plus Proches Voisins (k-NN)

Par: EL MHAMDI NOUHA

---

1. Introduction

Ce rapport présente les résultats d'une étude comparative de l'algorithme des **K plus Proches Voisins (k-NN)** appliqué à la classification d'images. Le projet utilise un sous-ensemble du jeu de données **CIFAR-10**, qui contient 10 classes d'objets (avion, voiture, oiseau, chat, cerf, chien, grenouille, cheval, bateau, camion).

L'objectif principal était d'évaluer l'impact de deux facteurs clés sur la performance du k-NN :

1. Le choix de la **métrique de distance** (Euclidienne L2 vs. Manhattan L1).
2. L'application de la **Réduction de Dimension par Analyse en Composantes Principales (PCA)**.

Une comparaison avec deux autres algorithmes de classification, le **Support Vector Machine (SVM)** et le **XGBoost**, a également été réalisée.

2. Méthodologie Simplifiée

Le jeu de données a été préparé comme suit :

- **Données d'entraînement :** 5000 images.
- **Données de test :** 1000 images.
- **Prétraitement :** Les images 32x32x3 (3072 pixels) ont été aplaties en vecteurs de 3072 dimensions.

2.1. Évaluation du k-NN (Sans PCA)

L'algorithme k-NN a été testé pour différentes valeurs de k (1, 3, 5, 7, 9) en utilisant les distances L1 et L2.

2.2. Impact de la PCA

L'ACP a été appliquée pour réduire la dimensionnalité des données de 3072 à **100 composantes principales**, dans le but d'accélérer l'entraînement et potentiellement d'améliorer la précision en filtrant le bruit.

2.3. Comparaison des Modèles

Le k-NN (avec PCA) a été comparé au SVM (avec noyau RBF) et au XGBoost, tous entraînés sur les données réduites par PCA.

---

3. Résultats et Analyse

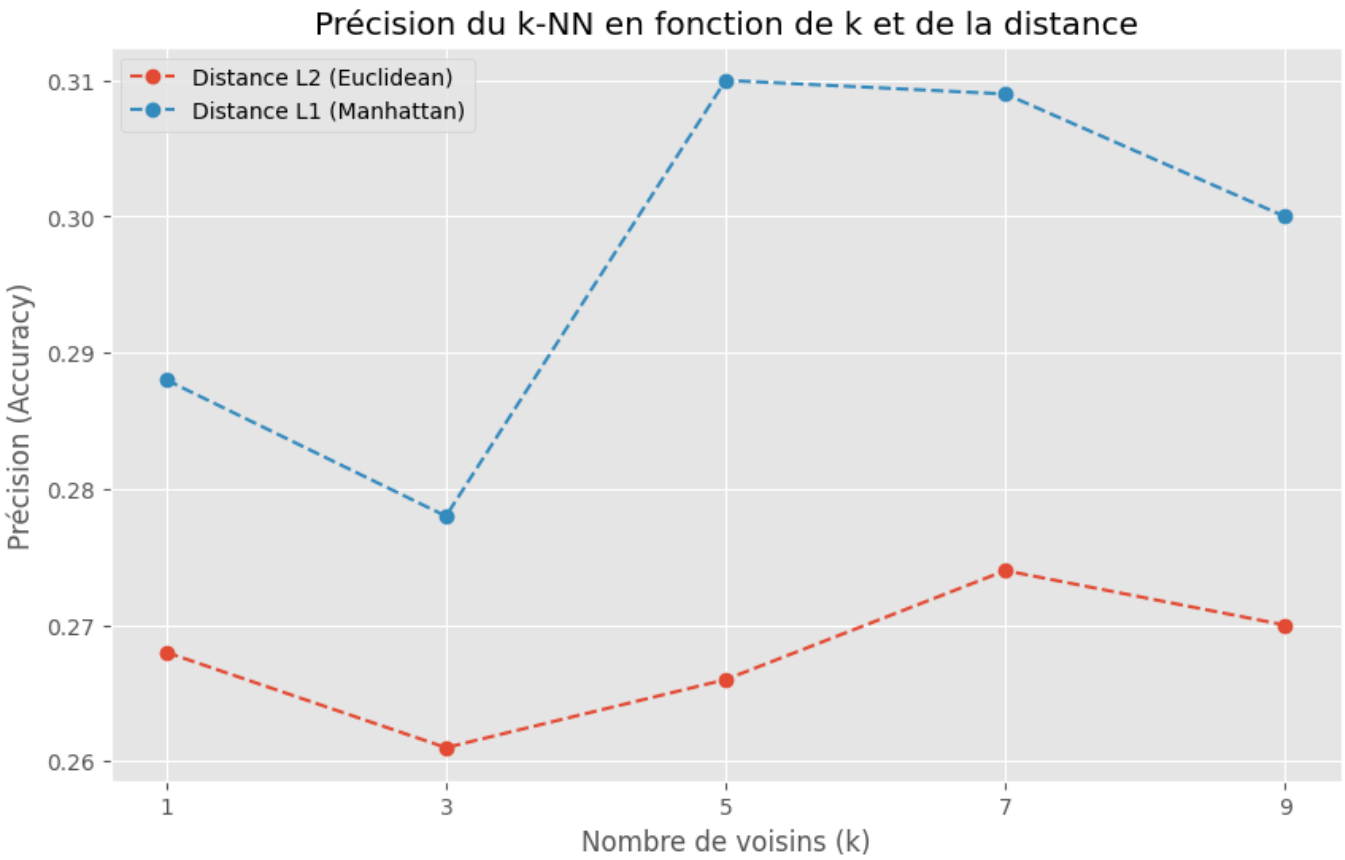
3.1. Impact de la Distance et du paramètre k

Le tableau ci-dessous résume les meilleures précisions obtenues pour chaque métrique de distance sans réduction de dimension :

Métrique de Distance	Meilleur k	Précision (Accuracy)
L2 (Euclidienne)	7	27.40%
L1 (Manhattan)	5	**31.00%**

Analyse :

- La **distance L1 (Manhattan)** a surpassé la distance L2 (Euclidienne) dans ce contexte, indiquant que la différence absolue entre les coordonnées des pixels est une meilleure mesure de similarité pour ces données.
- La précision globale reste faible (autour de 30%), ce qui est typique pour le k-NN sur des données d'images brutes et de haute dimension comme CIFAR-10.



3.2. Impact de la Réduction de Dimension (PCA)

L'application de la PCA a eu un impact significatif sur les performances du k-NN. En utilisant la distance L2 et $k = 7$ sur les données réduites à 100 dimensions :

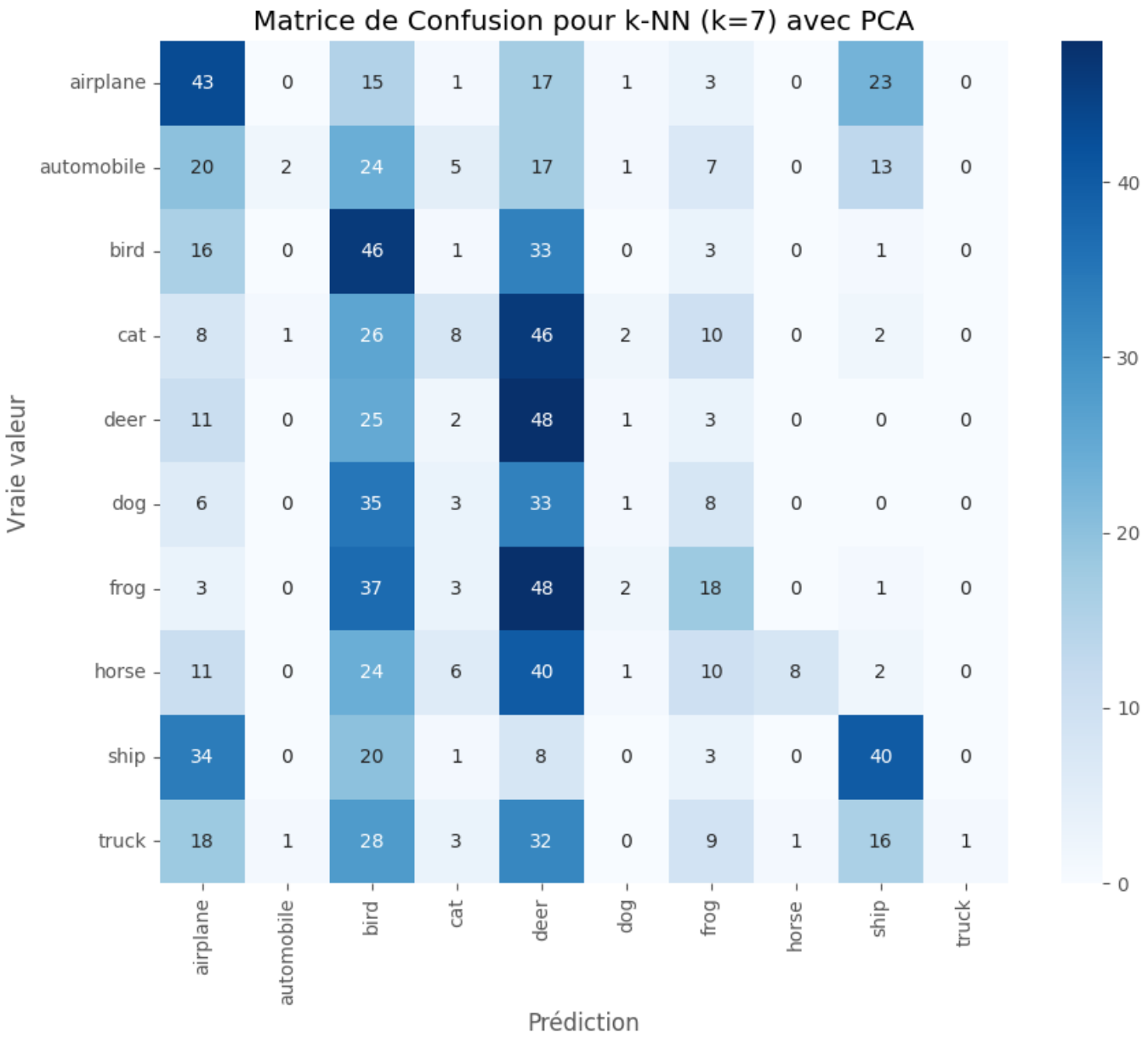
- **Précision (avec PCA) :** 21.50%
- **Temps de prédiction :** 0.09s

Analyse :

- **Précision :** Contrairement à l'attendu, la PCA a **diminué la précision** du k-NN (passant de 27.40% à 21.50% pour L2). Cela suggère que les 100 composantes principales n'ont pas réussi à capturer l'information discriminante essentielle pour la classification, ou que la perte d'information due à la réduction a été préjudiciable.
- **Vitesse :** La PCA a permis une **accélération massive** du temps de prédiction (0.09s contre 0.68s sans PCA), confirmant son rôle dans l'amélioration de l'efficacité computationnelle.

3.3. Matrice de Confusion

La Figure suivante présente la matrice de confusion pour le k-NN avec PCA ($k = 7$).



Analyse :

- La diagonale (prédictions correctes) est peu marquée, confirmant la faible précision.
- Les erreurs sont nombreuses, notamment la confusion entre les classes visuellement similaires comme **'cat' (chat)** et **'dog' (chien)**, ou **'deer' (cerf)** et **'horse' (cheval)**.

3.4. Exemples de Classification Erronée

La Figure 3 montre 10 exemples d'images mal classées par le k-NN avec PCA.

Figure 3 : Exemples d'images mal classées

Analyse :

- Ces exemples illustrent la difficulté de l'algorithme à distinguer des objets dans des images complexes, souvent en raison de l'arrière-plan, de la variation de pose ou de la faible résolution.

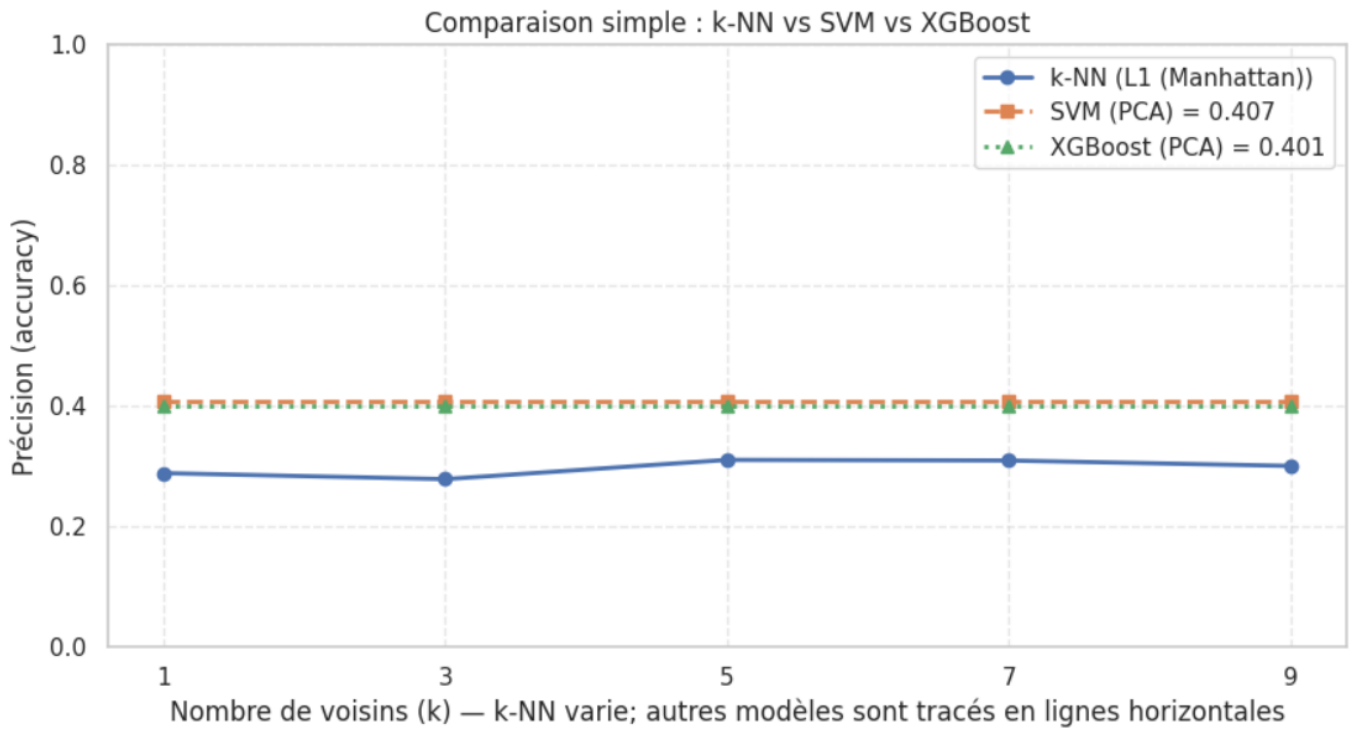
3.5. Comparaison avec d'autres Modèles

Le tableau récapitulatif ci-dessous compare les performances des trois modèles sur les données réduites par PCA :

Modèle	Précision (Accuracy)	Temps de Prédiction (s)
k-NN (PCA)	21.50%	**0.09**
SVM (PCA)	**40.10%**	4.73
XGBoost (PCA)	40.70%	13.49

Analyse :

- Le **SVM** a obtenu la meilleure précision (40.10%), démontrant une capacité supérieure à trouver une frontière de séparation dans l'espace de caractéristiques réduit par PCA.
- Le **k-NN** est de loin le plus rapide en prédiction, mais au prix d'une précision très faible.
- Le **XGBoost** offre une précision comparable au SVM, mais avec un temps de prédiction beaucoup plus long.



4. Conclusion

Ce mini-projet a permis de mettre en évidence les limites de l'algorithme k-NN sur des tâches de classification d'images complexes comme CIFAR-10, même après une tentative d'optimisation par PCA.

- **Le choix de la distance L1 (Manhattan) s'est avéré plus efficace** que la distance L2 (Euclidienne) pour les données brutes.
- **La PCA a échoué à améliorer la précision** du k-NN, mais a considérablement réduit le temps de calcul.
- **Le SVM a été le modèle le plus performant** en termes de précision, suggérant que des méthodes basées sur la recherche d'hyperplans (SVM) ou sur des ensembles d'arbres de décision (XGBoost) sont plus adaptées à ce type de problème que les méthodes basées sur la distance (k-NN).

Pour améliorer les résultats futurs, il serait recommandé d'explorer des techniques de *feature engineering* plus avancées ou d'utiliser des modèles d'apprentissage profond (CNN) spécifiquement conçus pour les données d'images.