Analyse du dataset MNIST

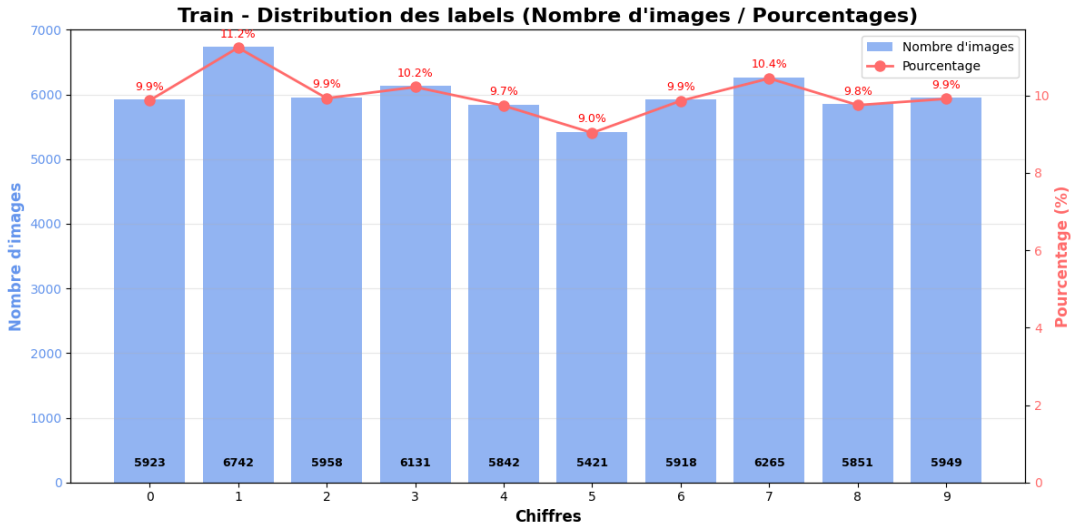
Ce document présente une analyse exploratoire du dataset MNIST, qui contient des images de chiffres manuscrits. L'objectif est de comprendre la distribution des données, les caractéristiques des images, et d'identifier des tendances ou des anomalies potentielles.

Statistiques de Base

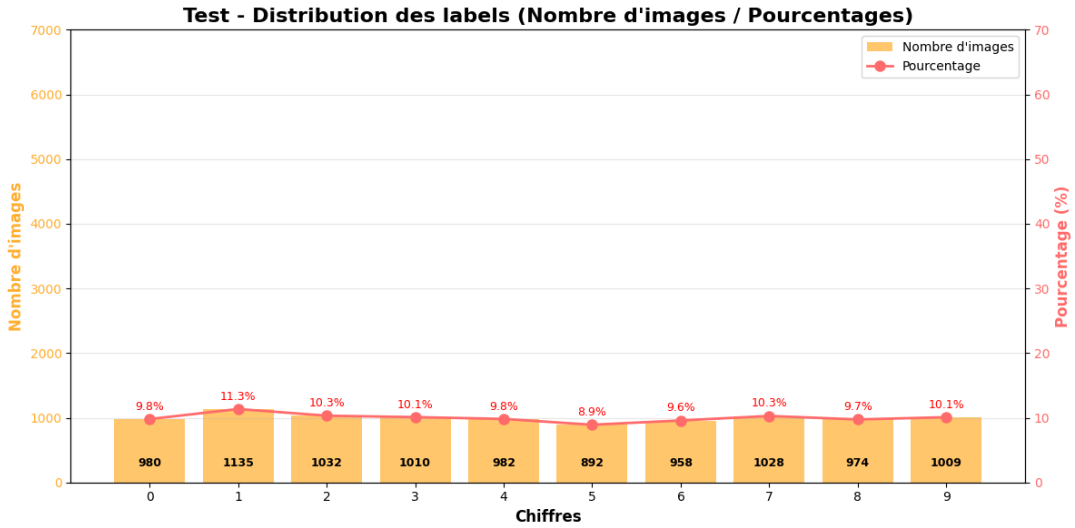
Voici quelques statistiques de base sur le dataset d'entraînement :

- Nombre total d'images train : 60 000
- Nombre total d'images test : 10 000
- Nombre de classes (chiffres) : 10 (0 à 9)
- Dimensions des images : 28x28 pixels (784 pixels aplatis)
- Valeur comprise entre 0 (noir) et 255 (blanc)

Distribution des labels (Train) :



Distribution des labels (Test) :

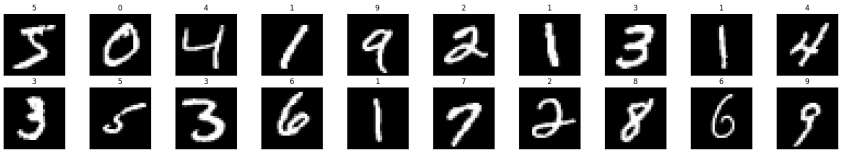


Normalisation des données

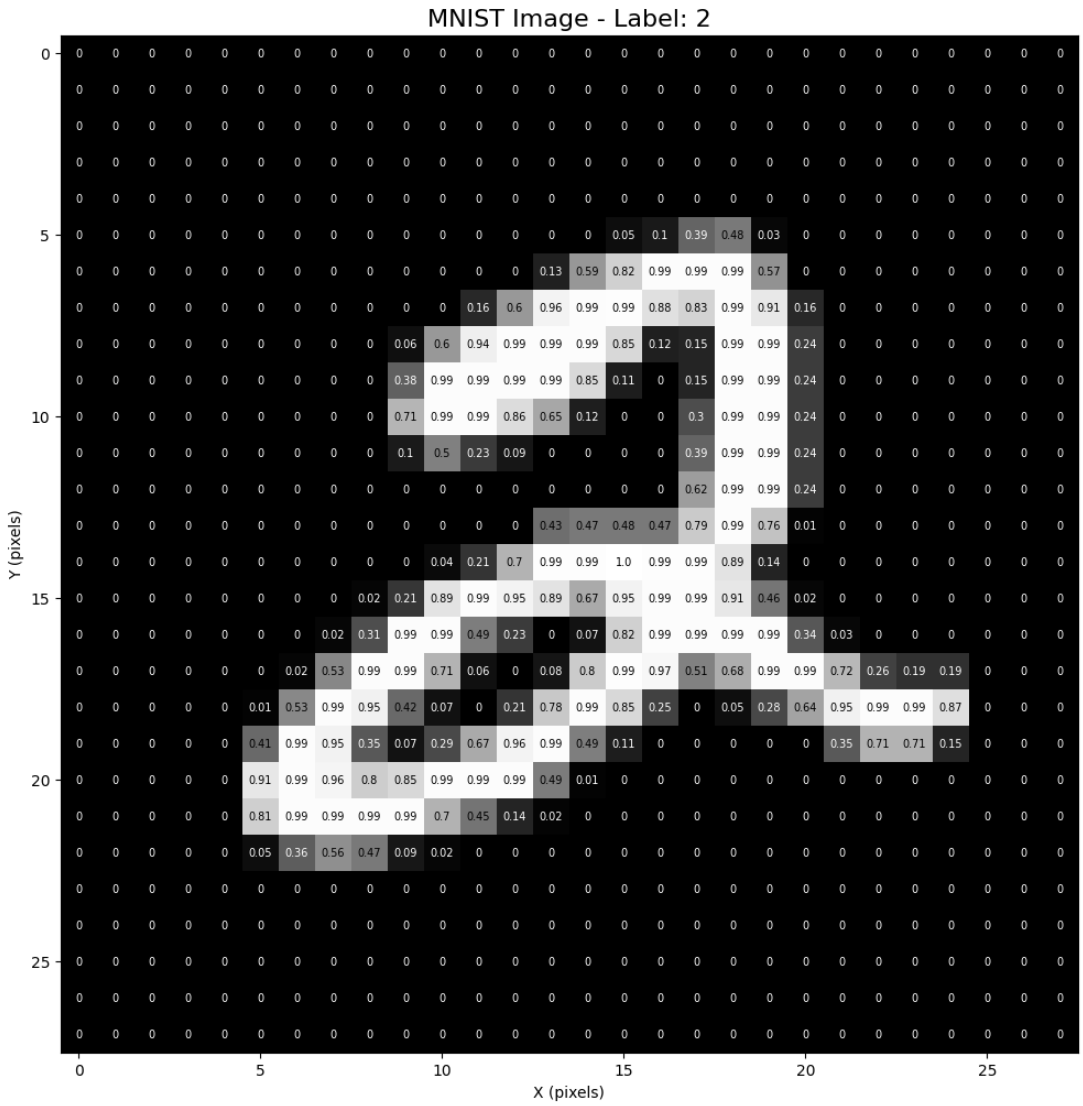
Nous avons normalisé les valeurs des pixels pour qu'elles soient comprises entre 0 et 1, ce qui est une pratique courante en apprentissage automatique pour améliorer la convergence des modèles. Les fichiers `mnist_train.csv` et `mnist_test.csv` contiennent les valeurs normalisées.

Représentation des images

Chaque image est représentée par une matrice de 28x28 pixels. Voici quelques exemples d'images du dataset :



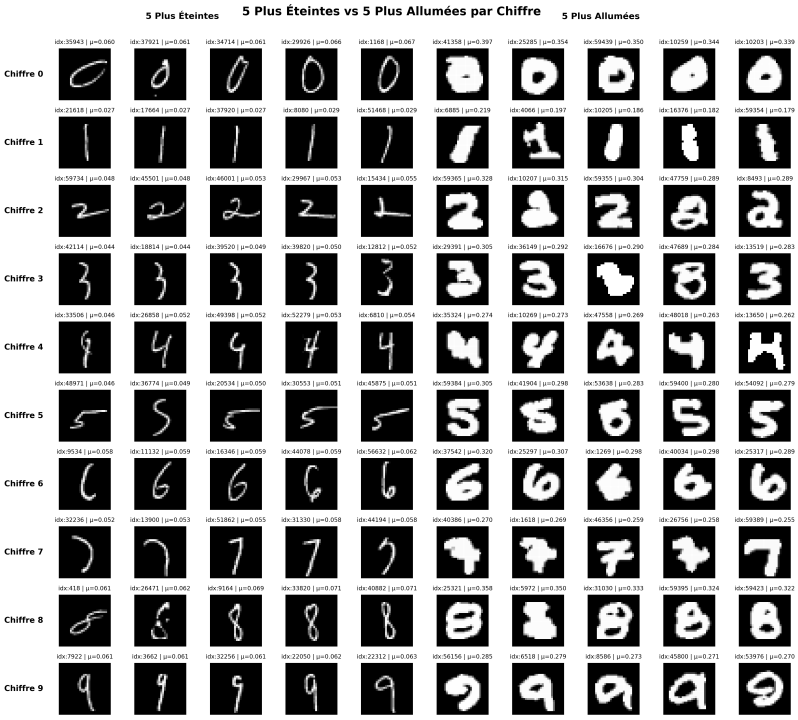
Pour comprendre simple les images, chaque pixel a une intensité qui contribue à la formation du chiffre manuscrit :



Outliers et Anomalies

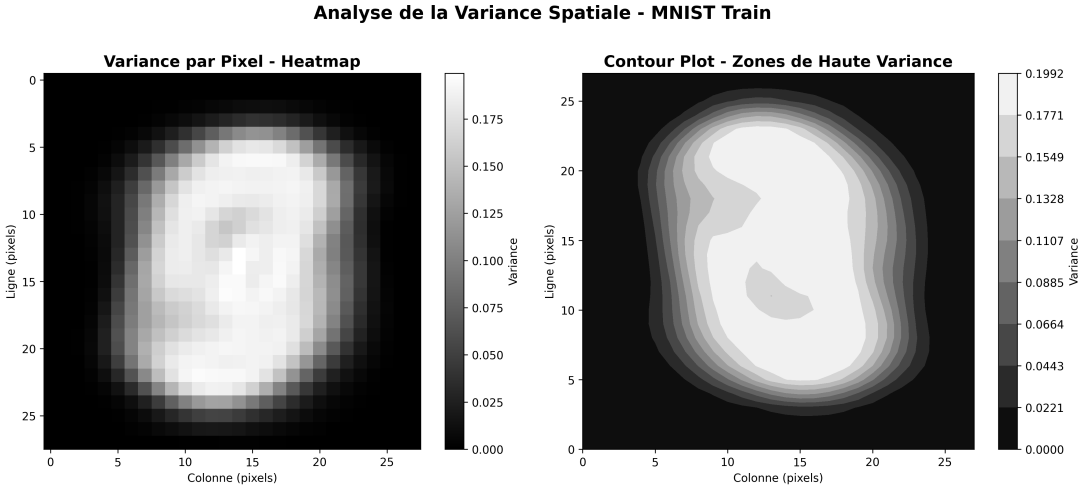
Après une analyse visuelle et statistique, nous avons identifié quelques images qui pourraient être considérées comme des outliers ou des anomalies, telles que des chiffres mal écrits ou des images floues. Cependant, ces cas sont rares et ne devraient pas affecter significativement la performance du modèle.

Voici un exemple d'image atypique :



Analyse de la variance par pixel

Nous avons analysé la variance des pixels à travers toutes les images pour identifier les zones les plus informatives. Les pixels avec une variance élevée sont ceux qui changent le plus entre les différentes images, ce qui peut indiquer des caractéristiques importantes pour la classification. Voici une visualisation de la variance des pixels :



Voici le top 10 des pixels avec la variance la plus élevée :

Position (x, y)	Variance
Pixel (13, 14)	0.199210

Position (x, y)	Variance
Pixel (14, 14)	0.198910
Pixel (16, 13)	0.196118
Pixel (22, 11)	0.195872
Pixel (16, 14)	0.195137
Pixel (15, 14)	0.195039
Pixel (15, 17)	0.195031
Pixel (15, 13)	0.195012
Pixel (22, 12)	0.194466
Pixel (14, 17)	0.194218

Analyse de la variance par chiffres

Nous avons également analysé la variance des pixels pour chaque chiffre individuellement. Cela nous permet de comprendre quelles parties de l'image sont les plus importantes pour différencier chaque chiffre. Voici quelques exemples de visualisations de la variance par chiffre :