Rapport d’Analyse : Détection de Pneumonies via PCA et Régression Logistique

1. Introduction

Ce projet vise à développer un modèle de classification d’images radiographiques pour distinguer trois classes : **Normal**, **Pneumonie Bactérienne** et **Pneumonie Virale**. Contrairement aux approches basées sur les réseaux convolutifs (CNN), nous utilisons ici une méthode classique combinant la réduction de dimension par **Analyse en Composantes Principales (PCA)** et un modèle de **Régression Logistique**. L’objectif est d’évaluer l’impact de plusieurs paramètres (taille des images, nombre de composantes PCA, réglages du modèle) sur la précision globale.

---

2. Description du Dataset

Le dataset est composé d’images radiographiques annotées, réparties en trois catégories : normal, pneumonie bactérienne, pneumonie virale. Les images sont

initialement en niveaux de gris et ont été redimensionnées pour uniformiser les entrées. Le découpage des données suit un schéma stratifié 80% entraînement et 20% test pour assurer la représentativité de chaque classe.

---

3. Prétraitement des Images

Les images ont subi les étapes suivantes :

- Conversion en niveaux de gris et aplatissement en vecteurs
- Redimensionnement initial à 400x400 pixels (baseline), puis expérimentations avec tailles plus petites (200x200, 128x128, 100x100)
- Normalisation des pixels entre 0 et 1 pour faciliter l'entraînement et la convergence des modèles

---

4. Réduction de Dimension avec PCA

La PCA est appliquée pour réduire la dimensionnalité des vecteurs d'images tout en conservant la majorité de la variance.Différentes configurations ont été testées :

- Nombre de composantes défini par la variance expliquée (`n_components=0.95`, `0.90`, `0.99`)
- Nombre fixe de composantes (100, 300)

L'objectif est de trouver un compromis entre richesse des données conservées et complexité du modèle.

---

5. Modélisation par Régression Logistique

Le classifieur utilisé est une régression logistique multiclasse, entraînée sur les données projetées par la PCA.

Plusieurs variantes ont été évaluées pour optimiser les performances :

- Nombre maximal d’itérations (`max_iter`)
- Type de pénalité (L1 ou L2)
- Méthode de résolution (`solver`)
- Force de régularisation (`C`)

---

6. Expérimentations et Résultats

6.1 Impact de la taille des images

ID	Taille Image	Description	Accuracy
Baseline	400x400	Réglage de base	82%
V1	200x200	Réduction de la taille, plus rapide	82%
V2	128x128	Taille intermédiaire	80%
V3	100x100	Compression agressive	85%

Observation : La réduction agressive à 100x100 améliore légèrement la précision, probablement par effet de régularisation ou réduction du bruit.

6.2 Influence du nombre de composantes PCA (avec image_size=100x100)

ID	n_components	Description	Accuracy
P0	0.95 (variance)	Baseline	85%
P1	0.90 (moins de comp.)	Moins de composantes, plus rapide	86%
P2	0.99 (plus riche)	Conserve davantage d’information	86%
P3	100 (fixe)	Nombre fixe de composantes	84%
P4	300 (très riche)	Risque de bruit ou surapprentissage	86%

6.3 Réglages du modèle de régression logistique

ID	Paramètres	Description	Accuracy
M0	max_iter=1000	Baseline	86%
M1	max_iter=2000	Plus d’itérations	84%
M2	solver='saga'	Optimisé pour grands jeux de données	85%
M3	penalty='l1', solver='saga'	Régularisation Lasso favorisant la parcimonie	86%
M4	C=0.1	Régularisation forte, modèle plus simple	78%
M5	C=10.0	Faible régularisation, modèle plus flexible	84%

7. Synthèse des Meilleures Configurations

Test ID	Accuracy	Commentaires
V3	85%	Taille image réduite améliore la précision
P1, P2, P4	86%	PCA avec 90%-99% variance conservée optimal
M0, M3	86%	Régression avec L1 et solver 'saga' performant

8. Conclusion

L’utilisation combinée de la réduction de dimension par PCA et d’un modèle de régression logistique permet d’atteindre une précision satisfaisante (~86%) pour la classification de radiographies en trois classes. Les résultats suggèrent qu’une réduction modérée de la taille des images ainsi qu’un choix judicieux du nombre de

composantes PCA améliorent les performances. La régularisation L1 avec solver ‘saga’ aide à obtenir un modèle plus parcimonieux sans perte de précision notable.

Pour aller plus loin, l’intégration de techniques de Deep Learning, notamment les CNN, serait la voie privilégiée pour exploiter pleinement la nature visuelle des images médicales.

---

9. Exemple de Code

```
from sklearn.decomposition import PCA
from sklearn.linear_model import LogisticRegression
from sklearn.pipeline import Pipeline
from sklearn.metrics import accuracy_score

pipeline = Pipeline([
    ('pca', PCA(n_components=0.99)),
    ('clf', LogisticRegression(
        solver='saga',
        penalty='l1',
        C=0.5,
        max_iter=1000
    ))
])

pipeline.fit(X_train, y_train)
y_pred = pipeline.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
```