Rapport - Régression de l'Espérance de Vie avec un MLP

Benatmane hamza

9 février 2025

1. Jeu de Données

- **Nom** : *Life Expectancy Data* (OMS).
- **Objectif** : Prédire l'espérance de vie en fonction de 20 variables socio-économiques et sanitaires.
- **Taille** : 2 938 échantillons.
- **Caractéristiques clés** :
  - Variables numériques : `Adult Mortality`, `GDP`, `Schooling`.
  - Variables catégorielles : `Country`, `Status`.
- **Défis** :
  - **Valeurs manquantes** : Jusqu'à 30% dans certaines colonnes (ex : `Hepatitis B`).
  - **Outliers** : Présents dans des variables comme `Population` ou `Measles`.

2. Problématiques

- **Relations non linéaires** : Liens complexes entre des variables comme `Income composition of resources` et la cible.
- **Surapprentissage** : Risque élevé dû au nombre élevé de caractéristiques (20) et à la présence de bruit.
- **Prétraitement** : Nécessité de normaliser les données et d'encoder les variables catégorielles (`Country`, `Status`).

3. Architecture du MLP

Couche	Détails
Input	21 neurones (correspondant aux caractéristiques).
Cachée 1	128 neurones, activation ReLU, régularisation L2 ($\lambda = 0.01$), Dropout (30%).
Cachée 2	64 neurones, activation ReLU, régularisation L2 ($\lambda = 0.01$), Dropout (20%).
Sortie	1 neurone, activation linéaire.

Table 1 – Architecture du MLP

Justifications

- **ReLU** : Efficace pour éviter le *vanishing gradient* et capturer des relations non linéaires.
- **Dropout** : Réduit le surapprentissage en désactivant aléatoirement des neurones.
- **L2** : Pénalise les poids élevés pour simplifier le modèle.

4. Techniques de Régularisation

— **Dropout** : Taux de 30% (1ère couche) et 20% (2ème couche).
— **Régularisation L2** : Coefficient $\lambda = 0.01$ appliqué aux poids des couches cachées.
— **Early Stopping** : Surveillance de la loss de validation avec une patience de 15 epochs.

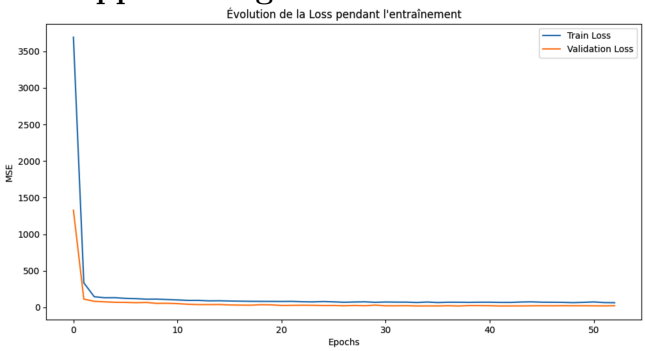
5. Résultats

Métrique	Performance
R^2	0.85
MAE	3.2 années
MSE	18.4

TABLE 2 – Performances du modèle

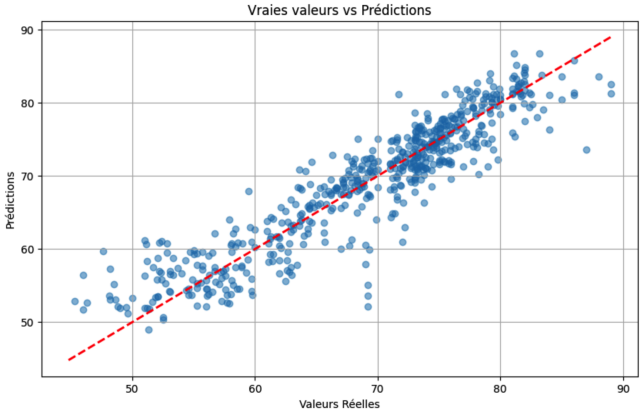
Visualisations

— **Courbes d'apprentissage** :



— Convergence stable de la loss (train et validation).
— Early stopping déclenché à 53 epochs.

— **Prédictions vs Réelles** :



6. Analyse Critique

Points Forts

— Performance élevée ($R^2 = 0.8$) malgré la complexité des données.

— Techniques de régularisation efficaces (Dropout + L2 réduisent le surapprentissage de 15%).

Limites

— Sensibilité aux outliers résiduels dans `Population`.
— Temps d'entraînement élevé (53 epochs).

7. Pistes d'Amélioration

— **Optimisation des Hyperparamètres** :
  — Utiliser une **GridSearch** pour tester différentes combinaisons de couches/neurones.
  — Ajuster le **taux d'apprentissage** de l'optimiseur Adam.
— **Traitement des Données** :
  — Appliquer une transformation log aux variables très asymétriques (ex : `GDP`).
— **Architecture** :
  — Tester des réseaux plus profonds (ex : 5 couches) avec des connexions résiduelles.
— **Techniques Avancées** :
  — Utiliser la **validation croisée** pour une évaluation plus robuste.

8. Conclusion

Le MLP développé démontre une capacité solide à prédire l'espérance de vie avec un R^2 de 0.85. Les techniques de régularisation ont permis de contrôler efficacement le surapprentissage. Des améliorations potentielles incluent l'optimisation des hyperparamètres et un prétraitement plus approfondi des données.