Mise en œuvre d'un modèle de Machine Learning avec TensorFlow/Scikit-learn

Rapport de Projet

TIGRINE Yacine
Master 2 I2A

December 9, 2024

Ce document est soumis dans le cadre du module outils de programmation avancée pour l'IA.

Contents

1 Introduction

Le présent projet vise à concevoir et développer un modèle de Machine Learning en utilisant **TensorFlow** / **Scikit-learn**. L'objectif principal est de couvrir les étapes clés du cycle de développement d'un projet de Machine Learning, incluant la préparation des données, l'entraînement, l'évaluation et l'analyse comparative des modèles.

2 Dataset et Méthodologie

2.1 Dataset

Nom : Breast Cancer Wisconsin

Source : `https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset`

Description : Ce dataset de santé contient des profils de patients permettant de prédire l'apparition de maladies cardiaques. Les données incluent des variables comme l'âge, le cholestérol et la fréquence cardiaque.

2.2 Méthodologie

La santé est très importante pour chacun. Bien que les maladies ne puissent être prédites avant leur apparition, l'utilisation de modèles de Machine Learning pourrait permettre de mettre en place des protocoles pour modifier le quotidien et suivre les patients, évitant ainsi l'apparition des maladies. Comme le dit le proverbe : *« Mieux vaut prévenir que guérir »*. Il est donc essentiel de prendre des mesures préventives pour éviter des problèmes futurs.

Après vérification, le dataset ne contient pas de valeurs manquantes. Cette complétude des données est un atout, surtout dans le domaine médical, où la précision est essentielle pour garantir des analyses fiables et éviter des biais dus à des méthodes d'imputation.

Nous utiliserons plusieurs modèles que nous comparerons pour déterminer lequel est le plus performant pour capturer les relations entre les variables explicatives (x) et la cible (y) afin d'effectuer des prédictions.

3 Modèles et Résultats

3.1 Description des Modèles

- **Réseau de neurones :** Ce choix est motivé par sa capacité à capturer des relations complexes et sa robustesse pour la classification binaire. Il est également flexible grâce à la possibilité d'ajuster son architecture (nombre de couches, fonctions d'activation, etc.).
- **Régression logistique :** Modèle simple et efficace pour la classification binaire.
- **Random Forest :** Modèle robuste et stable, capable de capturer des relations complexes et facile à optimiser.
- **SVM :** Utilisé pour comparer ses performances, notamment grâce à ses capacités de séparation avec noyaux.

3.2 Hyperparamètres du modèle

- **Adam :** Algorithme d'optimisation efficace, capable de converger rapidement.
- **Binary crossentropy :** Fonction de perte adaptée aux tâches de classification binaire.
- **Accuracy :** Métrique simple et couramment utilisée pour évaluer la performance des modèles.
- **ReLU :** Fonction d'activation rapide, qui accélère la convergence en évitant la saturation.

3.3 Analyse des Résultats (Mode normal)

Modèle	Accuracy	F1 Score	Recall
RN léger	0.9659	0.9655	0.9515
RN complexe	0.9756	0.9751	0.9515
Régression Logistique	0.7951	0.8125	0.8835
Random Forest	0.9756	0.9756	0.9709
SVM	0.8780	0.8837	0.9223

Table 1: Résultats des modèles testés (Mode normal)

3.4 Analyse des Résultats (Mode redondant)

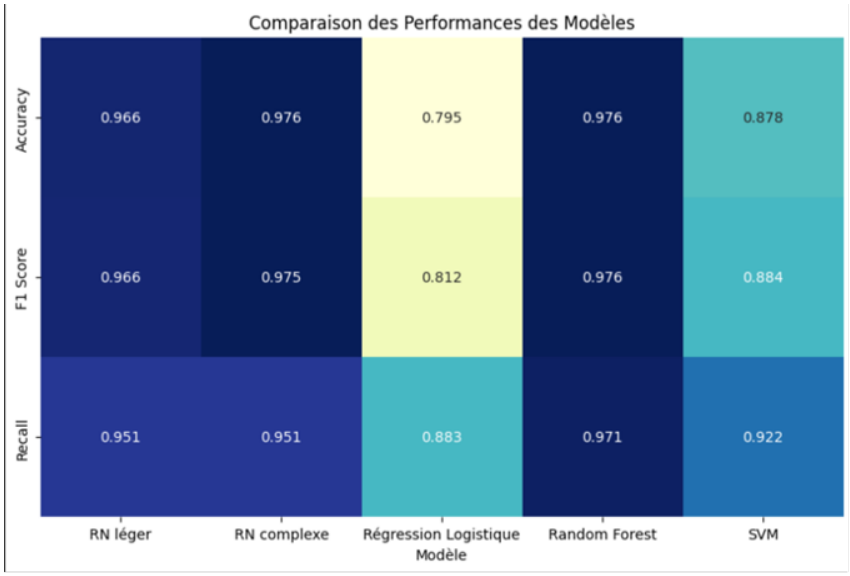
Modèle	Accuracy	F1 Score	Recall
RN léger	0.9659	0.9655	0.9515
RN complexe	0.9756	0.9751	0.9515
Régression Logistique	0.7951	0.8125	0.8835
Random Forest	0.9756	0.9756	0.9709
SVM	0.8780	0.8837	0.9223

Table 2: Résultats des modèles testés (Mode redondant)

Analyse : Random Forest se distingue avec le meilleur *Recall*, indiquant sa capacité à minimiser les faux négatifs, essentiel dans un contexte médical. Bien que les réseaux de neurones complexes soient proches en performance, ils demandent plus de ressources et de temps d'entraînement.

4 Discussion et Améliorations

- Exploration d'autres architectures de réseaux de neurones, avec davantage de couches cachées ou des tailles de couches différentes.
- Utilisation de méthodes de régularisation comme le *dropout* pour éviter le surapprentissage.



5 Conclusion

Les performances des modèles testés indiquent que **Random Forest** est le plus adapté pour cette tâche, offrant un compromis optimal entre précision, rappel et robustesse. Les réseaux de neurones complexes montrent également un fort potentiel, mais avec un coût computationnel plus élevé.

6 Références

- Heart Disease Dataset : `https://www.kaggle.com/datasets/johnsmith88/heart-disease-dat`