Rapport de Dépistage du Trouble du Spectre Autistique

Analyse Comparative de Modèles d'Apprentissage Supervisé

Auteurs :

KSIKSI NEDIA - FADWA BOUABID - LEILA MISSAOUI - SARRA HMERCHA

Date : 19 décembre 2024

Table des matières

INTRODUCTION

Le trouble du spectre autistique (TSA) est une condition neurodéveloppementale associée à des coûts de santé importants, et un diagnostic précoce peut réduire ces coûts de manière significative. Malheureusement, les délais d'attente pour un diagnostic de TSA sont longs et les procédures ne sont pas rentables. L'impact économique de l'autisme et l'augmentation du nombre de cas de TSA à travers le monde révèlent un besoin urgent de développer des méthodes de dépistage accessibles et efficaces. Par conséquent, un dépistage du TSA rapide et accessible est nécessaire pour aider les professionnels de la santé et informer les individus sur la nécessité de poursuivre un diagnostic clinique formel.

La croissance rapide du nombre de cas de TSA dans le monde nécessite des jeux de données liés aux traits comportementaux. Cependant, de tels jeux de données sont rares, ce qui rend difficile la réalisation d'analyses approfondies pour améliorer l'efficacité, la sensibilité, la spécificité et la précision prédictive du processus de dépistage du TSA. Actuellement, très peu de jeux de données sur l'autisme associés à des diagnostics cliniques ou de dépistage sont disponibles et la plupart d'entre eux sont d'ordre génétique. Ainsi, nous proposons un nouveau jeu de données relatif au dépistage de l'autisme chez les adultes, qui contient 20 caractéristiques à utiliser pour des analyses ultérieures, notamment pour déterminer les traits autistiques influents et améliorer la classification des cas de TSA. Ce jeu de données inclut dix caractéristiques comportementales (AQ-10-Adulte) ainsi que dix caractéristiques individuelles qui se sont avérées efficaces pour détecter les cas de TSA parmi les témoins dans les sciences du comportement.

Dans cette étude, nous appliquerons plusieurs algorithmes d'apprentissage supervisé pour améliorer la classification des cas de TSA. Les méthodes utilisées incluront les k plus proches voisins (KNN), la machine à vecteurs de support (SVM), Naïve Bayes, les arbres de décision, la forêt aléatoire (Random Forest) ainsi que la régression logistique. Nous comparerons ces modèles en termes de précision, de F1-Score et de rappel. L'objectif de cette comparaison est d'identifier le modèle offrant les meilleures performances pour le dépistage du TSA, facilitant ainsi la mise en œuvre de solutions plus rapides et efficaces pour les professionnels de santé.

I Régression logistique

I.1 Définition

La régression logistique est un modèle de classification binaire qui prédit la probabilité d'appartenance à une classe (0 ou 1) en utilisant la fonction sigmoïde. Elle ajuste ses paramètres pour minimiser la log-vraisemblance et est efficace pour les problèmes de classification discrète.

I.2 Analyse des résultats

```
[[89  2]
 [ 5 45]]
```

FIGURE 1 – Matrice de confusion

```
Accuracy : 0.950354609929078

Rapport de classification :
              precision    recall  f1-score   support

           0       0.95      0.98      0.96        91
           1       0.96      0.90      0.93        50

    accuracy                           0.95       141
   macro avg       0.95      0.94      0.94       141
weighted avg       0.95      0.95      0.95       141
```

FIGURE 2 – Rapport de classification

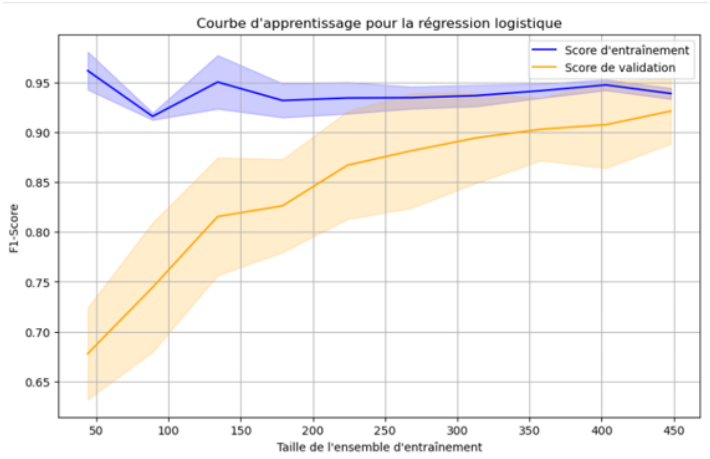

FIGURE 3 – Courbe d'apprentissage

I.3 Conclusion

Le modèle de régression logistique atteint 95 % de précision avec des scores équilibrés pour les deux classes. Bien que plus de données puissent améliorer les performances, il est déjà bien adapté à la tâche.

II MODELE KNN

II.1 Définition

Le modèle KNN (K-Nearest Neighbors) est un algorithme d'apprentissage supervisé qui classe un échantillon en fonction des étiquettes de ses K voisins les plus proches dans l'espace des caractéristiques.

II.2 Analyse des résultats

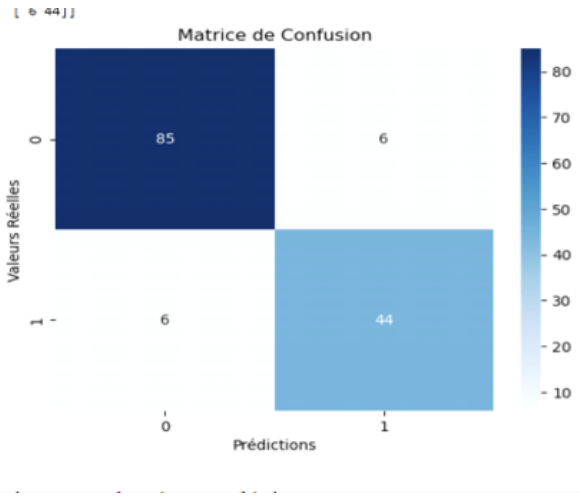

FIGURE 4 – Matrice de confusion

	precision	recall	f1-score	support
0	0.93	0.93	0.93	91
1	0.88	0.88	0.88	50
accuracy			0.91	141
macro avg	0.91	0.91	0.91	141
weighted avg	0.91	0.91	0.91	141

Accuracy : 0.9148936170212766

FIGURE 5 – Rapport de classification

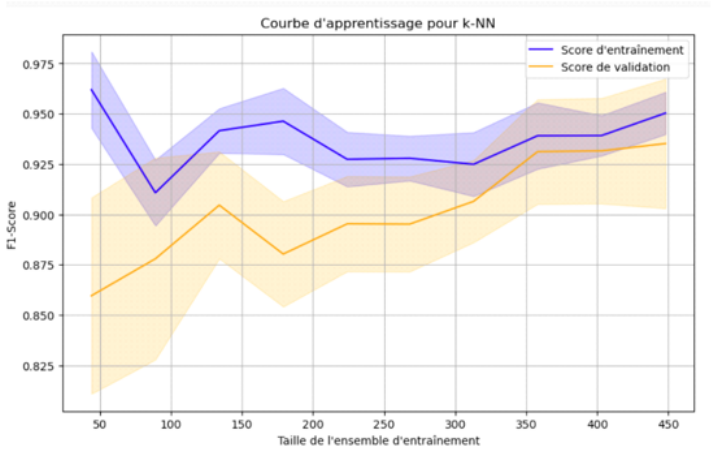

FIGURE 6 – Courbe d'apprentissage

II.3 Conclusion

Le modèle KNN avec des paramètres proposés par GridSearch `{'metric': 'manhattan', 'n_neighbors': 5}` est très performant pour la tâche de classification et fournit des résultats solides, avec un bon compromis entre précision, rappel et F1-score.

III MODeLE SVM

III.1 Définition

L'algorithme SVM (Support Vector Machine) est un modèle d'apprentissage supervisé utilisé pour la classification et la régression. Il cherche à trouver l'hyperplan qui sépare au mieux les classes de données tout en maximisant la marge entre les points les plus proches des différentes classes.

III.2 Analyse des résultats

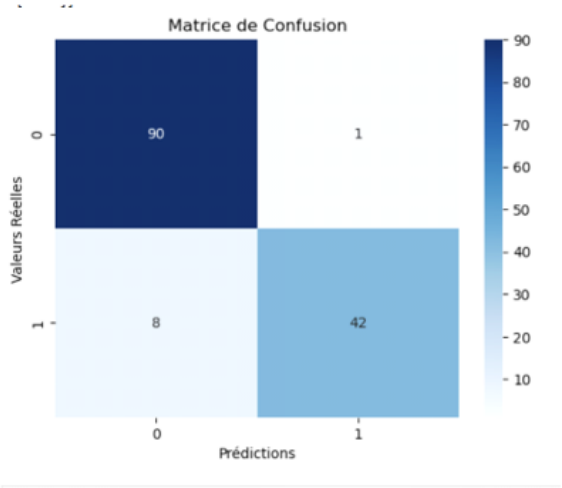

FIGURE 7 – Matrice de confusion

	precision	recall	f1-score	support
0	0.93	0.93	0.93	91
1	0.88	0.88	0.88	50
accuracy			0.91	141
macro avg	0.91	0.91	0.91	141
weighted avg	0.91	0.91	0.91	141

Accuracy : 0.9148936170212766

FIGURE 8 – Rapport de classification

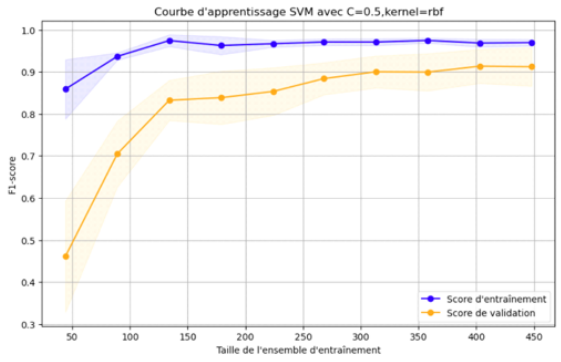

FIGURE 9 – Courbe d'apprentissage

III.3 Conclusion

En résumé, le modèle offre de bonnes performances. Cependant, les résultats sont légèrement meilleurs pour la classe 0 que pour la classe 1. Cela est dû au déséquilibre des deux classes.

IV Arbre de Décision (Decision Tree)

IV.1 Définition

Un arbre de décision est un modèle d'apprentissage supervisé pour la classification et la régression qui divise les données en sous-ensembles en fonction des conditions des caractéristiques jusqu'à ce qu'elles soient homogènes.

— **Nœuds internes** : Conditions basées sur une caractéristique.
— **Feuilles** : Nœuds finaux contenant les prédictions ou résultats.
— **Branches** : Chemins reliant les nœuds, représentant les décisions (vrai ou faux pour une condition).

IV.2 Analyse des résultats

Les meilleurs paramètres pour notre modèle, récupérés après l'exécution de la recherche en grille (GridSearch) sont (max depth : 7 ; min samples leaf : 1 min samples split : 5)

Les métriques obtenues sont les suivantes :

— **Accuracy (Précision globale)** : 0.82
— **Precision (Précision)** : 0.70
— **Recall (Rappel)** : 0.61
— **F1-score** : 0.65

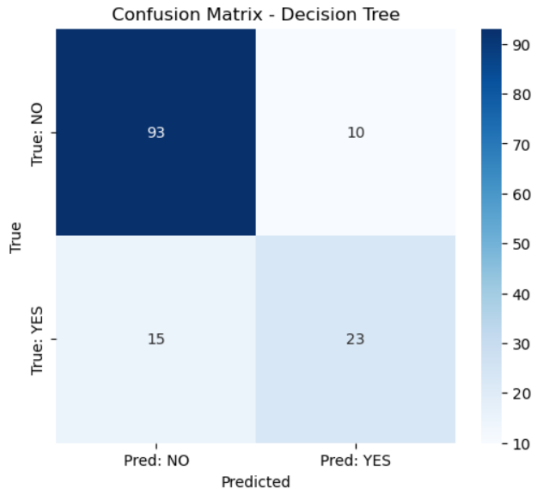

FIGURE 10 – Matrice de confusion

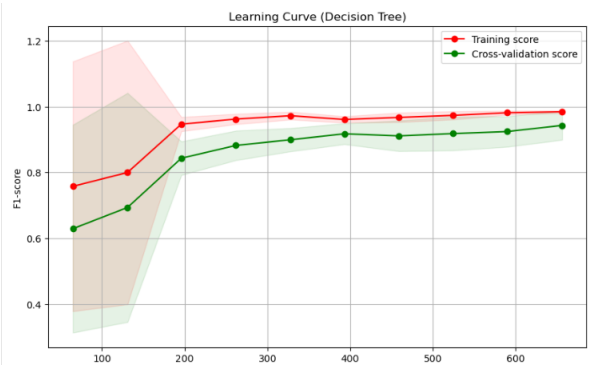

FIGURE 11 – Courbe d'apprentissage

IV.3 Conclusion

Globalement, le modèle d'arbre de décision fonctionne de manière raisonnable, avec une précision de 82%. Cependant, il présente un compromis notable entre la précision et le rappel.

V Modèle Random Forest

V.1 Définition

Le **Random Forest** est un ensemble d'arbres de décision où chaque arbre est construit sur un sous-ensemble aléatoire des données et des caractéristiques. Le modèle utilise un vote majoritaire des prédictions des arbres pour déterminer la prédiction finale.

— **Arbres** : Plusieurs arbres de décision sont construits en parallèle sur des sous-ensembles aléatoires des données et des caractéristiques.
— **Vote majoritaire** : La classe prédite par la majorité des arbres est considérée comme la prédiction finale.

V.2 Analyse des résultats

Les meilleurs paramètres pour notre modèle, obtenus après l'exécution de la **recherche en grille (GridSearch)**, sont (max depth : 10 ;max features : 'sqrt' ; min samples leaf : 1 ; min samples split : 5 ; n estimators : 50)

V.3 Performances du modèle

— **Accuracy (Précision globale)** : 0.82
— **Precision (Précision)** : 0.70
— **Recall (Rappel)** : 0.61
— **F1-score** : 0.65

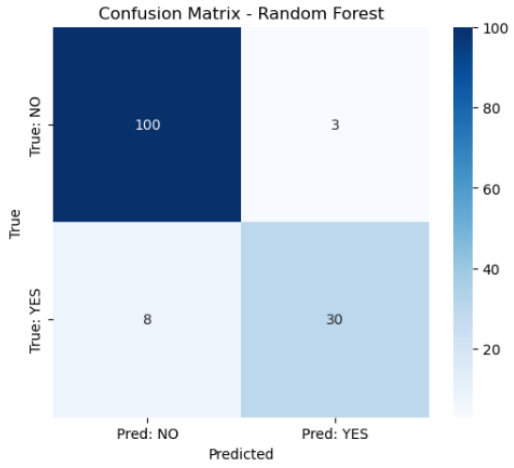

FIGURE 12 – Matrice de confusion

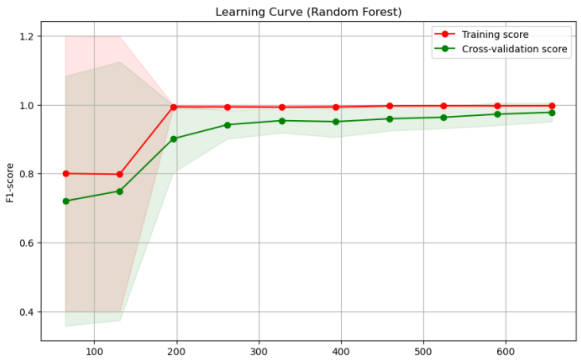

FIGURE 13 – Courbe d'apprentissage

V.4 Conclusion

Le modèle **Random Forest** semble très performant sur cette tâche, surtout en comparaison avec l'arbre de décision classique. Il offre un meilleur rappel ainsi qu'une précision élevée, ce qui démontre sa capacité à généraliser efficacement les données.

VI Modèle Naïve Bayes

VI.1 Définition

Le modèle Naïve Bayes est un algorithme de classification qui applique le théorème de Bayes pour calculer la probabilité d'un ensemble d'observations en utilisant les « features » afin de prédire les observations tout en supposant une indépendance conditionnelle. La formule du théorème de Bayes est la suivante :

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

VI.2 Types de Modèles Naïve Bayes

Il existe plusieurs types de modèles Naïve Bayes, parmi lesquels nous pouvons citer :

- **Naïve Bayes Gaussien** : Utilisé pour les données continues, il se base sur une distribution normale (gaussienne).
- **Naïve Bayes Multinomial** : Adapté aux données textuelles formées par les fréquences de mots.
- **Naïve Bayes avec lissage de Laplace (Additive Smoothing)** : Cette méthode ajoute une petite constante pour éviter l'apparition de zéros dans les probabilités conditionnelles.
- **Naïve Bayes de Bernoulli avec Smoothing** : Variante du Bernoulli Naïve Bayes qui applique des techniques de lissage pour réduire les erreurs dues aux valeurs manquantes.

VI.3 Analyse des résultats

L'outil **G**ridSearchCV : Analyse des meilleurs hyperparamètres principalement par «Var smoothing».

```
Accuracy: 0.9645390070921985
F1-score: 0.964616598018095
Recall: 0.9645390070921985
Précision sur l'ensemble de test: 0.9645390070921985
```

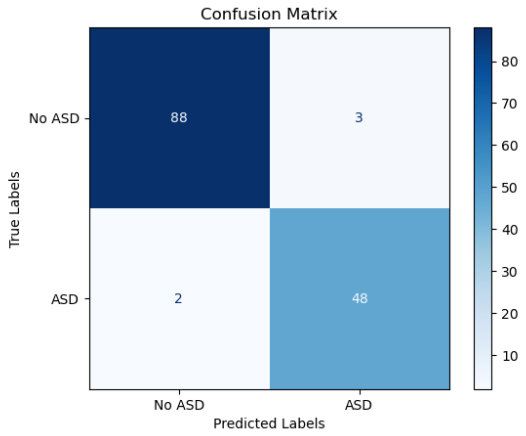

FIGURE 14 – Matrice de confusion.

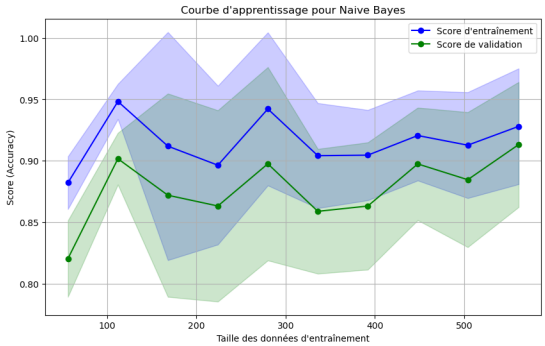

FIGURE 15 – Courbe d'apprentissage

VI.4 Conclusion

Les résultats montrent une performance élevée, ce qui indique que le modèle apprend bien et fait preuve de robustesse.

Conclusion générale

Après une analyse approfondie des six modèles de classification, il est clair que certains se démarquent par leur performance. Le modèle **Naive Bayes** s'impose comme le plus performant, affichant une excellente précision, un rappel élevé et un score F1 de 0.96. De plus, il présente une convergence parfaite des scores d'entraînement et de validation sans signe d'**overfitting**. La **régression logistique** se positionne

juste derrière, avec également une très bonne précision et un équilibre remarquable entre rappel et précision. Les modèles **k-NN** et **SVM** offrent des performances comparables, avec une précision et un rappel satisfaisants, ainsi qu'une convergence cohérente entre les scores d'entraînement et de validation. Le mo-

dèle **Random Forest**, bien que performant, présente un peu plus d'erreurs de classification comparé aux précédents modèles. Cependant, il conserve une bonne capacité de généralisation. En revanche, l'**arbre de décision** se révèle être le moins performant de tous les modèles, avec une précision et un rappel plus faibles, une faible capacité de généralisation, et des signes légers d'**overfitting**. En se basant sur les

métriques clés et les courbes d'apprentissage, le modèle **Naive Bayes** apparaît comme le choix optimal pour ce problème de classification. Toutefois, les performances de la **régression logistique** méritent également d'être prises en considération. Les modèles **k-NN** et **SVM** peuvent être utilisés si l'objectif est de simplifier le modèle par rapport à Naive Bayes, bien qu'ils soient légèrement moins performants. Le choix final du modèle dépendra des objectifs spécifiques du projet, ainsi que de la balance souhaitée

entre la **précision**, le **rappel**, et la **complexité** du modèle.