Master Bioinformatique

HAI817 – Machine learning 1 (méthodes classiques)

Professeurs : P. Poncelet, K. Todorov, E. Raoufi

Classification d'assertions venant d'X (Twitter) selon leur rapport à la science

---

Par : Tiziri Tamani (22415178), Ciarán Mahony (22400729), Raphaël Ribes (22401925), Dalia Belmadi (22208849)

Lien vers le projet : `https://github.com/RaphaelRibes/machine-learning/`

Table des matières

Résumé

1- Résultats

A) Pré-traitement

On regarde le nombre de `#`, `@` et lien par tweet selon le type de tweet scientifique (figure 1).

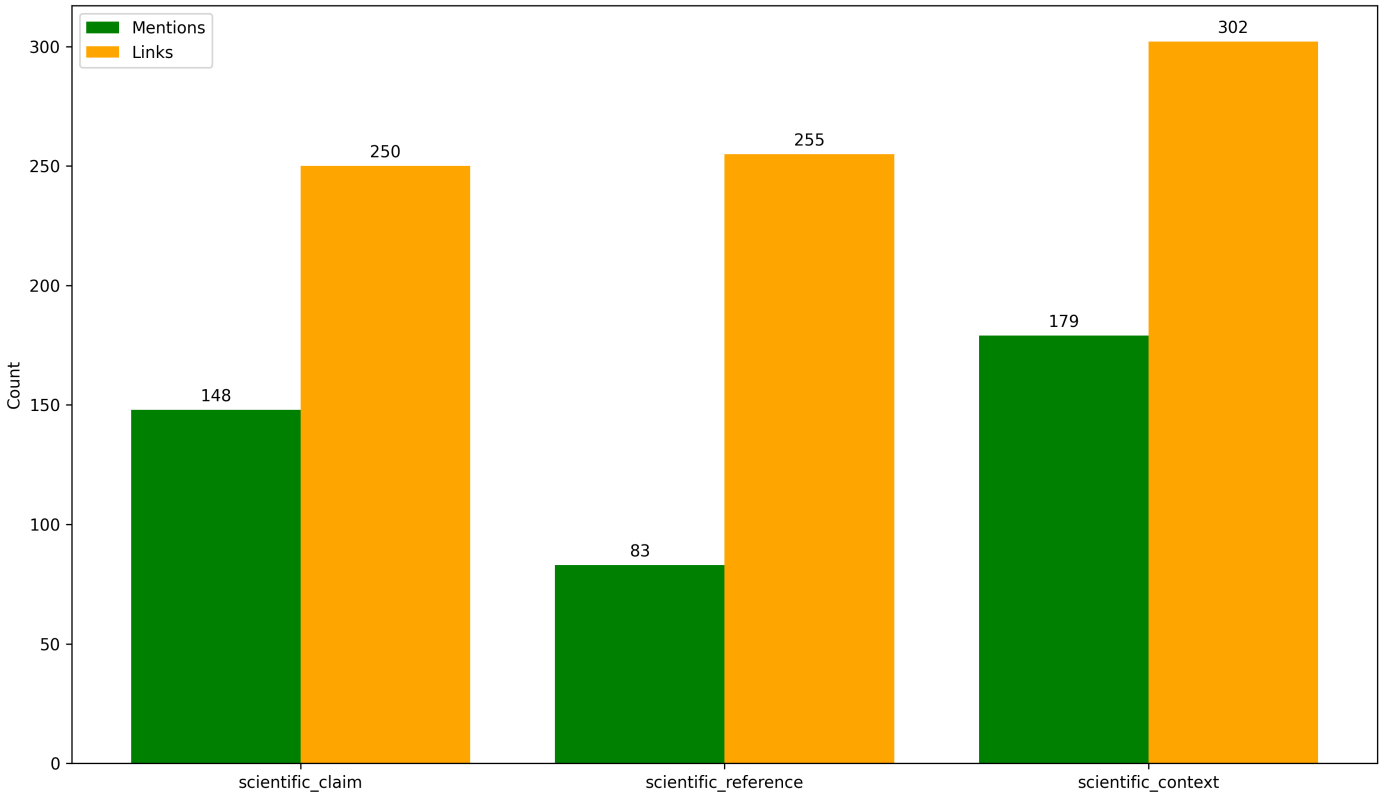

FIGURE 1 – Nombre de hashtags par tweet selon le type de tweet scientifique.

Après avoir réalisé des tests d'indépendance de student sur chaque variable, les *p-values* ne descendent pas en dessous de 0.18 donc on ne remarque pas de différence significative entre les tweets scientifiques et non scientifiques. On peut ainsi conclure que ces variables ne sont pas pertinentes pour la classification des tweets scientifiques, on les retirera de notre dataset. Malgrès tout, les `#` permettent une meilleur accuracy overall, on l'a donc gardé.

B) Modélisation

Pour chacunes des étapes, nous allons comparer différents modèles entre eux pour sélectionner le meilleur modèle. Par meilleur modèle, on entend la meilleure précision et le plus petit écart-type.

Chaque modèle est testé par cross-validation sur 10 itérations.

C) Modèle 1 : Sci vs Non-sci

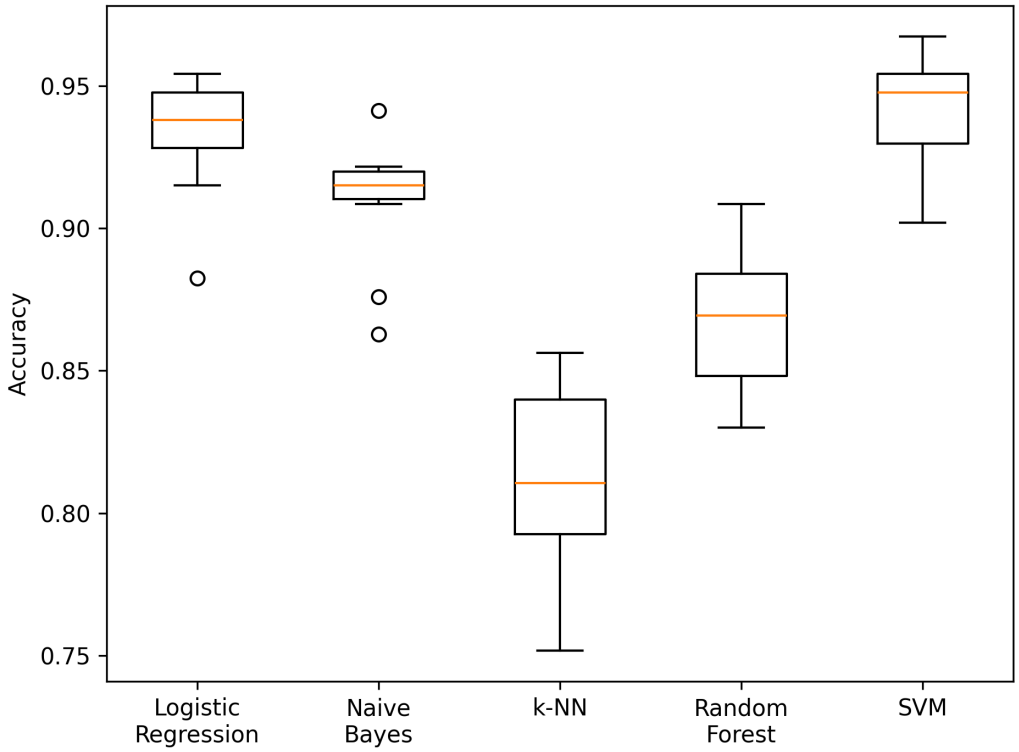

FIGURE 2 – Comparaison des modèles pour la classification des tweets scientifiques et non scientifiques.

Les deux meilleurs modèles sont le `SVM` et `Logistic Regression` mais on préfèrera cette dernière, car elle a un plus petit écart-type.

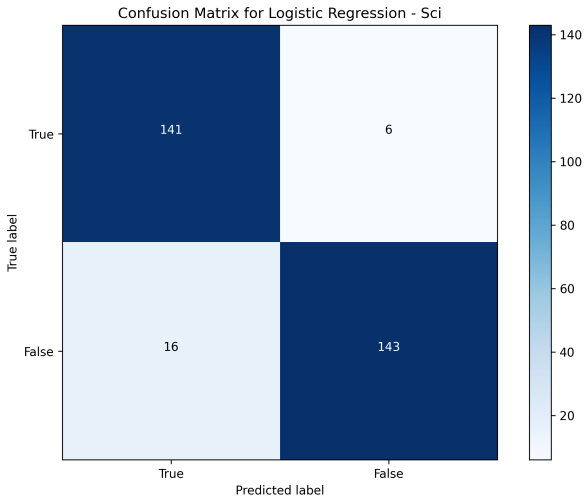

FIGURE 3 – Matrice de confusion du modèle Logistic Regression pour la classification des tweets scientifiques et non scientifiques.

D) Modèle 2 : claim et ref vs contexte

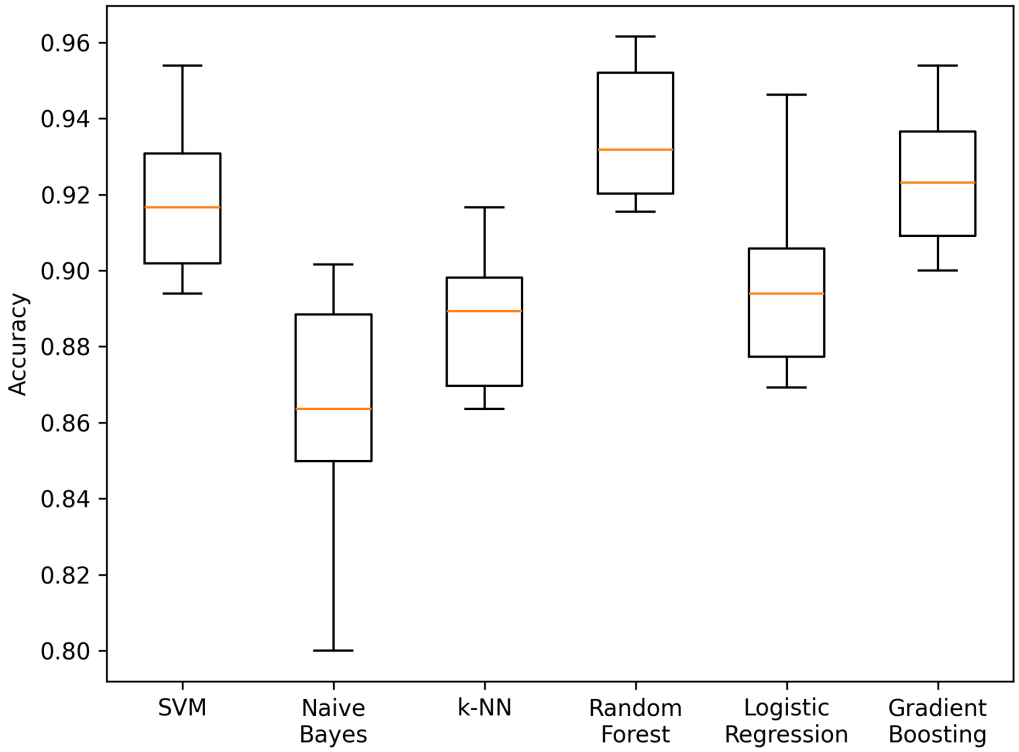

FIGURE 4 – Comparaison des modèles pour la classification des tweets claim et ref vs contexte.

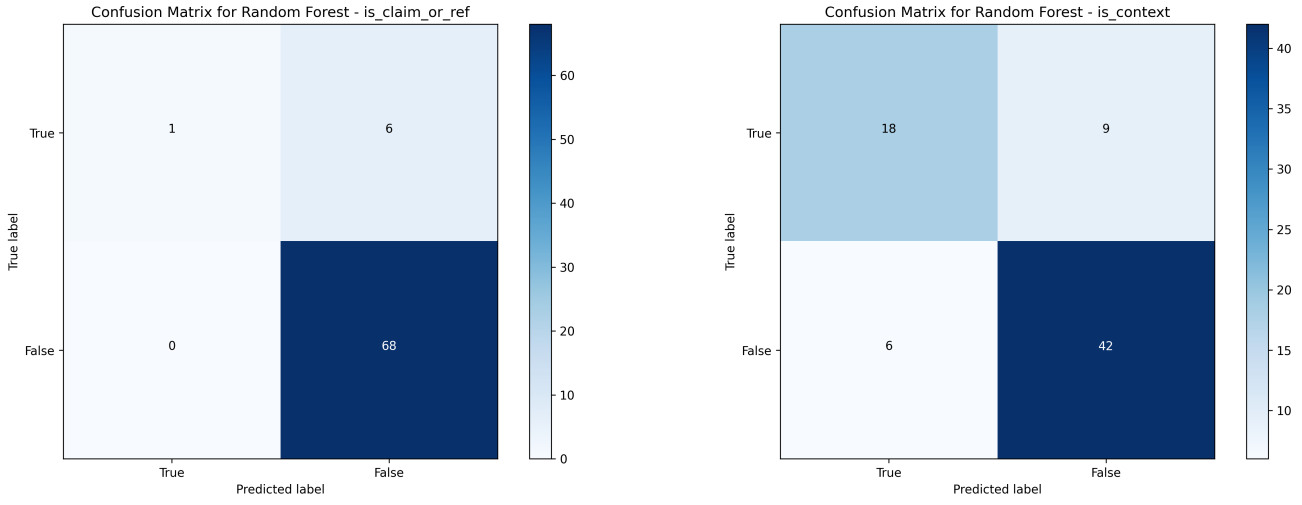

FIGURE 5 – Matrice de confusion du modèle Random Forest pour la classification des tweets claim et ref vs contexte.

E) Modèle 3 : claim vs ref vs contexte

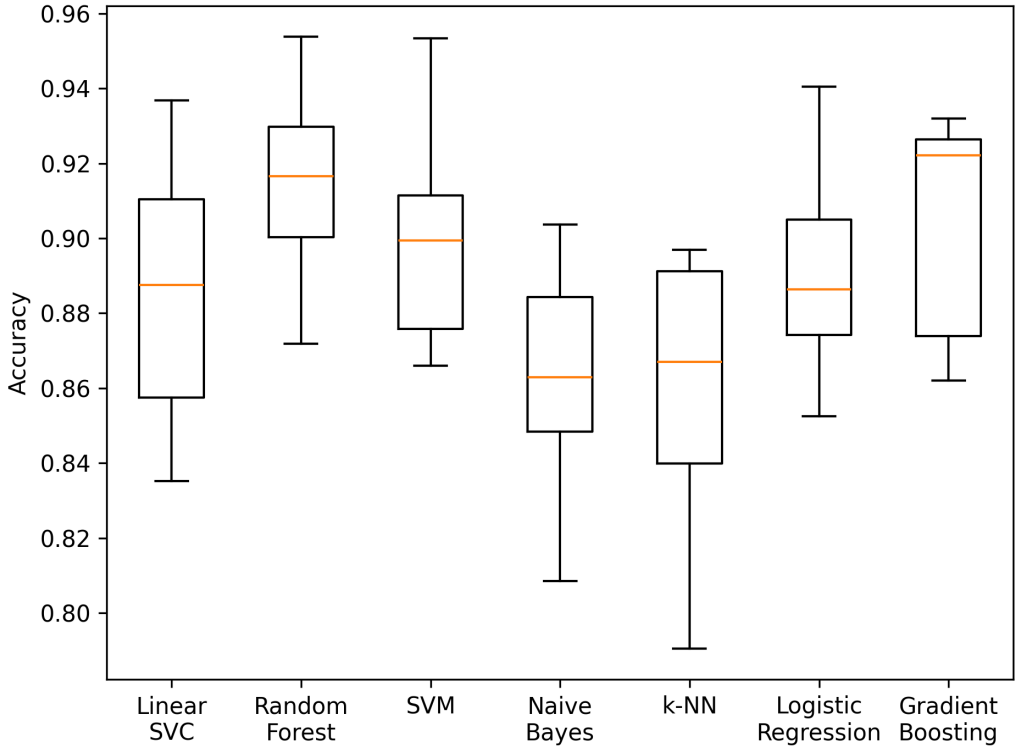

FIGURE 6 – Comparaison des modèles pour la classification des tweets claim vs ref vs contexte.

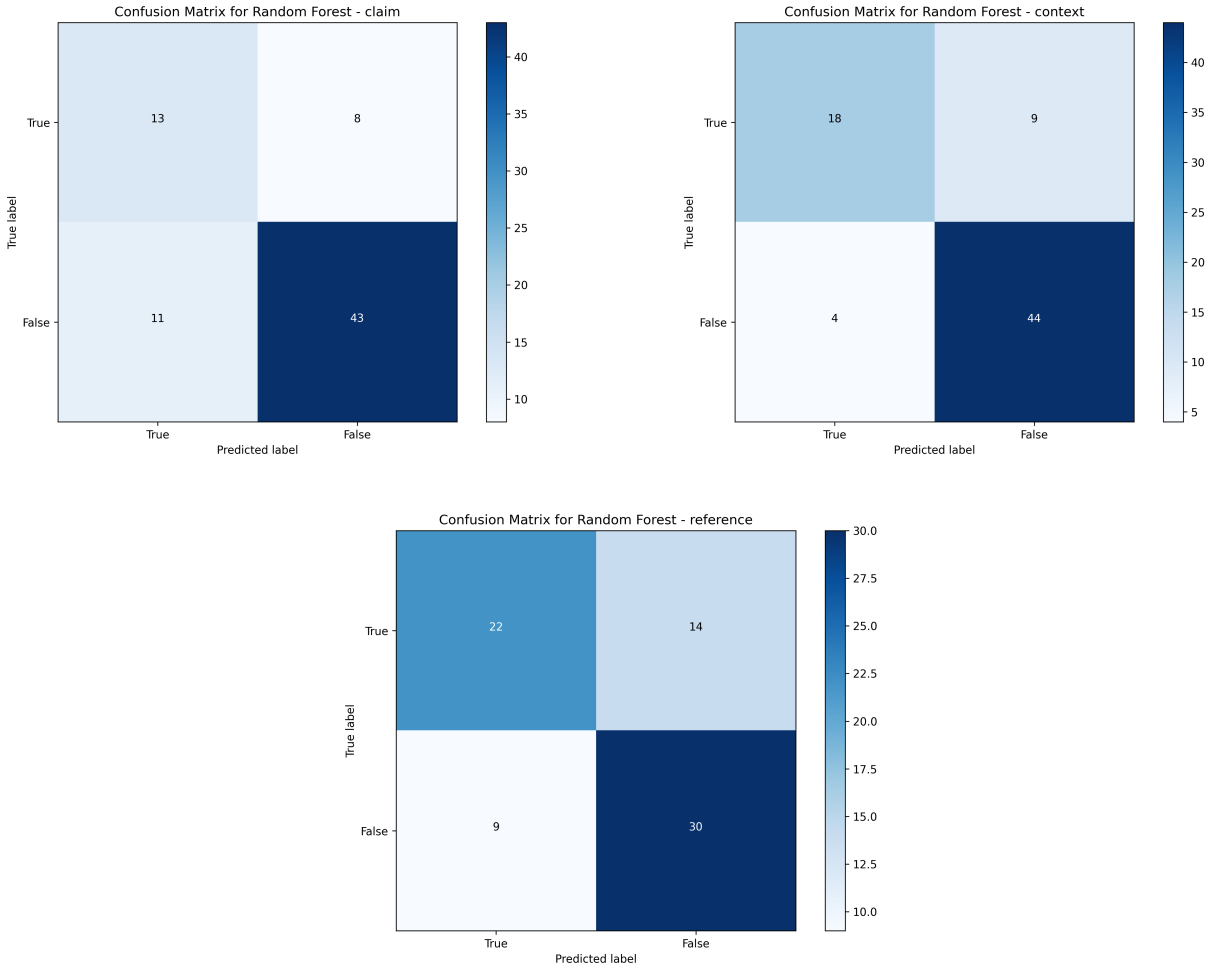

FIGURE 7 – Matrice de confusion du modèle Random Forest pour la classification des tweets claim vs ref vs contexte.

2- Discussion

3- Conclusion