**Rapport sur l'Interprétabilité des Modèles pour la Détection de Fake News sur Twitter**

La détection de fake news sur les réseaux sociaux, en particulier sur Twitter, est une problématique cruciale dans le contexte de l'information numérique. Une question importante se pose : comment évaluer si une information publiée sur Twitter est vraie ou fausse en utilisant un modèle de machine learning ? Dans ce rapport, nous explorerons l'utilisation de la méthode SHAP (SHapley Additive exPlanations) comme un outil puissant pour comprendre comment les modèles de machine learning prennent des décisions spécifiques.

**Compréhension de SHAP**

La méthode SHAP est une technique d'interprétabilité des modèles qui aide à comprendre comment les modèles de machine learning font des prédictions spécifiques. Elle repose sur un concept de la théorie des jeux appelé "valeurs de Shapley" pour attribuer une "contribution" à chaque caractéristique pour une prédiction donnée. Cette approche permet de déterminer quelles caractéristiques ont le plus d'influence sur la décision du modèle, ce qui est essentiel pour évaluer la crédibilité d'une information.

Documentation Shap : <https://shap.readthedocs.io/en/latest/>

**Avantages de l'utilisation de SHAP**

L'utilisation de SHAP présente plusieurs avantages pour notre étude de détection de fake news sur Twitter :

1. **Facilité d'Installation** : SHAP est facile à installer, ce qui simplifie son intégration dans notre flux de travail d'analyse de données.
2. **Library Python** : SHAP est une bibliothèque Python, ce qui signifie qu'elle est compatible avec de nombreuses autres bibliothèques couramment utilisées en data science et en machine learning. Cela facilite son utilisation dans nos projets.
3. **Facilité d’interprétabilité**: SHAP est en mesure de faire des représentations graphiques pour nous permettre de faire des graphiques explicite qui sont simple à comprendre et qui nous offre des réponses

**Catégories de Crédibilité**

Dans le cadre de notre étude, nous avons défini quatre catégories de crédibilité pour les informations publiées sur Twitter :

1. Complètement fiable
2. Fiable
3. Peu fiable
4. Fake

Il existe déjà une étude qui se base sur l’utilisation de SHAP et LIME pour la détection de fake news, nous sommes en mesure de s’inspirer ainsi : <https://link.springer.com/article/10.1007/s11277-023-10582-2>

L'article parle de la création d'un outil, I-FLASH, qui détecte et explique les fausses nouvelles sur les médias sociaux. Ils ont utilisé différents modèles d'apprentissage automatique et ont obtenu de bons résultats, en particulier sur les nouvelles liées à la COVID-19. L'outil peut également être efficace sur des sujets spécifiques. Ils ont utilisé des méthodes comme LIME et SHAP pour identifier les termes importants dans la classification des nouvelles.

Dans l'étude, les ensembles de données suivants ont été utilisés :

**Ensemble de données COVID-19** : Il s'agit d'un ensemble de données liées aux informations sur la COVID-19.

**Ensemble de données LIAR :** Cet ensemble de données est appelé LIAR, et il est utilisé pour évaluer la détection de mensonges dans les informations.

**Ensemble de données FactCheck :** Cet ensemble de données a été créé à partir des comptes officiels Twitter et des sites Web de divers médias bien connus. Il couvre une variété de domaines sociaux, notamment l'éducation, la criminalité et la technologie.

**Ensemble de données FactCheck2 :** Il s'agit d'un autre ensemble de données créé à partir de sources similaires, couvrant des sujets sociaux variés.

**Conclusion**

L'évaluation de la crédibilité des informations sur Twitter est un enjeu majeur, et la méthode SHAP offre une approche prometteuse pour comprendre les décisions de nos modèles de machine learning. En utilisant SHAP, nous pourrons déterminer quels paramètres et caractéristiques ont le plus d'impact sur la classification des informations en fonction de leur crédibilité. Cela nous permettra d'améliorer notre capacité à détecter les fakes news et à prendre des décisions éclairées en matière d'information sur les réseaux sociaux.

**Références**

- Interprétabilité des modèles – Medium :(<https://medium.com/towards-data-science/model-interpretability-and-explainability-27fe31cc0688>)

- SHAP (SHapley Additive exPlanations) - Christoph Molnar: (<https://christophm.github.io/interpretable-ml-book/shap.html>)