Détection de fraude dans les transactions financières en utilisant des techniques de deep learning et d'apprentissage automatique

FACULTE DES SCIENCES BEN M'SICK
UNIVERSITÉ HASSAN II DE CASABLANCA

PR.OUNACER SOUMAYA soumayaounacer@gmail.com
OUSSAMA EL HAFIDI Hfd.oussama@gmail.com

Université Hassan II de Casablanca faculté de science ben m'sik
Master Data Science & Big data

INTRODUCTION :

Le développement des systèmes de détection de fraude dans les transactions financières est devenu une priorité pour les institutions financières, face à l'augmentation des fraudes et des pertes qui en découlent. Ce projet se concentre sur l'application de techniques d'apprentissage automatique et de deep learning pour détecter les fraudes dans les transactions financières[1]. Nous avons utilisé des modèles tels que la régression logistique, les forêts aléatoires, les réseaux de neurones profonds et les autoencodeurs. Pour améliorer la performance des modèles face au déséquilibre des données, la méthode SMOTE a été appliquée[2]. Les résultats montrent que certains modèles, en particulier ceux basés sur le deep learning, offrent une meilleure capacité de détection des fraudes. Ce travail vise à fournir des outils efficaces pour aider les institutions à protéger leurs systèmes contre les attaques frauduleuses[1].

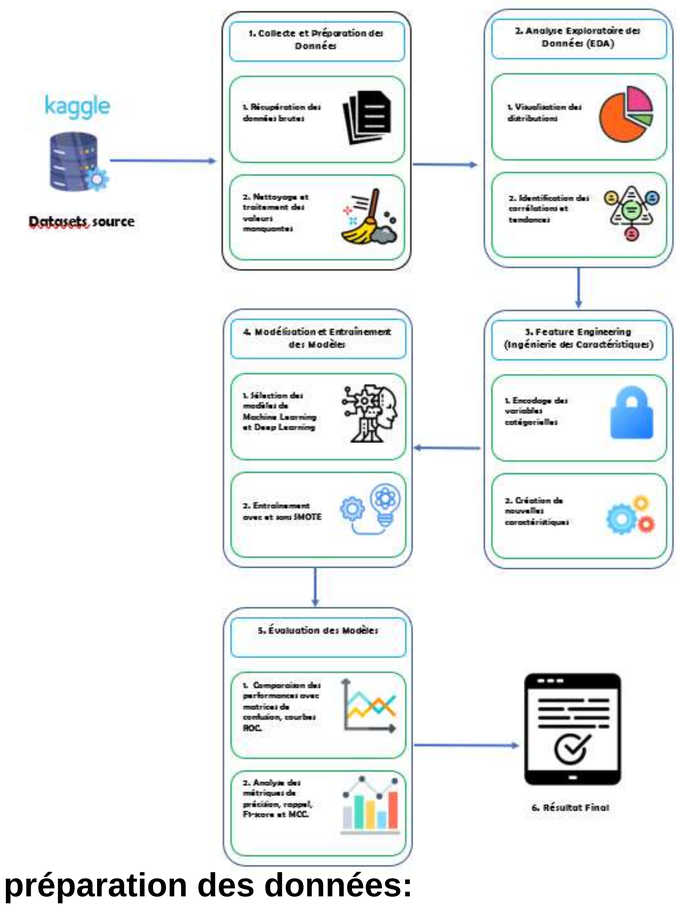
OBJECTIVES :

L'objectif principal de ce projet est de développer un système efficace de détection de fraude dans les transactions financières en utilisant des techniques d'apprentissage automatique et de deep learning[3]. Nous visons à :

- Identifier les caractéristiques clés des transactions frauduleuses : Extraire des variables pour distinguer fraudes et transactions légitimes.
- Comparer et évaluer la performance des modèles de machine learning et deep learning : Tester divers modèles pour trouver ceux avec la meilleure précision.
- Traiter le déséquilibre des classes à l'aide de techniques de suréchantillonnage : Utiliser SMOTE pour générer plus d'exemples de fraudes.
- Développer un modèle robuste et généralisable : Créer un modèle performant pour s'adapter à de nouvelles données.

METHOLOGIE :

La méthodologie de ce projet se compose de plusieurs étapes clés qui garantissent une analyse approfondie et structurée des données. Voici un aperçu des étapes illustrées dans le diagramme de workflow ci-dessous :



1. **Collecte et préparation des données:**
   - Source des données.
   - Nettoyage et préparation.
2. **Analyse exploratoire des données (EDA) :** Visualisation, Statistiques.
3. **Ingénierie des caractéristiques:** Transformation, Sélection, Normalisation
4. **Modélisation et entraînement:** Algorithmes, Optimisation, Validation, Entraînement.
5. **Évaluation des modèles:** Précision, Rappel, F1-score, AUC.
6. **Résultat final:** Conclusion, Comparaison, Performance, Amélioration.

RESULTATS :

Nous allons maintenant examiner les résultats détaillés des différents modèles que nous avons testés, en mettant l'accent sur leur précision, leur rappel et leur F1-score. Ces métriques nous permettront de déterminer l'approche la plus efficace pour la détection des fraudes.

Modèles	précision	Rappel	F1-score	AUC
Machine learning				
RL	0.855952	0.839	0.719	0.832
SVM	0.94.90	0.743	0.1025	0.706
ARBRE DE DECISION	0.9752	0.9987	0.9988	0.9976
FORET ALEATOIRE	0.9817	0.8017	0.8569	0.8038
KNN	0.9845	0.9036	9.7823	0.9401
XGBOOST	0.976	0.9121	0.5733	0.9054
Deep learning				
LSTM	0.9599	0.9513	0.9543	0.9216
RNN	0.9603	0.8285	0.1389	0.59
DNN	0.9167	0.8263	0.1576	0.7976

Les résultats montrent une variabilité dans les performances des modèles de machine learning et de deep learning. Parmi les modèles de machine learning, KNN et l'arbre de décision affichent des précisions élevées, tandis que le SVM présente des scores plus faibles en F1-score. Du côté du deep learning, LSTM se distingue par ses performances globales avec un F1-score élevé, tandis que RNN et DNN ont des résultats moins cohérents, particulièrement en F1-score. Enfin, les métriques comme le rappel et l'AUC varient également entre les modèles, reflétant les différences dans leur capacité à gérer la détection des fraudes.

CONCLUSION :

Dans ce projet de détection de fraude, plusieurs modèles de machine learning et de deep learning ont été comparés à l'aide de métriques telles que la précision, le rappel, le F1-score et l'AUC. L'arbre de décision et K-Nearest Neighbors (KNN) se sont distingués avec des précisions élevées de 97,52% et 98,45%, bien que l'arbre de décision puisse souffrir de surapprentissage. Parmi les modèles de deep learning, LSTM s'est avéré être le plus performant, avec une précision de 95,99% et un F1-score de 95,43%, en faisant le modèle le mieux adapté à la détection de fraudes. D'autres modèles comme SVM et RNN ont montré des performances moins satisfaisantes en raison de leur difficulté à gérer le déséquilibre des classes. En résumé, le LSTM est le modèle le plus recommandé pour cette tâche, tandis que KNN pourrait être une option pour des implémentations plus rapides.

REMERCIEMENTS :

Je souhaite exprimer ma profonde reconnaissance à Mme Soumaya Ounacer pour son encadrement précieux tout au long de ce projet. Son appui constant, ses conseils éclairés et son expertise ont joué un rôle clé dans la réalisation de ce travail. En tant qu'étudiant en Master Data Science & Big Data, j'ai énormément bénéficié de son mentorat, qui m'a permis de surmonter les obstacles et de progresser vers mes objectifs. Sa confiance en moi a été une source de motivation et d'inspiration tout au long de ce parcours. Je la remercie sincèrement pour son implication et son dévouement, qui ont grandement contribué à la réussite de ce projet.

REFERENCES :

[1] Böhmer, M., & Fry, B. (2019). Deep learning for fraud detection: A review. arXiv preprint arXiv:1907.06712

[2] He, H., Garcia, E. A., & Li, S. (2019). Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 32(7), 1323-1339.

[3] Hollmann et al. (2023). Large Language Models for Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering. (ArXiv preprint arXiv:2306.09055).