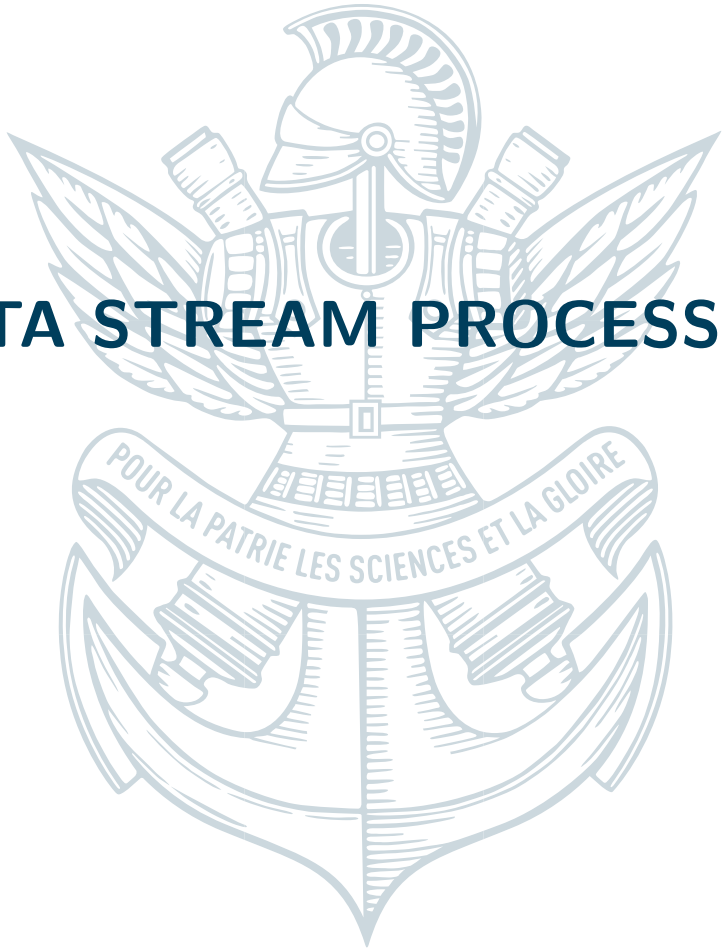

DATA STREAM PROCESSING

25 octobre 2025

El Vilaly Oumouhani, Fredo Alejos Arrieta

TABLE DES MATIÈRES

1
PARTIE I : EXPÉRIMENTATION AVEC RIVER

1.1 Étude 1

1.1.1 • Méthodologie

Pour cette partie, nous travaillons avec le jeu de données Credit Card Fraud Detection disponible dans la librairie River. Ce dataset contient des transactions réalisées par carte bancaire en septembre 2013 par des porteurs européens. Il regroupe les opérations effectuées sur une période de deux jours, pour un total de 284 807 transactions, dont 492 fraudes identifiées. Le jeu de données est donc fortement déséquilibré, la classe positive (fraude) ne représentant que 0,172 % de l'ensemble des observations.

L'objectif est de réaliser une classification binaire, afin de distinguer les transactions normales des fraudes.

Nous comparons plusieurs modèles issus de River, notamment :

— LogisticRegression et NaiveBayes, pour des approches probabilistes et linéaires,

— HoeffdingTreeClassifier et sa version adaptative HoeffdingAdaptiveTreeClassifier.

La principale différence entre ces deux derniers modèles réside dans leur capacité à s'adapter à la dérive conceptuelle (concept drift). En effet, le HoeffdingAdaptiveTreeClassifier intègre un mécanisme de détection et d'adaptation locale (basé sur ADWIN) qui lui permet de réagir automatiquement aux changements de distribution des données dans le temps, contrairement à la version classique qui apprend de manière statique.

1.1.2 • Résultats

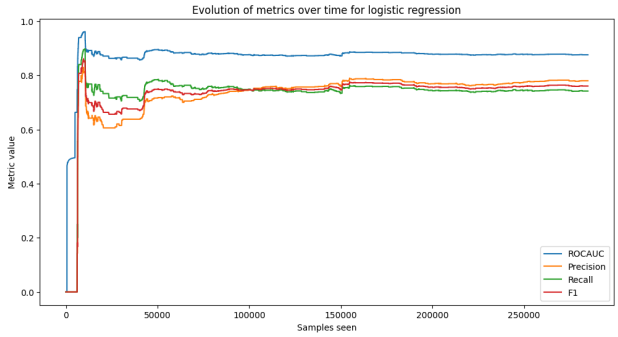

Figure 1 – Évolution des métriques ROC-AUC, Recall, F1 et Accuracy dans le temps pour la régression logistique

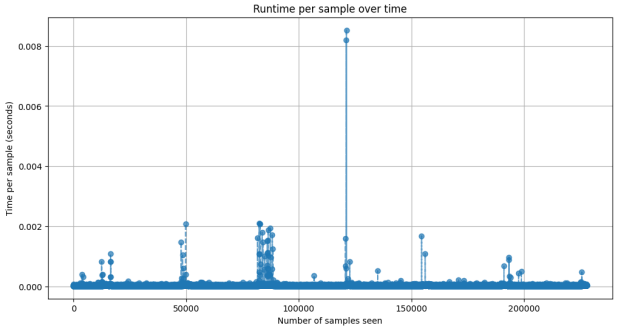

Figure 2 – Évolution du temps d'exécution par échantillon dans le temps pour la régression logistique

Pour la régression logistique de River, on remarque que les metriques évoluent rapidement au début, mais stagnent après les 50000 premiers échantillons. Concernant le temps d'exécution par échantillon, il est très variable avant le 120000 échantillon, avec un pic à ce niveau, puis se stabilise par la suite.

La figure 3 présente la comparaison des performances entre la version standard du HoeffdingTree et sa version adaptative (AdaptiveTree) en termes de précision au cours de l'apprentissage incrémental. On observe que le modèle standard atteint, dans un premier temps, une précision plus élevée, avoisinant 1, avant de se stabiliser progressivement autour de 99.85 %. En revanche, l'AdaptiveTree présente une précision légèrement inférieure (environ 99.83 %), mais plus stable au fil des échantillons.

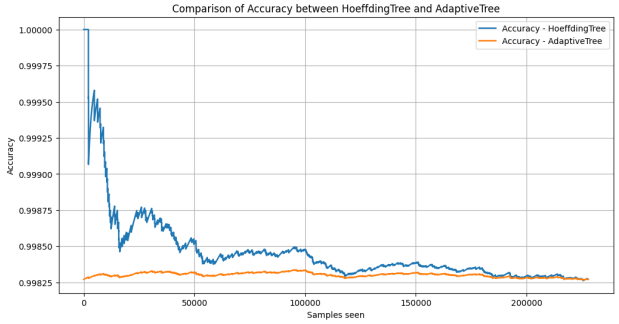

Figure 3 – Évolution de l'Accuracy dans le temps

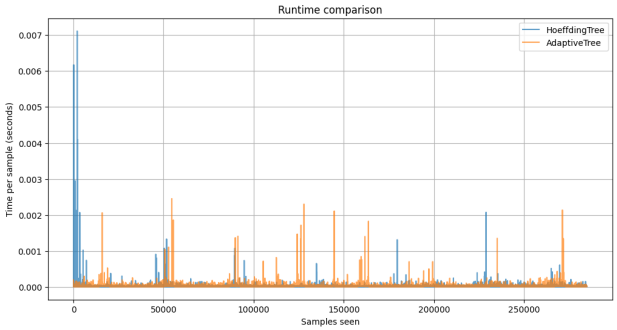

Figure 4 – Évolution du temps d'exécution par échantillon dans le temps

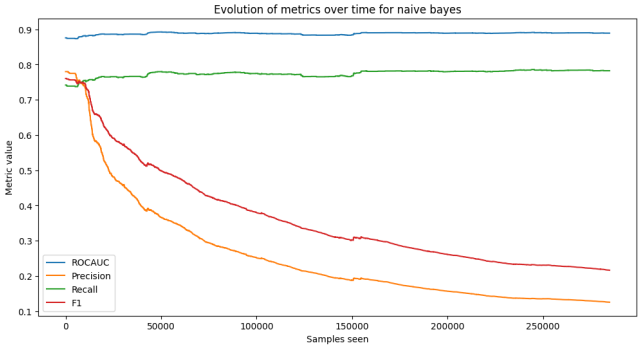

Figure 5 – Évolution des métriques pour le modèle Naive Bayes

Ces résultats suggèrent que le HoeffdingTree classique est plus performant sur des données stationnaires, tandis que la version adaptative se montre plus robuste face aux variations potentielles du flux de données. L'AdaptiveTree semble donc mieux convenir aux environnements où la distribution des données évolue dans le temps (concept drift), au prix d'une légère baisse de précision moyenne. Concernant le temps d'exécution par échantillon (Figure 4), on observe que la version adaptative du modèle présente une évolution plus stable au fil des données, avec un temps de traitement globalement constant. À l'inverse, le HoeffdingTree standard voit son temps d'exécution augmenter de manière significative après le seuil des 1000 échantillons. Cette différence peut s'expliquer par le fait que la version adaptative intègre des mécanismes de gestion dynamique de la structure de l'arbre, lui permettant de mieux contrôler sa croissance et d'éviter une complexité excessive lors de l'apprentissage continu.

Une quatrième méthode sur laquelle nous nous sommes focalisés est le Naive Bayes (figure 5). Cependant, ses performances se sont révélées inférieures à celles de la régression logistique.

Cela peut s'expliquer par la construction de chaque modèle. Le modèle Naive Bayes repose sur une hypothèse forte d'indépendance conditionnelle entre les variables explicatives, supposant que chaque caractéristique est indépendante des autres une fois la classe cible connue. Cependant, cette hypothèse est rarement respectée dans les jeux de données réels, où des corrélations entre variables sont fréquentes. Cette simplification peut entraîner une perte d'information et, par conséquent, une baisse de la performance prédictive.

En comparaison, la régression logistique ne repose pas sur une telle hypothèse. Elle estime directement les coefficients des variables en maximisant la vraisemblance des observations, ce qui lui permet de modéliser efficacement les relations linéaires entre les caractéristiques et la probabilité d'appartenance à une classe. Cette capacité à prendre en compte les interactions entre variables explique généralement ses performances supérieures en pratique.

1.2 Étude 2

1.2.1 • Méthodologie

Pour la deuxième étude, nous avons choisi le jeu de données HTTP provenant de River, qui correspond au HTTP dataset du KDD 1999 Cup. L'objectif est de prédire si une connexion HTTP est anormale ou non. Le jeu de données ne contient que 2 211 labels positifs (soit 0,4 % des échantillons), ce qui représente un problème fortement déséquilibré.

Nous avons ensuite testé les quatre modèles de classification déjà évalués sur le jeu de données de cartes bancaires, afin de comparer leurs performances sur ce nouveau jeu de données.

1.2.2 • Résultats

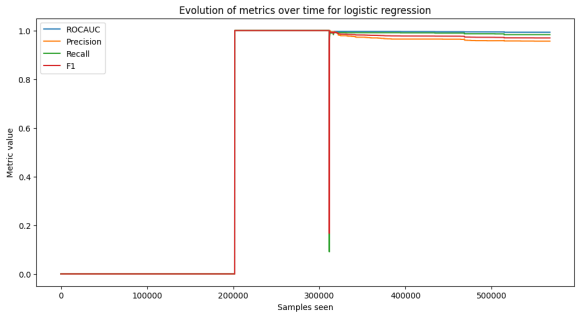

Figure 6 – Évolution de metriques dans le temps pour la regression logistique

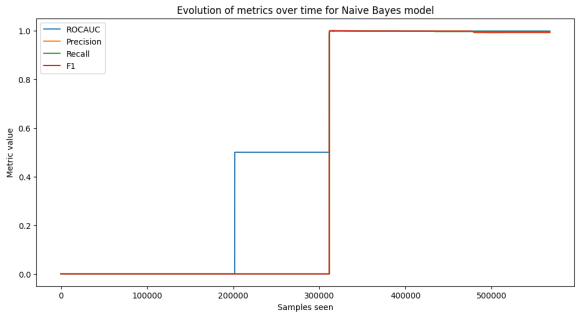

Figure 7 – Évolution de metriques dans le temps pour le naive bayes

Sur le jeu de données HTTP, les modèles Naive Bayes et régression logistique présentent des comportements distincts. La régression logistique met plus de temps à converger et atteint des performances stables mais légèrement inférieures, avec un ROCAUC proche de 1 mais une précision et un rappel légèrement plus faibles. En revanche, le Naive Bayes converge rapidement vers des valeurs de métriques proches de 1, traduisant une excellente séparation des classes.

Ainsi pour le jeu de données HTTP, le modèle Naive Bayes présente de meilleures performances que la régression logistique, contrairement au cas précédent où cette dernière obtenait les meilleurs résultats sur le jeu de données des cartes bancaires(figure 6, figure 7)

On remarque que le modèle Hoeffding Tree n'est pas adapté à ce jeu de données. En effet, ses performances restent très faibles, avec des valeurs de précision, rappel et F1-score nulles, tandis que le ROC AUC demeure à 0 durant les 200 000 premiers échantillons avant de se stabiliser autour de 0,5, ce qui traduit un comportement proche d'un classifieur aléatoire(figure 8)

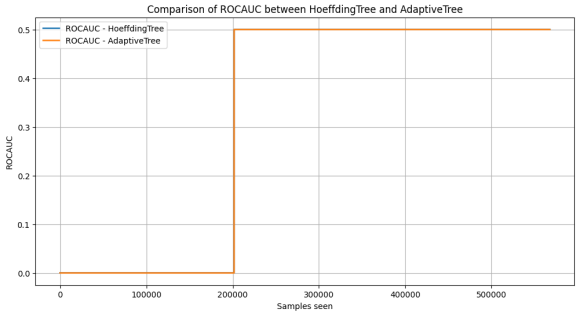

Figure 8 – Évolution Dr ROC AUC Hoeffding tree

2
PARTIE II : EXPÉRIMENTATION AVEC CAPYMOA

2.1 Étude 1

2.1.1 • Méthodologie

Dans cette partie, nous nous intéressons à CapyMOA. Nous avons choisi le jeu de données synthétique RBFm100k, qui contient 100 000 instances et 10 classes, générées à partir d'un mélange de fonctions radiales de base (RBF – Radial Basis Function). Nous avons décidé de tester quatre modèles : le Hoeffding Tree, le Hoeffding Adaptive Tree, l'Adaptive Random Forest et le Naive Bayes. Pour l'évaluation, nous nous concentrons sur l'accuracy cumulative dans le temps, calculée par fenêtre glissante (windowing), ainsi que sur le coefficient Kappa, afin de mesurer à la fois la performance globale et la robustesse du modèle face aux variations de distribution.

2.1.2 • Résultats

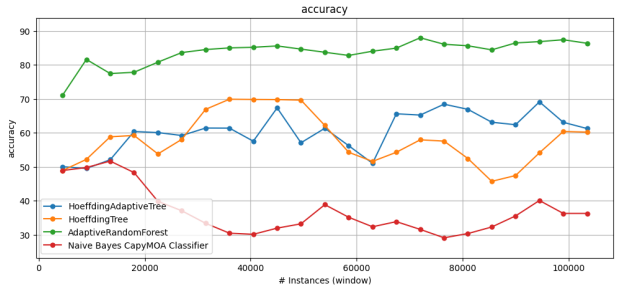

Figure 9 – Évolution de l'accuracy au cours du temps pour quatre modèles CapyMOA data rbfm100k : Hoeffding Tree, Hoeffding Adaptive Tree, Adaptive Random Forest et Naive Bayes

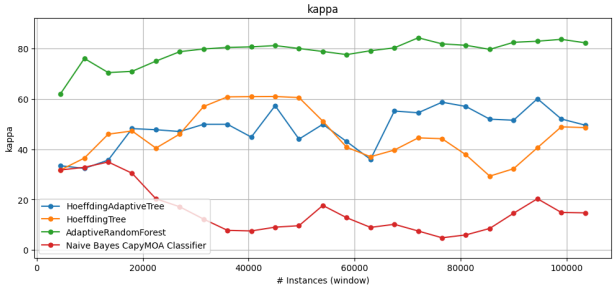

Figure 10 – Évolution de Kappa au cours du temps pour quatre modèles CapyMOA data rbfm100k : Hoeffding Tree, Hoeffding Adaptive Tree, Adaptive Random Forest et Naive Bayes

En termes d'accuracy (figure 9), les méthodes adaptatives montrent généralement de meilleures performances. Pour notre expérience sur le flux RBFm100k avec CapyMOA, nous obtenons en moyenne :

— Hoeffding Adaptive Tree (HAT) : 60,36%
— Hoeffding Tree (HT) : 57,93%
— Adaptive Random Forest (ARF) : 83,50%
— Naive Bayes (NB) : 36,79%

On observe donc une différence notable entre les modèles adaptatifs de CapyMOA et ceux de River. Par exemple, sur le dataset Credit Card avec River, le modèle adaptatif avait des performances inférieures au modèle non adaptatif. Cette différence peut s'expliquer par la nature des données : certains flux ne présentent pas de concept drift significatif, ou bien les distributions de classes et les corrélations sont différentes.

D'ailleurs, pour les deux métriques considérées, l'accuracy et le Kappa(figure 10), le Naive Bayes présente les performances les plus faibles, tandis que l'Adaptive Random Forest obtient les meilleurs résultats. Le Hoeffding Tree et le Hoeffding Adaptive Tree présentent des fluctuations dans le temps, bien que, de manière générale, le modèle adaptatif (HAT) reste légèrement supérieur au Hoeffding Tree classique

2.2 Étude 2

2.2.1 • Méthodologie

Dans cette section, nous conduisons notre étude sur le jeu de données Electricity disponible dans la bibliothèque Capymoa. Ce jeu de données correspond à un problème de classification issu du marché de l'électricité de la Nouvelle-Galles du Sud (Australie). Il contient 45 312 instances et 8 variables explicatives. L'objectif est de prédire la direction du prix de l'électricité, à savoir s'il augmente (up) ou diminue (down). Nous entraînons ensuite quatre modèles Hoeffding Tree, Hoeffding Adaptive Tree, Adaptive Random Forest et Naive Bayes, afin de comparer leurs performances en termes d'accuracy et de coefficient kappa.

2.2.2 • Résultats

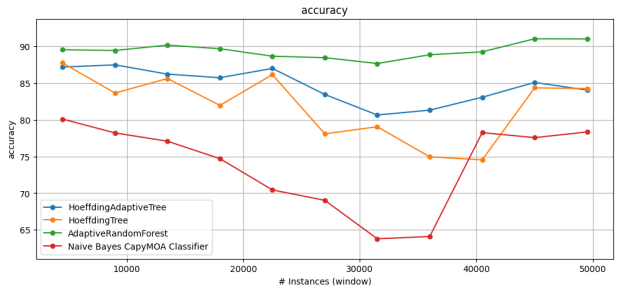

Figure 11 – Évolution de l'accuracy au cours du temps pour quatre modèles CapyMOA data electricity : Hoeffding Tree, Hoeffding Adaptive Tree, Adaptive Random Forest et Naive Bayes

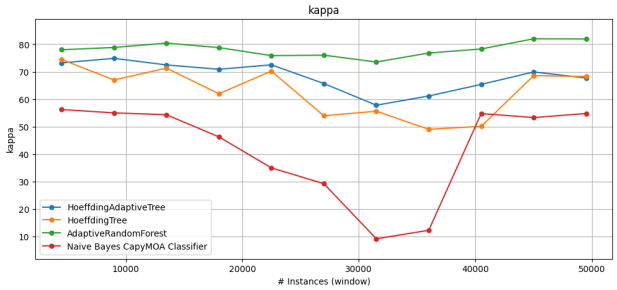

Figure 12 – Évolution de Kappa au cours du temps pour quatre modèles CapyMOA data electricity : Hoeffding Tree, Hoeffding Adaptive Tree, Adaptive Random Forest et Naive Bayes

Pour ce jeu de données, on observe que le comportement des modèles reste globalement similaire en termes d'accuracy et de coefficient kappa. L'Adaptive Hoeffding Tree demeure le modèle le plus performant, confirmant sa capacité d'adaptation aux variations du flux de données. Le seul modèle se distinguant notablement est le Naive Bayes, dont la performance s'améliore considérablement sur ce jeu de données par rapport au jeu

RBFM100k, bien qu'il reste globalement moins performant que les modèles arborescents. Enfin, comme précédemment, les performances du Hoeffding Tree standard et de sa version adaptative demeurent proches, cette dernière conservant toutefois un léger avantage(figure 11 et 12)