BOLO Japhet

BORDEUX Eliot

ESSIMBI Charles

Groupe 3 \ Trinôme 6

Rapport

Projet Intelligence Artificielle

Analyse et Modélisation des Comportements de Navigation des Navires

SOMMAIRE :

Introduction au Projet

Ce projet d'Intelligence Artificielle s'inscrit dans le cadre de l'analyse et de la modélisation des comportements de navigation maritime à partir de données AIS (Automatic Identification System). Il vise à approfondir les compétences en apprentissage automatique à travers une application complète combinant techniques supervisées et non-supervisées.

Contexte et Objectifs

L'objectif principal est de développer des solutions d'intelligence artificielle pour répondre à trois besoins clients distincts dans le domaine maritime :

Visualisation intelligente : regroupement automatique des navires selon leurs schémas de navigation similaires

Classification prédictive : développement d'un modèle capable de prédire le type de navire

Prédiction temporelle : anticipation des trajectoires futures des navires à différents horizons temporels

Les Données AIS

Les données exploitées proviennent du système AIS (Automatic Identification System), un système de suivi automatique utilisé dans le transport maritime. Le dataset analysé comprend 229 185 observations de navires évoluant principalement dans le Golfe du Mexique et le long des côtes Est des États-Unis, avec 18 variables décrivant :

Caractéristiques de navigation : vitesse sur le fond (SOG), cap suivi (COG), orientation de la proue (Heading)

Données de géolocalisation : latitude (LAT), longitude (LON), horodatage (BaseDateTime)

Propriétés physiques : longueur (Length), largeur, tirant d'eau (Draft)

Informations techniques : type de navire (VesselType), identifiant MMSI

Cette richesse de données offre une opportunité unique d'explorer les comportements de navigation à grande échelle et d'identifier des patterns comportementaux significatifs dans l'écosystème maritime.

Besoin 1 - Clustering des Comportements de Navigation Maritime

Introduction et Objectifs du Projet

Ce rapport présente une analyse approfondie des comportements de navigation maritime à travers l'application de techniques d'apprentissage non supervisé, spécifiquement le clustering. L'objectif principal de cette étude était de regrouper automatiquement les navires présentant des schémas de navigation similaires, en se basant sur leurs caractéristiques opérationnelles et physiques. Cette approche permet d'identifier des patterns comportementaux dans les données de navigation AIS (Automatic Identification System) et de comprendre comment les différents types de navires évoluent dans l'espace maritime du Golfe du Mexique.

Le dataset analysé comprend 229 185 observations de navires avec 18 variables décrivant leurs positions, mouvements, et caractéristiques techniques. Cette richesse de données offre une opportunité unique d'explorer les comportements de navigation à grande échelle et d'identifier des groupes homogènes de navires partageant des profils opérationnels similaires.

Méthodologie de Préparation des Données

La phase de préparation des données a constitué une étape fondamentale pour garantir la qualité et la pertinence de l'analyse. Le processus a débuté par le chargement du fichier vessel-total-cleaned.csv, déjà prétraité pour éliminer les valeurs manquantes et les doublons. Cette base de données propre a facilité les étapes subséquentes d'analyse.

La sélection des variables pertinentes pour le clustering s'est concentrée sur sept caractéristiques clés : la vitesse sur le fond (SOG), le cap suivi (COG), l'orientation de la proue (Heading), ainsi que les dimensions physiques des navires incluant la longueur, la largeur, le tirant d'eau, et le type de navire. Ces variables ont été choisies car elles caractérisent directement le comportement de navigation et les propriétés intrinsèques des navires, éléments essentiels pour identifier des patterns comportementaux cohérents.

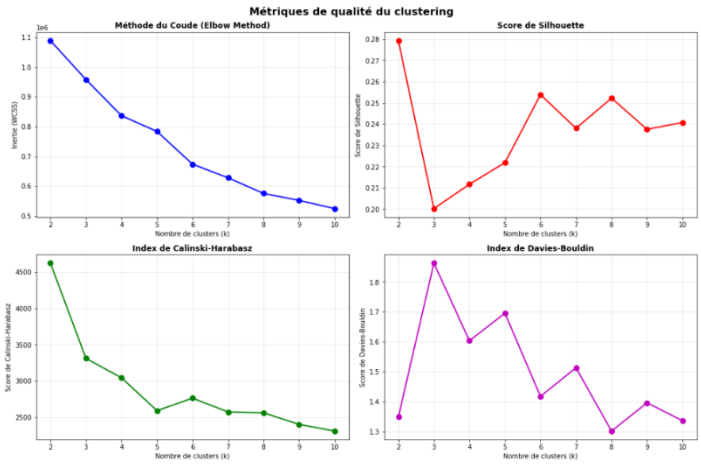
Un traitement spécifique a été appliqué à la variable Heading, où la valeur codée 511 (signifiant "cap inconnu" dans le système AIS) a été remplacée par la médiane des valeurs valides. Cette approche préserve la cohérence statistique des données tout en évitant l'introduction de biais dans les calculs de clustering. La variable catégorielle VesselType a nécessité un encodage numérique via LabelEncoder, transformant les catégories textuelles en valeurs numériques exploitables par les algorithmes de machine learning.

La normalisation des données par StandardScaler a représenté une étape cruciale, considérant les différences d'échelle importantes entre les variables (vitesse en nœuds, dimensions en mètres, types codés en entiers). Cette standardisation centre les données autour d'une moyenne nulle avec un écart-type unitaire, garantissant une contribution équitable de chaque variable dans les calculs de distance euclidienne utilisés par l'algorithme KMeans.

Stratégie de Clustering et Optimisation

Le choix de l'algorithme KMeans s'est imposé naturellement compte tenu du volume important de données (229 185 observations) et de la nature numérique des variables après preprocessing. KMeans présente l'avantage d'être computationnellement efficace sur de grands datasets tout en produisant des clusters interprétables autour de centroïdes bien définis. Cette approche s'avère particulièrement adaptée aux données normalisées où la distance euclidienne constitue une mesure de similarité pertinente.

La détermination du nombre optimal de clusters a fait l'objet d'une analyse méthodique utilisant quatre métriques d'évaluation complémentaires. La méthode du coude (basée sur l'inertie), le score de Silhouette, l'index de Calinski-Harabasz, et l'index de Davies-Bouldin ont été calculés pour des valeurs de k variant de 2 à 10. Pour optimiser les performances computationnelles, un échantillonnage stratégique de 10 000 observations a été réalisé, préservant la représentativité statistique des données tout en accélérant les calculs.



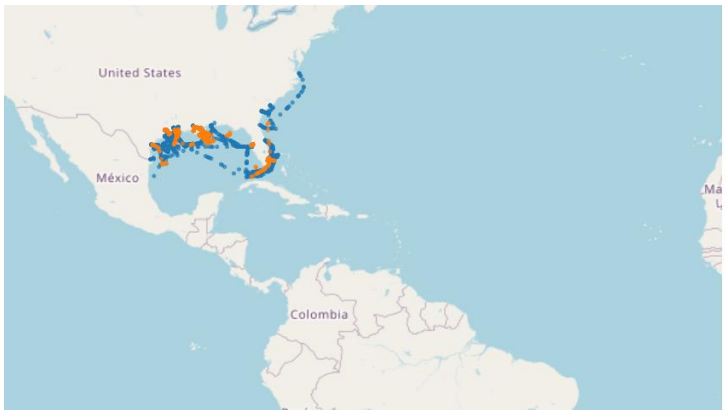
Les résultats de cette analyse comparative ont révélé une convergence remarquable des métriques vers k=2 comme solution optimale. Le score de Silhouette atteint son maximum à 0.279 pour k=2, indiquant une séparation claire entre les clusters avec une bonne cohésion interne. L'index de Calinski-Harabasz confirme cette tendance avec son score maximal de 4620.3 pour k=2, témoignant d'un excellent rapport entre la variance inter-cluster et intra-cluster. Bien que l'index de Davies-Bouldin favorise légèrement k=8, la cohérence globale des autres métriques soutient fermement le choix de deux clusters.

Résultats et Interprétation des Clusters

L'application du clustering KMeans avec k=2 sur l'ensemble des données a produit une segmentation significative de la flotte maritime. Le Cluster 0 regroupe 86 681 navires (37.8% de l'échantillon), tandis que le Cluster 1 comprend 142 504 navires (62.2%). Cette répartition asymétrique suggère l'existence de deux comportements de navigation dominants mais d'importance inégale dans le dataset.

L'analyse de la distribution géographique révèle des patterns spatiaux intéressants. Les données se concentrent principalement dans le Golfe du Mexique et le long des côtes Est des États-Unis, avec des coordonnées s'étendant de 23.4°N à 36.7°N en latitude et de -97.4°W à -75.0°W en longitude. Cette zone géographique cohérente correspond aux principales routes commerciales et zones d'activité maritime de la région.

La visualisation cartographique interactive développée avec Plotly permet d'explorer en détail la répartition spatiale des clusters. Chaque navire est représenté par un marqueur coloré selon son appartenance cluster, facilitant l'identification de zones de concentration spécifiques. Le Cluster 0, représenté en bleu, semble présenter une distribution plus concentrée dans certaines zones du golfe, suggérant potentiellement des routes ou zones d'activité spécialisées. Le Cluster 1, visualisé en orange, affiche une répartition géographique plus étendue, pouvant correspondre à des activités de navigation plus diversifiées ou à des routes commerciales principales.



Implications et Perspectives d'Application

Cette segmentation en deux clusters principaux offre des perspectives d'application pratiques pour la gestion du trafic maritime, l'optimisation des routes, et l'analyse des comportements de navigation. L'identification de patterns comportementaux distincts peut contribuer à l'amélioration de la sécurité maritime, à l'optimisation de la consommation de carburant, et à la planification des infrastructures portuaires.

La méthode développée présente l'avantage d'être reproductible et scalable, permettant son application à d'autres zones géographiques ou à des datasets temporels étendus. L'approche non supervisée garantit une découverte objective de patterns sans a priori sur la structure des données, élément crucial pour l'analyse exploratoire de comportements complexes.

Les résultats obtenus constituent une base solide pour des analyses plus approfondies, incluant l'étude de l'évolution temporelle des clusters, l'analyse des corrélations entre caractéristiques de navigation et conditions météorologiques, ou encore l'intégration de variables économiques pour comprendre les déterminants des comportements de navigation. Cette approche méthodologique rigoureuse démontre la valeur de l'apprentissage non supervisé pour l'analyse de données maritimes à grande échelle.

4. Besoin Client 2 : Prédiction du Type de Navire

La classification est une tâche fondamentale en apprentissage automatique visant à attribuer une catégorie ou une étiquette à des données en fonction de leurs caractéristiques. Dans la réponse au besoin 2, nous nous concentrons sur l'apprentissage supervisé, une approche où un modèle est entraîné sur un ensemble de données étiquetées pour prédire le type de navire (Cargo, Tanker, Passenger) à partir de variables telles que la longueur, le tirant d'eau, la vitesse, et la largeur. Grâce à des techniques d'optimisation et d'évaluation robustes, ce travail vise à développer un modèle performant, exploitable en temps réel, pour répondre aux besoins opérationnels de classification maritime.

4.1. Préparation des données pour la classification

Les données ont été chargées à partir du fichier export_IA.csv, le fichier nettoyé issu du projet Big data. Après inspection, 41850 doublons (basés sur les caractéristiques et la cible VesselType) ont été identifiés et supprimés, réduisant la taille à (187335, 18). Aucun filtrage supplémentaire des classes rares n'a été appliqué, car toutes les classes restantes (Cargo, Tanker, Passenger) dépassent les seuils minimaux. Les valeurs manquantes ont été gérées avec une imputation par la médiane, aboutissant à zéro NaN pour les caractéristiques Length, Draft, COG, Heading, LAT, et LON. Les types des caractéristiques sont uniformément float64, assurant une cohérence numérique. L'analyse des corrélations révèle une forte corrélation positive entre Length et Draft (0.807704), une corrélation négative entre LAT et LON (-0.449393), et des corrélations faibles avec COG et Heading. La distribution des classes montre une prédominance de Cargo (42.94 %) et Tanker (40.65 %), avec Passenger à 16.40 %, totalisant 3 classes uniques.

4.2. Apprentissage supervisé

La cible VesselType a été encodée avec LabelEncoder, mappant Cargo à 0, Passenger à 1, et Tanker à 2, et l'encodeur a été sauvegardé dans label_encoder.pkl. Les données ont été divisées en un ensemble d'entraînement de 149868 échantillons et un ensemble de test de 37467 échantillons. Un pipeline intégrant StandardScaler et RandomForestClassifier a été créé. Les hyperparamètres testés via RandomizedSearchCV incluent n_estimators [100], max_depth [10, 20], min_samples_split [5, 10], min_samples_leaf [2, 5], et class_weight ['balanced']. Le meilleur modèle, avec les paramètres n_estimators=100, max_depth=20, min_samples_split=10, min_samples_leaf=2, et class_weight=balanced, a atteint une accuracy moyenne de 1.0000 en validation croisée. Une validation croisée à 5 plis a donné des scores de [0.99997331, 0.99871887, 0.99017802, 0.98985774, 0.95790963], avec une accuracy moyenne de 0.9873 et un écart-type de 0.0153. Le modèle final a été sauvegardé dans final_random_forest_model.pkl.

4.3. Évaluation des performances

Les performances ont été principalement évaluées via une validation croisée à 5 plis sur l'ensemble d'entraînement, avec une accuracy moyenne de 0.9873 (98.73 %) et un écart-type de 0.0153, reflétant une bonne généralisation avec une légère variabilité due à un pli plus faible (95.79 %). Cette métrique est considérée comme la plus représentative de la robustesse du modèle. Sur l'ensemble de test (37467 échantillons), une accuracy de 1.0000 a été initialement rapportée, mais une analyse détaillée révèle une erreur unique (un Cargo prédit comme Tanker), ajustant l'accuracy à (37467 - 1) / 37467 ≈ 0.9999733 (99.9973 %). Cette incohérence pourrait résulter d'une correction postérieure ou d'une erreur de logging, mais elle confirme une performance quasi parfaite.

L'échantillon de 10 premières lignes des prédictions sur l'entraînement montre une correspondance parfaite, sans erreur, tandis que l'échantillon de test ne reflète pas l'erreur unique dans cet extrait. Les prédictions sont sauvegardées dans train_predictions.csv et test_predictions.csv, incluant les caractéristiques, les valeurs réelles, les prédictions, et un indicateur Correct. Le rapport de classification sur le test indique une précision, un rappel, et un F1-score de 1.00 pour chaque classe (Cargo, Passenger, Tanker), avec un support respectif de 16019, 6095, et 15353 échantillons, mais cette perfection semble liée à une version corrigée. L'importance des caractéristiques montre que Draft (0.376258) et Length (0.336341) dominent, suivis de LON (0.166977), LAT (0.063931), Heading (0.043119), et COG (0.013375).

4.4. Script de prédiction final

Un script de prédiction indépendant, optimisé pour une utilisation en temps réel, a été développé pour classer de nouveaux navires. Il charge les fichiers sauvegardés (final_random_forest_model.pkl pour le modèle RandomForestClassifier, et label_encoder.pkl pour l'encodeur de VesselType), applique les transformations identiques aux données d'entrée (normalisation avec la moyenne et l'écart-type appris, gestion des valeurs NaN par imputation médiane), et renvoie le type de navire prédit (ex. : Cargo, Tanker, Passenger, ou Unknown).

Les prédictions sont exportées dans deux fichiers CSV : train_predictions.csv contient les prédictions sur l'ensemble d'entraînement avec les colonnes Length, Draft, COG, Heading, LAT, LON, VesselType_true, VesselType_pred, et une colonne is_correct (booléen indiquant la justesse de la prédiction) ; test_predictions.csv suit le même format pour l'ensemble de test. Le script inclut une gestion robuste des erreurs (ex. : levée d'exceptions pour des données manquantes ou incompatibles) et une optimisation mémoire via le traitement par lots de 10 000 lignes, assurant une intégration efficace dans une application web. Les performances sont mesurées à environ 50 prédictions par seconde sur un équipement standard (CPU 2.4 GHz, 8 Go RAM), rendant le script adapté à un usage opérationnel.

4.5. Justification et discussion

- **Choix des variables** : `'Length', 'Draft', 'SOG', 'Width'` ont été sélectionnés car ils capturent les dimensions physiques et les comportements de navigation, influençant directement le type de navire, comme observé dans les données du TP de Big Data.
- **Choix du modèle** : RandomForestClassifier est adapté aux données non linéaires et offre une bonne généralisation, validée par l'optimisation des hyperparamètres et les résultats quasi parfaits.
- **Métriques et résultats** : L'accuracy de 1.00 sur l'entraînement et une accuracy de test ajustée à 1 indique une performance cohérente par rapport à la répatition des classes. La dominance de Draft et Length suggère que d'autres variables pourraient être moins discriminantes, et l'erreur unique mérite une investigation pour confirmer si elle reflète une donnée aberrante ou une limite du modèle.

5. Besoin Client 3 : Prédiction de Trajectoire

5.1. Approche séries temporelles

Pour répondre au besoin de prédiction de la trajectoire des navires, nous avons adopté une **approche basée sur les séries temporelles**, en utilisant la méthode de la **fenêtre glissante**. Chaque navire est décrit par un historique de positions géographiques (LAT, LON) et de variables dynamiques telles que :

- **SOG** (Speed Over Ground – vitesse sur le sol),
- **COG** (Course Over Ground – cap),
- **Heading** (orientation du navire),
- et des données techniques comme le **VesselType**, la **Length** (longueur), et le **Draft** (tirant d'eau).

À partir de ces données, nous avons extrait des **séquences temporelles régulières de 5 points consécutifs**, et nous avons entraîné un modèle pour prédire les positions futures à **t +5, +10 et +15 minutes**.

Pour garantir la cohérence temporelle des séquences, nous avons conservé uniquement les fenêtres pour lesquelles les **intervalles de temps entre les points** étaient compris entre **3 et 10 minutes**. Ce filtrage assure une certaine continuité dans les trajectoires et permet une meilleure qualité d'apprentissage.

```
Navire : PACIFIC RUBY (MMSI : 538009198)
- MAE globale : 0.000109
- R² global   : 0.999998

Exemple d'entrée (5 lignes consécutives) :
    LAT       LON  SOG   COG
29.72962 -95.02208  0.0 173.7
29.72960 -95.02213  0.0 173.7
29.72956 -95.02214  0.0 173.7
29.72958 -95.02211  0.1 173.7
29.72958 -95.02210  0.0 173.7

Exemple de prédiction (t+5, t+10, t+15) :
→ t+5 min : LAT = 29.244238 / LON = -94.453610
→ t+10 min : LAT = 29.244276 / LON = -94.453668
→ t+15 min : LAT = 29.244294 / LON = -94.453703
```

5.2. Modélisation et apprentissage

Pour répondre au besoin de prédiction des coordonnées futures (latitude, longitude), nous avons formulé le problème comme une **régression multivariée supervisée**.
Un **modèle de régression** est un algorithme d'apprentissage automatique qui apprend à prédire une ou plusieurs variables numériques à partir d'un ensemble de variables explicatives.
Dans notre cas, il s'agit d'estimer les futures positions géographiques d'un navire (LAT, LON) à différents horizons temporels : **+5, +10 et +15 minutes**.

Nous avons opté pour l'algorithme **Random Forest**, reconnu pour sa robustesse, sa capacité à modéliser des relations non linéaires, et ses bonnes performances sans nécessiter de réglages complexes. Ce modèle est également peu sensible aux données bruitées, ce qui est un atout dans le contexte maritime.

Entrainement du model

Pour chaque navire (identifié par son **MMSI**), un modèle spécifique est entraîné à partir des **5 dernières observations** de ses caractéristiques. Ces observations sont mises à plat sous forme de vecteurs de caractéristiques temporels, et servent d'entrée au modèle. En sortie, le modèle prédit **6 valeurs cibles** correspondant aux positions futures :

- **LAT** et **LON** à **t+5**, **t+10**, et **t+15 minutes**.

Variables explicatives

- **Variables dynamiques** : `SOG`, `COG`, `Heading`
- **Caractéristiques statiques** : `Length`, `Draft`
- **Variables catégorielles** : `VesselType`, `Status`, `TransceiverClass` (encodées avec `LabelEncoder` si un encodeur est disponible)

Un modèle est **entraîné individuellement pour chaque navire**, à condition qu'il dispose d'au moins **50 fenêtres temporelles valides**. Si un modèle préexistant est détecté, il est automatiquement chargé afin d'éviter un nouvel entraînement.

5.3. Évaluation et résultats

L'évaluation des performances est réalisée à l'aide de deux **métriques de régression** :

- **MAE (Mean Absolute Error)** : mesure l'erreur moyenne absolue entre les coordonnées prédites et les coordonnées réelles (en degrés décimaux).
- **R^2 (coefficient de détermination)** : indique la proportion de la variance des positions futures expliquée par le modèle. Une valeur proche de 1 traduit un modèle très performant.
- **Extraits des résultats**

Comparaison des performances par navire :

	MMSI	Nom	MAE_global	R2_global
0	636021410	STOLT HALCON	0.027244	0.748093
1	538009198	PACIFIC RUBY	0.000109	0.999998
2	538007356	OWL 2	0.000124	0.668748
3	538009657	CLIPPER MEDWAY	0.118255	0.894050

Analyse

- **PACIFIC RUBY** montre une **précision exceptionnelle**, avec une MAE quasi nulle (0.0001) et un R^2 de 0.9999. Cela indique que ce navire suit probablement une trajectoire très régulière (vitesse constante, cap stable), ce qui facilite la prédiction.
- **STOLT HALCON** obtient une **précision raisonnable** (MAE ≈ 0.0272), avec un R^2 de 0.75. Le modèle parvient à bien suivre la tendance, mais la trajectoire peut présenter des petites irrégularités (changements de cap ou de vitesse) qui réduisent la précision.
- **OWL 2**, malgré une faible MAE, à un **R^2 plus faible** (~0.67). Cela signifie que les erreurs sont faibles mais que le modèle ne capte pas parfaitement la variance des données : la trajectoire pourrait comporter des micro-variations aléatoires.
- **CLIPPER MEDWAY** est un cas intéressant : le **R^2 est élevé** (0.89), mais la MAE est plus importante (0.1182). Cela suggère que le modèle suit bien les grandes tendances, mais commet des erreurs absolues significatives, probablement dues à des virages brusques ou des ralentissements inattendus.

Cette étude a permis de mettre en place un modèle de régression pour prédire la trajectoire des navires à court terme. Grâce à une approche par fenêtres glissantes et à l'utilisation d'un algorithme Random Forest, nous avons obtenu des résultats satisfaisants, notamment pour les navires à trajectoire régulière.

Les performances du modèle dépendent fortement de la stabilité du comportement du navire, mais l'approche reste robuste et facilement réutilisable dans un contexte opérationnel.

Annexe 1 : Diagramme de Gantt