# TP2 — Analyse des messages ADS-B (MSG)

## 1. Préparation du jeu d'apprentissage

On souhaite analyser un avion particulier (ICAO) disposant d'au moins 500 points valides. À partir de sa trajectoire, répondre aux questions suivantes :

### 1.a — Extraction des variables dynamiques

Pour cet avion, calculer les grandeurs suivantes entre deux messages consécutifs :

1. L'accélération instantanée (variation de la vitesse sol).
2. La variation de cap (heading).
3. La vitesse verticale en ft/min.
4. La distance parcourue (méthode Haversine recommandée).

Justifier brièvement l'intérêt physique de chacune de ces grandeurs pour la modélisation de l'évolution future de l'altitude.

### 1.b — Horodatage

Expliquer pourquoi il est indispensable de convertir les champs *date* et *heure* en un timestamp datetime unique avant toute manipulation temporelle.

### 1.c — Construction du label

Construire la variable cible :

altitude_future_t+10s\text{altitude_future_t+10s}altitude_future_t+10s

Expliquer clairement :

- la méthode utilisée pour aligner les données à t + 10 s,
- comment les points sans correspondance temporelle sont gérés (interpolation, suppression ou autre).

## 2. Modélisation supervisée

En utilisant le jeu d'apprentissage préparé :

### 2.a — Régression Linéaire

Entraîner un modèle de régression linéaire pour prédire l'altitude dans 10 secondes.

### 2.b — Random Forest

Entraîner un modèle de Random Forest pour la même tâche.

### 2.c — Comparaison numérique des performances

Comparer les deux modèles à l'aide des métriques suivantes :

- RMSE
- MAE
- Coefficient de détermination $R^2$
- Analyse visuelle des résidus (différence $y - \hat{y}$)

Discuter la signification de ces métriques dans le contexte d'un vol réel.

## 3. Visualisation et analyse graphique des performances

Les deux modèles entraînés (Régression Linéaire et Random Forest) doivent maintenant être comparés visuellement.

### 3.a — Comparaison visuelle des valeurs réelles et prédites

Tracer sur une même figure :

- la valeur réelle de l'altitude future (t+10s),
- la prédiction du modèle Régression Linéaire,
- la prédiction du Random Forest.

Questions d'analyse :

3.a.1. Quel modèle reproduit le mieux la dynamique réelle de l'altitude ?
3.a.2. Dans quelles zones (segments temporels) les écarts sont-ils les plus importants ?
3.a.3. La Régression Linéaire présente-t-elle un comportement "rigide" ? Justifier.
3.a.4. Le Random Forest semble-t-il mieux capter les variations brusques ? Pourquoi ?

## 3.b — Analyse des erreurs : erreur absolue et erreur quadratique

Tracer, pour les deux modèles :

- la courbe de l'erreur absolue $|y - \hat{y}|$,
- la courbe de l'erreur quadratique $(y - \hat{y})^2$.

Questions d'analyse :

3.b.1. Durant quelles phases du vol observe-t-on les erreurs les plus élevées ?
3.b.2. Comparer la stabilité temporelle des erreurs entre les deux modèles.
3.b.3. Les erreurs semblent-elles corrélées à des événements particuliers (montées rapides, descentes, virages serrés) ?
3.b.4. Globalement, quel modèle commet le moins d'erreurs ?
3.b.5. Lequel gère le mieux les changements rapides d'altitude ?

## 3.c — Diagramme des résidus

Tracer, pour chaque modèle, un nuage de points :

- en abscisse : $\hat{y}$ (prédiction),
- en ordonnée : $y - \hat{y}$ (résidu).

Questions :

3.c.1. Les résidus sont-ils globalement centrés autour de 0 ?
3.c.2. Observe-t-on un biais systématique (prédictions trop hautes ou trop basses) ?
3.c.3. Quel modèle présente la dispersion la plus faible ?
3.c.4. Comment ce nuage illustre-t-il les limites structurelles d'un modèle linéaire ?

## 3.d — Histogramme des erreurs

Tracer un histogramme (ou KDE) des erreurs absolues pour les deux modèles.

Questions :

3.d.1. Quel modèle présente la distribution la plus concentrée autour de 0 ?
3.d.2. Quel modèle génère les erreurs extrêmes les plus fréquentes ?
3.d.3. Expliquer physiquement la présence éventuelle de grosses erreurs :

- bruit ADS-B,
- virages serrés,
- turbulences,
- montée ou descente rapide,
- pertes temporaires de précision.

### 3.e — Courbes d'apprentissage (Random Forest uniquement)

Tracer, pour différentes tailles de dataset :

- l'erreur d'entraînement,
- l'erreur de validation.

Questions :

3.e.1. Le modèle présente-t-il des signes de surapprentissage ?
3.e.2. Les courbes convergent-elles vers une erreur stable ?
3.e.3. Le Random Forest semble-t-il particulièrement dépendant du volume de données ?
3.e.4. Comparer brièvement ce comportement à celui attendu pour un modèle linéaire.

### 3.f — Importance des variables (Random Forest)

Tracer un graphique en barres représentant l'importance relative des variables.

Questions :

3.f.1. Quelles sont les variables les plus influentes dans la prédiction de l'altitude future ?
3.f.2. Les variables dynamiques (accélération, variation de cap, variation d'altitude) dominent-elles les variables absolues ?

3.f.3. Expliquer pourquoi la Régression Linéaire ne permet pas une interprétation directe comparable.

## 4. Analyse et interprétation finale

### 4.a — Choix du meilleur modèle et justification

Quel modèle apparaît le plus adapté pour prédire l'altitude future ? Justifier clairement.

### 4.b — Localisation des erreurs importantes

Identifier les segments de trajectoire où les modèles échouent le plus.

### 4.c — Explication physique des erreurs

Proposer au moins deux explications plausibles basées sur :

- phénomènes météorologiques,
- virages serrés,
- phases de montée ou de descente,
- bruit ADS-B ou pertes de précision.