Projet Data Science

Analyse Prédictive des Vols — AirNova

✈ **AirNova**
Master 1 Intermédiaire
Durée : 3 semaines

Individuel ou binôme | Notebook Jupyter commenté | 100 points

ℹ️ **Contexte métier**
AirNova est une compagnie aérienne régionale européenne. Vous êtes mandaté comme Data Scientist junior pour analyser l'historique de 300 vols et construire un modèle prédictif de retard. Votre travail servira à la direction opérationnelle pour prendre des décisions d'optimisation.

ℹ️ **Fichiers fournis**
Deux fichiers de données sont disponibles :
• airnova_flights.csv — source principale (300 lignes, 27 colonnes)
• airnova_flights.json — source secondaire (même données, format JSON)

Dictionnaire des variables

Le dataset contient les 20 variables suivantes. Référez-vous à ce tableau tout au long du projet.

Variable	Type	Description
`flight_id`	str	Identifiant unique du vol
`flight_number`	str	Numéro de vol commercial
`aircraft_type`	catégorielle	Type d'appareil (5 modèles)
`origin_code / destination_code`	str	Code IATA aéroport
`origin_city / destination_city`	str	Ville et pays
`scheduled_departure / arrival`	datetime (str)	Horaires programmés
`flight_duration_min`	int	Durée du vol en minutes
`status`	catégorielle	À l'heure / Retardé / Annulé / En avance
`delay_minutes`	int	Retard en minutes (négatif = en avance)
`delay_reason`	catégorielle	Météo / Technique / Trafic / Opérationnel
`passengers / seat_capacity`	int	Passagers embarqués et capacité
`load_factor_pct`	float	Taux de remplissage (%)
`cabin_class`	catégorielle	Économique / Premium / Affaires
`ticket_price_eur`	float	Prix moyen du billet (€)

`fuel_consumed_kg`	float	Carburant consommé (kg)
`distance_km`	float	Distance estimée (km)
`is_international`	binaire	1 = international, 0 = domestique
`season`	catégorielle	Hiver / Printemps / Été / Automne
`day_of_week`	catégorielle	Jour de la semaine
`is_weekend`	binaire	1 = samedi ou dimanche

Partie 1 — Chargement et prétraitement (20 pts)

Livrables : cellules de code + justifications rédigées en Markdown pour chaque choix.

1.1 Ingestion des données

Chargez les deux fichiers de données avec les bibliothèques appropriées.

1. Chargez airnova_flights.csv dans un DataFrame pandas. Affichez les 5 premières lignes et la forme du dataset.
2. Chargez airnova_flights.json. Vérifiez que les deux sources contiennent le même nombre d'enregistrements.
3. Fusionnez les deux DataFrames sur la clé primaire flight_id. Justifiez le type de jointure choisi (inner, left, outer).

1.2 Audit de qualité

4. Identifiez les colonnes contenant des valeurs manquantes. Pour chacune, proposez et appliquez une stratégie de traitement adaptée à la nature de la colonne, et justifiez votre choix.
5. Vérifiez les types de chaque colonne. Convertissez les colonnes de dates en type datetime. Expliquez pourquoi cette conversion est importante pour la suite.
6. Repérez au moins deux incohérences sémantiques dans les données (exemples : valeurs impossibles, contradictions entre colonnes). Décrivez-les et corrigez-les.

> ⚠️ **Point de vigilance**
> Une incohérence sémantique n'est pas une valeur manquante. C'est une valeur présente mais logiquement impossible ou contradictoire avec une autre colonne.

1.3 Feature engineering

Créez au minimum 4 nouvelles variables pertinentes pour la prédiction. Pour chacune :

- Donnez son nom et sa formule ou sa règle de construction
- Justifiez son intérêt métier pour AirNova
- Précisez son type (numérique, binaire, catégorielle)

Parmi vos 4 variables, vous devez inclure obligatoirement :

- Une variable temporelle (issue de scheduled_departure)
- Une variable binaire cible is_delayed indiquant si un vol a subi un retard

Partie 2 — Statistiques descriptives (20 pts)

Pour chaque résultat statistique, rédigez une interprétation en langage clair dans une cellule Markdown.

2.1 Statistiques univariées

7. Calculez les indicateurs de tendance centrale et de dispersion (moyenne, médiane, écart-type, min, max) pour les variables delay_minutes, ticket_price_eur, load_factor_pct et flight_duration_min.
8. Analysez la distribution de delay_minutes : est-elle symétrique ? Quelle en est l'asymétrie ? Que signifie cela pour AirNova ?
9. Calculez la fréquence de chaque modalité de status. Quel statut est le plus rare ? Quel problème cela peut-il poser pour la modélisation ?

2.2 Statistiques bivariées

10. Calculez la matrice de corrélation des variables numériques. Identifiez les deux paires les plus corrélées et expliquez pourquoi cette corrélation est logique d'un point de vue métier.
11. Comparez le retard moyen selon la saison, le jour de la semaine et le type d'appareil. Rédigez une synthèse de 5 à 8 lignes sur ce que ces comparaisons révèlent.
12. Y a-t-il une relation entre le taux de remplissage (load_factor_pct) et le retard ? Justifiez votre réponse par un calcul.

Partie 3 — Visualisation (20 pts)

Chaque graphique doit avoir un titre, des légendes et être suivi d'un commentaire Markdown d'au moins 3 lignes.

Produisez les 5 graphiques suivants. Pour chacun, choisissez le type de graphique le plus adapté et justifiez ce choix.

Graphique 1 — Distribution des retards

Visualisez la distribution de delay_minutes pour les vols retardés uniquement. Mettez en évidence la moyenne et la médiane sur le graphique. Qu'observe-t-on sur la forme de la distribution ?

Graphique 2 — Retard selon la saison et le statut

Comparez simultanément le retard moyen et la répartition des statuts de vols (À l'heure, Retardé, Annulé, En avance) pour chaque saison. Quelle saison pose le plus de problèmes à AirNova ?

Graphique 3 — Relation prix / taux de remplissage

Visualisez la relation entre ticket_price_eur et load_factor_pct. Différenciez les points par classe cabine. Quelle conclusion peut-on tirer sur la stratégie tarifaire d'AirNova ?

Graphique 4 — Carte thermique de corrélation

Produisez une heatmap annotée de la matrice de corrélation des variables numériques. Identifiez visuellement les zones de forte corrélation positive et négative et expliquez leur signification.

Graphique 5 — Graphique libre au choix

Produisez un graphique de votre choix qui apporte une information nouvelle non visible dans les 4 graphiques précédents. Justifiez le choix du type de graphique et la question à laquelle il répond.

Partie 4 — Machine Learning (30 pts)

Variable cible : is_delayed (classification binaire — 0 = non retardé, 1 = retardé)

> ⚠️ **Point de vigilance**
> Veillez à ne jamais inclure dans vos features des variables qui 'trichent' : delay_minutes, delay_reason ou cancellation_reason ne seraient pas disponibles avant le départ du vol.

4.1 Préparation des données

13. Sélectionnez et justifiez au minimum 8 variables pertinentes comme features. Expliquez pourquoi chacune pourrait contribuer à la prédiction d'un retard.
14. Gérez les variables catégorielles : encodez-les de manière appropriée. Expliquez la différence entre Label Encoding et One-Hot Encoding, et justifiez votre choix pour chaque colonne.
15. Divisez les données en ensemble d'entraînement (80%) et de test (20%). Pourquoi utilise-t-on deux ensembles séparés ? Qu'est-ce que le data leakage et comment l'évite-t-on ici ?

4.2 Entraînement et comparaison de modèles

Entraînez les trois modèles suivants sur les données d'entraînement :

- Arbre de Décision
- Random Forest
- Régression Logistique

16. Pour chaque modèle, calculez et présentez dans un tableau synthétique : Accuracy, Precision, Recall et F1-score.
17. Affichez la matrice de confusion du meilleur modèle. Identifiez les types d'erreurs (faux positifs, faux négatifs) et expliquez lequel est le plus coûteux pour AirNova et pourquoi.
18. Analysez l'importance des variables du meilleur modèle. Quelles sont les 3 features les plus influentes ? Est-ce cohérent avec vos observations en Parties 2 et 3 ?

4.3 Amélioration du modèle

19. Appliquez au moins une technique pour améliorer les performances : ajustement des hyperparamètres, gestion du déséquilibre de classes, ou ingénierie de features supplémentaires. Mesurez l'impact sur les métriques.
20. Réalisez une validation croisée à 5 plis (k=5) sur le meilleur modèle. Quelle est la moyenne et l'écart-type du F1-score ? Que révèle l'écart-type sur la stabilité du modèle ?

Partie 5 — Synthèse et recommandations (10 pts)

Rédigez une synthèse finale de 400 à 600 mots dans une cellule Markdown. Elle doit impérativement couvrir les 4 points suivants :

21. Profil type du vol retardé : décrivez en quelques phrases le type de vol le plus susceptible d'être retardé chez AirNova, en vous appuyant sur vos analyses.
22. Bilan des modèles : comparez les trois modèles, expliquez lequel vous recommandez pour AirNova et pourquoi — en termes métier, pas seulement en termes de métriques.
23. Recommandations opérationnelles : formulez 3 recommandations concrètes et argumentées pour qu'AirNova réduise ses retards, basées sur vos résultats.
24. Limites de l'analyse : identifiez au moins 2 limites de votre étude (taille des données, biais potentiels, variables manquantes...) et proposez comment les surmonter avec plus de ressources.

Grille d'évaluation

Partie	Critère principal	Points
1 – Chargement & prétraitement	Données propres, types corrects, justifications	20
2 – Statistiques descriptives	Calculs justes, interprétations rédigées	20
3 – Visualisation	Graphiques appropriés, titrés, commentés	20
4 – Machine Learning	Pipeline complet, métriques calculées et interprétées	30
5 – Synthèse	Conclusions claires, recommandations argumentées	10
TOTAL		**100**

ℹ️ **Critères transversaux évalués dans toutes les parties**

- Qualité des commentaires Markdown — chaque décision est expliquée en langage clair
- Cohérence entre les parties — vos conclusions ML concordent avec vos analyses descriptives
- Reproductibilité — le notebook s'exécute de A à Z sans erreur (Kernel > Restart & Run All)
- Honnêteté scientifique — les limites et erreurs sont identifiées, pas cachées