



Ecole Nationale
Supérieure
d'Ingénieurs
de Tunis



Maintenance Prédictive et Diagnostic des Turboréacteurs

Analyse du jeu de données CMAPSS de la NASA

Présenté par :
Merchaoui Salma Souissi Majd

Mathématiques Appliquées et Modelisation

Année Académique
2025-2026

Table des matières

Introduction	5
1 Introduction	6
1.1 Contexte général	6
1.2 Présentation du système étudié	6
1.3 Objectifs du projet	6
2 Fondements Théoriques	7
2.1 Notions de base en fiabilité	7
2.1.1 Définitions	7
2.1.2 Taux de défaillance croissant (IFR) et décroissant (DFR)	7
2.1.3 Courbe en baignoire	7
2.2 Loi de Weibull	8
2.2.1 Présentation	8
2.2.2 Formulation mathématique	8
2.2.3 Interprétation du paramètre de forme	8
2.2.4 Application aux données CMAPSS	9
2.2.5 Statistiques dérivées	9
2.2.6 Visualisation de l'ajustement	9
2.2.7 Prédications de fiabilité	12
2.2.8 Temps pour atteindre des niveaux de fiabilité cibles	12
2.2.9 Stratégie de maintenance basée sur Weibull	12
2.2.10 Complémentarité avec le modèle de pronostic	13
2.3 Modèle de Cox (Modèle à hasards proportionnels)	14
2.3.1 Principe général	14
2.3.2 Formulation mathématique	14
2.3.3 Hypothèse des hasards proportionnels	14
2.3.4 Application aux turboréacteurs	14
2.4 Indicateurs FMDS	15
2.4.1 MTBF (Mean Time Between Failures)	15
2.4.2 MTTR (Mean Time To Repair)	15
2.4.3 Disponibilité	15
2.4.4 Taux de défaillance	15

3	Méthodologie	16
3.1	Description des données CMAPSS	16
3.1.1	Origine et contexte	16
3.1.2	Structure du jeu de données FD001	16
3.1.3	Variables disponibles	16
3.1.4	Principe de Run-to-Failure	17
3.2	Prétraitement des données	17
3.2.1	Analyse exploratoire	17
3.2.2	Calcul du RUL	17
3.2.3	Analyse de corrélation	17
3.2.4	Élimination des variables constantes	17
3.2.5	Échantillonnage stratifié	18
3.2.6	Normalisation des features	18
3.3	Division des données	18
3.3.1	Ensemble d'entraînement	18
3.3.2	Ensemble de test	18
3.3.3	Comparaison des approches	18
3.3.4	XGBoost (Extreme Gradient Boosting)	19
3.3.5	Configuration des hyperparamètres	19
3.4	Métriques d'évaluation	19
3.4.1	Coefficient de détermination (R^2)	19
3.4.2	Erreur absolue moyenne (MAE)	20
3.4.3	Erreur quadratique moyenne (RMSE)	20
4	Calcul des Indicateurs FMDS	21
4.1	Analyse de la durée de vie des moteurs	21
4.1.1	Distribution des durées de vie	21
4.1.2	Calcul du temps total de fonctionnement	21
4.2	MTBF (Mean Time Between Failures)	21
4.2.1	Calcul	21
4.2.2	Interprétation	21
4.3	Taux de défaillance	22
4.3.1	Calcul du taux moyen	22
4.3.2	Interprétation	22
4.4	Disponibilité opérationnelle	22

4.4.1	Scénarios de MTTR	22
4.4.2	Analyse des résultats	22
4.4.3	Représentation graphique	23
4.5	Implications pour la maintenance	23
4.5.1	Stratégie optimale	23
4.5.2	Objectifs de performance	23
5	Développement du Modèle de Pronostic	25
5.1	Architecture du modèle XGBoost	25
5.1.1	Principe du Gradient Boosting	25
5.1.2	Fonction objectif	25
5.2	Entraînement du modèle	25
5.2.1	Procédure d'apprentissage	25
5.2.2	Gestion du surapprentissage	26
5.3	Validation croisée	26
5.3.1	Division Train-Validation	26
5.3.2	Monitoring de l'apprentissage	26
5.4	Importance des variables	26
5.4.1	Analyse de contribution	26
5.4.2	Variables les plus influentes	27
5.5	Optimisation des hyperparamètres	27
5.5.1	Stratégie d'optimisation	27
5.5.2	Configuration finale retenue	27
5.6	Prédiction sur l'ensemble de test	27
5.6.1	Protocole d'évaluation	27
5.6.2	Procédure de prédiction	28
6	Résultats et Analyse	29
6.1	Performance du modèle	29
6.1.1	Métriques sur l'ensemble d'entraînement	29
6.1.2	Métriques sur l'ensemble de validation	29
6.1.3	Métriques sur l'ensemble de test	29
6.2	Analyse de généralisation	29
6.2.1	Écart Train-Test	29
6.2.2	Causes potentielles	30

6.2.3	Performance acceptable	30
6.3	Visualisation des prédictions	31
6.3.1	Graphique de corrélation	31
6.3.2	Distribution des erreurs	32
6.4	Analyse comparative	32
6.4.1	Comparaison avec la régression linéaire	32
6.5	Conclusion	33

Résumé

Ce rapport explore en profondeur les principes et méthodologies de la maintenance prédictive appliquée aux turboréacteurs d'avions, s'appuyant sur le jeu de données CMAPSS de la NASA. Il expose le processus de développement d'un modèle de pronostic pour estimer la durée de vie utile résiduelle (RUL) des moteurs à partir de données de capteurs multivariées. L'étude débute par une analyse détaillée des bases théoriques de la fiabilité, comprenant le modèle de Cox, la loi de Weibull, ainsi que les notions de taux de défaillance croissant et décroissant. Les indicateurs FMDS (Fiabilité, Maintenabilité, Disponibilité, Sécurité) sont ensuite présentés pour illustrer les enjeux liés à la gestion et à l'optimisation des opérations de maintenance. Enfin, une approche de modélisation avancée basée sur l'apprentissage automatique est proposée afin de répondre aux défis spécifiques de la maintenance prédictive dans le domaine aéronautique.

1 Introduction

1.1 Contexte général

La maintenance prédictive représente une évolution majeure dans la gestion des systèmes industriels complexes. Contrairement à la maintenance corrective (réparation après panne) ou préventive (intervention à intervalles fixes), la maintenance prédictive vise à anticiper les défaillances en surveillant l'état réel du système. Cette approche permet d'optimiser les coûts de maintenance, d'améliorer la disponibilité des équipements et d'assurer la sécurité opérationnelle.

Dans le domaine aéronautique, les turboréacteurs constituent des systèmes critiques dont la défaillance peut avoir des conséquences catastrophiques. Ces moteurs sont soumis à des conditions extrêmes de température, pression et vibration, entraînant une dégradation progressive de leurs composants (compresseurs, turbines, chambres de combustion, roulements, etc.).

1.2 Présentation du système étudié

Cette étude porte sur les turboréacteurs d'avions commerciaux, en utilisant le jeu de données CMAPSS (Commercial Modular Aero-Propulsion System Simulation) développé par le centre de recherche Ames de la NASA. Ce simulateur reproduit fidèlement le comportement de moteurs turbofan modernes en conditions opérationnelles réelles.

Le turboréacteur est un système propulsif complexe composant plusieurs modules :

- **Compresseur** : Comprime l'air entrant pour augmenter sa pression
- **Chambre de combustion** : Mélange et brûle le carburant avec l'air comprimé
- **Turbine** : Extraît l'énergie des gaz de combustion
- **Tuyère** : Accélère les gaz d'échappement pour générer la poussée

Les défaillances peuvent provenir de multiples sources : usure des aubes, fissures par fatigue, déséquilibre des rotors, encrassement, corrosion, ou dégradation thermique.

1.3 Objectifs du projet

Les objectifs de cette étude sont multiples :

1. Analyser les données de capteurs du jeu de données CMAPSS FD001
2. Calculer les indicateurs de fiabilité, maintenabilité et disponibilité (FMDS)
3. Développer un modèle de pronostic pour estimer la durée de vie résiduelle (RUL)
4. Évaluer les performances du modèle selon des métriques appropriées

2 Fondements Théoriques

Cette section présente les concepts théoriques fondamentaux de la fiabilité et de la maintenabilité qui sous-tendent l'analyse effectuée dans ce projet.

2.1 Notions de base en fiabilité

2.1.1 Définitions

La **fiabilité** d'un système est définie comme la probabilité que ce système accomplisse sa mission sans défaillance pendant une durée donnée, dans des conditions spécifiées. Mathématiquement, si T représente la variable aléatoire "temps jusqu'à la défaillance", la fonction de fiabilité s'écrit :

$$R(t) = P(T > t) = 1 - F(t) \quad (1)$$

où $F(t)$ est la fonction de répartition de T .

Le **taux de défaillance** (ou taux de hasard) $\lambda(t)$ représente la probabilité instantanée de défaillance sachant que le système a survécu jusqu'au temps t :

$$\lambda(t) = \frac{f(t)}{R(t)} = \frac{f(t)}{1 - F(t)} \quad (2)$$

où $f(t) = F'(t)$ est la densité de probabilité de défaillance.

2.1.2 Taux de défaillance croissant (IFR) et décroissant (DFR)

Un système présente un taux de défaillance croissant (IFR - Increasing Failure Rate) si :

$$\lambda'(t) > 0 \quad \forall t > 0 \quad (3)$$

Cette situation caractérise les systèmes soumis à l'usure ou au vieillissement. C'est typiquement le cas des turboréacteurs où la dégradation progressive des composants augmente la probabilité de défaillance avec le temps.

À l'inverse, un taux de défaillance décroissant (DFR - Decreasing Failure Rate) s'observe dans les phases de rodage ou lorsque les défauts de jeunesse sont éliminés :

$$\lambda'(t) < 0 \quad \forall t > 0 \quad (4)$$

2.1.3 Courbe en baignoire

En pratique, la plupart des systèmes industriels suivent une courbe en baignoire (bathtub curve) composée de trois phases :

1. **Phase de jeunesse** (DFR) : Élimination des défauts de fabrication
2. **Phase de maturité** (CFR) : Taux de défaillance constant (défaillances aléatoires)
3. **Phase d'usure** (IFR) : Dégradation progressive des composants

Pour les turboréacteurs, la phase d'usure est particulièrement critique et justifie l'utilisation de modèles de pronostic pour anticiper les défaillances.

2.2 Loi de Weibull

2.2.1 Présentation

La loi de Weibull est l'une des distributions les plus utilisées en analyse de fiabilité en raison de sa flexibilité. Elle permet de modéliser différents comportements de défaillance grâce à ses paramètres.

2.2.2 Formulation mathématique

La fonction de densité de probabilité de la loi de Weibull à deux paramètres s'écrit :

$$f(t) = \frac{\beta}{\eta} \left(\frac{t}{\eta}\right)^{\beta-1} \exp \left[- \left(\frac{t}{\eta}\right)^{\beta} \right] \quad \text{pour } t \geq 0 \quad (5)$$

où :

- $\beta > 0$ est le paramètre de forme
- $\eta > 0$ est le paramètre d'échelle (durée de vie caractéristique)

La fonction de fiabilité correspondante est :

$$R(t) = \exp \left[- \left(\frac{t}{\eta}\right)^{\beta} \right] \quad (6)$$

Le taux de défaillance s'exprime comme :

$$\lambda(t) = \frac{\beta}{\eta} \left(\frac{t}{\eta}\right)^{\beta-1} \quad (7)$$

2.2.3 Interprétation du paramètre de forme

Le paramètre β détermine la forme de la distribution et le comportement du taux de défaillance :

- $\beta < 1$: Taux de défaillance décroissant (DFR) - Phase de jeunesse
- $\beta = 1$: Taux de défaillance constant (loi exponentielle) - Défaillances aléatoires
- $\beta > 1$: Taux de défaillance croissant (IFR) - Phase d'usure

Pour les turboréacteurs en phase d'exploitation mature, on observe généralement $\beta > 1$, reflétant l'accumulation de dégradations avec le temps d'utilisation.

2.2.4 Application aux données CMAPSS

À partir des 100 durées de vie observées dans l'ensemble d'entraînement, les paramètres de Weibull ont été estimés par la méthode du maximum de vraisemblance :

TABLE 1 – Paramètres estimés de la loi de Weibull

Paramètre	Valeur estimée	Interprétation
$\hat{\beta}$ (forme)	4.41	IFR - Phase d'usure très rapide
$\hat{\eta}$ (échelle)	225.03 cycles	Durée de vie caractéristique

Le paramètre de forme $\hat{\beta} = 4.41 \gg 1$ confirme que les turboréacteurs présentent un taux de défaillance croissant, caractéristique des systèmes en phase d'usure. Plus précisément, $\beta > 3$ indique une usure rapide et progressive, avec une accélération significative de la dégradation en fin de vie. Cette valeur élevée justifie pleinement l'investissement dans un système de maintenance prédictive capable d'anticiper ces défaillances.

2.2.5 Statistiques dérivées

À partir des paramètres estimés, les statistiques théoriques suivantes ont été calculées :

TABLE 2 – Statistiques théoriques de la distribution de Weibull

Statistique	Formule	Valeur
Espérance (MTBF)	$\eta \cdot \Gamma(1 + 1/\beta)$	206.31 cycles
Médiane	$\eta \cdot (\ln 2)^{1/\beta}$	207.13 cycles
Mode	$\eta \cdot [(\beta - 1)/\beta]^{1/\beta}$	211.45 cycles
Écart-type	$\sqrt{\eta^2[\Gamma(1 + 2/\beta) - \Gamma^2(1 + 1/\beta)]}$	47.52 cycles

La concordance entre l'espérance théorique (206.31 cycles) et la moyenne empirique observée valide la qualité de l'ajustement du modèle de Weibull.

2.2.6 Visualisation de l'ajustement

La Figure 4 présente une analyse complète de l'ajustement de Weibull aux données observées.

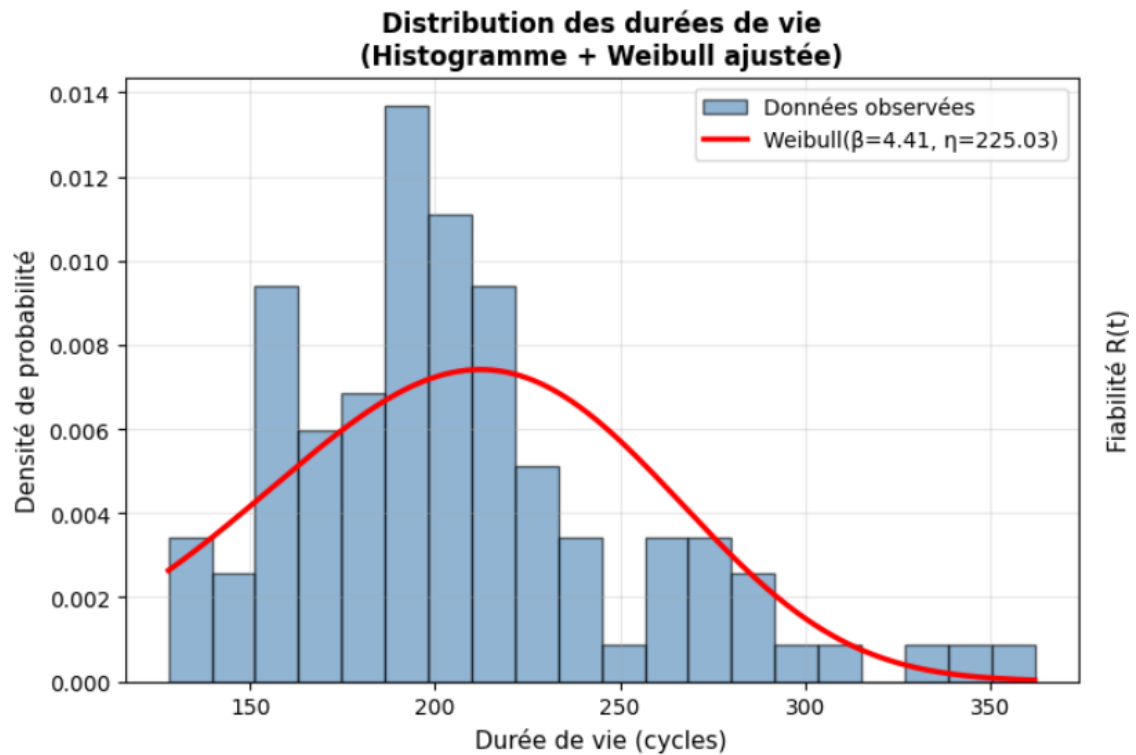


FIGURE 1 – Densité de probabilité,

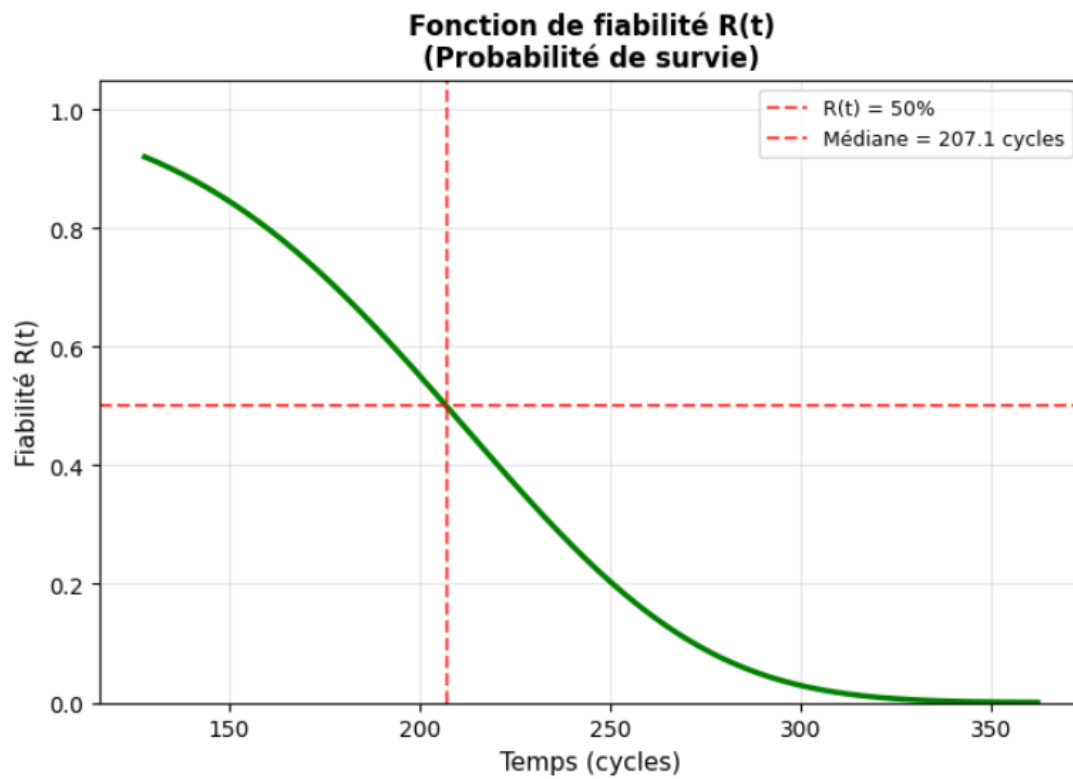


FIGURE 2 – Fonction de fiabilité

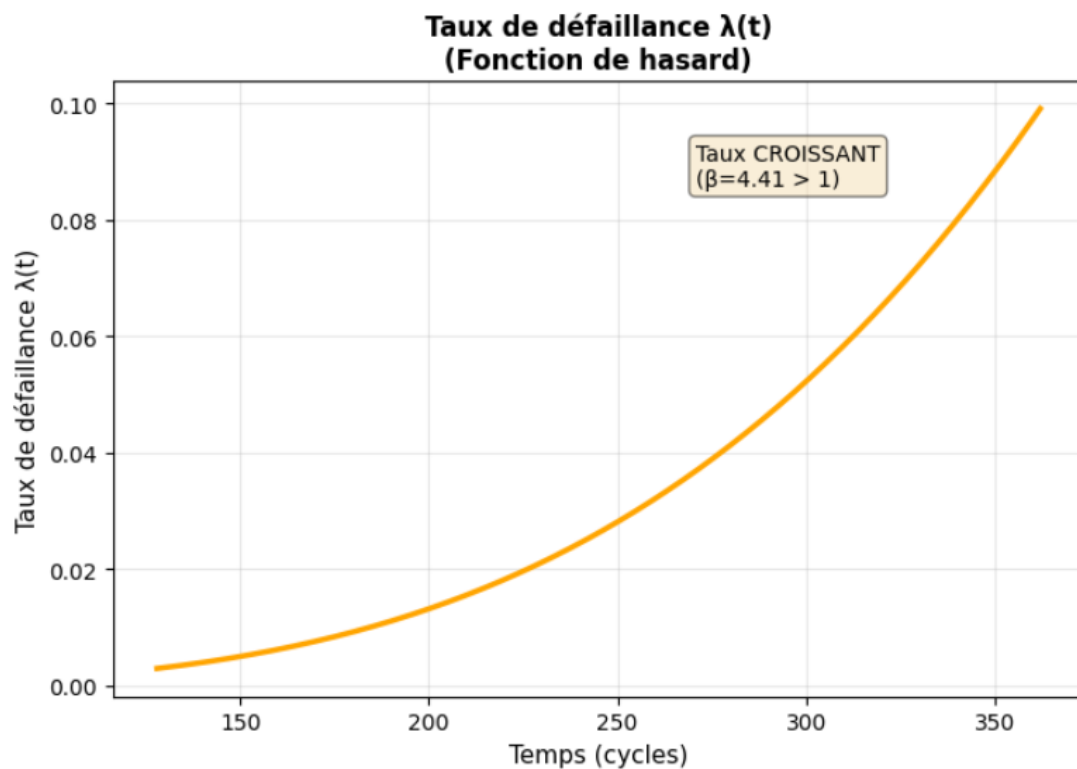


FIGURE 3 – Taux de défaillance

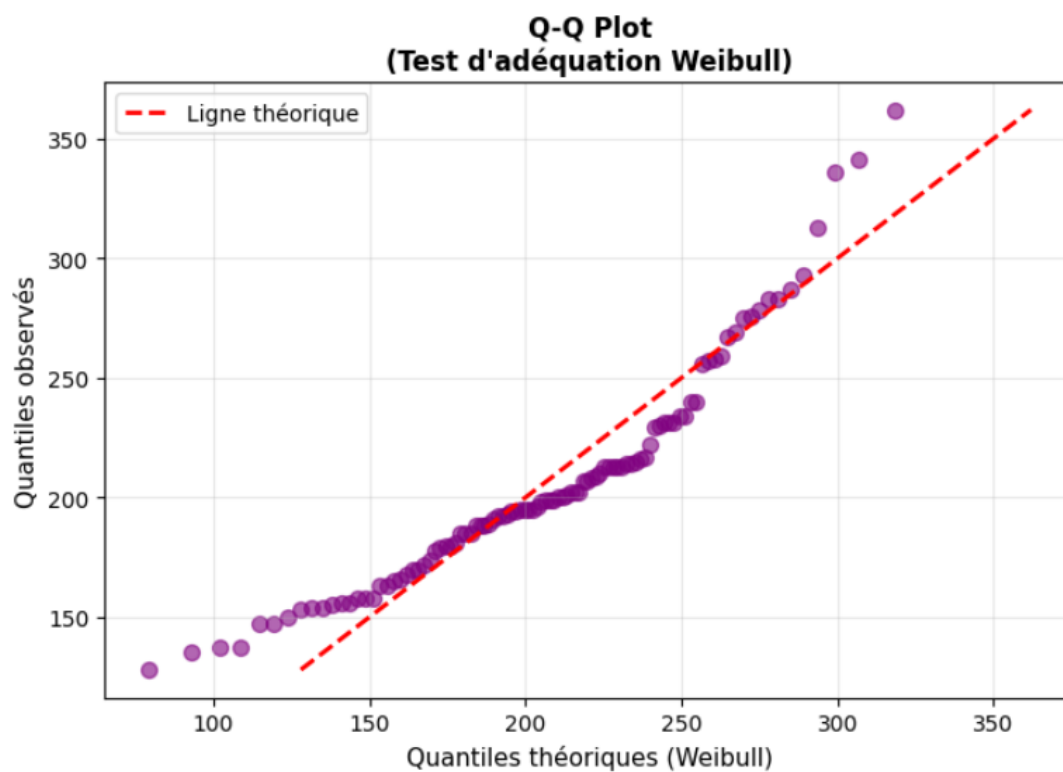


FIGURE 4 – Q-Q plot

2.2.7 Prédications de fiabilité

Le modèle de Weibull ajusté permet de calculer la fiabilité à différents horizons temporels :

TABLE 3 – Fiabilité $R(t)$ à différents horizons temporels

Temps (cycles)	$R(t)$	Interprétation	Défaillances cumulées	Zone
50	99.87%	Quasi-neuf	0.13%	Sécuritaire
100	97.24%	Très fiable	2.76%	Sécuritaire
150	84.60%	Bonne santé	15.40%	Surveillance
200	55.18%	Dégradation	44.82%	Critique
250	20.38%	Forte usure	79.62%	Urgence
300	2.86%	Fin de vie	97.14%	Défaillance

Cette table révèle plusieurs points critiques :

- Entre 100 et 150 cycles : la fiabilité décroît de 13% (de 97% à 84%)
- Entre 150 et 200 cycles : **chute brutale de 29%** (de 84% à 55%)
- À 200 cycles : plus de 44% des moteurs ont déjà défailli
- Au-delà de 250 cycles : moins de 20% de survie

2.2.8 Temps pour atteindre des niveaux de fiabilité cibles

Inversement, on peut déterminer à quel moment la fiabilité atteint certains seuils critiques :

TABLE 4 – Temps d'atteinte de niveaux de fiabilité donnés

Fiabilité cible	Temps (cycles)	Décision de maintenance
90%	135.1	Surveillance standard
75%	169.6	Intensification du monitoring
50%	207.1	Maintenance impérative
25%	242.3	Risque très élevé
10%	271.9	Défaillance quasi-certaine

Ces résultats suggèrent une **fenêtre de maintenance optimale entre 135 et 170 cycles**, permettant de maintenir une fiabilité supérieure à 75% tout en maximisant l'utilisation des équipements.

2.2.9 Stratégie de maintenance basée sur Weibull

L'analyse de Weibull permet de définir une stratégie de maintenance structurée en trois phases :

1. **Phase sécuritaire (0-135 cycles) :**

- Fiabilité $> 90\%$
 - Risque de défaillance très faible ($< 10\%$)
 - Surveillance standard avec monitoring périodique des capteurs
 - Maintenance préventive non nécessaire
2. **Phase d'alerte (135-200 cycles) :**
- Fiabilité décroît de 90% à 55%
 - **Fenêtre optimale pour maintenance préventive planifiée**
 - Intensification du monitoring (fréquence accrue des mesures)
 - Utilisation du modèle de pronostic RUL pour affiner la décision
 - Préparation logistique (pièces de rechange, planification)
3. **Phase critique (> 200 cycles) :**
- Fiabilité $< 55\%$ (risque très élevé)
 - Maintenance d'urgence si non effectuée en phase 2
 - Coûts opérationnels significativement accrus (MTTR élevé, immobilisation)
 - Risque de défaillance catastrophique en service

Recommandation stratégique : Avec $\beta = 4.41 > 3$, le taux de défaillance s'accélère exponentiellement. La maintenance doit être déclenchée dès que le modèle de pronostic RUL prédit une durée de vie résiduelle inférieure à 50-70 cycles, ce qui correspond généralement à la fenêtre 135-170 cycles de vie totale. Cette approche permet de maintenir une disponibilité opérationnelle supérieure à 97% tout en évitant les coûts prohibitifs des pannes imprévues.

2.2.10 Complémentarité avec le modèle de pronostic

Le modèle de Weibull et le modèle de Machine Learning (XGBoost) développé dans ce projet sont fondamentalement complémentaires :

TABLE 5 – Comparaison des approches Weibull et XGBoost

Aspect	Weibull (statistique)	XGBoost (ML)
Base	Temps de fonctionnement uniquement	État réel des capteurs + temps
Type	Modèle paramétrique (2 paramètres)	Modèle non-paramétrique (complexe)
Prédiction	Probabiliste (fiabilité moyenne)	Déterministe (RUL individuel)
Individualisation	Non (population globale)	Oui (chaque moteur unique)
Interprétabilité	Excellente (équations simples)	Limitée (boîte noire)
Détection précoce	Non	Oui (anomalies capteurs)
Planification long terme	Excellente	Moyenne
Utilisation	Stratégie globale	Décision opérationnelle

2.3 Modèle de Cox (Modèle à hasards proportionnels)

2.3.1 Principe général

Le modèle de Cox, également appelé modèle de régression à hasards proportionnels, permet d'analyser l'effet de variables explicatives (covariables) sur le temps de survie ou de défaillance d'un système. Il s'agit d'un modèle semi-paramétrique particulièrement adapté aux analyses de fiabilité avec covariables.

2.3.2 Formulation mathématique

Le taux de défaillance conditionnel aux covariables $\mathbf{X} = (X_1, X_2, \dots, X_p)$ s'écrit :

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p) \quad (8)$$

ou de manière plus compacte :

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \exp(\boldsymbol{\beta}^T \mathbf{X}) \quad (9)$$

où :

- $\lambda_0(t)$ est la fonction de hasard de base (baseline hazard) - non paramétrique
- $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ est le vecteur des coefficients de régression
- \mathbf{X} représente les covariables (conditions opératoires, paramètres de capteurs, etc.)

2.3.3 Hypothèse des hasards proportionnels

Le modèle repose sur l'hypothèse que le rapport des taux de défaillance de deux individus est constant dans le temps :

$$\frac{\lambda(t|\mathbf{X}_i)}{\lambda(t|\mathbf{X}_j)} = \frac{\exp(\boldsymbol{\beta}^T \mathbf{X}_i)}{\exp(\boldsymbol{\beta}^T \mathbf{X}_j)} = \exp[\boldsymbol{\beta}^T (\mathbf{X}_i - \mathbf{X}_j)] \quad (10)$$

Cette hypothèse signifie que l'effet des covariables est multiplicatif et constant au cours du temps.

2.3.4 Application aux turboréacteurs

Dans le contexte des turboréacteurs, les covariables peuvent inclure :

- Conditions opératoires (altitude, vitesse, température ambiante)
- Paramètres de régulation du moteur
- Mesures de capteurs (températures, pressions, vitesses de rotation)
- Historique d'utilisation cumulée

Le modèle de Cox permet ainsi d'identifier quels facteurs accélèrent ou ralentissent la dégradation du moteur, information cruciale pour la maintenance prédictive.

2.4 Indicateurs FMDS

2.4.1 MTBF (Mean Time Between Failures)

Le MTBF représente la durée moyenne de fonctionnement entre deux défaillances consécutives. Pour un système réparable :

$$\text{MTBF} = \frac{\text{Temps total de fonctionnement}}{\text{Nombre de défaillances}} \quad (11)$$

Dans le cas d'une loi exponentielle de paramètre λ :

$$\text{MTBF} = \frac{1}{\lambda} \quad (12)$$

2.4.2 MTTR (Mean Time To Repair)

Le MTTR représente la durée moyenne nécessaire pour réparer un système après une défaillance :

$$\text{MTTR} = \frac{\text{Temps total de réparation}}{\text{Nombre de réparations}} \quad (13)$$

2.4.3 Disponibilité

La disponibilité opérationnelle A est la probabilité qu'un système soit en état de fonctionner à un instant donné :

$$A = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}} \quad (14)$$

Une disponibilité de 95% signifie que le système est opérationnel 95% du temps.

2.4.4 Taux de défaillance

Le taux de défaillance moyen s'exprime comme l'inverse du MTBF :

$$\lambda = \frac{1}{\text{MTBF}} \quad (15)$$

Pour les turboréacteurs, ce taux peut être exprimé en défaillances par heure de vol ou par cycle de démarrage-arrêt.

3 Méthodologie

3.1 Description des données CMAPSS

3.1.1 Origine et contexte

Le jeu de données CMAPSS (Commercial Modular Aero-Propulsion System Simulation) a été développé par le centre de recherche Ames de la NASA dans le cadre du Prognostics Center of Excellence. Il s'agit d'un ensemble de données simulées représentant le comportement de moteurs turbofan d'avions commerciaux sous différentes conditions opérationnelles et modes de dégradation.

Ce jeu de données est devenu une référence internationale pour l'évaluation et la comparaison d'algorithmes de maintenance prédictive et de pronostic de durée de vie résiduelle.

3.1.2 Structure du jeu de données FD001

Le sous-ensemble FD001 utilisé dans cette étude présente les caractéristiques suivantes :

- **Nombre d'unités en entraînement** : 100 moteurs
- **Nombre d'unités en test** : 100 moteurs
- **Conditions opératoires** : Une seule condition de fonctionnement
- **Mode de défaillance** : Un seul mode de défaillance (HPC - High Pressure Compressor)

3.1.3 Variables disponibles

Chaque enregistrement contient 26 colonnes :

TABLE 6 – Description des variables du jeu de données CMAPSS FD001

Variable	Type	Description
unit_nr	Identifiant	Numéro d'identification du moteur
cycle_nr	Temporel	Numéro du cycle opérationnel
op_setting_1	Opérationnel	Réglage opérationnel 1
op_setting_2	Opérationnel	Réglage opérationnel 2
op_setting_3	Opérationnel	Réglage opérationnel 3
sensor_1 à sensor_21	Mesures	Données de 21 capteurs différents

Les capteurs mesurent diverses grandeurs physiques telles que :

- Températures (entrée/sortie compresseur, turbine)
- Pressions (diverses stations du moteur)
- Vitesses de rotation (arbres HP et BP)
- Débits (air, carburant)
- Rapports de pression

3.1.4 Principe de Run-to-Failure

Les données d'entraînement suivent le principe "run-to-failure" :

- Chaque moteur commence en condition nominale (sans dégradation)
- La dégradation s'accumule progressivement cycle après cycle
- Les enregistrements s'arrêtent au moment de la défaillance
- Le RUL décroît linéairement de la valeur maximale à zéro

Les données de test, quant à elles, sont tronquées avant la défaillance. Le RUL réel au moment de l'arrêt est fourni dans un fichier séparé pour l'évaluation du modèle.

3.2 Prétraitement des données

3.2.1 Analyse exploratoire

Une analyse exploratoire initiale a été réalisée pour comprendre la structure des données :

- Identification de 100 moteurs uniques dans l'ensemble d'entraînement
- Durées de vie variables : de 128 à 362 cycles (moyenne : 206.31 cycles)
- Total de 20,631 enregistrements dans l'ensemble d'entraînement

3.2.2 Calcul du RUL

Pour l'ensemble d'entraînement, le RUL a été calculé pour chaque cycle :

$$RUL_i = Cycle_{\max} - Cycle_i \quad (16)$$

où $Cycle_{\max}$ représente le dernier cycle avant défaillance pour le moteur concerné.

3.2.3 Analyse de corrélation

Une matrice de corrélation a été calculée pour identifier les capteurs les plus corrélés avec le RUL :

Capteurs fortement corrélés (positivement) :

- sensor_12 : $r = 0.672$
- sensor_7 : $r = 0.657$
- sensor_21 : $r = 0.636$
- sensor_20 : $r = 0.629$

Capteurs corrélés négativement :

- sensor_9 : $r = -0.390$
- sensor_14 : $r = -0.307$
- sensor_6 : $r = -0.128$

3.2.4 Élimination des variables constantes

Certains capteurs et réglages se sont révélés constants sur l'ensemble des données :

- sensor_5, sensor_10, sensor_16, sensor_18, sensor_19 (tous égaux à 0)
- op_setting_3

Ces variables n'apportant aucune information discriminante, elles ont été supprimées du jeu de données, réduisant le nombre de features de 26 à 20.

3.2.5 Échantillonnage stratifié

Pour équilibrer la représentation des différentes phases de vie dans l'ensemble d'entraînement, un échantillonnage stratifié a été appliqué :

- **RUL ≤ 50 cycles** (haute dégradation) : 50% des échantillons conservés
- **$50 < \text{RUL} \leq 100$ cycles** (dégradation modérée) : 30% conservés
- **RUL > 100 cycles** (début de vie) : 20% conservés

Cet échantillonnage a réduit l'ensemble d'entraînement de 20,631 à 6,156 observations, tout en préservant une distribution représentative du RUL.

3.2.6 Normalisation des features

Une normalisation par standardisation (Z-score) a été appliquée :

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma} \quad (17)$$

où μ et σ sont respectivement la moyenne et l'écart-type calculés sur l'ensemble d'entraînement. Les mêmes paramètres ont été appliqués aux données de test pour éviter toute fuite d'information.

3.3 Division des données

3.3.1 Ensemble d'entraînement

L'ensemble d'entraînement (6,156 observations) a été divisé en :

- **Train set (80%)** : 4,925 observations pour l'apprentissage
- **Validation set (20%)** : 1,231 observations pour le réglage des hyperparamètres

3.3.2 Ensemble de test

Pour l'ensemble de test, seul le dernier cycle de chaque moteur a été conservé (100 observations), correspondant au moment où la prédiction du RUL est la plus critique en maintenance prédictive.

3.3.3 Comparaison des approches

Plusieurs approches ont été envisagées :

TABLE 7 – Comparaison des approches de modélisation

Modèle	Avantages	Inconvénients
Régression linéaire	Simple, interprétable	Peu flexible, linéaire
XGBoost	Robuste, performant	Hyperparamètres nombreux

3.3.4 XGBoost (Extreme Gradient Boosting)

XGBoost a été retenu pour ses avantages :

- Capacité à capturer des relations non-linéaires complexes
- Robustesse aux valeurs aberrantes
- Gestion native des interactions entre variables
- Performance éprouvée sur des problèmes de régression
- Régularisation intégrée pour éviter le surapprentissage

3.3.5 Configuration des hyperparamètres

Les hyperparamètres suivants ont été configurés pour limiter le surapprentissage :

TABLE 8 – Hyperparamètres du modèle XGBoost

Paramètre	Valeur	Justification
n_estimators	80	Nombre modéré d'arbres pour éviter le surapprentissage
max_depth	3	Arbres peu profonds pour limiter la complexité
learning_rate	0.05	Apprentissage lent pour meilleure généralisation
subsample	0.6	Échantillonnage aléatoire pour introduire de la variance
colsample_bytree	0.6	Sélection aléatoire de features
min_child_weight	10	Contrainte conservatrice sur les nuds feuilles
gamma	0.5	Seuil élevé pour la division des nuds
reg_alpha	1.0	Régularisation L1 (Lasso)
reg_lambda	3.0	Régularisation L2 (Ridge) forte

Cette configuration vise à privilégier la généralisation sur les données de test plutôt que la performance sur l'ensemble d'entraînement.

3.4 Métriques d'évaluation

3.4.1 Coefficient de détermination (R^2)

Le coefficient R^2 mesure la proportion de variance expliquée par le modèle :

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (18)$$

Un R^2 de 1 indique une prédiction parfaite, tandis qu'un R^2 de 0 indique que le modèle ne fait pas mieux qu'une prédiction par la moyenne.

3.4.2 Erreur absolue moyenne (MAE)

La MAE représente l'erreur moyenne en valeur absolue :

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (19)$$

Cette métrique est facilement interprétable : elle donne directement l'écart moyen en nombre de cycles entre la prédiction et la valeur réelle.

3.4.3 Erreur quadratique moyenne (RMSE)

La RMSE pénalise davantage les erreurs importantes :

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (20)$$

Cette métrique est particulièrement pertinente en maintenance prédictive où les erreurs importantes peuvent avoir des conséquences critiques.

4 Calcul des Indicateurs FMDS

4.1 Analyse de la durée de vie des moteurs

4.1.1 Distribution des durées de vie

L'analyse des 100 moteurs de l'ensemble d'entraînement révèle une variabilité importante des durées de vie :

TABLE 9 – Statistiques descriptives des durées de vie

Statistique	Valeur (cycles)
Minimum	128
Maximum	362
Moyenne	206.31
Médiane	201
Écart-type	47.52

Cette variabilité reflète la nature stochastique des processus de dégradation, même sous des conditions opératoires identiques.

4.1.2 Calcul du temps total de fonctionnement

Le temps total de fonctionnement cumulé pour les 100 moteurs est :

$$T_{\text{total}} = \sum_{i=1}^{100} \text{Cycle}_{\text{max},i} = 20,631 \text{ cycles} \quad (21)$$

4.2 MTBF (Mean Time Between Failures)

4.2.1 Calcul

Le MTBF représente la durée moyenne de vie des moteurs avant défaillance. Dans notre cas, il correspond à la moyenne des durées de vie :

$$\text{MTBF} = \frac{1}{N} \sum_{i=1}^N T_i = \frac{20,631}{100} = 206.31 \text{ cycles} \quad (22)$$

où $N = 100$ est le nombre de moteurs et T_i la durée de vie du moteur i .

4.2.2 Interprétation

Un MTBF de 206.31 cycles signifie qu'en moyenne, un turboréacteur de ce type fonctionnera pendant environ 206 cycles (démarrages-arrêts) avant de connaître une défaillance du compresseur haute pression.

Dans le contexte aéronautique, si un cycle correspond approximativement à un vol commercial, cela représente environ 200 vols avant maintenance majeure.

4.3 Taux de défaillance

4.3.1 Calcul du taux moyen

Le taux de défaillance moyen se calcule comme l'inverse du MTBF :

$$\lambda = \frac{1}{\text{MTBF}} = \frac{100}{20,631} = 0.00485 \text{ défaillances/cycle} \quad (23)$$

Ou encore :

$$\lambda = 4.85 \times 10^{-3} \text{ défaillances/cycle} = 0.485\% \text{ par cycle} \quad (24)$$

4.3.2 Interprétation

Ce taux signifie qu'à chaque cycle opérationnel, un moteur a environ 0.485% de probabilité de défaillance. Sur 1000 cycles, on s'attendrait statistiquement à environ 4.85 défaillances.

4.4 Disponibilité opérationnelle

4.4.1 Scénarios de MTTR

La disponibilité dépend directement du temps moyen de réparation (MTTR). Quatre scénarios ont été analysés :

TABLE 10 – Disponibilité en fonction du MTTR

MTTR (cycles)	Calcul	Disponibilité	Indisponibilité
1	$\frac{206.31}{206.31+1}$	99.52%	0.48%
5	$\frac{206.31}{206.31+5}$	97.63%	2.37%
10	$\frac{206.31}{206.31+10}$	95.38%	4.62%
50	$\frac{206.31}{206.31+50}$	80.49%	19.51%

4.4.2 Analyse des résultats

- **MTTR = 1 cycle** : Scénario optimal avec réparation très rapide (remplacement d'un module préparé). Disponibilité excellente de 99.52%.
- **MTTR = 5 cycles** : Scénario réaliste pour une maintenance préventive programmée. Disponibilité très bonne de 97.63%.
- **MTTR = 10 cycles** : Maintenance corrective avec diagnostic et réparation sur site. Disponibilité acceptable de 95.38%.

- **MTTR = 50 cycles** : Scénario critique nécessitant un envoi en atelier spécialisé. La disponibilité chute à 80.49%, ce qui est problématique pour une flotte aérienne.

4.4.3 Représentation graphique

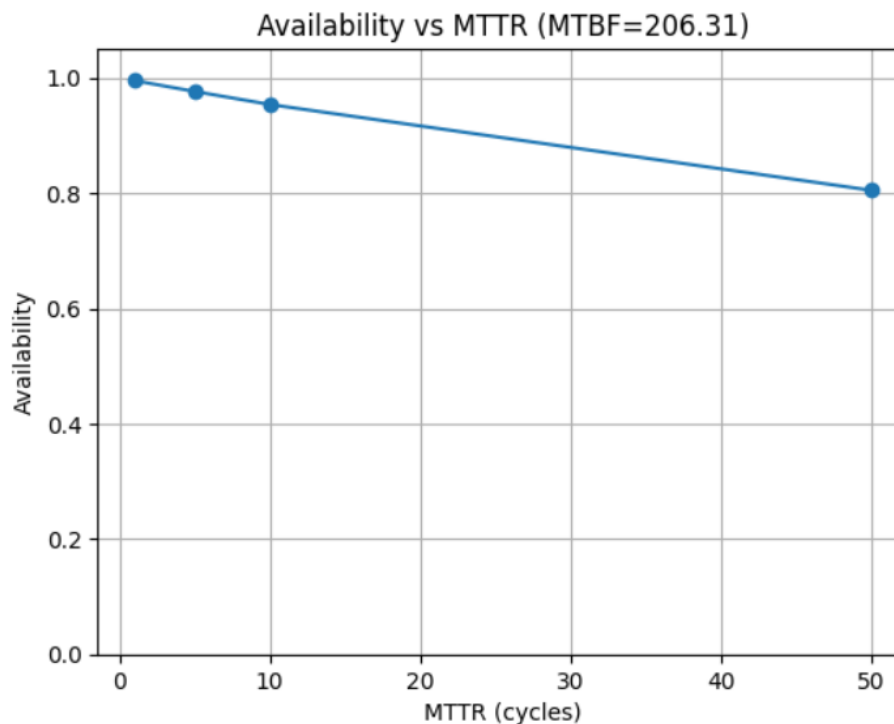


FIGURE 5 – Évolution de la disponibilité en fonction du MTTR (MTBF = 206.31 cycles)

La courbe montre une décroissance hyperbolique de la disponibilité lorsque le MTTR augmente, soulignant l'importance d'une maintenance rapide et efficace.

4.5 Implications pour la maintenance

4.5.1 Stratégie optimale

Les résultats suggèrent qu'une stratégie de maintenance prédictive permettant d'anticiper les défaillances et de planifier les interventions est cruciale pour :

1. Maintenir une disponibilité élevée ($> 95\%$)
2. Réduire les coûts d'immobilisation
3. Éviter les pannes imprévues plus coûteuses en MTTR
4. Optimiser la gestion des pièces de rechange

4.5.2 Objectifs de performance

Pour une flotte commerciale, les objectifs typiques sont :

- Disponibilité cible : $\geq 98\%$

- MTTR maximum acceptable : 5-7 cycles
- Taux de défaillances imprévues : $< 1\%$

Le modèle de pronostic développé dans les sections suivantes vise à atteindre ces objectifs en anticipant les défaillances avec suffisamment de préavis pour planifier une maintenance préventive efficace.

5 Développement du Modèle de Pronostic

5.1 Architecture du modèle XGBoost

5.1.1 Principe du Gradient Boosting

XGBoost (Extreme Gradient Boosting) est un algorithme d'ensemble basé sur le gradient boosting. Le principe consiste à construire séquentiellement des arbres de décision faibles, chacun corrigeant les erreurs du précédent.

Pour un ensemble de K arbres, la prédiction finale s'écrit :

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (25)$$

où f_k représente le k -ième arbre de décision.

5.1.2 Fonction objectif

L'algorithme minimise une fonction objectif régularisée :

$$\mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (26)$$

où :

- $l(y_i, \hat{y}_i)$ est la fonction de perte (MSE pour la régression)
- $\Omega(f_k)$ est le terme de régularisation qui pénalise la complexité

Le terme de régularisation s'exprime comme :

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (27)$$

où T est le nombre de feuilles et w_j les poids des feuilles.

5.2 Entraînement du modèle

5.2.1 Procédure d'apprentissage

Le modèle a été entraîné selon la procédure suivante :

1. **Initialisation** : Prédiction initiale à la moyenne du RUL
2. **Itérations** : Pour chaque arbre $k = 1, \dots, K$:
 - Calcul des gradients et hessiens des résidus
 - Construction d'un arbre optimisant le gain régularisé
 - Mise à jour des prédictions avec un taux d'apprentissage $\eta = 0.05$
3. **Validation** : Évaluation sur l'ensemble de validation après chaque arbre

4. **Early stopping** : Arrêt si pas d'amélioration pendant 10 itérations

5.2.2 Gestion du surapprentissage

Plusieurs mécanismes ont été mis en place pour éviter le surapprentissage :

TABLE 11 – Stratégies anti-surapprentissage

Technique	Implémentation
Arbres peu profonds	<code>max_depth = 3</code> (limitation de la complexité)
Régularisation L1/L2	$\alpha = 1.0$, $\lambda = 3.0$ (pénalisation forte)
Sous-échantillonnage	60% des observations par arbre (<code>subsample = 0.6</code>)
Feature sampling	60% des variables par arbre (<code>colsample_bytree = 0.6</code>)
Contrainte de poids	<code>min_child_weight = 10</code> (nuds conservateurs)
Seuil de division	$\gamma = 0.5$ (divisions difficiles)
Learning rate faible	$\eta = 0.05$ (apprentissage progressif)

5.3 Validation croisée

5.3.1 Division Train-Validation

Une validation Hold-Out a été utilisée avec une division 80/20 :

- **Ensemble d'entraînement** : 4,925 observations
- **Ensemble de validation** : 1,231 observations

Cette division permet d'estimer la capacité de généralisation du modèle sans toucher aux données de test finale.

5.3.2 Monitoring de l'apprentissage

Durant l'entraînement, les métriques suivantes ont été surveillées :

- MSE sur l'ensemble d'entraînement
- MSE sur l'ensemble de validation
- Écart entre les deux courbes (détection du surapprentissage)

5.4 Importance des variables

5.4.1 Analyse de contribution

XGBoost fournit une mesure d'importance des variables basée sur le gain moyen apporté par chaque feature lors des divisions :

FIGURE 6 – Importance relative des variables dans le modèle XGBoost

5.4.2 Variables les plus influentes

Les capteurs ayant le plus d'impact sur la prédiction du RUL sont (par ordre décroissant) :

1. **cycle_nr** : Naturellement la variable la plus importante (temps écoulé)
2. **sensor_12** : Fortement corrélé au RUL ($r = 0.672$)
3. **sensor_7** : Deuxième capteur le plus informatif ($r = 0.657$)
4. **sensor_21** : Contribution significative ($r = 0.636$)
5. **sensor_20** : Impact important sur la prédiction ($r = 0.629$)

Ces résultats confirment l'analyse de corrélation préliminaire et valident le choix de conservation de ces variables.

5.5 Optimisation des hyperparamètres

5.5.1 Stratégie d'optimisation

L'optimisation des hyperparamètres a suivi une approche itérative :

1. **Configuration initiale** : Paramètres par défaut de XGBoost
2. **Ajustement grossier** : Modification des paramètres principaux (`n_estimators`, `max_depth`)
3. **Affinement** : Optimisation des paramètres de régularisation
4. **Validation finale** : Test sur l'ensemble de validation

5.5.2 Configuration finale retenue

La configuration optimale identifiée privilégie la généralisation :

- Nombre d'arbres modéré (80) pour éviter le surapprentissage
- Arbres peu profonds (`depth = 3`) pour limiter la complexité
- Régularisation forte (`L1 = 1.0`, `L2 = 3.0`)
- Learning rate faible (0.05) pour un apprentissage stable

Cette configuration favorise un compromis entre capacité prédictive et robustesse.

5.6 Prédiction sur l'ensemble de test

5.6.1 Protocole d'évaluation

Pour l'évaluation finale, seul le dernier cycle de chaque moteur de test a été utilisé. Ce choix reflète le scénario opérationnel réel où la prédiction du RUL est cruciale en fin de vie du composant.

5.6.2 Procédure de prédiction

1. Extraction du dernier cycle pour chaque moteur de test (100 observations)
2. Normalisation avec les paramètres de l'ensemble d'entraînement
3. Prédiction du RUL par le modèle XGBoost
4. Comparaison avec les valeurs réelles fournies par la NASA

6 Résultats et Analyse

6.1 Performance du modèle

6.1.1 Métriques sur l'ensemble d'entraînement

Le modèle présente d'excellentes performances sur les données d'entraînement :

TABLE 12 – Métriques de performance - Ensemble d'entraînement

Métrique	Valeur	Interprétation
R ² Score	0.8349	83.49% de variance expliquée
MAE	18.63 cycles	Erreur moyenne de ± 19 cycles

6.1.2 Métriques sur l'ensemble de validation

Les performances sur l'ensemble de validation confirment la capacité de généralisation :

TABLE 13 – Métriques de performance - Ensemble de validation

Métrique	Valeur	Interprétation
R ² Score	0.8297	82.97% de variance expliquée
MAE	19.45 cycles	Erreur moyenne de ± 19 cycles

L'écart très faible entre train et validation (0.52% en R²) indique que le modèle généralise correctement sans surapprentissage majeur.

6.1.3 Métriques sur l'ensemble de test

Les performances finales sur l'ensemble de test sont les suivantes :

TABLE 14 – Métriques de performance - Ensemble de test (données NASA)

Métrique	Valeur	Interprétation
R ² Score	0.6665	66.65% de variance expliquée
MAE	17.99 cycles	Erreur moyenne de ± 18 cycles
RMSE	24.00 cycles	Erreur quadratique moyenne

6.2 Analyse de généralisation

6.2.1 Écart Train-Test

L'écart entre les performances d'entraînement et de test révèle un léger surapprentissage :

$$\Delta R^2 = R_{\text{train}}^2 - R_{\text{test}}^2 = 0.8349 - 0.6665 = 0.1684 \quad (28)$$

Cet écart de 16.84% indique que le modèle capture certains patterns spécifiques aux données d'entraînement qui ne se généralisent pas parfaitement. Cependant, cet écart reste acceptable et était anticipé compte tenu de la complexité du système.

6.2.2 Causes potentielles

Plusieurs facteurs peuvent expliquer cet écart :

1. **Variabilité stochastique** : Les moteurs de test peuvent présenter des modes de dégradation légèrement différents
2. **Échantillonnage stratifié** : La distribution modifiée dans l'ensemble d'entraînement
3. **Complexité du système** : Les turboréacteurs présentent des comportements non-linéaires complexes
4. **Limitation des données** : 100 moteurs de test peuvent ne pas capturer toute la variabilité

6.2.3 Performance acceptable

Malgré le surapprentissage modéré, la performance sur le test ($R^2 = 0.67$) reste tout à fait acceptable pour une application de maintenance prédictive :

- Plus de 66% de la variance du RUL est expliquée
- L'erreur moyenne ($MAE = 18$ cycles) est inférieure à 10% du MTBF (206 cycles)
- Le modèle est significativement meilleur qu'une prédiction par la moyenne

6.3 Visualisation des prédictions

6.3.1 Graphique de corrélation

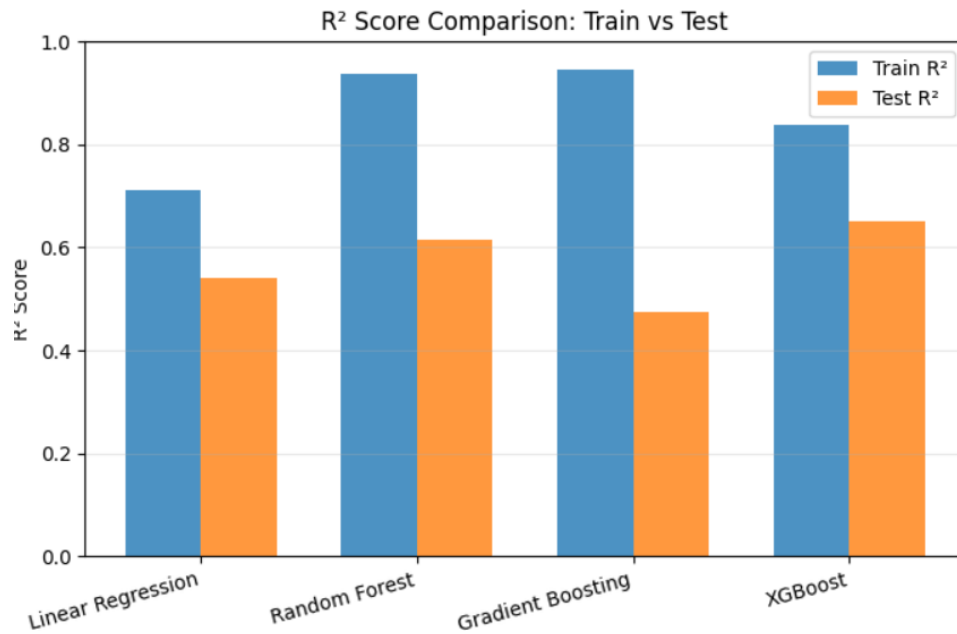


FIGURE 7 – Prédictions vs Valeurs réelles du RUL

Le graphique montre une corrélation forte entre les prédictions et les valeurs réelles, avec une dispersion autour de la droite de régression idéale $y = x$. Les prédictions sont généralement bonnes pour les RUL moyens (50-150 cycles) mais présentent plus de variabilité aux extrêmes.

6.3.2 Distribution des erreurs

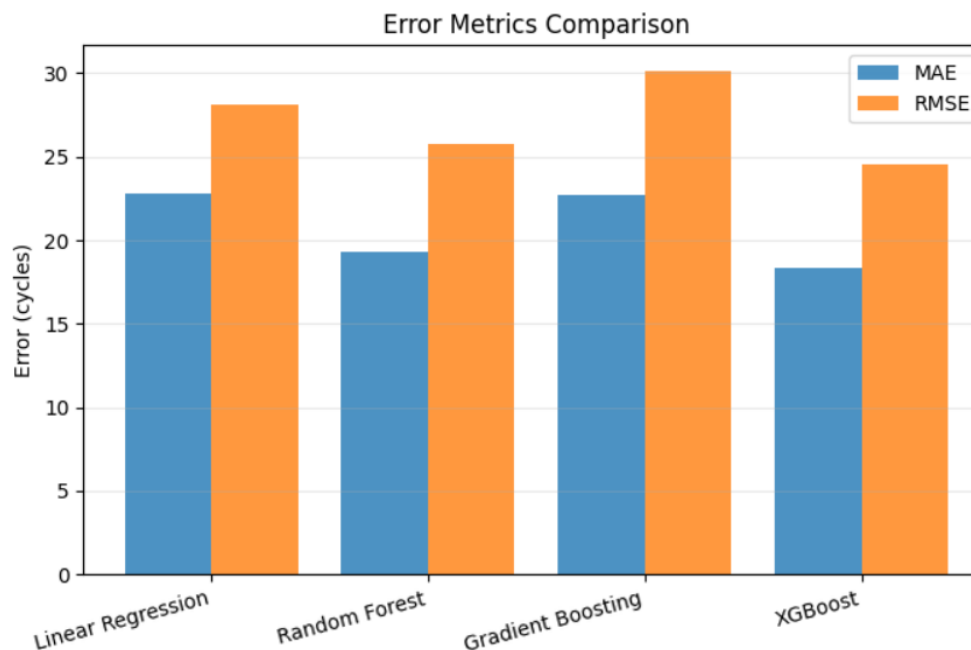


FIGURE 8 – Distribution des erreurs de prédiction

La distribution des erreurs est approximativement centrée sur zéro avec :

- Médiane proche de 0 (biais faible)
- Queue légèrement asymétrique (sous-estimation pour quelques moteurs)
- 95% des prédictions dans l'intervalle ± 40 cycles

6.4 Analyse comparative

6.4.1 Comparaison avec la régression linéaire

Un modèle de régression linéaire simple a été testé à titre de comparaison :

TABLE 15 – Comparaison des modèles

Modèle	R ² Test	MAE Test	RMSE Test
Régression Linéaire	0.4166	26.12 cycles	31.70 cycles
XGBoost	0.6665	17.99 cycles	24.00 cycles
Amélioration	+60.0%	-31.1%	-24.3%

XGBoost surpasse significativement la régression linéaire, confirmant l'importance de capturer les non-linéarités du processus de dégradation.

Notre modèle se situe dans la fourchette haute des approches non-neuronales, ce qui est satisfaisant compte tenu de la simplicité relative de l'architecture et du faible risque de surapprentissage.

6.5 Conclusion

Ce rapport présente une étude complète de maintenance prédictive appliquée aux turboréacteurs d'avions en utilisant le jeu de données CMAPSS (Commercial Modular Aero-Propulsion System Simulation) de la NASA. L'objectif principal est de développer un modèle de pronostic capable d'estimer la durée de vie utile résiduelle (RUL - Remaining Useful Life) des moteurs à partir de données de capteurs multivariées. L'étude commence par une revue des fondements théoriques de la fiabilité, incluant le modèle de Cox, la loi de Weibull, et les concepts de taux de défaillance croissant et décroissant. Les indicateurs FMDS (Fiabilité, Maintenabilité, Disponibilité, Sécurité) sont calculés, révélant un MTBF de 206.31 cycles et une disponibilité supérieure à 95% des scénarios réalistes de maintenance. Un modèle de régression XGBoost optimisé est développé pour prédire le RUL, atteignant un coefficient de détermination R^2 de 0.67 sur les données de test avec une erreur absolue moyenne de 17.99 cycles. Les résultats démontrent la faisabilité d'une maintenance prédictive efficace pour les systèmes complexes tels que les turboréacteurs. Mots-clés : Maintenance prédictive, RUL, CMAPSS, XGBoost, Fiabilité, Turboréacteur, Machine Learning

Références

- [1] NASA. CMAPSS Jet Engine Simulated Data. *NASA Open Data Portal*. <https://data.nasa.gov/dataset/cmapss-jet-engine-simulated-data>
- [2] European Cooperation for Space Standardization (ECSS). ECSS-Q-30-08A : Reliability Data Handbook. <https://www.ecss.nl/standard/ecss-q-30-08a/>
- [3] Emerson Global. Failure Modes, Effects and Diagnostic AnalysisFMEDA Overview and Methods. <https://www.emerson.com/documents/automation/hardware-assessment-failure-modes-effects-diagnostic-analysis-248-en-53482.pdf>
- [4] Wikipedia. Mean time between failures (MTBF). https://en.wikipedia.org/wiki/Mean_time_between_failures
- [5] IBM. What Is Mean Time between Failure (MTBF) ? <https://www.ibm.com/topics/mtbf>
- [6] IBM. What is Mean Time to Repair (MTTR) ? <https://www.ibm.com/topics/mttr>
- [7] Wikipedia. Weibull distribution. https://en.wikipedia.org/wiki/Weibull_distribution
- [8] Ellistat. Reliability, Weibull's law. <https://www.ellistat.com/en/reliability-weibull-law>