Université Ibn Zohr

École Nationale Supérieure de l'Intelligence Artificielle et Sciences des Données

Filière : Sciences des Données, Big Data & IA

PROJET ACADÉMIQUE

Prédiction de la défaillance des machines industrielles

Réalisé par :

Badereddine El Alouani

Ayoub Idlhaj

Lahcen Ahtat

Encadré par :

Pr. Sara El-Ateif

Année Universitaire : 2025 - 2026

Résumé

Ce rapport présente une étude complète sur la prédiction de défaillances des machines industrielles à l'aide de données de capteurs (température, vitesse de rotation, couple, usure de l'outil). L'objectif est de développer un modèle de classification capable d'anticiper les pannes rares mais critiques.

L'analyse porte sur un jeu de données de 10 000 observations caractérisé par un fort déséquilibre de classes (3,4% de défaillances). Nous avons comparé plusieurs algorithmes (Régression Logistique, SVM, Random Forest) et évalué l'impact de techniques de ré-échantillonnage (SMOTE). Les résultats montrent que le modèle **Random Forest**, après optimisation des hyperparamètres, offre les meilleures performances globales avec une Average Precision (AP) de 0,85, surpassant significativement les autres approches. L'analyse d'importance des variables identifie le couple et la vitesse de rotation comme les précurseurs les plus fiables des défaillances.

Mots-clés : Machine Learning, Maintenance prédictive, Random Forest, SMOTE, Déséquilibre de classes.

Table des matières

Table des figures

Chapitre 1

Introduction et problématique

1.1 Contexte industriel

La transformation numérique du secteur industriel (Industrie 4.0) permet désormais la collecte massive de données issues des machines de production. Ces données, incluant des températures de procédé et des mesures cinématiques, constituent une source précieuse pour anticiper les comportements anormaux.

1.2 Enjeux économiques

La défaillance imprévue d'une machine entraîne des arrêts de production coûteux. L'enjeu est de passer d'une maintenance corrective (réparer après la panne) ou préventive (remplacer à intervalles fixes) à une maintenance **prédictive**, qui intervient juste avant l'occurrence de la panne, optimisant ainsi la durée de vie des composants et la disponibilité des équipements.

1.3 Objectifs du projet

L'objectif de ce projet est de concevoir un modèle de classification binaire supervisé. Le travail vise à :

1. Analyser les corrélations entre les paramètres physiques (température, couple, etc.) et les pannes.
2. Gérer le défi technique du déséquilibre des classes (peu de pannes par rapport au fonctionnement normal).
3. Comparer les performances de modèles linéaires et non-linéaires.
4. Optimiser le modèle retenu pour maximiser la détection des pannes (Rappel) tout en limitant les fausses alarmes (Précision).

Chapitre 2

Analyse exploratoire des données

2.1 Présentation du jeu de données

2.1.1 Description générale

Le jeu de données contient 10 000 enregistrements. Chaque observation représente un état machine décrit par les variables suivantes :

- **Air temperature [K]** : Température ambiante (convertie en °C pour l'analyse).
- **Process temperature [K]** : Température du procédé (convertie en °C).
- **Rotational speed [rpm]** : Vitesse de rotation de la broche.
- **Torque [Nm]** : Couple exercé.
- **Tool wear [min]** : Durée d'utilisation de l'outil en minutes.
- **Type** : Qualité du produit (Low, Medium, High).

La variable cible **Target** est binaire : 0 pour un fonctionnement normal, 1 pour une défaillance.

2.1.2 Déséquilibre des classes

L'analyse de la variable cible révèle un fort déséquilibre :

- **Normal (0)** : 9 661 observations (96,6 %)
- **Défaillance (1)** : 339 observations (3,4 %)

Ce déséquilibre (visualisé Figure 2.1) impose l'utilisation de métriques adaptées (F1-score, Average Precision) plutôt que la simple exactitude (Accuracy).

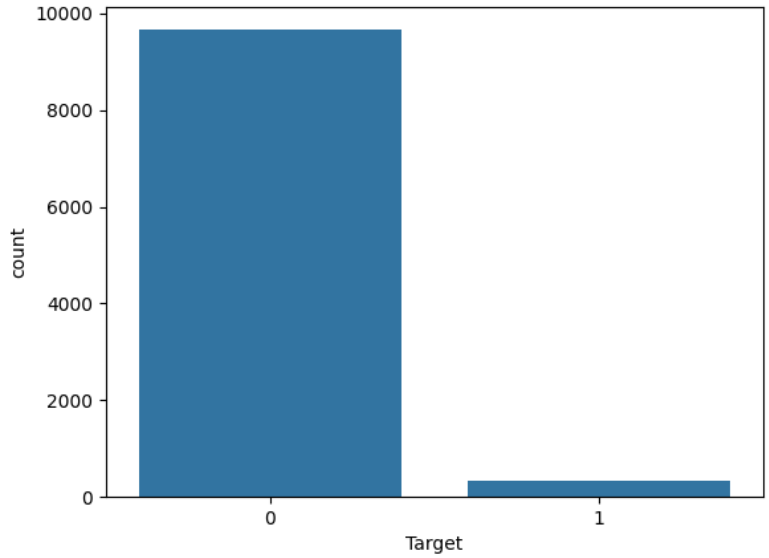

FIGURE 2.1 – Distribution de la variable cible : prédominance des cas normaux

2.2 Analyse des variables et corrélations

2.2.1 Distributions

Les distributions des températures (Air et Process) suivent des lois approximativement normales. La variable *Type* montre que la majorité des machines produisent des produits de qualité "Low" (L), suivis de "Medium" (M) et "High" (H) (voir Figure 2.2).

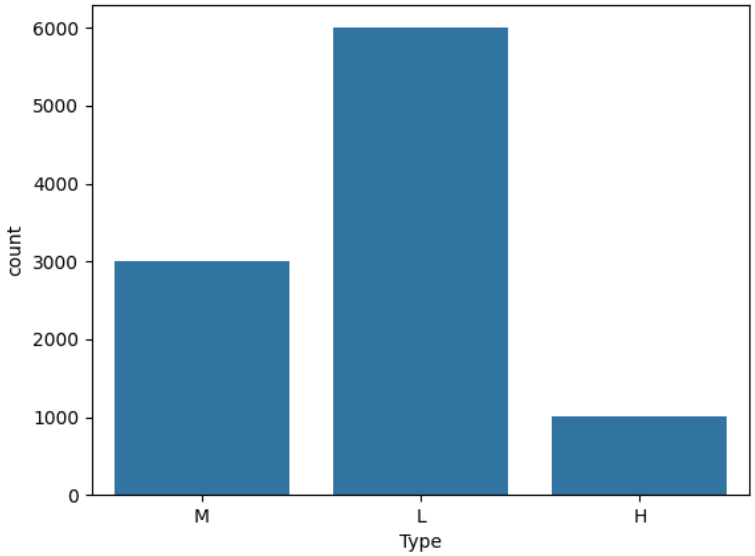

FIGURE 2.2 – Répartition des types de machines (Qualité produit L, M, H)

2.2.2 Matrice de corrélation

La Figure 2.3 présente les corrélations entre variables numériques. On observe :

— Une très forte corrélation positive entre la température de l'air et celle du procédé.

— Une corrélation négative marquée entre la vitesse de rotation et le couple (relation physique $P = C \times \omega$).

— L'usure de l'outil (*Tool wear*) montre une corrélation positive faible mais existante avec la cible (*Target*), suggérant son rôle dans les pannes.

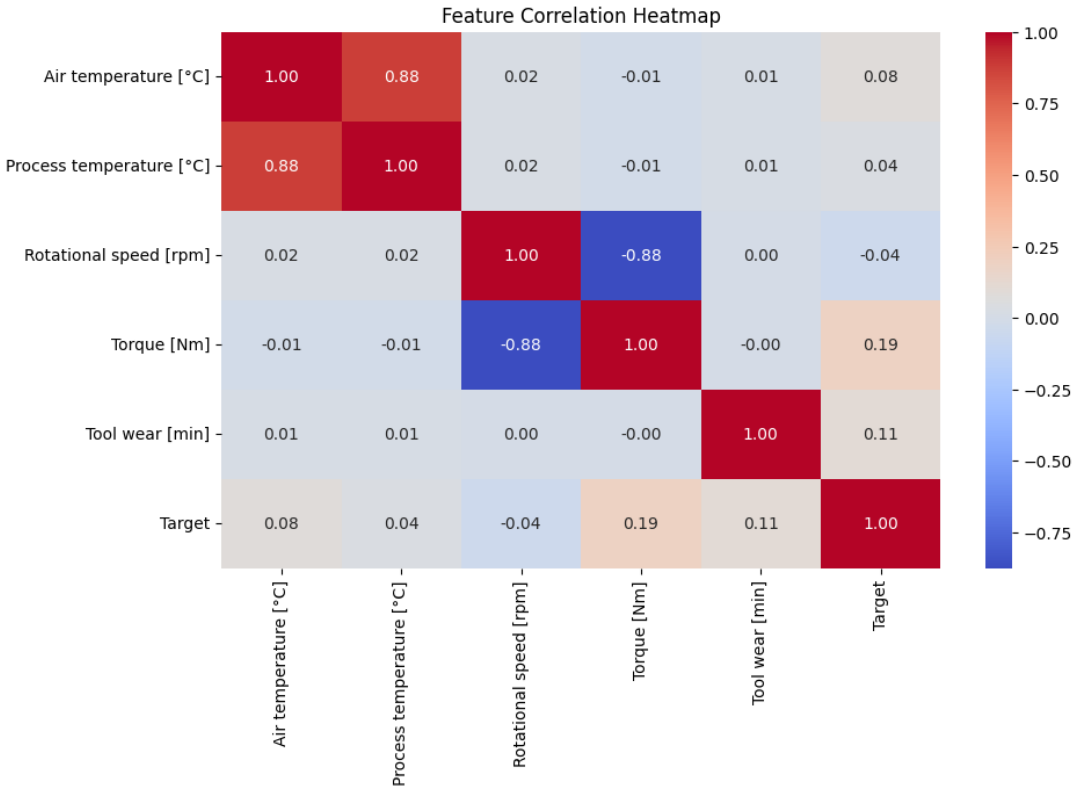

FIGURE 2.3 – Matrice de corrélation des variables numériques

Chapitre 3

Méthodologie de Machine Learning

3.1 Prétraitement des données

Avant la modélisation, les transformations suivantes ont été appliquées :

1. **Nettoyage** : Suppression des colonnes identifiants inutiles (*UDI, Product ID*).
2. **Conversion** : Les températures en Kelvin ont été converties en Celsius.
3. **Encodage** : La variable catégorielle *Type* a été transformée via *One-Hot Encoding* (création de variables binaires $Type\_L$, $Type\_M$).
4. **Séparation** : Division du jeu de données en ensemble d'entraînement (80%) et de test (20%), avec stratification pour préserver la proportion de pannes.
5. **Normalisation** : Application du *StandardScaler* pour mettre les variables à la même échelle (moyenne 0, écart-type 1), essentiel pour les modèles comme la Régression Logistique et les SVM.

3.2 Stratégie de modélisation

3.2.1 Modèles testés

Nous avons comparé trois familles d'algorithmes :

— **Régression Logistique (LR)** : Modèle linéaire de référence.
— **Support Vector Machine (SVM)** : Efficace pour les marges complexes.
— **Random Forest (RF)** : Méthode d'ensemble robuste aux non-linéarités et ne nécessitant pas obligatoirement de mise à l'échelle.

3.2.2 Gestion du déséquilibre (SMOTE)

Pour améliorer la détection de la classe minoritaire (pannes), nous avons testé la technique **SMOTE** (Synthetic Minority Over-sampling Technique) qui génère des exemples synthétiques de pannes dans l'ensemble d'entraînement.

Chapitre 4

Résultats et évaluation

4.1 Performances des modèles de base

Les modèles ont été évalués initialement sans rééchantillonnage. Le tableau ci-dessous résume les performances sur l'ensemble de test :

Modèle	Précision (Classe 1)	Rappel (Classe 1)	Average Precision (AP)
Régression Logistique	0.64	0.10	0.42
SVM	0.87	0.19	0.59
Random Forest	**0.88**	**0.53**	**0.80**

TABLE 4.1 – Comparaison des modèles sans SMOTE

Le **Random Forest** domine largement avec une AP de 0.80. La Régression Logistique échoue à détecter la majorité des pannes (Rappel de 10%).

4.2 Impact du SMOTE

L'introduction du SMOTE a permis d'augmenter significativement le Rappel (capacité de détection), mais au prix d'une chute importante de la Précision (plus de fausses alarmes).

- **RF + SMOTE** : Le Rappel monte à 72%, mais la Précision chute à 45%. L'AP global descend à 0.63.
- **Conclusion** : Bien que le SMOTE aide à "voir" plus de pannes, il introduit trop de bruit pour ce dataset spécifique. Le modèle Random Forest standard (avec gestion de poids ou optimisation) reste préférable.

4.3 Optimisation du modèle final

Une recherche par grille (Grid Search) a été effectuée sur le Random Forest pour optimiser les hyperparamètres (*n_estimators*, *max_depth*, *min_samples_leaf*).

4.3.1 Meilleurs paramètres

Le modèle optimal possède les caractéristiques suivantes :

— Nombre d'arbres : 200

— Profondeur maximale : 10

— Critère de poids : Aucun (le modèle apprend naturellement les motifs sans pondération artificielle forte).

4.3.2 Performance finale

Ce modèle optimisé atteint une **Average Precision de 0.85** sur le jeu de test. En ajustant le seuil de décision pour viser un rappel minimal, nous obtenons un compromis viable pour l'industrie (détecter 50-60% des pannes avec une confiance de 60%).

4.4 Importance des variables

L'analyse de l'importance des caractéristiques (Feature Importance) du modèle Random Forest révèle les facteurs critiques :

1. **Torque [Nm]** : Le couple est l'indicateur le plus fort.
2. **Rotational speed [rpm]** : La vitesse de rotation est le second facteur.
3. **Tool wear [min]** : L'usure de l'outil joue un rôle secondaire mais significatif.

Les températures ont un impact moindre sur la prédiction immédiate des pannes dans ce jeu de données.

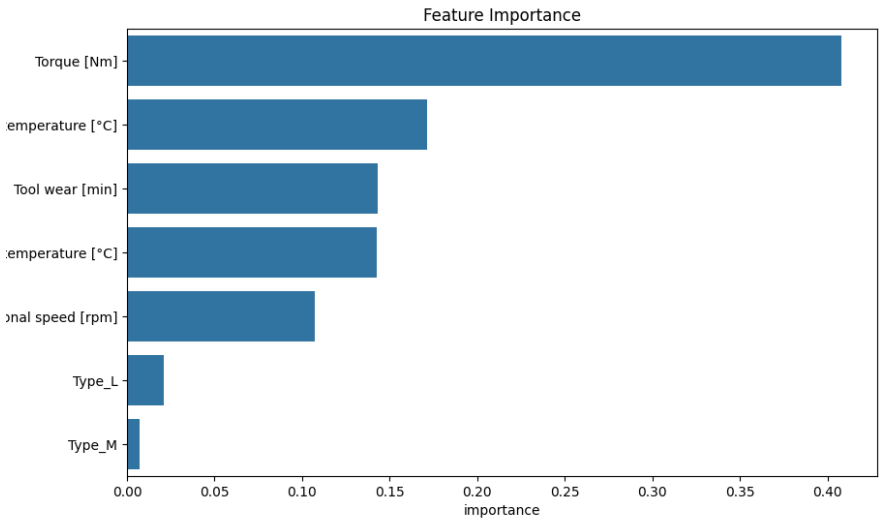

FIGURE 4.1 – Importance des variables dans le modèle Random Forest optimisé

Chapitre 5

Conclusion et perspectives

Ce projet a permis de développer un modèle de maintenance prédictive efficace basé sur l'algorithme **Random Forest**. L'analyse a démontré que :

- Le déséquilibre des classes est le défi majeur, mieux géré par les méthodes d'ensemble (Random Forest) que par les modèles linéaires.
- Les techniques de sur-échantillonnage comme SMOTE, bien qu'utiles pour le rappel, dégradent trop la précision globale dans ce contexte précis.
- Le couple et la vitesse de rotation sont les signaux précurseurs clés à surveiller.

Le modèle final offre une Average Precision de **0.85**, ce qui permet un déploiement industriel pertinent pour filtrer les machines à risque et planifier les interventions avant la panne critique. Pour les travaux futurs, l'intégration de données temporelles (séries chronologiques) pourrait permettre de prédire non seulement l'état (panne/normal) mais aussi le temps restant avant défaillance (RUL).

Code source : L'intégralité du projet est consultable sur GitHub à l'adresse https://github.com/Ahtat204/Machine-Failure-Prediction.