

Documentation – Méthodologie de Prédiction et d'Évaluation des Dysfonctionnements des Équipements

□ Introduction

Ce document explique, de manière pédagogique, le fonctionnement de l'algorithme de prédiction de dysfonctionnements utilisé dans l'application Streamlit développée pour **UMS PASTA**.

Il décrit notamment :

- comment sont calculées les **probabilités prédites**,
 - comment sont obtenues les **probabilités empiriques observées**,
 - et pourquoi la **comparaison entre les deux** est essentielle pour évaluer les performances des modèles.
-

1. □ Probabilité Empirique Observée

? Définition :

Il s'agit de la **probabilité réelle** qu'un équipement tombe en panne, **calculée à partir des données des 6 derniers mois** (c'est-à-dire après la phase d'entraînement des modèles).

🔍 Étapes de calcul :

✓ Étape 1 : Regroupement par "bins"

Les équipements sont classés dans des groupes (appelés **bins**) selon deux critères :

- Leur **récurrence** de pannes (Recurrence)
- Leur **temps d'arrêt total** (Total_Time)

Exemple de code :

```
agg_derniers["Recurrence_Bin"] = pd.cut(agg_derniers["Recurrence"], bins=5)
```

```
agg_derniers["Total_Time_Bin"] = pd.cut(agg_derniers["Total_Time"], bins=5)
```

✓ Étape 2 : Calcul de la probabilité moyenne observée

Dans chaque groupe formé, on calcule la **proportion réelle d'équipements qui ont subi un dysfonctionnement** (Dysfunction = 1).

Exemple :

- Un bin contient 10 équipements
 - 4 d'entre eux ont eu un Dysfunction = 1
- 👉 Probabilité observée = $4 / 10 = 0.4$

Code :

```
observed_probs = agg_derniers.groupby(  
    ["Recurrence_Bin", "Total_Time_Bin"]  
)[  
    "Dysfunction".mean()
```

✓ Étape 3 : Attribution à chaque équipement

Chaque équipement se voit attribuer la **probabilité moyenne** de son bin, selon sa récurrence et son temps d'arrêt. Cela donne la **probabilité empirique observée** pour cet équipement.

2. 🧠 Probabilité Prédite par les Modèles

Les prédictions sont générées à l'aide de modèles de machine learning entraînés sur les **6 premiers mois de données**.

🔗 Étapes détaillées :

✓ Étape 1 : Construction des variables

Pour chaque équipement :

- **Recurrence** = nombre de pannes
- **Total_Time** = temps total d'arrêt
- **Dysfunction** = 1 si la récurrence dépasse un seuil (ex : 3), sinon 0

✓ Étape 2 : Entraînement des modèles

Trois modèles sont utilisés :

- Régression Logistique
- Forêt Aléatoire (Random Forest)
- Gradient Boosting

Ces modèles apprennent à prédire la probabilité qu'un équipement ait Dysfunction = 1 en fonction de Recurrence et Total_Time.

✓ Étape 3 : Génération des probabilités

Chaque modèle produit une **probabilité entre 0 et 1** représentant le risque estimé de dysfonctionnement.

Exemple :

Équipement Logistic Regression Random Forest

Moteur A	0.82	0.74
Convoyeur B	0.25	0.40
Pression C	0.65	0.62

3. 🔄 Comparaison : Écarts entre Prédictions et Réalité

□ Formule :

Écart = Probabilité observée - Probabilité prédite

- Si l'écart est **proche de 0**, le modèle prédit bien.
 - Si l'écart est **négatif**, le modèle **surestime** le risque.
 - Si l'écart est **positif**, le modèle **sous-estime** le risque.
-

4. ☐ Exemple Complet avec 3 Équipements

☑ 6 premiers mois (entraînement) :

Équipement Réccurrence Temps_total Cible (Dysfunction)

Moteur A	5	240	1
Convoyeur B	2	60	0
Pression C	4	130	1

🧠 Prédictions des modèles :

Équipement Logistic Regression Random Forest

Moteur A	0.82	0.74
Convoyeur B	0.25	0.40
Pression C	0.65	0.62

☑ 6 derniers mois (observations réelles) :

Équipement Réccurrence Temps_total Réalité (Dysfunction)

Moteur A	3	150	1
Convoyeur B	1	30	0
Pression C	2	70	0

🌀 Probabilités Observées (par bin) :

Équipement Probabilité Observée

Moteur A	0.50
Convoyeur B	0.00

Équipement Probabilité Observée



Pression C 0.00

Écarts (observé - prédit) :

Équipement	Modèle	Prédit	Observé	Écart
Moteur A	Logistic Regression	0.82	0.50	-0.32
Convoyeur B	Logistic Regression	0.25	0.00	-0.25
Pression C	Logistic Regression	0.65	0.00	-0.65

5. Visualisation des Résultats

Les écarts sont représentés dans l'application avec :

-  **Graphique en barres** (Altair) : Écart absolu par équipement et modèle
-  **Courbe interactive** : évolution des écarts par modèle

Ces visualisations permettent :

- d'évaluer la performance des modèles,
 - d'identifier les équipements à risque mal prédits,
 - d'ajuster le modèle ou le seuil si besoin.
-

Conclusion

Ce processus permet d'assurer une prédiction fiable des pannes en :

- exploitant les données historiques (6 mois),
- construisant des modèles prédictifs robustes,
- les comparant aux résultats réels du futur proche (6 mois suivants).

L'écart entre prédiction et réalité est un **indicateur essentiel** pour améliorer la maintenance préventive et optimiser les interventions sur les équipements de production.
