

Projet Final : MAINTENANCE PRÉdictive

Simplon.co - École Microsoft IA

Ngachili Maëlle

20/11/2020

Résumé

Ce rapport présente les étapes de la mise en place d'une application web qui affiche les prédictions de pannes faites en continu et en temps réel à partir de la simulation d'un flux de données télémétriques.

Table des matières

Table des matières	1
I Introduction	3
1 La Maintenance Industrielle	4
1.1 La Maintenance Traditionnelle	4
1.2 La Maintenance Prédictive	4
1.3 Applications	5
1.4 Outils pour la Maintenance Prédictive	6
II Recherche de la solution	7
2 Le Jeu De Données	8
2.1 Présentation	8
2.2 Dataset Machines	8
2.3 Dataset Maintenance	10
2.4 Dataset Errors	11
2.5 Dataset Failures	12
2.6 Dataset Telemetry	15

3 Feature Engineering	16
3.1 Telemetry	16
3.2 Errors	16
3.3 Maintenance	16
3.4 Finalisation	17
4 Algorithme de Classification	20
4.1 Modalités	20
4.2 Comparaison des modèles	21
4.3 Feature Importance	23
4.4 Ajustement de la Target	24
III Mise en Place de la Solution	25
5 Web Application	26
5.1 Flux de Données	26
5.2 Application	27
5.3 Base De Données	28
5.4 Interface Utilisateur	29
6 Bilan	30
6.1 Gestion de Projet	30
6.2 Version Control	31
6.3 Conclusion	31
6.4 Améliorations	31
Bibliographie	32
Appendices	33
A Fréquence des Interventions Sur Une Machine	34
B Distributions des Features	35
C XGBoost - Bayesian Optimisation	37
D Temps Computationnel de l'Apprentissage et des Prédictions	38
E Matrices de Confusion	39
F Scores	44
G Feature Importance	47
G.1 Par Catégorie de Features	47
G.2 Par Sous-Catégorie de Features	48
H Schéma Relationnel de la Base de Données	49
I Structure de l'Application Web	51

Première partie

Introduction

Chapitre 1

La Maintenance Industrielle

1.1 LA MAINTENANCE TRADITIONNELLE

Jusqu'ici dans l'industrie on pratiquait essentiellement la maintenance curative qui consiste à attendre de se faire surprendre par une panne pour agir. Chaque panne ayant une incidence directe sur la production et donc en général un coût élevé surtout en cas d'interruption. Pour palier à ce problème les usines y ont associé la maintenance préventive. Il s'agit alors de programmer des interventions de maintenances pour contrôler et éventuellement remplacer régulièrement le matériel industriel. Ce qui peut être également coûteux, il faut constituer des équipes de maintenance ou incorporer ce type de contrôle dans la charge de travail des ouvriers. Les remplacements automatiques peuvent aussi s'avérer prématurés.

Une autre des problématiques récurrentes pour les industriels est l'excédent de stock. En effet, pour pouvoir intervenir le plus vite possible en cas de panne et éviter une interruption trop longue de la production, les usines sont dans l'obligation de conserver des stocks très conséquents en pièces et machines de remplacement. Cela a pour effet d'engendrer énormément de gaspillages et de déchets industriels qui sont un véritable frein pour toute entreprise.

La maintenance prédictive fait cette belle promesse de résoudre les principaux inconvénients liés aux méthodes de maintenance traditionnelles dans l'industrie.

1.2 LA MAINTENANCE PRÉDICTIVE

Présentation

Les économistes s'accordent à dire que la maintenance prédictive a un avenir radieux devant elle. Selon une étude du cabinet McKinsey d'ici 2025 elle devrait permettre aux entreprises américaines d'économiser 630 milliards de dollars.

Selon [Fero Labs](#) les entreprises industrielles se débarrassent de 98% des données qu'elles collectent par manque de connaissances, de compétences ou de capacité pour les exploiter. Or en exploitant certaines de ces données correctement, ces entreprises pourraient efficacement prévenir toute sortes de pannes sans avoir à supporter le coût d'interventions préprogrammées parfois inutiles.

Au cours de ces dernières années, cette forme de maintenance industrielle à gagner en popularité puisque les équipes ont rapidement compris qu'elles pourraient, grâce à elle, réduire les pannes, éviter les imprévus et anticiper le moindre arrêt dans la production.

Plus concrètement, elle va permettre de gagner en fiabilité et de surveiller le plus précisément possible les performances des machines. Les problèmes imminents sont détectés et résolus, si bien que les équipes en charge peuvent procéder aux réparations et remplacements avant que les pannes ne surviennent ou n'entraînent de problèmes plus graves ou coûteux.

Pré-requis

Les mesures récupérées par tous types de capteurs s'avèrent être une ressource en or pour les algorithmes de machine learning. Avec l'essor de l'IoT (Internet of things) puis de l'IIoT (Industrial Internet of Things), les entreprises industrielles ont de plus en plus de possibilité pour accéder à toute sorte de données télémétriques sur leurs machines et leur environnement.

Les capteurs sur les machines fonctionnent comme un système de surveillance des outils de production en temps réel. Ils transmettent les données à un logiciel pour qu'un technicien de maintenance les analyse et ainsi il peut :

- déterminer la probabilité d'un défaut sur une machine,
- déterminer le type de défaut possible,
- anticiper une panne,
- prévoir l'entretien nécessaire à effectuer sur une machine pour éviter la panne et ne pas bloquer la production.

Avant de se lancer dans la maintenance prédictive il faut donc au préalable s'assurer de disposer des divers capteurs qui viendront alimenter une base de données. Ensuite il faut avoir un historique suffisamment grand de ces données pour pouvoir entraîner les modèles.

Avantages

Par rapport à la maintenance préventive, la maintenance prédictive permet de :

- diminuer le nombre d'interruptions des machines pour des opérations de maintenance,
- diminuer le nombre de pannes,
- mieux planifier les interventions,
- mieux préparer les équipes d'intervention,
- mieux échanger entre les professionnels de maintenance et les équipes de production,
- mieux anticiper et gérer les besoins de pièces détachées des outils.

Pour résumé, une maintenance prédictive permet d'économiser, entre autres, du temps, de la main d'œuvre ou encore de l'argent, et permet aussi d'identifier des problèmes qui n'aurait pu être résolus avec de simples inspections de routine.

Limites

Il est communément admis que la maintenance prédictive peut coûter très cher. Les procédés à mettre en place qu'elle implique sont assez onéreux, mais ces coûts sont très rapidement amortis par les résultats obtenus. Les industriels peuvent en effet réaliser d'importantes économies, puisqu'ils seront en mesure de prévoir les pannes et d'agir en amont.

Cependant, si l'usine ne recense pas beaucoup de machines, il est préférable de s'en tenir à des routines de maintenance préventive conventionnelle. C'est généralement plus économique.

1.3 APPLICATIONS

Le premier exemple d'application de la maintenance prédictive est assez intuitif, l'industrie. Les usines sont a priori toutes de bonne candidate pour ce système.

Les applications sont cependant bien plus vastes et variées. Théoriquement tout objet disposant de capteurs devrait pouvoir profiter de ce qu'offre la maintenance prédictive. On pourrait renforcer les signaux d'alertes des tableaux de bords afin qu'ils mettent en garde contre une défaillance à venir. Libre aux conducteurs d'ignorer également ses signaux. Les smartphones, ordinateurs, tablettes sont d'autres exemples. Il serait extrêmement pratique de pouvoir utiliser la maintenance prédictive sur les disques durs.

D'un point de vue algorithmique la maintenance prédictive se rapproche assez de la détection d'anomalies. Bien souvent - et idéalement - les pannes sont des évènements rares, cela revient donc bien à rechercher la ou les anomalies dans les logs de maintenance et les données de télé-métrie ayant pu causer une panne.

1.4 OUTILS POUR LA MAINTENANCE PRÉDICTIVE

Il existe déjà sur le marché plusieurs solutions de maintenance prédictive pour les industriels. Elles consistent le plus souvent à proposer un outil connecté qui combine la GMAO (gestion de maintenance assistée par ordinateur) et l'analyse des données.

Voici quelques exemples de ces outils :

- [InUse](#)
- [MAINTI4](#)
- [Mobility Work](#)
- [XMaint](#)

La plupart de ces solutions offrent un service en ligne avec des versions web ou application mobile.

Deuxième partie

Recherche et Expérimentation

Chapitre 2

Le Jeu De Données

2.1 PRÉSENTATION

Choix du Dataset

Sans grande surprise, les entreprises communiquent peu sur leurs problèmes de production et leurs pannes, elles partagent donc difficilement leurs données de maintenance. Il est impossible de trouver un dataset de logs de maintenance et de données télémétriques industrielles réelles publique. C'est pourquoi le jeu de données utilisé sur ce projet est fictif. Microsoft Azure met à disposition un ensemble de cinq datasets [1] pour la maintenance prédictive faits à partir de données simulées. Le choix s'est arrêté sur cet ensemble pour ce projet car c'est le plus volumineux qui a pu être récupéré avec 5Go de données. Il y a environ un an d'historique.

Vue d'Ensemble

Machines	Telemetry	Maintenance	Errors	Failures
Informations sur les 1000 machines 1 000 x 3 MachineID Age Model [1:5]	Moyenne horaire des mesures 8 761 000 x 6 Datetime MachineID Volt Rotation Pressure Vibration	Logs: interventions programmées ou non > Sur 1.5 an 32 592 x 6 Datetime MachineID Comp [1:4]	Log : erreurs sans interruption (pas de panne) 11 967 x 3 Datetime MachineID ErrorID [1:5]	Logs: remplacements de composant (panne) 6 726 x 3 Datetime MachineID Failure {complID}

FIGURE 2.1 – Le Jeu de Données

2.2 DATASET MACHINES

Ce dataset contient les informations sur chacune des 1000 machines constituant le parc industriel fictif. Il donne trois informations :

1. l'ID de la machine
2. l'âge de la machine
3. Le modèle de la machine

	machinID	model	age
0	1	model2	18
1	2	model4	7
2	3	model3	8
3	4	model3	7
4	5	model2	2

Modèle des Machines

Il y a cinq modèle de machine numérotés de 1 à 5. Les modèle 3 et 4 représentent à eux deux près de 80% du parc. Le modèle de machine le moins répandu et le modèle 1 avec seulement 8.5% des machines.

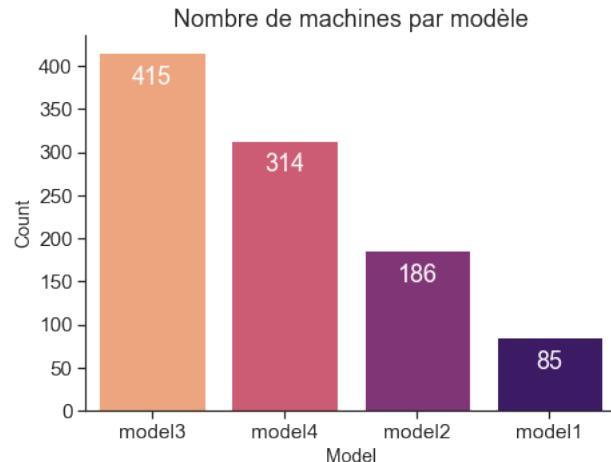
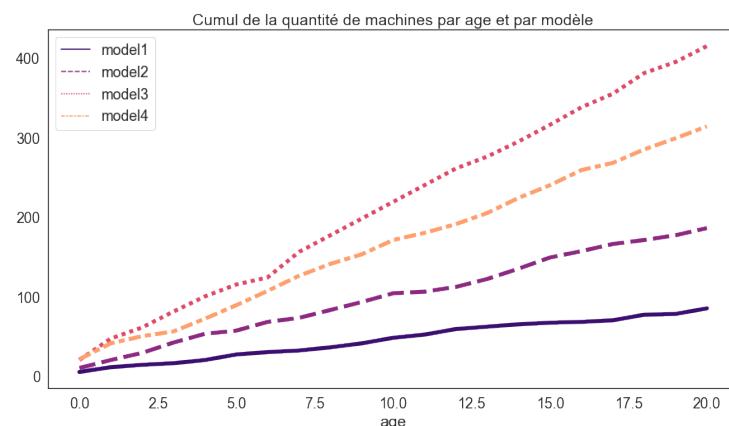


FIGURE 2.2 – Répartition des Modèles

Age des Machine

Les machines ont un âge compris entre 0 et 20 ans. Le nombre et la répartition des machine est globalement stable en fonction des âges. On peut en déduire que la construction du parc s'est faite de manière linéaire comme le montre la figure 2.3

FIGURE 2.3 – Évolution du Parc de Machines



Le nombre indiqué sur chacune des portions des barres de la figure 2.2 représente le pourcentage de machines pour le modèle donné à cet âge.

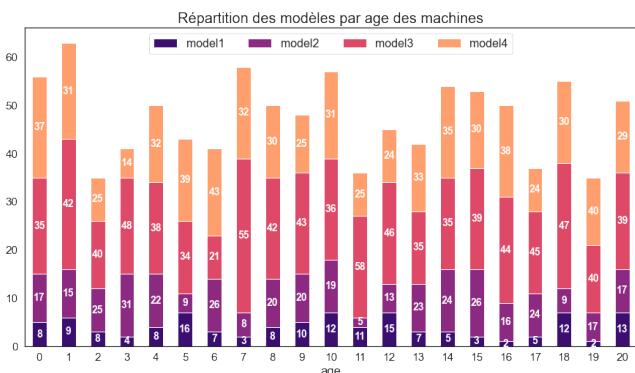


FIGURE 2.4 – Répartition des modèles par âges

2.3 DATASET MAINTENANCE

Il contient les logs de maintenance sur un an et demi, du 1er Juin 2014 à 6 heures au 1er Janvier 2016 à 6h. Chaque entrée correspond au remplacement d'un composant sur une machine pour un jour donné. Il peut s'agir d'un remplacement programmé ou non. Si l'intervention de maintenance s'est faite suite à une panne, elle est enregistrée à la fois de ce dataset et dans le dataset `failure`. Il contient six mois de données de plus que les autres logs car on (Azure) considère que ces logs permettent d'inférer sur le cycle de vie des composants.

Les logs sont enregistrés à 6h. Il y a 32 592 entrées.

Chacune d'elles donne les informations suivantes :

- l'ID de la machine concernée
- la date et l'heure du contrôle
- Le composant remplacé

	datetime	machineID	comp
0	2014-07-01 06:00:00	1	comp4
1	2014-09-14 06:00:00	1	comp1
2	2014-09-14 06:00:00	1	comp2
32590	2015-12-26 06:00:00	1000	comp3
32591	2015-12-26 06:00:00	1000	comp1

FIGURE 2.5 – Maintenance

Composants

Il y a en tout quatre types de composant, ils sont numérotés de 1 à 4. Chaque machine possède les 4 types de composants quel que soit son modèle.

```
df_maint.groupby('machineID')['comp'].nunique().unique()
>> array([4])
```

Fréquence des logs

En analysant la fréquence des entrées (figures 2.7a et 2.6) on voit que pour le deuxième semestre 2014 les interventions de maintenance sont programmées une fois toutes les deux semaines et environ 25% des machines sont inspectées à chaque fois. Tandis que pour l'année 2015, les interventions sont quotidiennes mais seulement 10% du parc environ est contrôlé chaque jour.

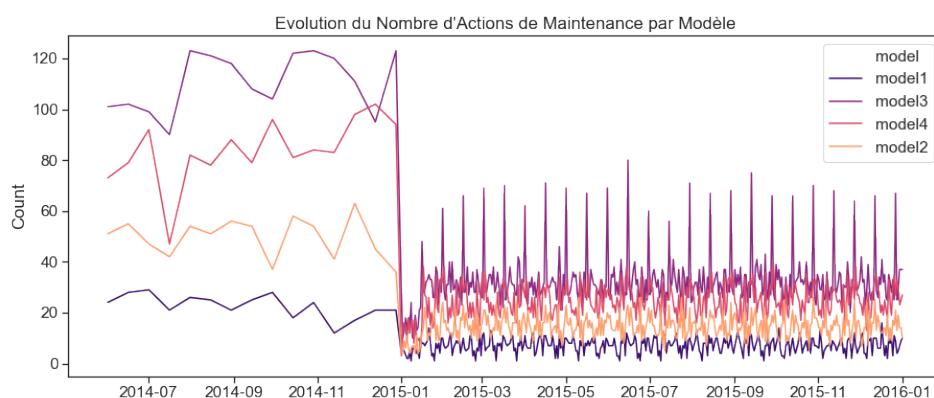
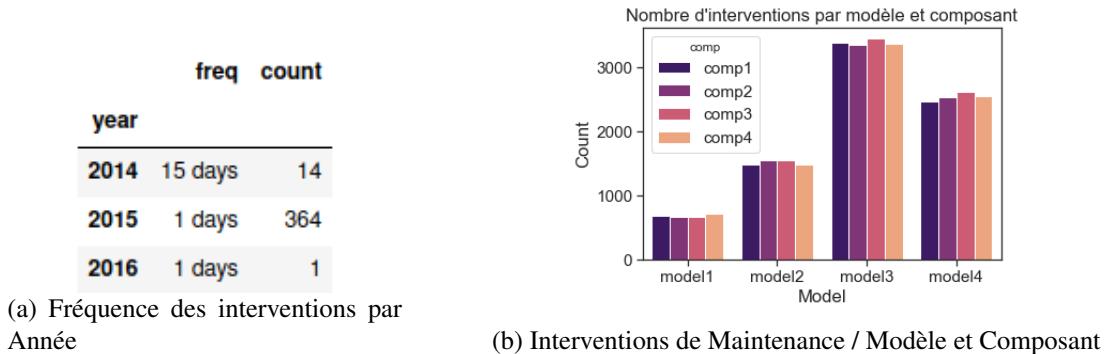


FIGURE 2.6 – Interventions de Maintenance par Modèle

Le choix des machines contrôlées est fait de telle sorte que chaque modèle a la même proportion de ses machines inspectées.

Tous les composants des machines ne sont pas remplacés à chaque intervention mais pour chaque machine, chacun des composants est remplacé en moyenne deux fois par mois (cf. annexe A).

FIGURE 2.7 – Visualisation des logs de maintenance



2.4 DATASET ERRORS

Ce dataset contient les logs pour l'année 2015 des erreurs qui ne causent aucune interruption dans la production et qui ne nécessite pas de remplacer un composant.

Il y a 11 967 lignes. Chaque entrée donne :

- l'heure et la date de l'erreur
- l'ID de la machine concernée
- l'ID du type d'erreur

	datetime	machineID	errorID
0	2015-01-06 03:00:00	1	error3
1	2015-02-03 06:00:00	1	error4
2	2015-02-21 11:00:00	1	error1
11965	2015-09-11 06:00:00	1000	error1
11966	2015-10-11 14:00:00	1000	error3

FIGURE 2.8 – Errors

Les Types d'Erreurs

Il y a cinq types d'erreurs. Les types 1, 2 et 3 représentent chacun environ un quart des erreurs. Le type 5 est le moins courant mais la figure 2.9 montre qu'elle n'occurre que pour les machines de 14 ans ou plus.

Bien qu'elles soit présente sur tous les modèles, la proportion de chaque modèle diffère de leurs proportions globales sur l'erreur de type 4. Elle est en effet plus courante sur les modèles 1 et 2 (cf. figure 2.10b). Pour ces deux modèles, elle représente environ 30% des erreurs. Les erreurs de type 1, 2 et 3 font chacune environ 20% des erreurs pour les modèles 1 et 2.

En revanche, l'erreur de type 4 est très rares sur les machines des modèles 3 et 4. Pour ces modèles ce sont les erreurs de type 1, 2 et 3 qui surviennent le plus souvent, avec en moyenne 24% à 29% d'occurrences chacune.

Les chiffres sur chacune des barres de la figure 2.10b représentent le pourcentage d'occurrence de l'erreur de donnée pour chacun des modèles.

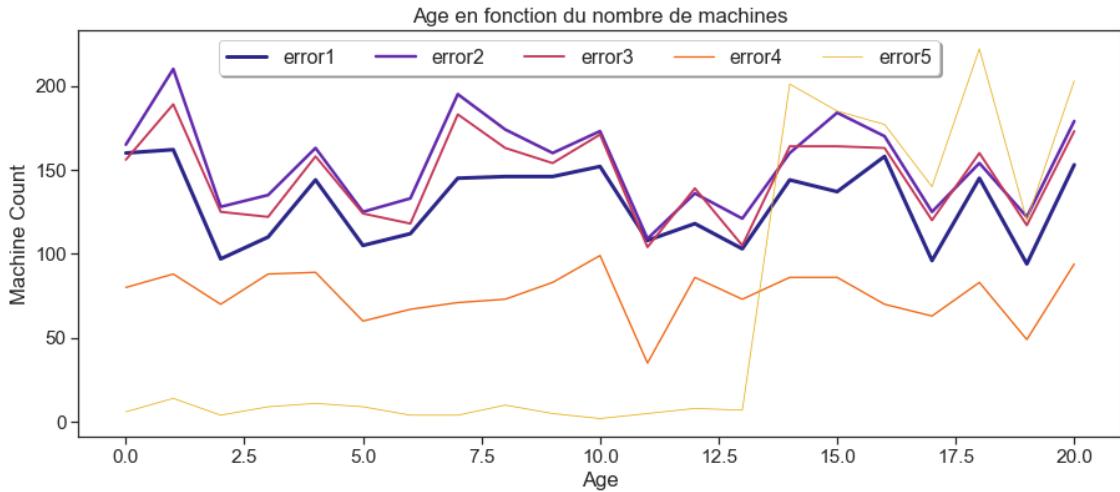
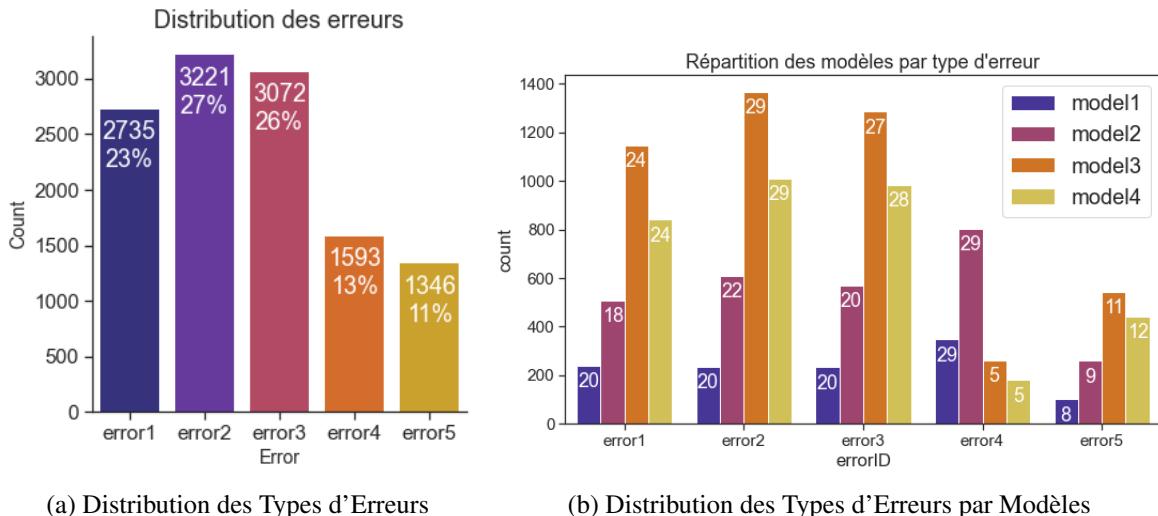


FIGURE 2.9 – Nombre de logs d’erreurs en fonction des Ages des Machines

FIGURE 2.10 – Maintenance - Distribution



2.5 DATASET FAILURES

Ce dataset contient les logs des pannes pour l’année 2015. On considère qu’il y a une panne dès lors qu’un composant doit être remplacé. Les pannes sont des erreurs qui impliquent un arrêt de la machine.

Il y a 6 726 lignes. Les pannes sont toujours enregistrées à 6h. Pour une machine donnée et une date donnée (heure et jour) on peut donc avoir plusieurs entrées, c’est le type du composant qui différera.

Chaque log donne les informations suivantes :

- l’heure et la date de la panne
- l’ID de la machine concernée
- l’ID du composant remplacé

	datetime	machineID	failure
0	2015-02-04 06:00:00	1	comp3
1	2015-03-21 06:00:00	1	comp1
2	2015-04-05 06:00:00	1	comp4
6724	2015-08-13 06:00:00	1000	comp2
6725	2015-09-12 06:00:00	1000	comp1

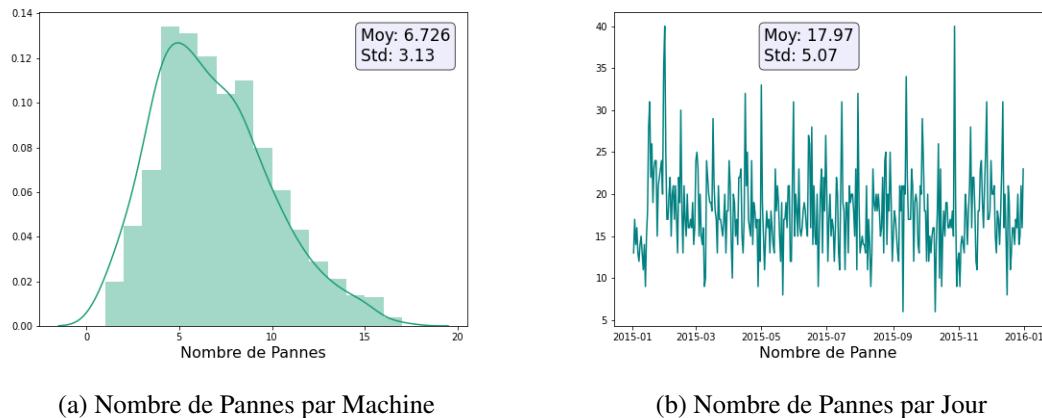
FIGURE 2.11 – Failures

Fréquence des Pannes

Il y a plus de 6 000 pannes enregistrées sur un an pour 1000 machines, on peut facilement supposer que chaque machine connaît en moyenne six pannes par an. La figure 2.12a confirme cela.

Ça peut sembler peu mais avec 1000 machines dans le parc industriel cela veut dire qu'il y a en moyenne 18 pannes par jour (figure 2.12b).

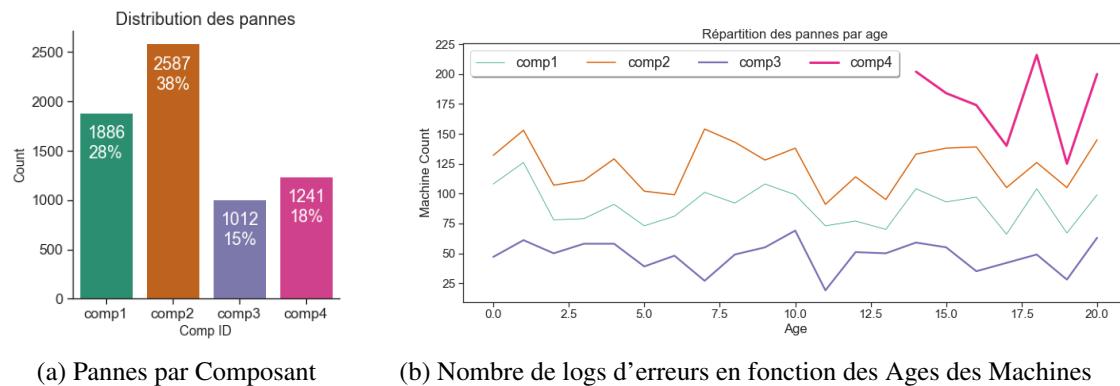
FIGURE 2.12 – Failures - Distribution



Répartition des Pannes

Les composants 1 et 2 représentent à eux deux les deux tiers des pannes, avec respectivement 28% et 38% des pannes chacun. Le composant 3 génère le moins de pannes avec 15% d'occurrence.

FIGURE 2.13 – Failures



Panne par Age

Là encore, il existe un type de composant qui ne crée des pannes que pour les machines de 14ans ou plus, il s'agit du composant 4. On peut déjà supposer que l'erreur de type 5 est liée au composant 4. Les trois autres composants sont globalement répartis de la même façon pour tous les âges (figure 2.13b)).

Pannes et Type d'Erreurs

En rapprochant les logs d'erreurs aux pannes on a pu lier dans la plupart des cas le type d'erreur associée à la panne. Puisque les erreurs sont enregistrées au moment précis où elles occurrent mais les pannes, tout comme pour le logs de maintenance, sont enregistrées une fois par jour à 6h, il a fallu trouver pour chaque panne si une erreur avait été enregistré pour la machine concernée dans les dernières 12 heures. Il existe 182 pannes (soit 2.7% des pannes), pour lesquelles une erreur n'a pu être identifiée d'après les logs.

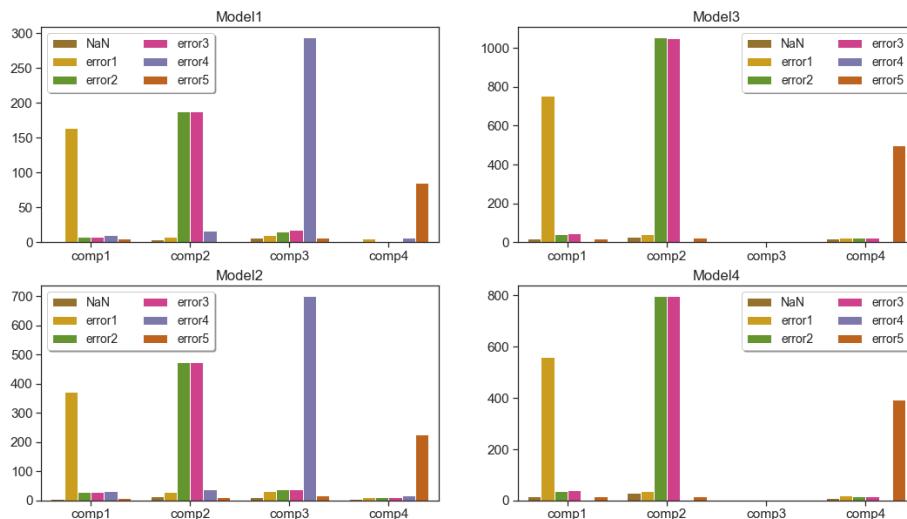


FIGURE 2.14 – Répartition des Pannes en fonction des Composants et des Erreurs pour chaque modèle

La figure 2.14 nous permet de constater les types d'erreurs et les ID des composants causant des pannes sont corrélés. Voici ce qu'on peut observer :

- Le composant 1 génère presque toujours une erreur de type 1
- Le composant 2 génère presque toujours des erreurs de type 2 et 3 dans des proportions égales.
- Le composant 3 génère presque toujours une erreur de type 4
- Le composant 4 génère presque toujours une erreur de type 5
- Les modèles 1 et 2 et les modèles 3 et 4 forment des groupes.

2.6 DATASET TELEMETRY

C'est le dataset le plus large avec plus de 5Go de données et 8,7 millions de lignes. La moyenne des quatre types de mesures envoyées par les capteurs est calculée toutes les heures pour chaque machine sur toute l'année 2015.

Il y a 8 761 000 lignes. Chaque log donne les informations suivantes :

- l'heure et la date
- l'ID de la machine
- la tension ('volt')
- la rotation
- la pression ('pressure')
- la vibrations

	datetime	machineID	volt	rotate	pressure	vibration
0	2015-01-01 06:00:00	1	151.919999	530.813578	101.788175	49.604013
1	2015-01-01 07:00:00	1	174.522001	535.523532	113.256009	41.515905
2	2015-01-01 08:00:00	1	146.912822	456.080746	107.786965	42.099694
8760998	2016-01-01 05:00:00	1000	160.007424	462.740287	108.397268	47.206940
8760999	2016-01-01 06:00:00	1000	164.590354	415.431358	102.142896	37.919958

FIGURE 2.15 – Failures

L'étude en série temporelle n'apporte aucune information de plus. Quel que soit la mesure, il n'y a pas de tendance, la légère périodicité observée toutes les deux semaines est sûrement liée à la façon dont ces données fictives ont été générées.

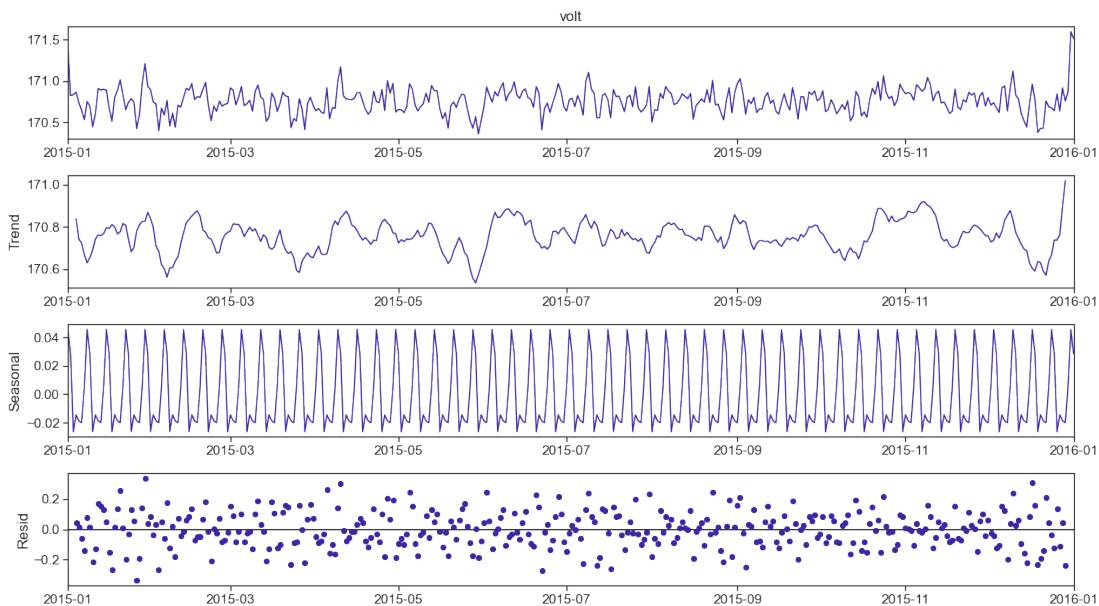


FIGURE 2.16 – Décomposition de la série temporelle Volt

Dans la partie suivant nous verrons comment extraire des ces datasets des features exploitables pour les algorithmes de machine learning.

Chapitre 3

Feature Engineering

3.1 TELEMETRY

Pour travailler avec ce dataset, chaque mesure est décomposée en moyennes mobiles et en écart-types mobiles. Dans un premier temps, les moyennes et écarts-types sont construits sur des fenêtre de 6, 12, 24 et 36 heures. Ils sont bien entendu calculés par machine et exclusivement à partir des données antérieures.

Ainsi, chacune des quatre mesures est remplacée par huit nouvelles colonnes. Ensuite les données sont regroupées par machine et par intervalle de 12 heures en faisant une moyenne sur l'intervalle.

Finalement, on obtient un dataset de 731 000 lignes et 34 colonnes.

3.2 ERRORS

On commence par *dummifier* la colonne `errorID` pour avoir une colonne booléenne par type d'erreur. Puis une somme mobile est faite pour chaque erreur pour des fenêtres de 6, 12, 24 et 36 heures.

Pour pouvoir aligner ces nouvelles features avec les features de télémétrie, elles sont aussi regroupées par machine et par intervalle de 12 heures en faisant une moyenne.

On obtient un dataset de 731 000 lignes et 22 colonnes.

3.3 MAINTENANCE

Ici aussi la première étape consiste à *dummifier* la colonne `comp` pour avoir une colonne booléenne par composant.

Ensuite on construit deux nouvelles colonnes par composant :

- La première comptabilise le nombre de jours depuis le dernier contrôle
- la deuxième comptabilise le nombre de contrôles effectué depuis le début

De même les données sont regroupées par machine et par intervalle de 12 heures en faisant une moyenne.

Le dataset de sortie contient 731 000 lignes et 10 colonnes

3.4 FINALISATION

Preprocessing Pipeline

Les colonnes `datetime` et `machineID` sont utilisées comme index pour chacun des trois transformations décrites précédemment.

Un *One Hot Encoder* est utilisé sur la colonne `model` du dataset `machines` et la colonne `age` est laissée telle quelle.

Les résultats des transformations sont enfin joints ensemble. La pipeline finale de preprocessing contient les étapes suivantes :

- *Transformer* pour le dataset `telemetry`
- *Transformer* pour le dataset `errors`
- *Transformer* pour le dataset `maintenance`
- *Transformer* pour le dataset `maintenance`
- *One Hot Encoder* pour le dataset `machines`
- *Merger* pour les 4 transformations

Standardisation

Les données sont standardisées avec un *StandardScaler*. Les features suivent en grande majorité une distribution gaussienne (cf. annexe B) et ce type de standardisation permet de ne pas "effacer" les outliers, ce qui est primordial ici car ce sont sûrement les outliers qui permettront de détecter les pannes.

Programmation Orientée Objet

Pour chacune des trois transformations majeures (`telemetry`, `errors`, `maintenance`) une classe qui hérite du *TransformerMixin* de *Scikit-Learn* a été créée de sorte à avoir de véritables *transformer* personnalisés.

Chaque classe contient au moins une méthode *fit* et une méthode *transform*. Elles héritent toutes d'une classe mère qui fournit entre autres les méthodes pour agréger les données par intervalle de 12h et pour faire les moyennes, écarts-types ou sommes mobiles.

Historique

La particularité de ces features est que chaque échantillon est lié à un certain nombre des échantillons qui le précèdent directement, en fonction des fenêtres choisies pour les moyennes, écarts-types et sommes mobiles.

C'est pourquoi l'une des méthodes les plus importantes est celle qui permet de garder un historique des données brutes pour pouvoir avoir une continuité sur les moyennes (ou écarts-types ou sommes) mobiles lorsque la méthode *transform* est appelée sur de nouvelles données (c'est-à-dire différentes de celles utilisées sur le *fit*). Le strict minimum est historisé pour ne pas alourdir inutilement la pipeline finale. On se base donc sur la plus grande des fenêtres utilisées pour les moyennes (ou écarts-types ou sommes) mobiles pour construire l'historique. Par exemple, à chaque fois que la méthode *transform* est appelé, les 35 entrées les plus récentes (donc les 35 dernières si elles sont classées par ordre croissant) sont conservées pour chaque machine.

Ainsi, que tout le jeu de données soit transformé directement ou que ça soit fait intervalle par intervalle, les résultats des fonctions d'agrégation mobiles sont bien les mêmes.

Création des labels

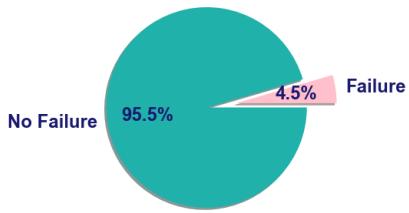
Le dataset `failures` va servir à créer les labels. La *target* consiste dans un premier temps en une colonne qui donne la liste des composants qui vont générer une panne dans les K_{period} prochains jours ou "no failure" si aucune panne n'est prévue. K_{period} correspond en réalité à un nombre d'intervalles de temps. Puisque nos données sont regroupées par intervalles de 12h, si K_{period} est égal à 4 cela signifie que les labels donnent les risques de pannes pour les deux jours (48 heures complètes) précédents la panne.

Cette colonne est ensuite transformée avec un *MultiLabelBinarizer* afin de pouvoir être comprise comme étant multi-label par les modèles. Finalement la *target* finale est composé de cinq colonnes booléennes.

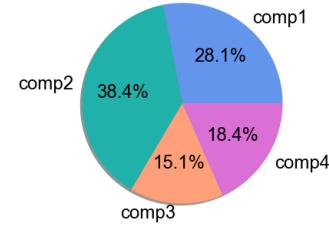
machineID	datetime	comp1	comp2	comp3	comp4	no failure
		0	0	0	1	0
890	2015-01-01 12:00:00	0	0	0	1	0
	2015-01-02 00:00:00	0	0	0	1	0
	2015-01-02 12:00:00	1	0	0	1	0
	2015-01-03 00:00:00	1	0	0	0	0
	2015-01-03 12:00:00	1	0	0	0	0
	2015-01-04 00:00:00	1	0	0	0	0
	2015-01-04 12:00:00	1	0	0	0	0
	2015-01-05 00:00:00	0	0	0	0	1
	2015-01-05 12:00:00	0	0	0	0	1
	2015-01-06 00:00:00	0	0	0	0	1

FIGURE 3.1 – Cible Finale (y) pour $K_{period} = 4$

FIGURE 3.2 – Target - Distribution



(a) Proportions "Failure" vs "No Failure"



(b) Proportions des Composants si "Failure"

Voici le schéma récapitulatif de la pipeline de feature engineering qui sera utilisée :

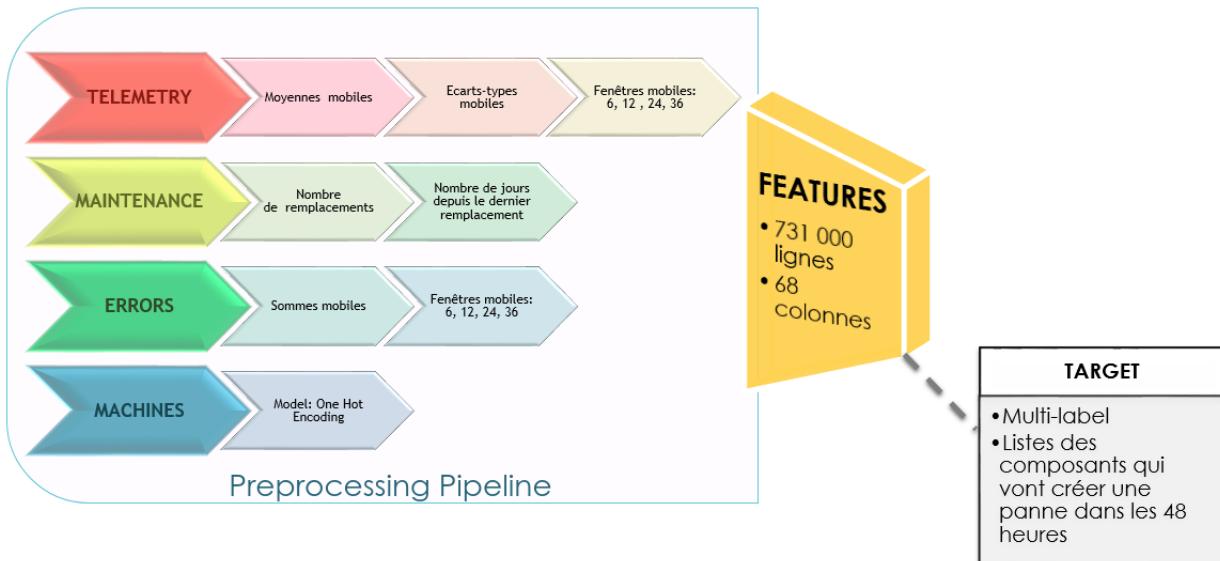


FIGURE 3.3 – Pre-processing Pipeline + Labels

On verra dans le prochain chapitre que notre meilleur modèle ne fera pas mieux que 2 jours d’anticipation pour les pannes.

Chapitre 4

Algorithme de Classification

4.1 MODALITÉS

Multi-Label

Encore une grande majorité des classificateurs disponibles dans la librairie *scikit-learn* ne gère pas les cas multi-labels. Parmi les modèles mis en concurrence ici, les modèles de régression logistique, Naive Bayes et SVM ne supportent pas les cas multi-label. La librairie offre tout de même la possibilité de travailler avec ces classificateurs sur ce type de cas en les combinant à la classe [OneVsRest](#). Cette classe permet d'entraîner un estimateur par label de sorte à ce que chaque estimateur soit une classification binaire où la target vaut 1 si le label en question vaut 1 et 0 sinon.

Ajustement des Prédictions

Dans certains cas, le label `no_failure` vaut 1 lorsque l'un des autres labels vaut aussi 1. Or il faut que ce label soit mutuellement-exclusif avec tous les autres. Puisque dans un premier temps il est d'abord essentiel de bien identifier chaque panne, chaque prédition est ajustée afin que le label `no_failure` soit égal à 0 si au moins un des autres labels vaut 1. De même il existe de rares cas où tous les labels valent 0. Dans ce cas la valeur du label `no_failure` est remplacé par 1.

Hyperparamètres

Par défaut une *Random Search* a été utilisée pour trouver les meilleurs hyperparamètres de chaque modèle. Pour le XGBM (librairie XGBoost) c'est une méthode d'optimisation bayésienne [2] [3] avec la librairie [bayesian-optimisation](#) qui a été utilisée car elle est plus rapide. Cette méthode d'optimisation requiert moins d'itérations pour des résultats plus ou moins équivalents.

En effet, Les modèles ensemblistes nécessitent un temps d'entraînement plus long en général puisqu'un certain nombre de *learners* (petits estimateurs) sont construits pour chaque modèle. Dans ce cas précis, il peut être intéressant d'utiliser une optimisation bayésienne pour deux raisons :

1. Large volume de donnée : 731 000 échantillons pour 68 features
2. Target multi-label : Un estimateur doit être créé pour chaque label. Dans le cas présent c'est donc cinq estimateurs qui sont entraînés par modèle.

Même en utilisant la parallélisation les temps de *fit* sont assez longs sur ce type de modèle (environ 25-30 min heure pour XGBoost). Une Random Search avec seulement 50 itérations et $k = 3$ pour la cross-validation prendrait environ 3 jours.

La meilleure combinaison d'hyper-paramètres a été obtenue au bout de six itérations seulement pour XGBoost (cf. annexe C).

Un résumé des temps computationnel pour l'apprentissage et les prédictions est disponible en annexe D

Choix des Scores

La cible est extrêmement déséquilibrée. La classe no failure est attribuée à 95.5% des échantillons voir figure 3.2a. C'est pourquoi l'accuracy et le ROC-AUC ne sont donnés qu'à titre informatif. Ces deux mesures sont influencées par la classe prédominante. Si le modèle se contentait de classer tous les échantillons dans la classe no failure il ferait des déjà 95% en accuracy et en ROC-AUC.

Un bon exemple est donné avec le modèle de Naive Bayes. Sur le test-set, pour le label comp1 il donne une accuracy et un ROC-AUC d'environ 94% (cf. annexe F.2b). Son taux de vrais positifs est de 67.5% avec un taux de faux positifs très faible à 4.73% (voir en annexe E.3). En réalité il a 1 198 échantillons bien classés en positif pour 6 872 mal classés en positif. La précision est en effet très mauvaise avec un score de 14.8% seulement.

Pour évaluer les véritables performances des modèles le recall, la précision et le F1-Score sont privilégiés. Pour une usine on peut supposer que les coût d'une panne causant une interruption dans la production est plus important que celui de remplacer un composant alors qu'il est encore fonctionnel. Il est donc important d'avoir un bon recall sur chacun des labels de type failure. Cependant, si la précision est trop faible et qu'il y a alors beaucoup de faux négatif, cela veut dire que des composant sont remplacé prématurément ce qui revient à peu de chose près à faire de la maintenance prévisionnelle. On perd tout l'intérêt de la maintenance prédictive. Il faut donc un bon F1-score sur chacun des labels en s'assurant qu'il n'y a pas un trop grand déséquilibre entre le recall et la précision.

Pour le calcul des scores moyens, une moyenne macro est utilisée afin que les labels soient considérés comme étant équipondérés. Elle annule le déséquilibre des classes et permet ainsi de facilement comparer les modèles sur leurs score moyens.

Train / Test Sets

Les échantillons étant liés à leurs plus proches prédecesseurs, le split entre les sets train et test ne peut être aléatoire. Dans la pratique ce type de donnée est produite de façon ordonnée. Elles doivent rester groupées par machineID et ordonnées par datetime. C'est pourquoi le dataset est découpé avec le ratio 80/20 à de tel sorte que le train-set regroupe les données du 1er Janvier 2015 6h au 20 octobre 2015 minuit et le test-set celles du 20 octobre 2015 12h au 1er Janvier 2016 12h.

4.2 COMPARAISON DES MODÈLES

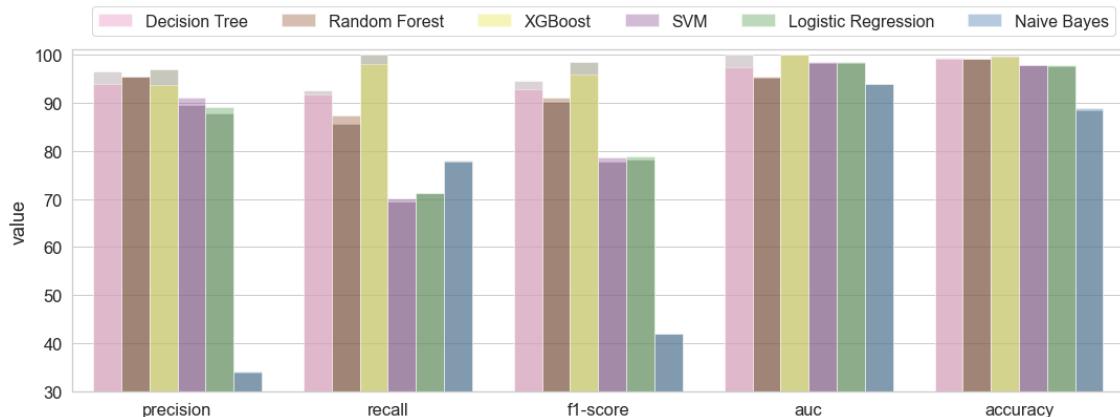
Six algorithmes de machine learning ont été mis en concurrence. Le modèle de K-Nearest-Neighbors a été disqualifié car il lui faut plus d'une heure pour faire les prédictions sur tout le

test-set (contre quelques secondes pour tous les autres modèles).

Voici ce qui ressort des résultats visibles sur la figure 4.1 :

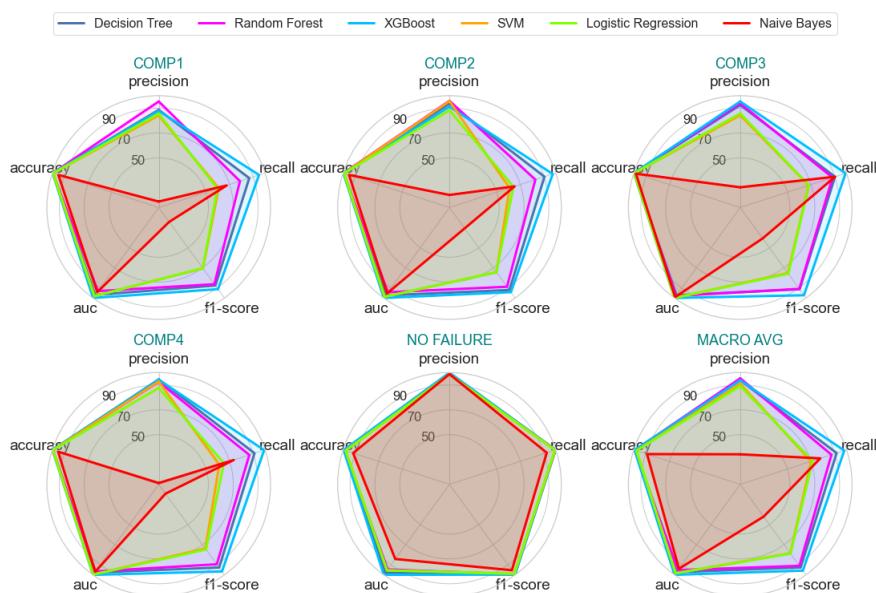
- les modèles basés sur des arbres de décision donnent de très bons résultats
- il y a peu ou pas d'overfitting quel que soit le modèle
- Le classificateur de XGBoost obtient les meilleurs scores de recall et F1-Score
- Le Naive Bayes classe très mal avec une précision de 34% (beaucoup de faux positif)

FIGURE 4.1 – Score des Moyens Modèles - Train/Test



Chaque barre du graphique donne le score sur le train-set (couleur désaturée) et le test-set (couleur saturée) pour un modèle et une mesure de performance.

FIGURE 4.2 – Score des Modèles par Label - Test-Set



Modèle Gagnant

Le classificateur de XGBoost est de toute évidence le meilleur modèle. Il a des recalls et des F1-Scores supérieurs quelque soit le label. La Random Forest a une précision un peu plus forte (95.3% contre 93.6% pour XGBoost en moyenne macro) mais c'est au coût d'un recall et d'un F1-Score bien plus faible (87% contre 98% pour XGBoost sur le recall). Les détails des scores sont disponibles en annexe [F](#).

Voici les caractéristiques du modèle

```
OneVsRestClassifier(estimator=XGBClassifier(alpha=5.73,
    base_score=None,
    booster=None,
    colsample_bylevel=None,
    colsample_bynode=None,
    colsample_bytree=0.7, gamma=2.48,
    gpu_id=None, importance_type='gain',
    interaction_constraints=None,
    learning_rate=0.15,
    max_delta_step=None, max_depth=17,
    min_child_weight=None, missing=nan,
    monotone_constraints=None,
    n_estimators=150, n_jobs=6,
    num_parallel_tree=None,
    random_state=None, reg_alpha=None,
    reg_lambda=None,
    scale_pos_weight=None,
    subsample=0.68, tree_method=None,
    validate_parameters=None,
    verbosity=None),
    n_jobs=5)
```

4.3 FEATURE IMPORTANCE

Afin de déterminer quelles sont les features qui sont le plus déterminantes pour la prédiction, un nouveau modèle XGBoost a été entraîné sur des features dont les fenêtres des fonctions d'agrégations sont étendues à 6, 12, 18, 24, 36 et 48.

En regroupant les features par catégorie (voir annexe [G.1](#)), on peut dire ceci :

- Les fenêtres 18 et 48 sont les moins utiles
- Les fenêtres 24 et 36 importantes surtout pour le label `no failure`
- les fenêtres 6, 12 et 18 sont importantes pour les autres labels
- pour prédire une panne les features de `telemetry` sont les plus importantes
- pour le label `no failure` les features de `errors` sont les plus importantes
- l'age des machines est déterminant pour prédire une panne du composant 4.
- chacun des 4 types de composant est relié à un l'une des mesures télémétriques (voir annexe [G.2](#)) :

```
comp1 : volt
comp2 : rotate
comp3 : pressure
```

comp4 : vibration

Finalement les fenêtres 6, 12, 24 et 36 semblent optimales. Le classificateur avec les fenêtres étendues a une précision similaire mais un recall très légèrement inférieur (97.8 contre 98.07, cf. annexe F.3). Il n'y a donc aucun intérêt augmenter le nombre de features (qui est de 91 avec les fenêtres étendues) d'autant plus que le classificateur est deux fois plus long à entraîner.

4.4 AJUSTEMENT DE LA TARGET

Après avoir changé la valeur de K_{period} à 6 lors de la construction de la cible, on voit que le modèle ne parvient pas à bien classifier les échantillons à plus de 48 heures de la panne. En étendant cette valeur, il commence même à faire plus d'erreurs de classification sur les échantillons à moins de 48 heures de la panne. C'est pourquoi la valeur du K_{period} reste fixé à 4.

Troisième partie

Mise en Place de la Solution

Chapitre 5

Web Application

Dans le but de pouvoir démontrer l'efficacité du modèle en temps réel, une application web a été conçue à partir du framework *Django*. Elle permet de voir les prédictions de pannes faites à partir d'une simulation d'un flux de données (logs et télémetrie).

5.1 FLUX DE DONNÉES

Emission des Données

Les données brutes ont préalablement été découpées en train-set et test-set chacune à partir de la même utilisé pour entraîner le modèle, c'est-à-dire le 20 octobre 2015 12h.

Le flux de données est simulé à partir des test-sets des données brutes en utilisant *Kafka*.

Un **Kafka Producer** s'occupe de prendre les données à partir des quatre datasets `telemetry`, `errors`, `maintenance` et `failures` pour les envoyer dans un topic dédié de *Kafka*. Il simule le travail de transmission des mesures faites par les capteurs et l'enregistrement des logs de maintenance, d'erreurs et de pannes.

Pour simplifier les choses, il y a une partition par intervalle de 12 heures. Ainsi la pipeline de preprocessing n'est déclenché que lorsqu'une partition est créée et complète.

Récupération et Traitements des Données

Un **Kafka Consumer** récupère les données de chaque partition et stocke les données brutes en base de données.

Dès qu'un lot de données (soit toutes les données sur un intervalle de 12 heures) est intégralement récupéré, elles sont ensuite passé dans la pipeline de preprocessing pour en extraire les features qui nous intéressent. Ces features sont enfin envoyées au modèle de classification et les prédictions sont stockées en base de données.

Une version agrégée par intervalle de 12 heures du dataset `failures` est stockée telle quelle de la base de données. Elle est utilisée à titre informatif dans l'application afin de pouvoir juger la qualité des prédictions faites.

Suivi de la Progression de la Simulation

Des tables uniquement utiles à l'application existent aussi. Elles permettent notamment de savoir quel lot de données a déjà été envoyé, extrait, preprocessé ou prédit.

Mise à jour de la Pipeline

Comme on l'a vu la pipeline garde un petit historique des données pour pouvoir agréger les données en fonctions mobiles. Elle est donc mise à jour à chaque fois qu'elle est utilisée. Chaque lot de données utilise la pipeline mise à jour par le lot précédent.

5.2 APPLICATION

File de Tâches et Tâches Récurrentes

Celery [4] est un gestionnaire de tâches asynchrone. Dans cette application il permet de faire deux chose :

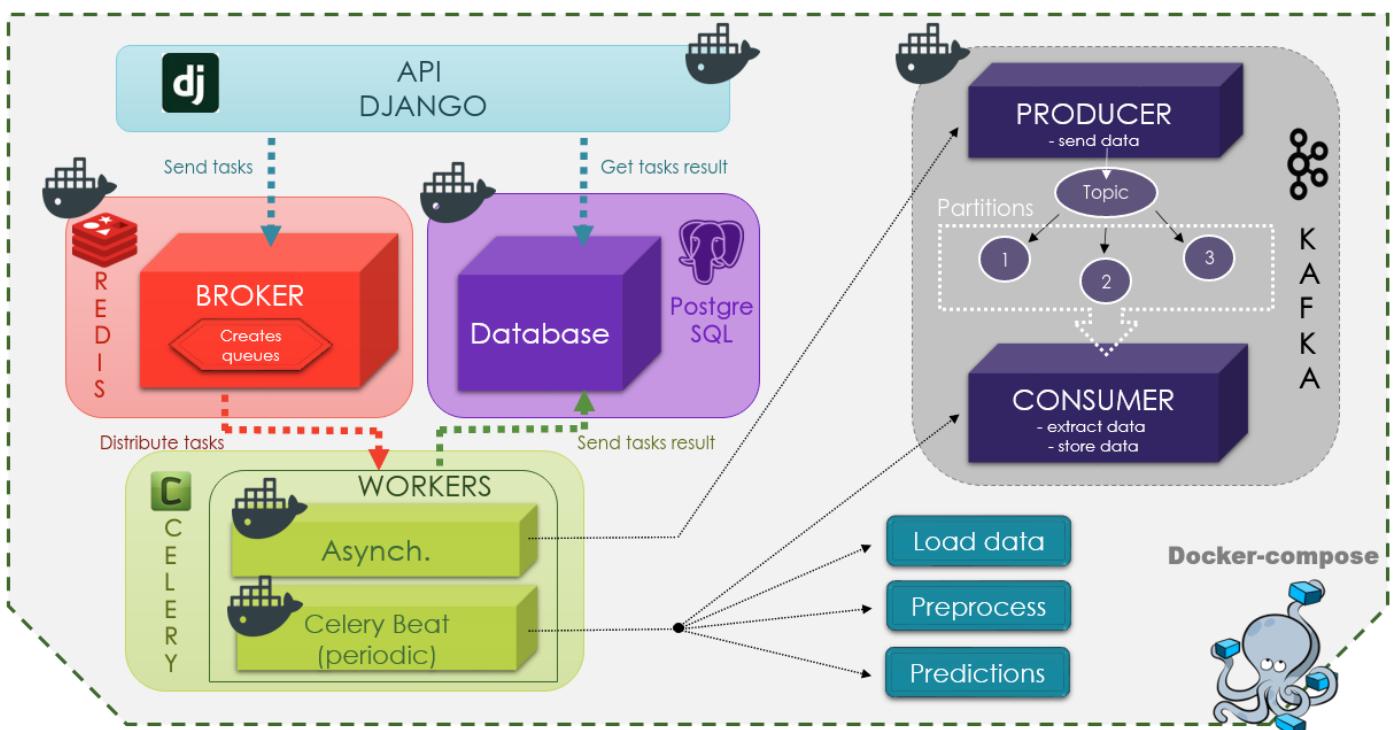
1. Démarrer la simulation du flux de données.
2. Programmer l'exécution automatique de la tâche qui lance le *kafka consumer* de façon périodique [5].

Dans un cas réel, le *kafka consumer* serait exécuté toutes les 12 heures puisque c'est l'intervalle utilisé par la pipeline de preprocessing. Pour la simulation ce serait trop long, on utilise donc un intervalle de cinq secondes seulement.

Docker

L'application web et tous les services qui lui sont utiles sont dockerisés et regroupés ensemble dans un docker-compose qui contient six conteneurs :

FIGURE 5.1 – Simulation du Flux de Données - Workflow



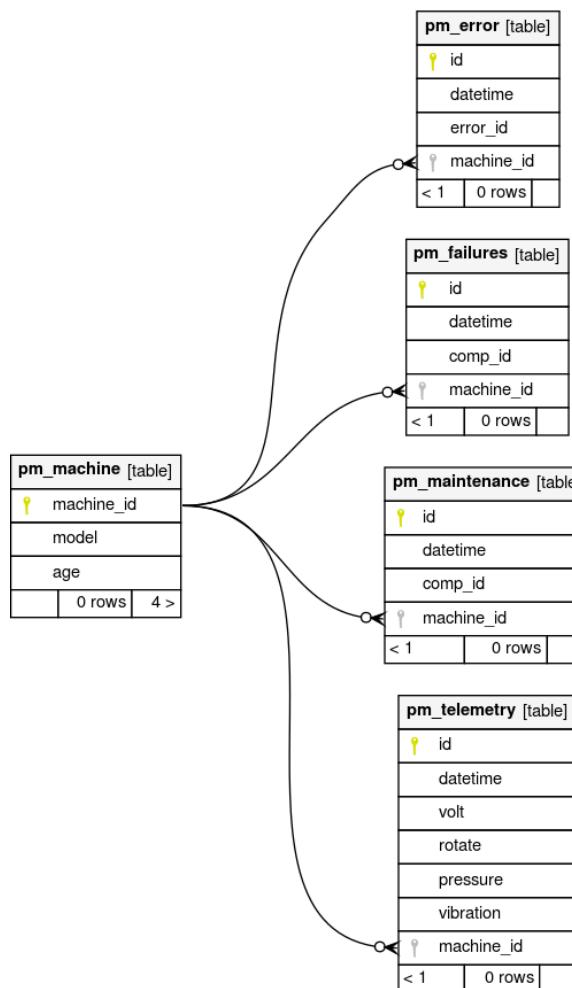
5.3 BASE DE DONNÉES

Au-delà des tables nécessaires à *Django*, *Celery* et *CeleryBeat*, deux autres groupes de tables ont été créés. Le premier groupe contient les données brutes (et quasi-brutes) des cinq datasets (figure 5.2a). Le deuxième groupe contient les tables qui permettent de suivre l'évolution de la simulation (figure 5.2b).

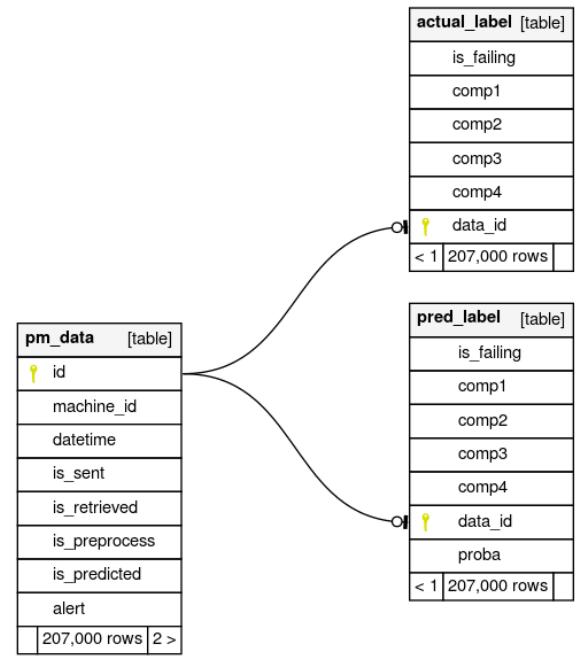
Une version condensée de l'intégralité du schéma relationnel de la base de données est visible en annexe

FIGURE 5.2 – Schema Relationnel Détaillé

(a) Groupe 1



(b) Groupe 2

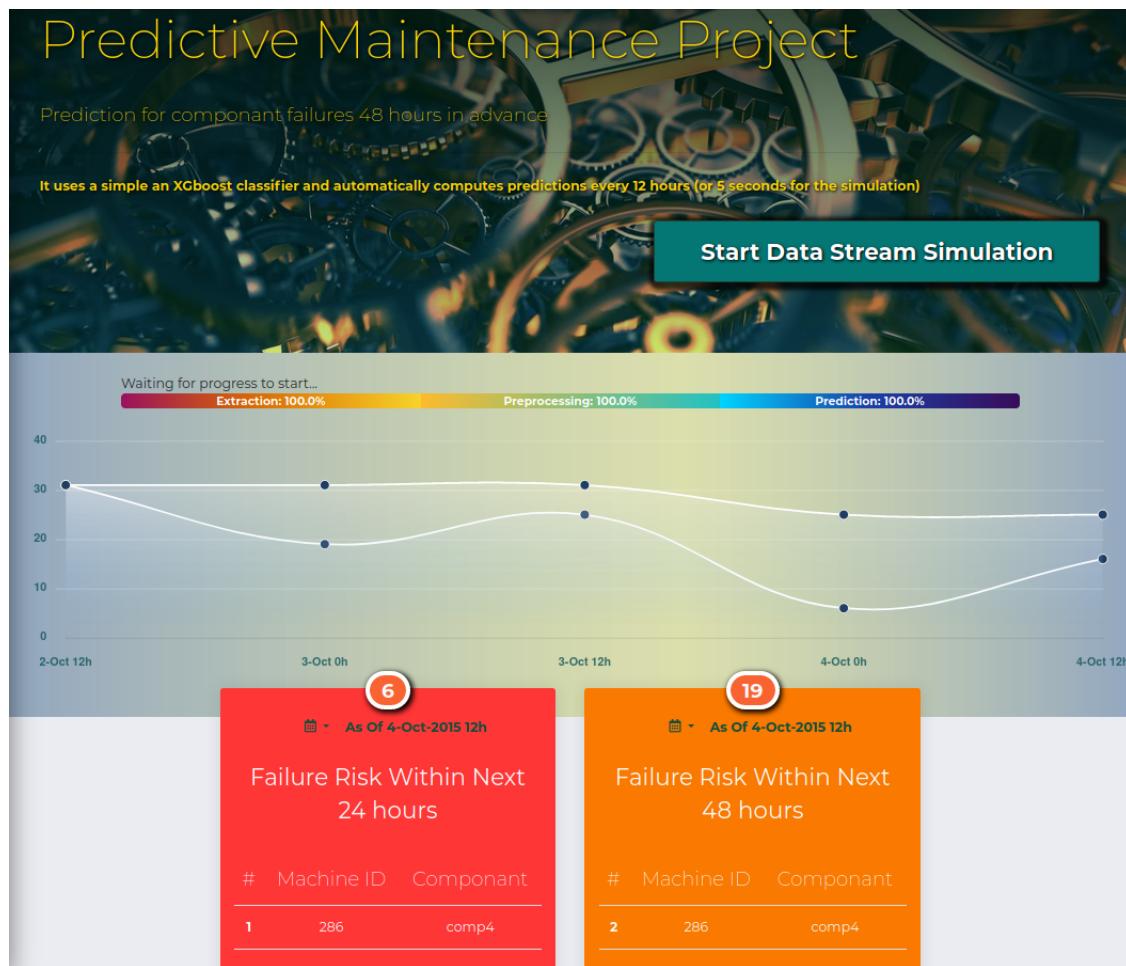


5.4 INTERFACE UTILISATEUR

L'interface est composée d'une première partie pour le lancement et le suivi de la simulation de donnée avec un graphique montrant le nombre de pannes prédictes et réelles et diverses barres de progression pour le suivi de l'état de la génération, de l'extraction, du pré-traitement et de la classification des données.

La deuxième partie montre quelles machines et quels composants risquent de tomber en panne à une date choisie. Il y a deux niveaux d'alertes. Orange si la panne est prévue dans les 48 heures et rouge si elle est prévue dans les 24 heures.

FIGURE 5.3 – Screenshot de l'application



Chapitre 6

Bilan

6.1 GESTION DE PROJET

Les différentes étapes de la mise en place du projet ont été organisées avec Trello.

The screenshot shows a Trello board titled "Projet Simplon - Prédictive Maintenance". The board is organized into several sections:

- Infos:** Objectif: Créer une application de maintenance prédictive avec simulation d'un flux de données. Github: https://github.com/Bapuch/Predictive_Maintenance. + Ajouter une autre carte.
- BIG STEPS:** Data Vizualisation (5/5), Pipeline Preprocessing (4/4), Machine Learning (4/4), Deep Learning (0/3), Django Application (9/11). + Ajouter une autre carte.
- TO DO:** Deployer l'API sur une plateforme Cloud (1), Utiliser Fastai (1). + Ajouter une autre carte.
- IN PROGRESS:** Faire un dashboard de visualisation des predictions (1). + Ajouter une autre carte.
- REVIEW:** Entrainé un LSTM (1), Entrainer un RNN (1). + Ajouter une autre carte.
- DONE:** Créer un Transformer pour Telemetry (1), Créer un Transformer pour Errors (1), Créer un Transformer pour Maintenance (1), Créer une classe pour construire les labels avec Failures (1), Dataviz pour le dataset de telemetry (1), Dataviz pour le dataset de machines (1), Dataviz pour le dataset de errors (1).

6.2 VERSION CONTROL

Tout le projet est versionné avec git et hébergé sur [GitHub](#) (repo privé).

The screenshot shows a GitHub repository page for 'Bapuch / Predictive_Maintenance'. The repository is private, has 6 branches, and 0 tags. The commit history shows 183 commits from 'c3530ed' 6 hours ago. The 'About' section notes 'No description, website, or topics provided.' The 'Readme' section links to 'README.md'. The 'Releases' section indicates 'No releases published' and 'Create a new release'. The 'Packages' section shows 'No packages published' and 'Publish your first package'. The 'Languages' section shows Jupyter Notebook (99.6%), HTML (0.2%), TeX (0.1%), Python (0.1%), JavaScript (0.0%), and CSS (0.0%). Below the repository details, there is a preview of the 'README.md' file content:

```
Predictive Maintenance - Projet Final Simplon

Using virtual data on machinery

Data Stream Simulation


```

6.3 CONCLUSION

Il a été démontré qu'il est a priori possible de construire une méthode efficace pour prédire les pannes 48 heures avant qu'elles ne se produisent. Pour cela il est primordial de travailler les données pour en extraire des features déterminantes. Parmi les features les importantes se trouvent le nombre de remplacements des composants et les calculs faits sur les données télémétriques.

Les algorithmes basés sur des arbres de décision sont les plus efficaces sur ce type de données. En réalité un simple arbre de décision donne déjà de très bon résultats.

Il est important d'avoir une bonne vision des problèmes de coûts réels engendrés par la maintenance et les pannes au sein d'une usine afin de savoir comment optimiser le modèle, c'est-à-dire en fonction du taux de faux positifs ou plutôt en fonction du taux de faux négatifs.

6.4 AMÉLIORATIONS

Pour aller plus loin, on pourrait utiliser des réseaux de neurones et éventuellement vérifier si la marge d'anticipation (les 48 heures) peut être allongée. Pouvoir tester le modèle sur des données réelles serait aussi intéressant. On pourrait aussi l'adapter pour des données non industrielles, pour les avions ou les voitures connectées par exemple.

Bibliographie

- [1] Microsoft Azure Advanced Scenario: General Predictive Maintenance
- [2] Implementing Bayesian Optimization On XGBoost: A Beginner's Guide
- [3] Bayesian Optimization For XGBoost
- [4] Asynchronous Tasks With Django and Celery
- [5] Django-celery-beat
- [6] kafka-python

Appendices

Annexe A

Fréquence des Interventions Sur Une Machine

FIGURE A.1 – Nombre d'interventions par mois pour chaque composant sur la machine ID=977

		comp1	comp2	comp3	comp4
year	month				
2014	June	1	1	1	1
	August	1	1	1	1
	September	1	1	1	1
	October	1	1	1	1
2015	January	1	1	1	1
	February	2	2	2	2
	March	2	2	2	2
	April	2	2	2	2
	May	2	2	2	2
	June	3	3	3	3
	July	1	1	1	1
	August	2	2	2	2
	September	2	2	2	2
	October	2	2	2	2
	November	2	2	2	2
	December	3	3	3	3

Annexe B

Distributions des Features

FIGURE B.1 – Distribution des Features de TELEMETRY

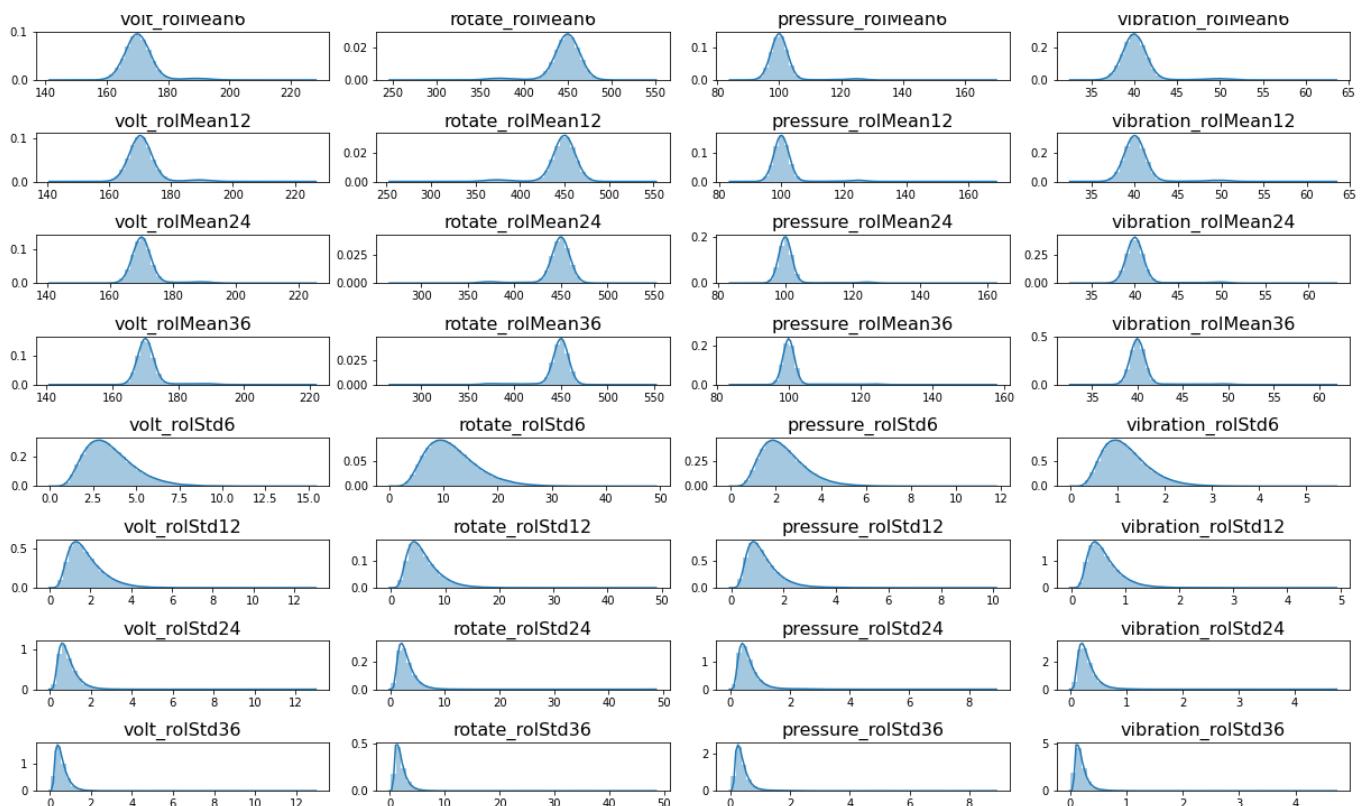


FIGURE B.2 – Distribution des Features de MAINTENANCE

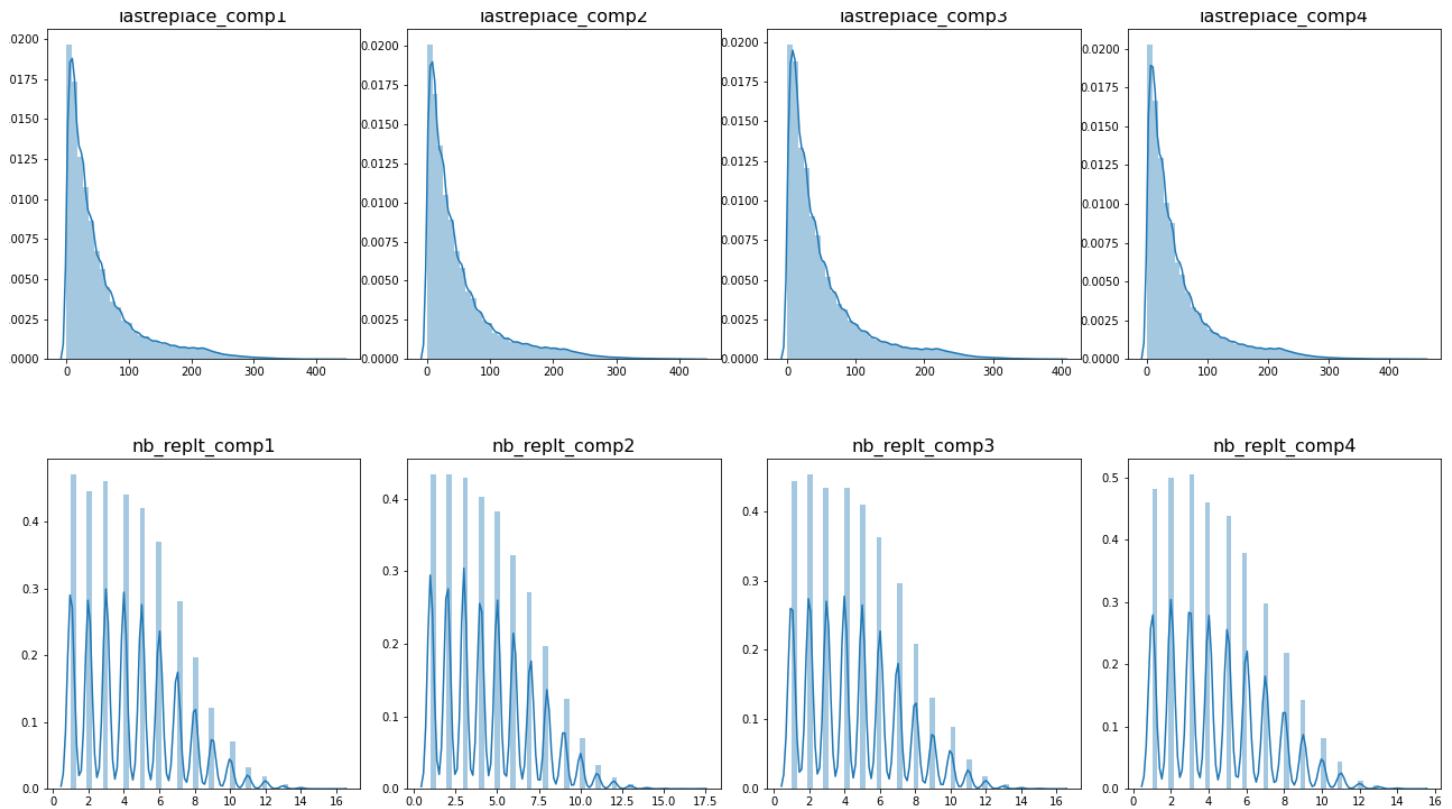
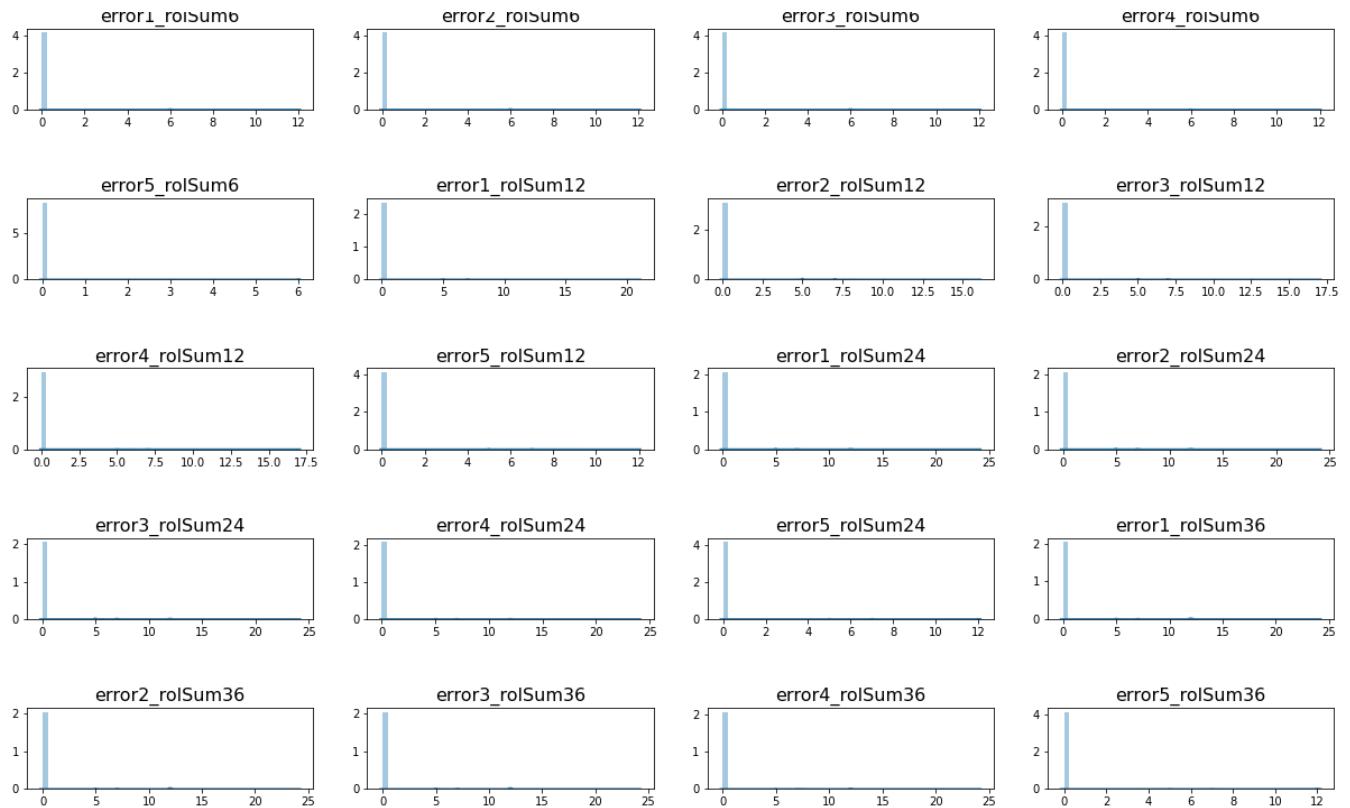


FIGURE B.3 – Distribution des Features de ERRORS



Annexe C

XGBoost - Bayesian Optimisation

La recherche a pris trois heures pour 18 itérations (faisant chacune un k-fold où $k = 3$).
Comme on peut le voir les meilleurs hyper-paramètres ont été trouvé dès la sixième itération.

FIGURE C.1 – Résultats de l'Optimisation Bayesienne pour XGBoost

iter	target	alpha	colsam...	gamma	learni...	max_depth	n_esti...	subsample
<hr/>								
n_jobs set to "5"								
1 0.9725	18.38	0.5643	8.586	0.9104	9.75	348.4	0.4494	
n_jobs set to "5"								
2 0.9907	12.69	0.3347	1.11	0.156	9.191	395.0	0.6995	
n_jobs set to "5"								
3 0.9882	16.43	0.9351	1.686	0.64	5.052	146.7	0.675	
n_jobs set to "5"								
4 0.9709	16.56	0.8475	0.7391	0.9915	3.445	146.7	0.6966	
n_jobs set to "5"								
5 0.9684	12.02	0.4158	9.537	0.9882	23.68	302.3	0.514	
n_jobs set to "5"								
6 0.9912	5.735	0.9789	2.483	0.1531	17.71	350.8	0.6785	
n_jobs set to "5"								
7 0.9902	10.01	0.6284	2.128	0.7271	17.01	447.0	0.5185	
n_jobs set to "5"								
8 0.9902	14.0	0.4037	1.245	0.126	9.543	277.6	0.769	
n_jobs set to "5"								
9 0.9848	9.329	0.3775	8.733	0.7541	24.94	207.6	0.8699	
n_jobs set to "5"								
10 0.9729	15.28	0.6789	5.226	0.8578	9.003	384.1	0.7687	
n_jobs set to "5"								
11 0.9905	15.54	0.3246	1.471	0.3574	21.64	421.7	0.5415	
n_jobs set to "5"								
12 0.9897	8.647	0.8623	9.938	0.5628	9.122	480.1	0.4115	
n_jobs set to "5"								
13 0.9899	9.42	0.8246	8.003	0.6364	23.71	135.5	0.366	
n_jobs set to "5"								
14 0.9899	11.32	0.4219	7.569	0.3426	21.34	267.6	0.6908	
n_jobs set to "5"								
15 0.9897	17.58	0.3241	4.305	0.1085	11.49	457.5	0.463	
n_jobs set to "5"								
16 0.9815	19.55	0.9405	8.764	0.01816	11.38	245.5	0.358	
n_jobs set to "5"								
17 0.9903	8.679	0.5309	3.054	0.4485	8.004	152.1	0.6475	
n_jobs set to "5"								
18 0.9895	17.52	0.8667	6.772	0.2107	9.291	302.4	0.5671	

Annexe D

Temps Computationnel de l'Apprentissage et des Prédictions

Classifier	Fit Time	Predict Time
Decision Tree	59 sec	0.6 sec
Random Forest	13 min 15	11 sec
XGBoost	31 min 55	32 sec
SVM	13 sec	0.7 sec
KNN	19 min	INFINITY
Logistic Regression	1 min 44	0.7 sec
Naive Bayes	2 sec	5 sec

Annexe E

Matrices de Confusion

FIGURE E.1 – Decision Tree - Test

CART: Confusion Matrix - Test					
COMP1		COMP3		NO FAILURE	
Actual	False	145,028 (TN)	198 (FP)	False	5,610 (TN)
	True	TNR: 99.86%	FPR: 0.14%	TNR: 99.95%	FPR: 0.05%
True	False	237 (FN)	1,537 (TP)	False	96 (FN)
	True	FNR: 13.36%	TPR: 86.64%	True	884 (TP)
Prediction		Prediction		Prediction	
COMP2		COMP4		COMP5	
Actual	False	144,354 (TN)	150 (FP)	False	145,772 (TN)
	True	TNR: 99.90%	FPR: 0.10%	TNR: 99.96%	FPR: 0.04%
True	False	240 (FN)	2,256 (TP)	False	62 (FP)
	True	FNR: 9.62%	TPR: 90.38%	True	1,062 (TP)
Prediction		Prediction		Prediction	

FIGURE E.2 – Decision Tree - Train

CART: Confusion Matrix - Train					
COMP1		COMP3		NO FAILURE	
Actual	False	576,107 (TN)	341 (FP)	23,706 (TN)	1,640 (FP)
	True	TNR: 99.94%	FPR: 0.06%	TNR: 93.53%	FPR: 6.47%
True	False	891 (FN)	6,661 (TP)	1,233 (FN)	557,421 (TP)
	True	FNR: 11.80%	TPR: 88.20%	FNR: 0.22%	TPR: 99.78%
Prediction		Prediction		Prediction	
COMP2					
Actual	False	573,356 (TN)	371 (FP)	578,838 (TN)	207 (FP)
	True	TNR: 99.94%	FPR: 0.06%	TNR: 99.96%	FPR: 0.04%
True	False	823 (FN)	9,450 (TP)	439 (FN)	4,516 (TP)
	True	FNR: 8.01%	TPR: 91.99%	FNR: 8.86%	TPR: 91.14%
Prediction		Prediction		Prediction	
COMP4					

FIGURE E.3 – Random Forest - Test

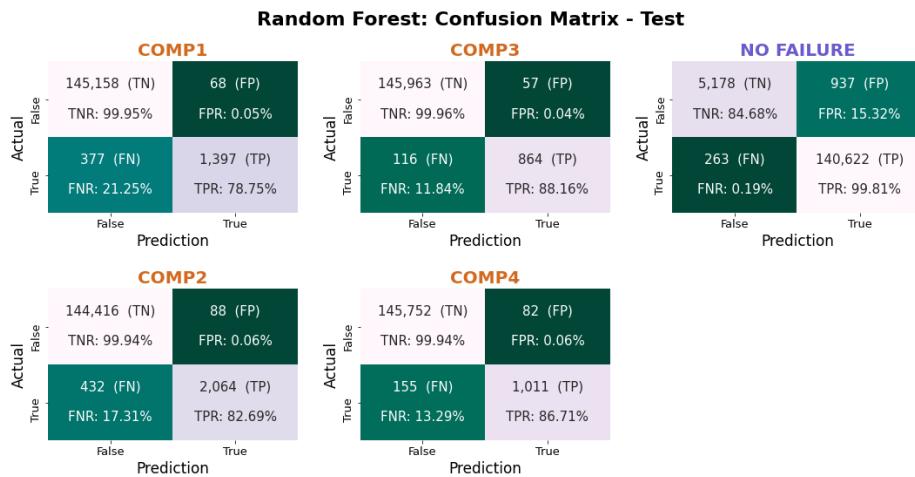


FIGURE E.4 – Random Forest - Train

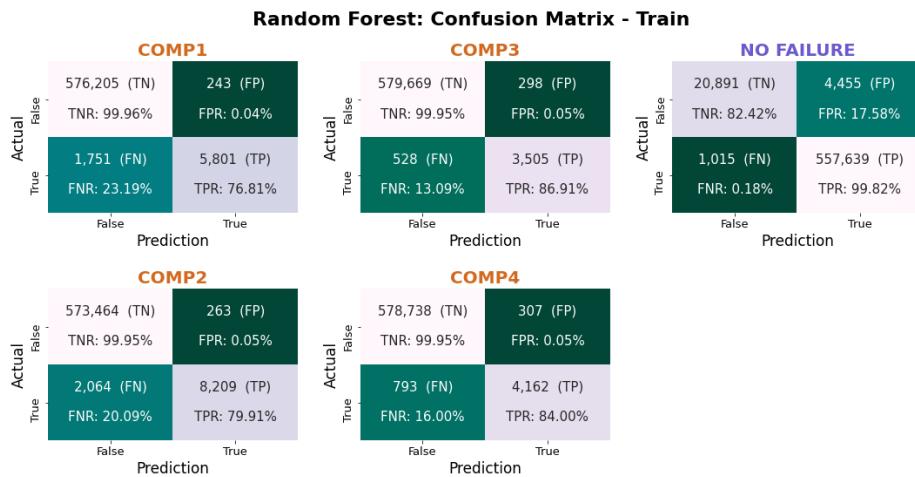


FIGURE E.5 – XGBoost - Test

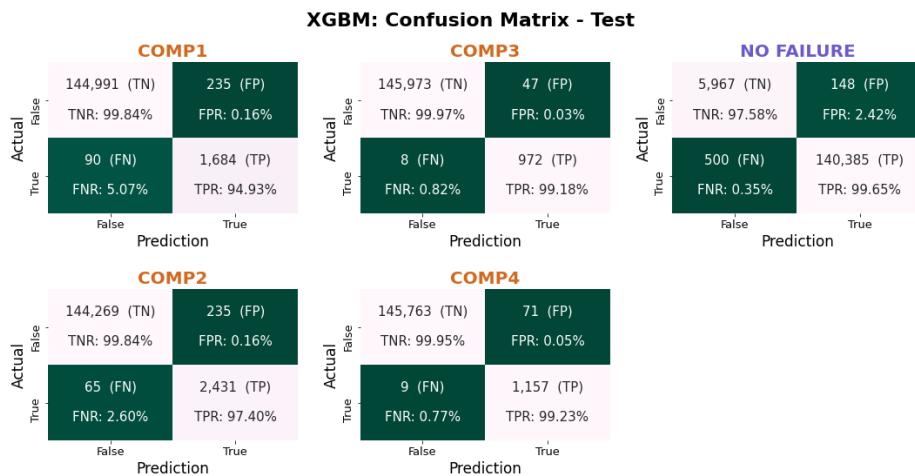


FIGURE E.6 – XGBoost - Train

XGBM: Confusion Matrix - Train							
		COMP1		COMP3		NO FAILURE	
Actual	False	COMP1		COMP3		NO FAILURE	
		576,066 (TN)	382 (FP)	579,838 (TN)	129 (FP)	25,346 (TN)	0 (FP)
	True	TNR: 99.93%	FPR: 0.07%	TNR: 99.98%	FPR: 0.02%	TNR: 100.00%	FPR: 0.00%
	False	0 (FN)	7,552 (TP)	0 (FN)	4,033 (TP)	949 (FN)	557,705 (TP)
	True	FNR: 0.00%	TPR: 100.00%	FNR: 0.00%	TPR: 100.00%	FNR: 0.17%	TPR: 99.83%
	False		Prediction		Prediction		Prediction

XGBM v2: Confusion Matrix - Train							
		COMP1		COMP3		NO FAILURE	
Actual	False	COMP1		COMP3		NO FAILURE	
		144,994 (TN)	232 (FP)	145,975 (TN)	45 (FP)	5,953 (TN)	162 (FP)
	True	TNR: 99.84%	FPR: 0.16%	TNR: 99.97%	FPR: 0.03%	TNR: 97.35%	FPR: 2.65%
	False	96 (FN)	1,678 (TP)	10 (FN)	970 (TP)	483 (FN)	140,402 (TP)
	True	FNR: 5.41%	TPR: 94.59%	FNR: 1.02%	TPR: 98.98%	FNR: 0.34%	TPR: 99.66%
	False		Prediction		Prediction		Prediction

XGBM v2: Confusion Matrix - Train							
		COMP1		COMP3		NO FAILURE	
Actual	False	COMP1		COMP3		NO FAILURE	
		144,273 (TN)	231 (FP)	145,774 (TN)	60 (FP)	5,953 (TN)	162 (FP)
	True	TNR: 99.84%	FPR: 0.16%	TNR: 99.96%	FPR: 0.04%	TNR: 97.35%	FPR: 2.65%
	False	71 (FN)	2,425 (TP)	13 (FN)	1,153 (TP)	483 (FN)	140,402 (TP)
	True	FNR: 2.84%	TPR: 97.16%	FNR: 1.11%	TPR: 98.89%	FNR: 0.34%	TPR: 99.66%
	False		Prediction		Prediction		Prediction

FIGURE E.7 – XGBoost *Extended Windows* - Test

XGBM v2: Confusion Matrix - Test							
		COMP1		COMP3		NO FAILURE	
Actual	False	COMP1		COMP3		NO FAILURE	
		144,994 (TN)	232 (FP)	145,975 (TN)	45 (FP)	5,953 (TN)	162 (FP)
	True	TNR: 99.84%	FPR: 0.16%	TNR: 99.97%	FPR: 0.03%	TNR: 97.35%	FPR: 2.65%
	False	96 (FN)	1,678 (TP)	10 (FN)	970 (TP)	483 (FN)	140,402 (TP)
	True	FNR: 5.41%	TPR: 94.59%	FNR: 1.02%	TPR: 98.98%	FNR: 0.34%	TPR: 99.66%
	False		Prediction		Prediction		Prediction

XGBM v2: Confusion Matrix - Train							
		COMP1		COMP3		NO FAILURE	
Actual	False	COMP1		COMP3		NO FAILURE	
		144,273 (TN)	231 (FP)	145,774 (TN)	60 (FP)	5,953 (TN)	162 (FP)
	True	TNR: 99.84%	FPR: 0.16%	TNR: 99.96%	FPR: 0.04%	TNR: 97.35%	FPR: 2.65%
	False	71 (FN)	2,425 (TP)	13 (FN)	1,153 (TP)	483 (FN)	140,402 (TP)
	True	FNR: 2.84%	TPR: 97.16%	FNR: 1.11%	TPR: 98.89%	FNR: 0.34%	TPR: 99.66%
	False		Prediction		Prediction		Prediction

FIGURE E.8 – XGBoost *Extended Windows* - Train

XGBM v2: Confusion Matrix - Train							
		COMP1		COMP3		NO FAILURE	
Actual	False	COMP1		COMP3		NO FAILURE	
		576,217 (TN)	231 (FP)	579,878 (TN)	89 (FP)	25,346 (TN)	0 (FP)
	True	TNR: 99.96%	FPR: 0.04%	TNR: 99.98%	FPR: 0.02%	TNR: 100.00%	FPR: 0.00%
	False	0 (FN)	7,552 (TP)	0 (FN)	4,033 (TP)	610 (FN)	558,044 (TP)
	True	FNR: 0.00%	TPR: 100.00%	FNR: 0.00%	TPR: 100.00%	FNR: 0.11%	TPR: 99.89%
	False		Prediction		Prediction		Prediction

XGBM v2: Confusion Matrix - Train							
		COMP1		COMP3		NO FAILURE	
Actual	False	COMP1		COMP3		NO FAILURE	
		573,457 (TN)	270 (FP)	578,916 (TN)	129 (FP)	25,346 (TN)	0 (FP)
	True	TNR: 99.95%	FPR: 0.05%	TNR: 99.98%	FPR: 0.02%	TNR: 100.00%	FPR: 0.00%
	False	0 (FN)	10,273 (TP)	0 (FN)	4,955 (TP)	610 (FN)	558,044 (TP)
	True	FNR: 0.00%	TPR: 100.00%	FNR: 0.00%	TPR: 100.00%	FNR: 0.11%	TPR: 99.89%
	False		Prediction		Prediction		Prediction

FIGURE E.9 – Logistic Regression - Test

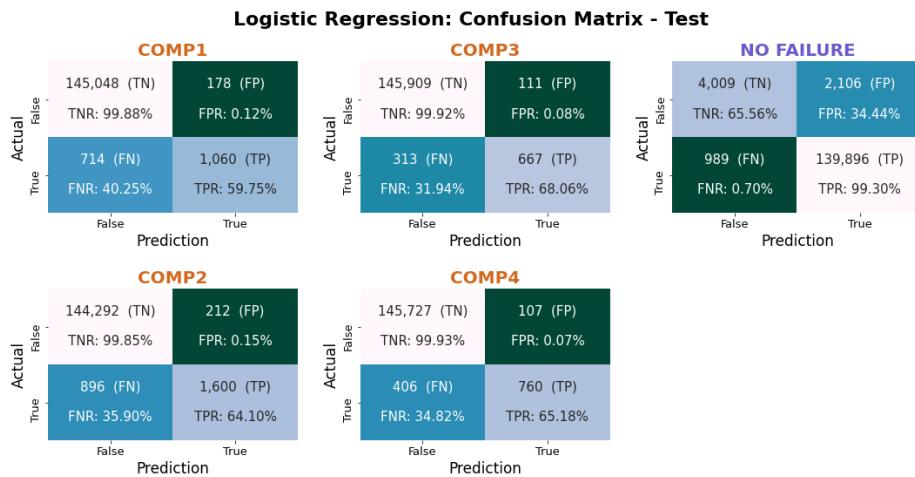


FIGURE E.10 – Logistic Regression - Train

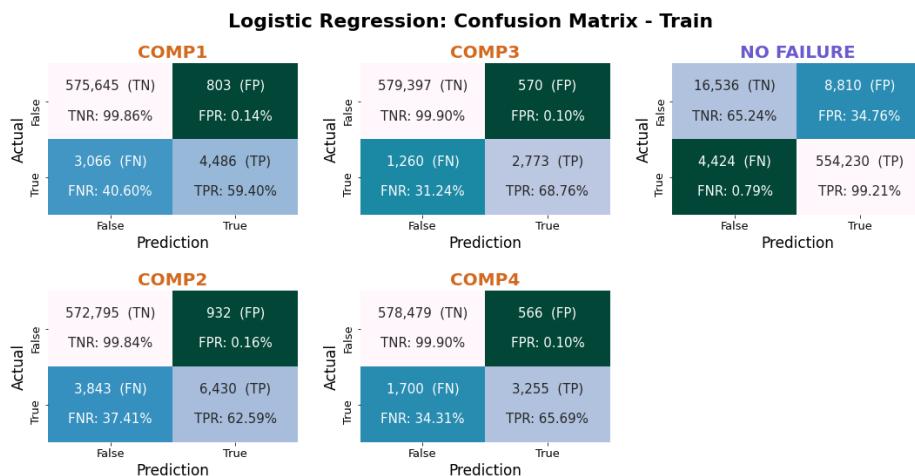


FIGURE E.11 – SVM - Test

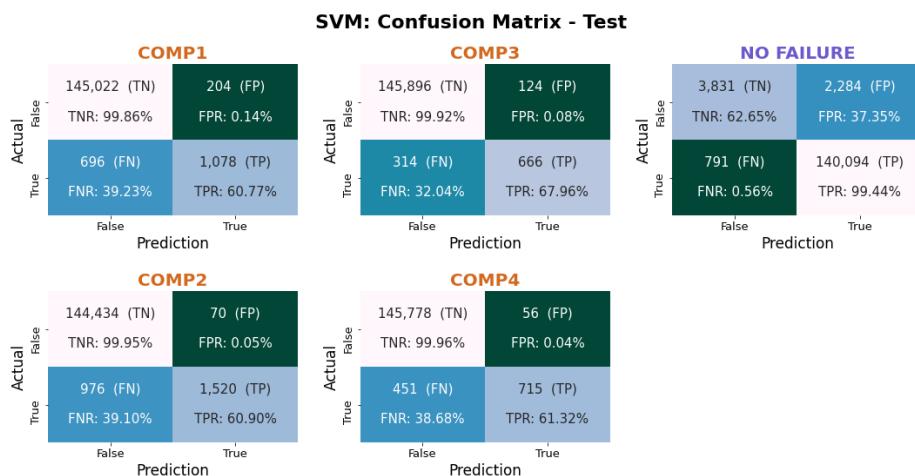


FIGURE E.12 – SVM - Train

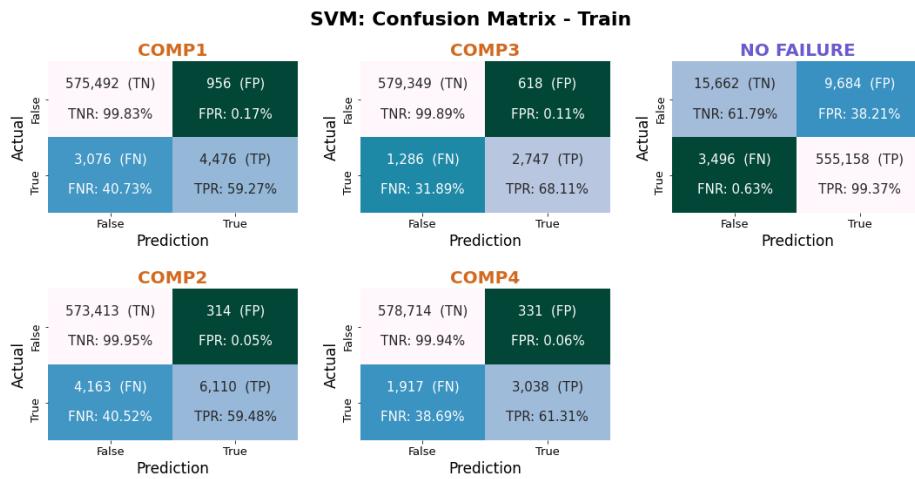


FIGURE E.13 – Naive Bayes - Test

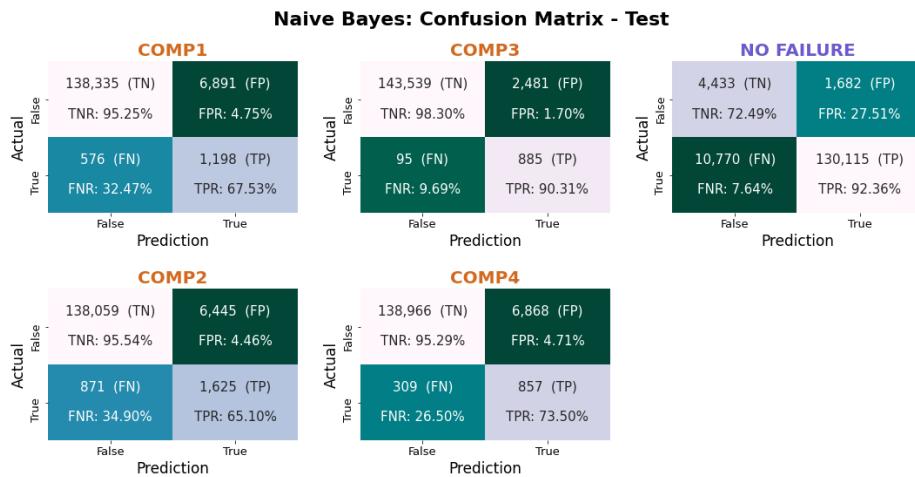
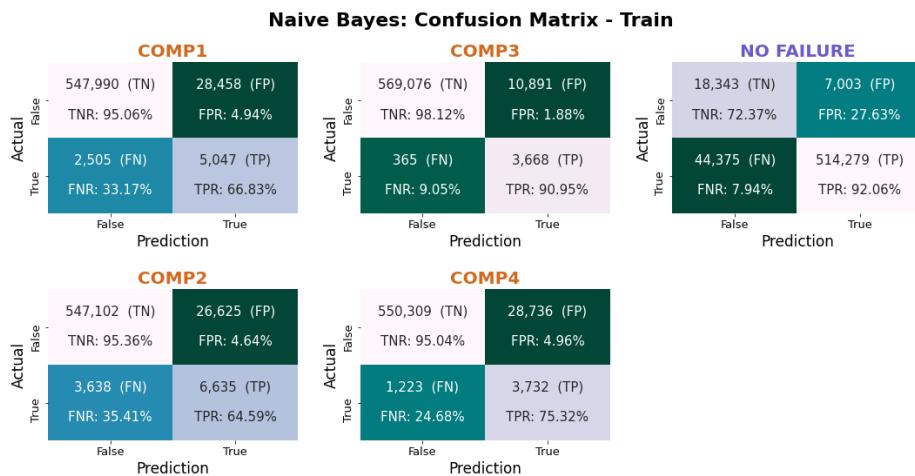


FIGURE E.14 – Naive Bayes - Train



Annexe F

Scores

FIGURE F.1 – Scores - Première Partie

NO FAILURE		accuracy	auc	f1-score	precision	recall	model		NO FAILURE		accuracy	auc	f1-score	precision	recall
Decision Tree	99.293	97.772	99.631	99.641	99.621		Decision Tree	99.508	99.959	99.743	99.707	99.779			
Logistic Regression	97.895	95.593	98.906	98.517	99.298		Logistic Regression	97.734	95.184	98.82	98.435	99.208			
Naive Bayes	91.529	84.247	95.434	98.724	92.355		Naive Bayes	91.202	83.903	95.242	98.657	92.057			
Random Forest	99.184	94.61	99.575	99.338	99.813		Random Forest	99.063	94.488	99.512	99.207	99.818			
SVM	97.908	95.534	98.914	98.396	99.439		SVM	97.743	95.142	98.827	98.286	99.374			
XGBoost	99.559	99.942	99.77	99.895	99.645		XGBoost	99.838	99.998	99.915	100	99.83			
MACRO AVG		accuracy	auc	f1-score	precision	recall	model		MACRO AVG		accuracy	auc	f1-score	precision	recall
Decision Tree		99.171	97.408	92.67	93.79	91.586	Decision Tree	99.359	99.977	94.445	96.427	92.571			
Logistic Regression		97.797	98.428	78.844	89.166	71.279	Logistic Regression	97.619	98.224	78.197	87.746	71.13			
Naive Bayes		88.874	93.936	42.098	34.211	77.759	Naive Bayes	88.467	93.766	41.943	34.071	77.948			
Random Forest		99.107	95.318	91.011	95.383	87.225	Random Forest	98.986	95.175	90.044	95.475	85.489			
SVM		97.867	98.379	78.589	91.024	70.077	SVM	97.696	98.182	77.643	89.522	69.509			
XGBoost		99.495	99.973	95.813	93.688	98.076	XGBoost	99.813	100	98.429	96.952	99.966			

(a) TEST

(b) TRAIN

FIGURE F.2 – Scores - Deuxième Partie

(a) TEST

COMP1	accuracy	auc	f1-score	precision	recall
model					
Decision Tree	99.704	96.933	87.603	88.588	86.64
Logistic Regression	99.393	98.165	70.385	85.622	59.752
Naive Bayes	94.92	94.119	24.293	14.81	67.531
Random Forest	99.697	93.164	86.261	95.358	78.749
SVM	99.388	98.059	70.55	84.087	60.767
XGBoost	99.779	99.961	91.2	87.754	94.927

(c) TRAIN

COMP1	accuracy	auc	f1-score	precision	recall
model					
Decision Tree	99.789	99.969	91.535	95.13	88.202
Logistic Regression	99.338	97.957	69.87	84.818	59.401
Naive Bayes	94.698	93.936	24.585	15.063	66.83
Random Forest	99.659	93.094	85.334	95.979	76.814
SVM	99.31	97.905	68.946	82.401	59.269
XGBoost	99.935	100	97.533	95.185	100

COMP2	accuracy	auc	f1-score	precision	recall
model					
Decision Tree	99.735	97.403	92.044	93.766	90.385
Logistic Regression	99.246	98.806	74.28	88.3	64.103
Naive Bayes	95.023	95.371	30.759	20.136	65.104
Random Forest	99.646	94.277	88.812	95.911	82.692
SVM	99.288	98.729	74.4	95.597	60.897
XGBoost	99.796	99.972	94.188	91.185	97.396

COMP2	accuracy	auc	f1-score	precision	recall
model					
Decision Tree	99.796	99.981	94.058	96.222	91.989
Logistic Regression	99.182	98.538	72.923	87.34	62.591
Naive Bayes	94.818	95.285	30.483	19.949	64.587
Random Forest	99.602	94.26	87.586	96.896	79.908
SVM	99.233	98.448	73.187	95.112	59.476
XGBoost	99.929	100	98.034	96.144	100

COMP3	accuracy	auc	f1-score	precision	recall
model					
Decision Tree	99.886	97.225	91.322	92.469	90.204
Logistic Regression	99.712	99.888	75.882	85.733	68.061
Naive Bayes	98.248	99.056	40.727	26.292	90.306
Random Forest	99.882	97.973	90.9	93.811	88.163
SVM	99.702	99.901	75.254	84.304	67.959
XGBoost	99.963	99.996	97.249	95.388	99.184

COMP3	accuracy	auc	f1-score	precision	recall
model					
Decision Tree	99.913	99.991	93.564	95.459	91.743
Logistic Regression	99.687	99.849	75.19	82.949	68.758
Naive Bayes	98.073	98.967	39.458	25.194	90.95
Random Forest	99.859	97.756	89.459	92.164	86.908
SVM	99.674	99.852	74.263	81.634	68.113
XGBoost	99.978	100	98.426	96.901	100

COMP4	accuracy	auc	f1-score	precision	recall
model					
Decision Tree	99.887	97.705	92.751	94.484	91.081
Logistic Regression	99.651	99.689	74.766	87.659	65.18
Naive Bayes	95.118	96.888	19.278	11.094	73.499
Random Forest	99.839	96.566	89.509	92.498	86.707
SVM	99.655	99.674	73.826	92.737	61.321
XGBoost	99.946	99.994	96.658	94.218	99.228

COMP4	accuracy	auc	f1-score	precision	recall
model					
Decision Tree	99.889	99.988	93.325	95.617	91.14
Logistic Regression	99.612	99.593	74.18	85.187	65.691
Naive Bayes	94.87	96.74	19.945	11.494	75.318
Random Forest	99.812	96.275	88.328	93.13	83.996
SVM	99.615	99.563	72.994	90.175	61.312
XGBoost	99.97	100	98.236	96.532	100

(b) TEST

(d) TRAIN

	comp1		comp2		comp3		comp4		no failure		macro avg	
	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train
precision	87.853	97.032	91.303	97.439	95.567	97.841	95.054	97.463	99.885	100	93.932	97.955
recall	94.589	100	97.155	100	98.98	100	98.885	100	99.657	99.891	97.853	99.978
f1-score	91.097	98.494	94.138	98.703	97.243	98.909	96.931	98.715	99.771	99.945	95.836	98.953
auc	99.96	100	99.971	100	99.99	100	99.996	100	99.941	100	99.972	100
accuracy	99.777	99.96	99.795	99.954	99.963	99.985	99.95	99.978	99.561	99.896	99.494	99.877

FIGURE F.3 – Scores pour XGBoost avec Fenêtres Étendues

Annexe G

Feature Importance

G.1 PAR CATÉGORIE DE FEATURES

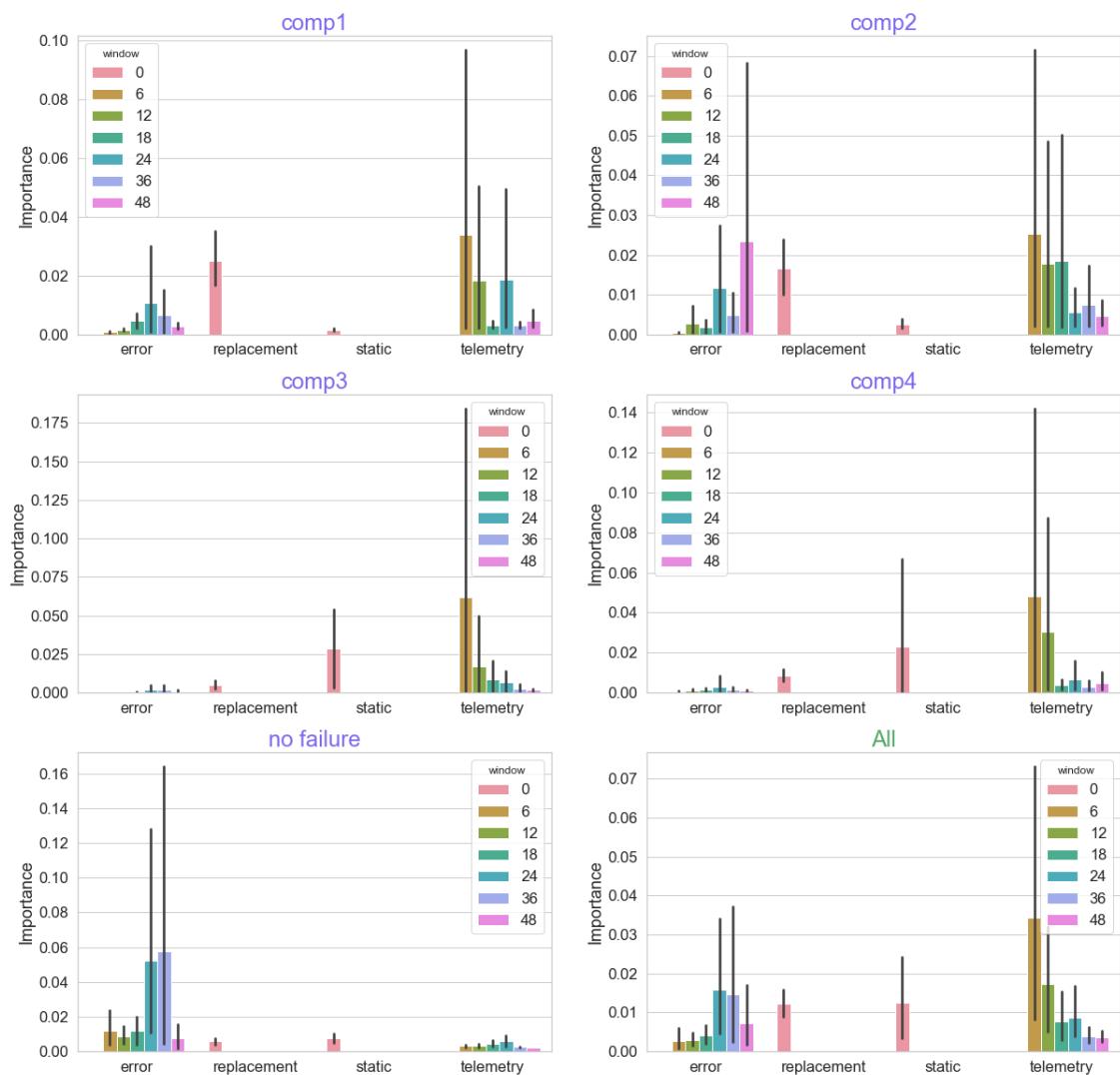


FIGURE G.1 – Feature Importance par Type et par Fenêtre

G.2 PAR SOUS-CATÉGORIE DE FEATURES

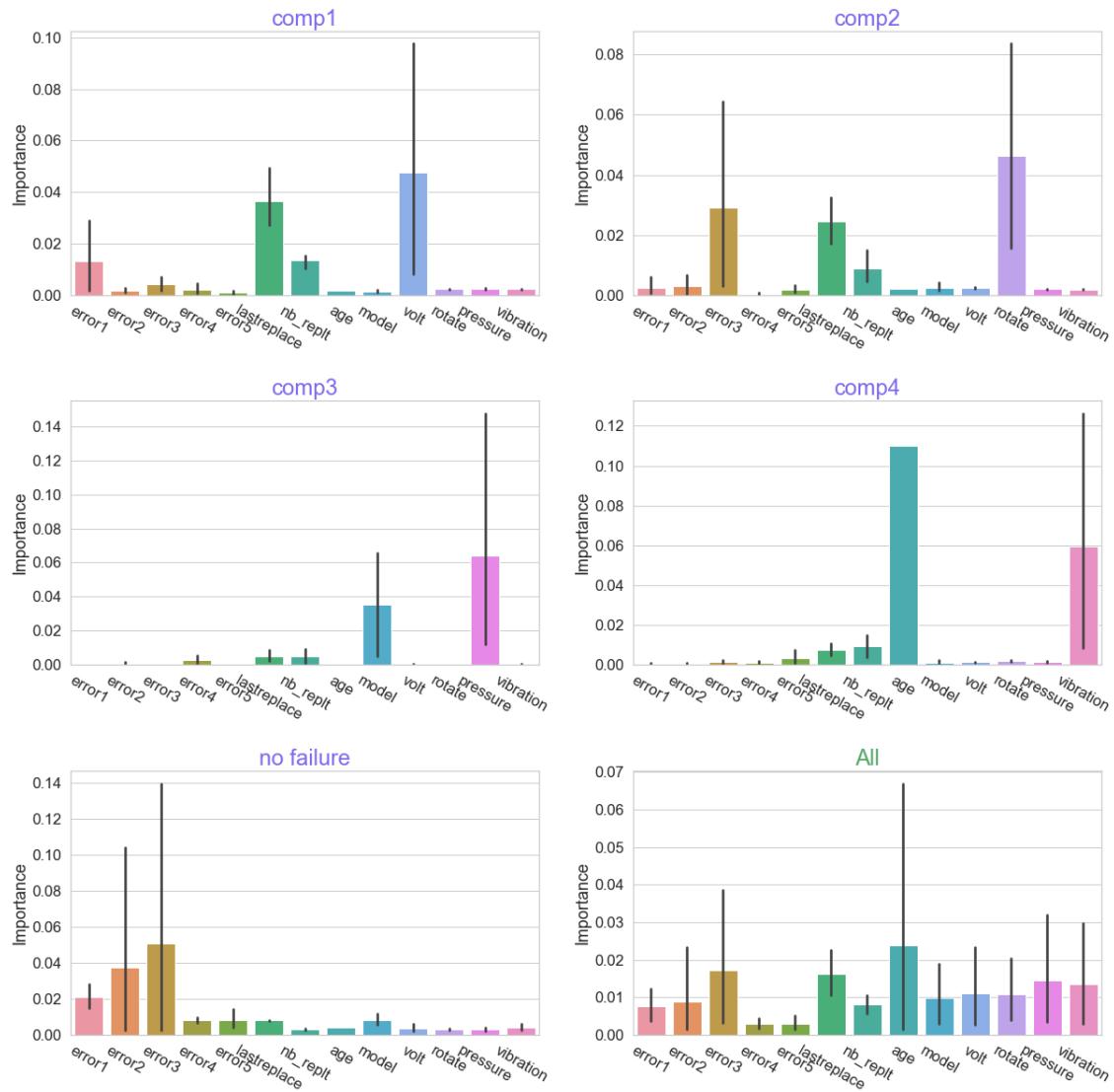


FIGURE G.2 – Feature Importance par Sous-Catégorie

Annexe H

Schéma Relationnel de la Base de Données

FIGURE H.1 – BDD Schéma Relationnel - Première Partie

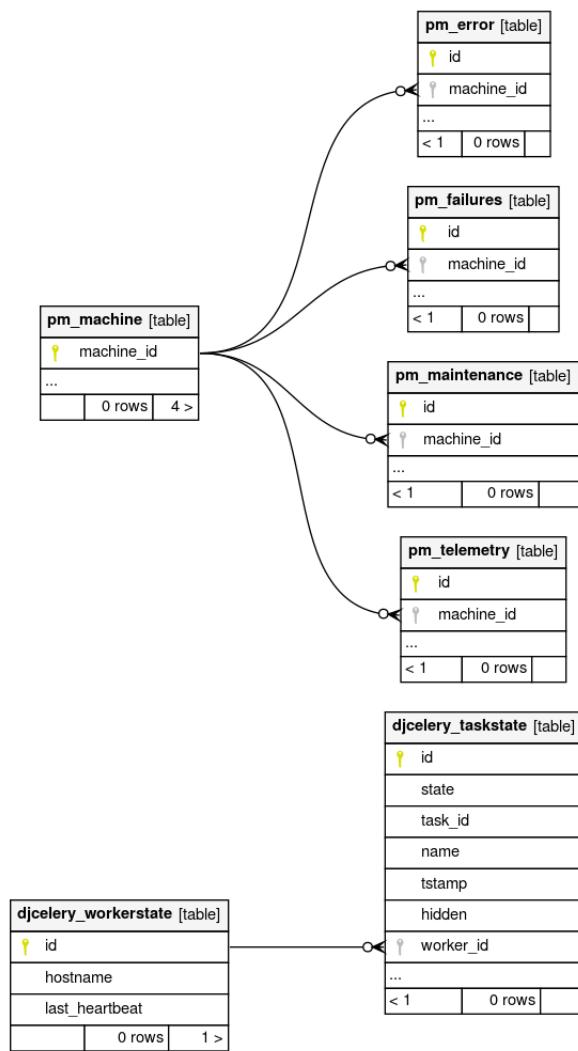
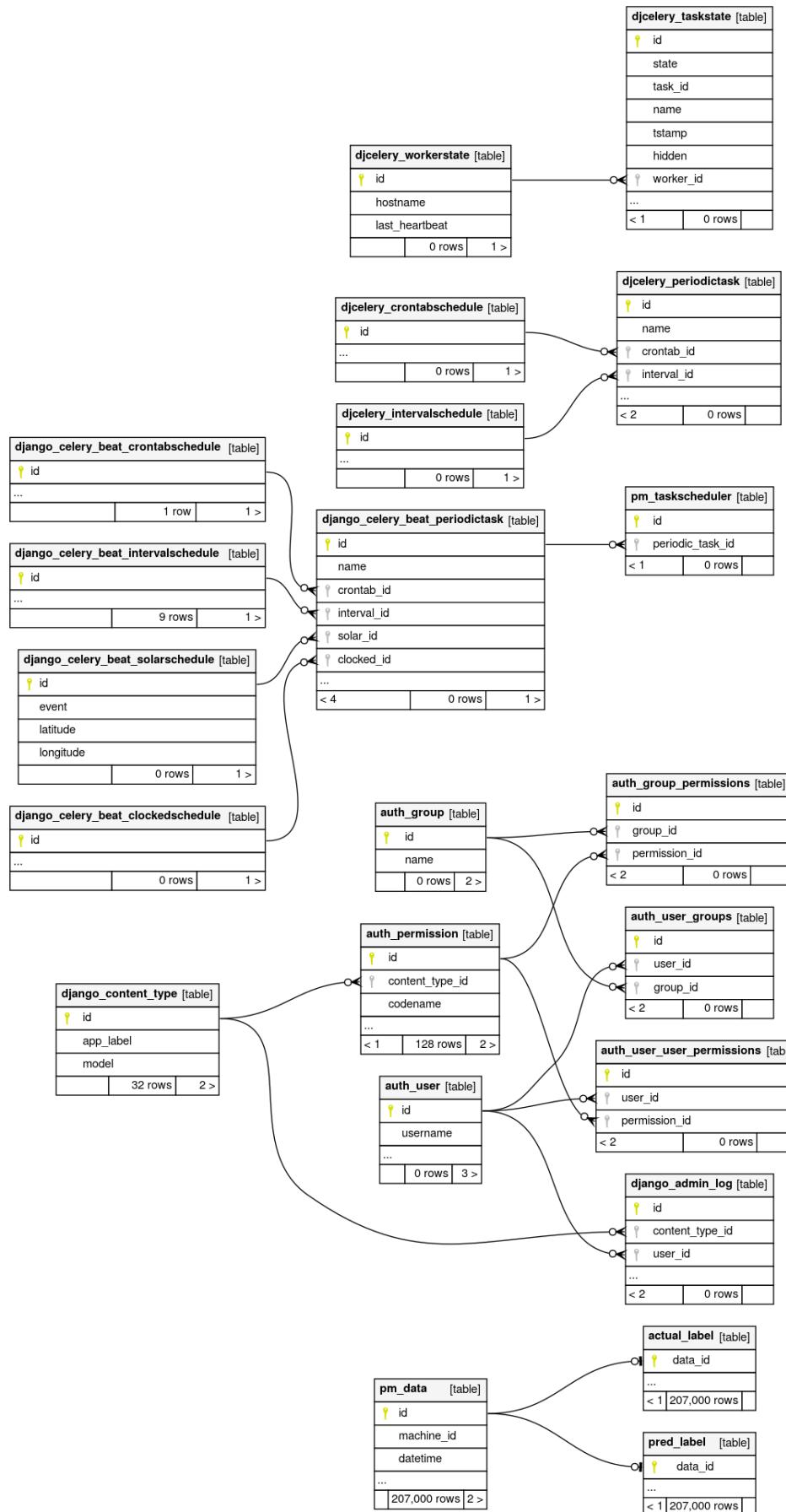


FIGURE H.2 – BDD Schéma Relationnel - Seconde Partie



Annexe I

Structure de l'Application Web

```
docker-compose-django/
├── log/
│   ├── access.log
│   ├── error.log
│   └── uwsgi.log
├── nginx/
│   ├── Dockerfile
│   ├── nginx-app.conf
│   └── nginx.conf
└── web/
    ├── api/
    │   ├── __init__.py
    │   ├── asgi.py
    │   ├── celery.py
    │   ├── keyconfig.py
    │   ├── settings.py
    │   ├── urls.py
    │   └── wsgi.py
    ├── data/
    │   ├── initial_model/
    │   ├── stream_samples/
    │   │   ├── stream_data_labels.csv
    │   │   ├── stream_errors.csv
    │   │   ├── stream_maint.csv
    │   │   └── stream_telemetry.csv
    │   ├── __init__.py
    │   ├── binarizer
    │   ├── current_pipeline.z
    │   ├── initial_pipeline.z
    │   └── std_scaler.p
    └── logs/
        ├── celery.log
        └── celery_beat.log
```

```
└── modules/
    ├── __init__.py
    ├── create_pipeline.py
    ├── kafka_consumer.py
    ├── kafka_producer.py
    ├── prediction.py
    ├── preprocess.py
    ├── preprocessing_pipeline.py
    └── utils.py
└── pm/
    ├── migrations/
        ├── 0001_initial.py
        ├── 0002_auto_20200805_1213.py
        ├── 0003_actual_data_partition_prediction.py
        ├── 0004_partition_topic.py
        ├── 0005_error_failures_machine_maintenance_telemetry.py
        └── __init__.py
    ├── __init__.py
    ├── admin.py
    ├── apps.py
    ├── models.py
    ├── signals.py
    ├── tasks.py
    ├── tests.py
    ├── urls.py
    └── views.py
└── settings/
    ├── __init__.py
    ├── base.py
    ├── dev.py
    └── prod.py
└── static/
    ├── celery_progress/
        ├── celery_progress.js
        ├── celery_progress_websockets.js
        └── websockets.js
    ├── css/
        ├── bootstrap.min.css
        ├── now-ui-dashboard.css
        └── style.css
    └── fonts/
        ├── nucleo-license.md
        ├── nucleo-outline.eot
        ├── nucleo-outline.ttf
        ├── nucleo-outline.woff
        └── nucleo-outline.woff2
```

```
|- img/
|   |__ apple-icon.png
|   |__ bg5.jpg
|   |__ default-avatar.png
|   |__ favicon.png
|   |__ header.jpg
|   |__ mechanics1.jpg
|   |__ mike.jpg
|   |__ now-logo.png
|   |__ now-ui-dashboard.gif
|- js/
|   |- core/
|   |   |__ bootstrap.min.js
|   |   |__ jquery.min.js
|   |   |__ popper.min.js
|   |- plugins/
|   |   |__ bootstrap-notify.js
|   |   |__ chartjs.min.js
|   |   |__ perfect-scrollbar.jquery.min.js
|   |   |__ now-ui-dashboard.js
|- templates/
|   |__ alerts.html
|   |__ base.html
|   |__ dashboard.html
|   |__ icons.html
|   |__ main_chart.html
|   |__ map.html
|   |__ notifications.html
|   |__ progressbar.html
|   |__ tables.html
|   |__ typography.html
|   |__ upgrade.html
|   |__ user.html
|- Dockerfile
|- entrypoint.sh
|- manage.py
|- requirements.txt
|- uwsgi.ini
|- uwsgi_params
|- app.env
|- docker-compose.yml
```