

Jedha

CERTIFICATION BLOC 6 - CDSD

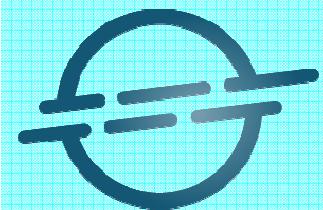
# PRÉDIRE LE DÉPART DES SALARIÉS PAR MACHINE LEARNING

28 AVRIL 2025

HEINRY ELODIE

DATA ESSENTIALS





## I. INTRODUCTION

- II. PROBLÉMATIQUE
- III. ANALYSE EXPLORATOIRE
  - DES DONNÉES
- IV. MACHINE LEARNING
- V. RECOMMANDATIONS
- VI. CONCLUSION
- VII. PERSPECTIVES

- Attrition entreprise

=> **Tous les départs d'une volontaires ou non**

- Taux d'attrition

=> « normal » ≈ 10 %

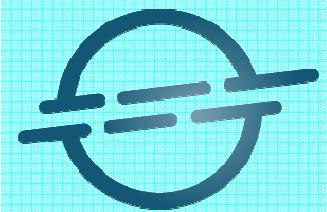
- Enjeu majeur

=> **Fidélisation des talents**

- Attrition élevée

=> ↓ productivité  
=> ↓ stabilité des équipes  
=> ↑ coûts (recrutement et formation)

- Analyse de l'attrition volontaire chez IBM



I. INTRODUCTION

## II. PROBLÉMATIQUE

III. ANALYSE EXPLORATOIRE

DES DONNÉES

IV. MACHINE LEARNING

V. RECOMMANDATIONS

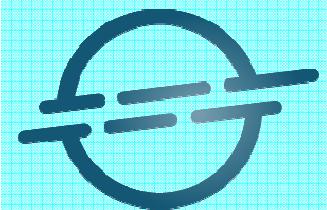
VI. CONCLUSION

VII. PERSPECTIVES

**Déterminer les facteurs qui jouent un rôle sur la décision de départ des salariés**

**Proposer des solutions pour diminuer l'attrition :**

- Recommandations pour limiter l'attrition volontaire
- Construction d'un modèle de Machine Learning pour prédire le départ d'un salarié



I. INTRODUCTION

II. PROBLÉMATIQUE

### III. ANALYSE EXPLORATOIRE DES DONNÉES

IV. MACHINE LEARNING

V. RECOMMANDATIONS

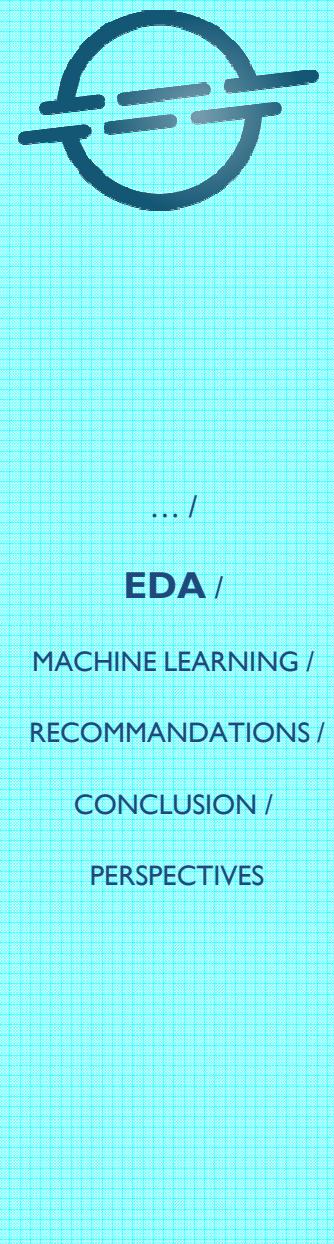
VI. CONCLUSION

VII. PERSPECTIVES

- Téléchargement sur [kaggle](#)
- Description annuelle des employés travaillant chez IBM
- **1 470 employés** (=lignes)
- **35 variables** (=colonnes) :
  - 26 numériques
  - 9 catégorielles
- Variable cible (catégorielle) = **Attrition**
- Légende

#### Attrition

- 0 - Restés
- 1 - Partis



## Préparation du jeu de données

### ➤ NETTOYAGE

- Élimination des doublons
- Suppression de variables (inutiles ou redondantes)
- Traitement des valeurs manquantes
- Traitement des valeurs atypiques ou aberrantes

### ➤ OBSERVATIONS AVEC TABLEAU

### ➤ ANALYSES STATISTIQUES

- Corrélogramme
- Variation Inflation Factor
- Corrélation du point bisérial
- Test du Chi<sup>2</sup>



TABLEAU

## Exploratory Data Analysis

PYTHON

TESTS STATISTIQUES



Employee Count  
Marital Status => RGPD

TABLEAU

## Exploratory Data Analysis

PYTHON

TESTS STATISTIQUES

Over18  
Standard Hours  
Employee Number  
Performance Rating

Monthly Rate  
Hourly Rate  
Daily Rate  
Education  
Gender

Num Companies Worked  
Percent Salary Hike  
Years Since Last Promotion  
Relationship Satisfaction



Employee Count  
Marital Status =>RGPD

TABLEAU

## Exploratory Data Analysis

PYTHON

TESTS STATISTIQUES

Over18  
Standard Hours  
Employee Number  
Performance Rating

Monthly Rate  
Hourly Rate  
Daily Rate  
Education  
Num Companies Worked  
Percent Salary Hike  
Years Since Last Promotion  
Relationship Satisfaction  
Gender

## VARIABLES SÉLECTIONNÉES

### Variables numériques (9)

Age  
Distance From Home  
Monthly Income  
Stock Option Level  
Total Working Years  
Training Time Last Year  
Years At Company  
Years In Current Role  
Years With Current Manager

### Variables catégorielles (10)

Business Travel  
Department  
Education Field  
Environment Satisfaction  
Job Involvement  
Job Level  
Job Role  
Job Satisfaction  
Over Time  
Work Life Balance



Employee Count  
Marital Status => RGPD

TABLEAU

## Exploratory Data Analysis

PYTHON

TESTS STATISTIQUES

Over18  
Standard Hours  
Employee Number  
Performance Rating

Monthly Rate  
Hourly Rate  
Daily Rate  
Education  
Num Companies Worked  
Percent Salary Hike  
Years Since Last Promotion  
Relationship Satisfaction  
Gender

## VARIABLES SÉLECTIONNÉES

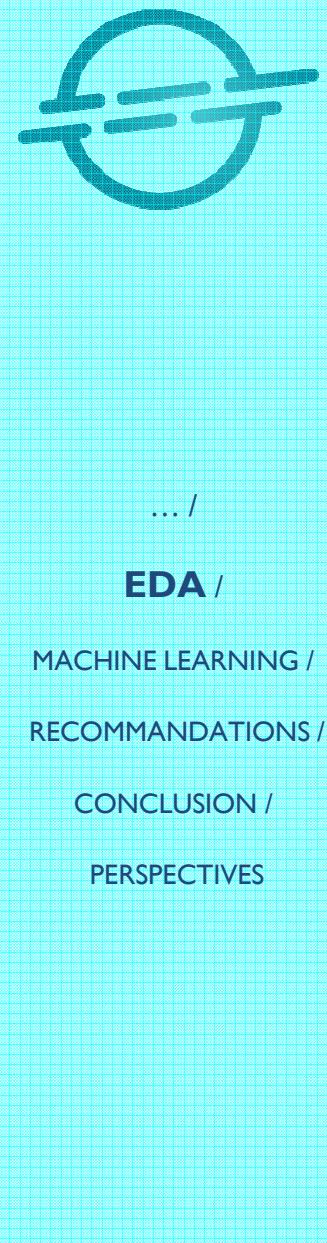
### Variables numériques (9)

Age  
Distance From Home  
Monthly Income  
Stock Option Level  
Total Working Years  
Training Time Last Year  
Years At Company  
Years In Current Role  
Years With Current Manager

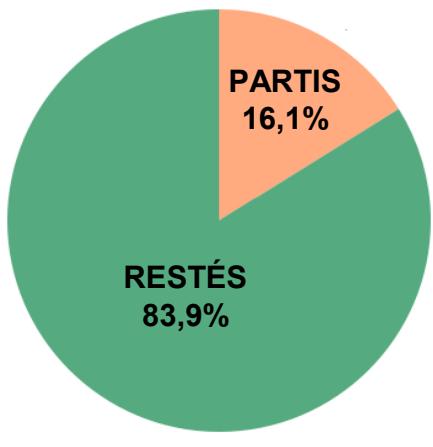
ML

### Variables catégorielles (10)

Business Travel  
Department  
Education Field  
Environment Satisfaction  
Job Involvement  
Job Level  
Job Role  
Job Satisfaction  
Over Time  
Work Life Balance



## Observations générales

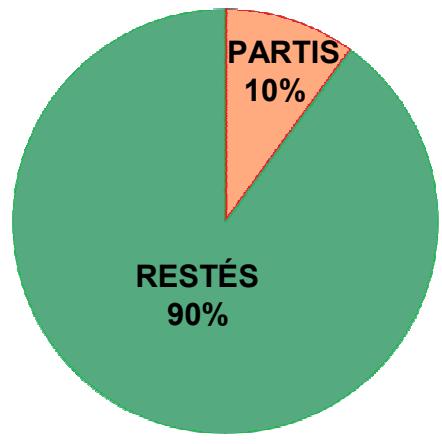


Attrition en pourcentage

237 départs sur l'année

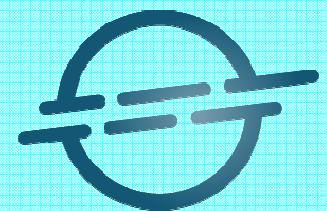
=> 16,1 % des employés

**+ 6,1 points**  
qu'une attrition « normale »



Attrition NORMALE

Nécessité :  
Mise en place de mesures pour réduire  
l'attrition



... /

EDA /

MACHINE LEARNING /

RECOMMANDATIONS /

CONCLUSION /

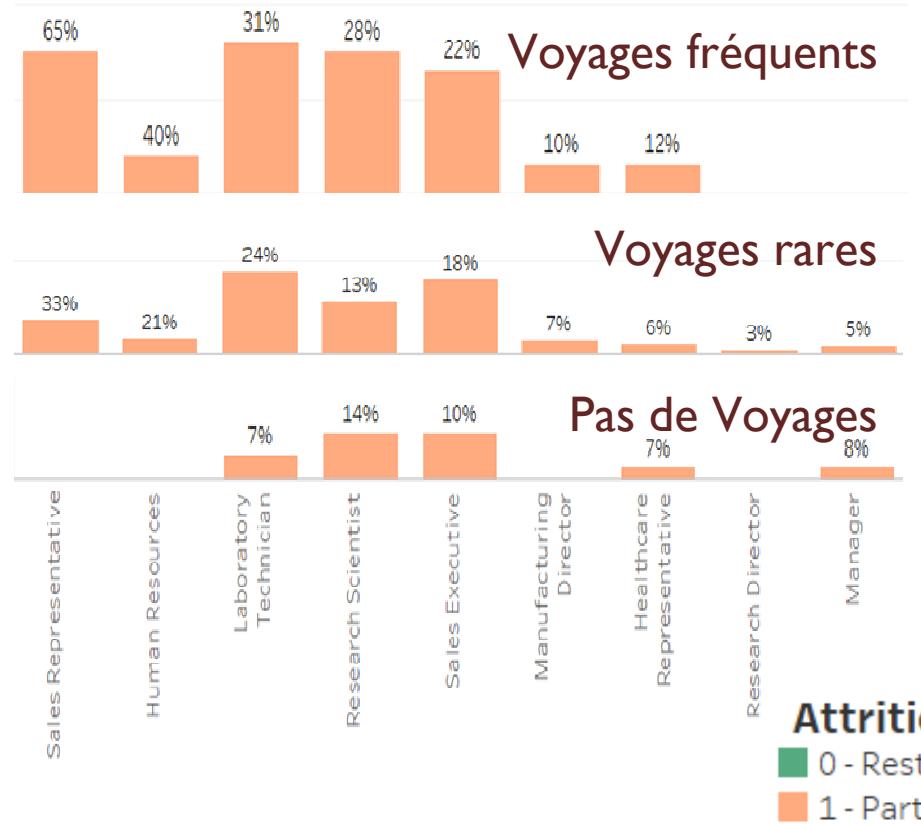
PERSPECTIVES

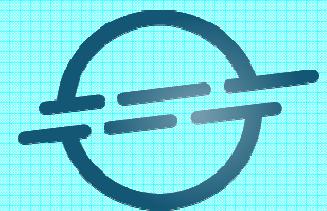
## Facteurs augmentant l'attrition

### Fréquence des déplacements professionnels

- La majorité des employés voyagent rarement pour le travail
- Figure : Proportion d'attritionnistes
  - pour chaque métier
  - selon la fréquence des déplacements
- + les voyages sont fréquents  
+ l'attrition ↗

Pourcentage d'**ATTRITIONNISTES** par métier et en fonction des déplacements professionnels





... /

EDA /

MACHINE LEARNING /

RECOMMANDATIONS /

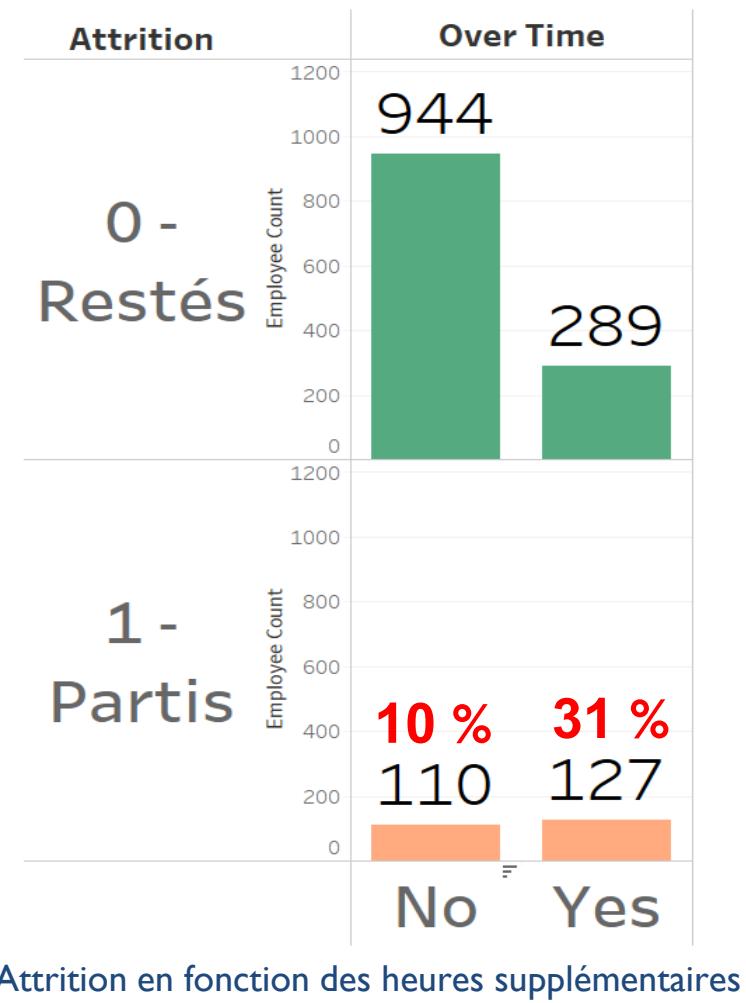
CONCLUSION /

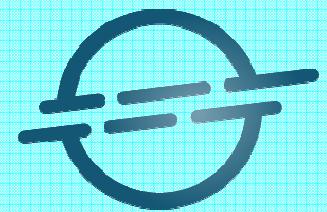
PERSPECTIVES

## Facteurs augmentant l'attrition

### Heures supplémentaires

- Seuls 416 employés font des heures supplémentaires < 30%
- **127 d'entre eux ont quitté l'entreprise => 31% d'attrition**





... /

EDA /

MACHINE LEARNING /

RECOMMANDATIONS /

CONCLUSION /

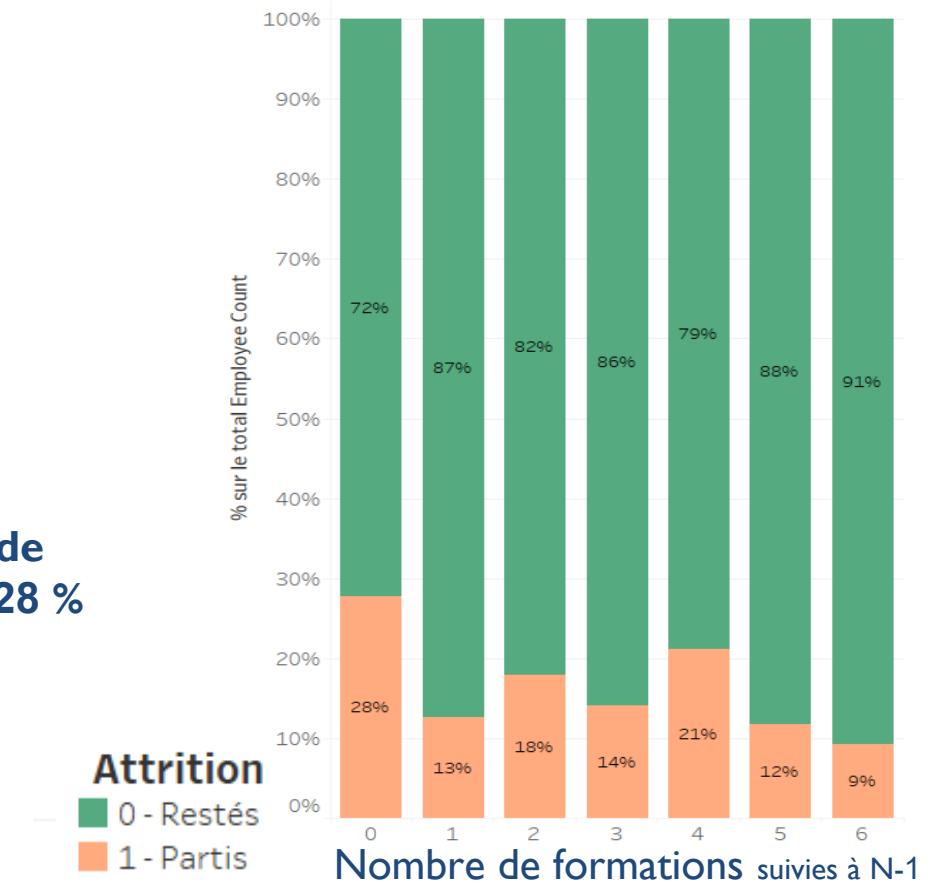
PERSPECTIVES

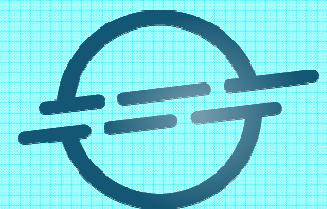
## Facteurs diminuant l'attrition

### Formations

- 70 % de la masse salariale a suivi 2 ou 3 formations l'année passée
- Les employés n'ayant pas suivi de formation ont une attrition à 28 %

Répartition de l'attrition en fonction du nombre de formation suivies à N-1





... /

EDA /

MACHINE LEARNING /

RECOMMANDATIONS /

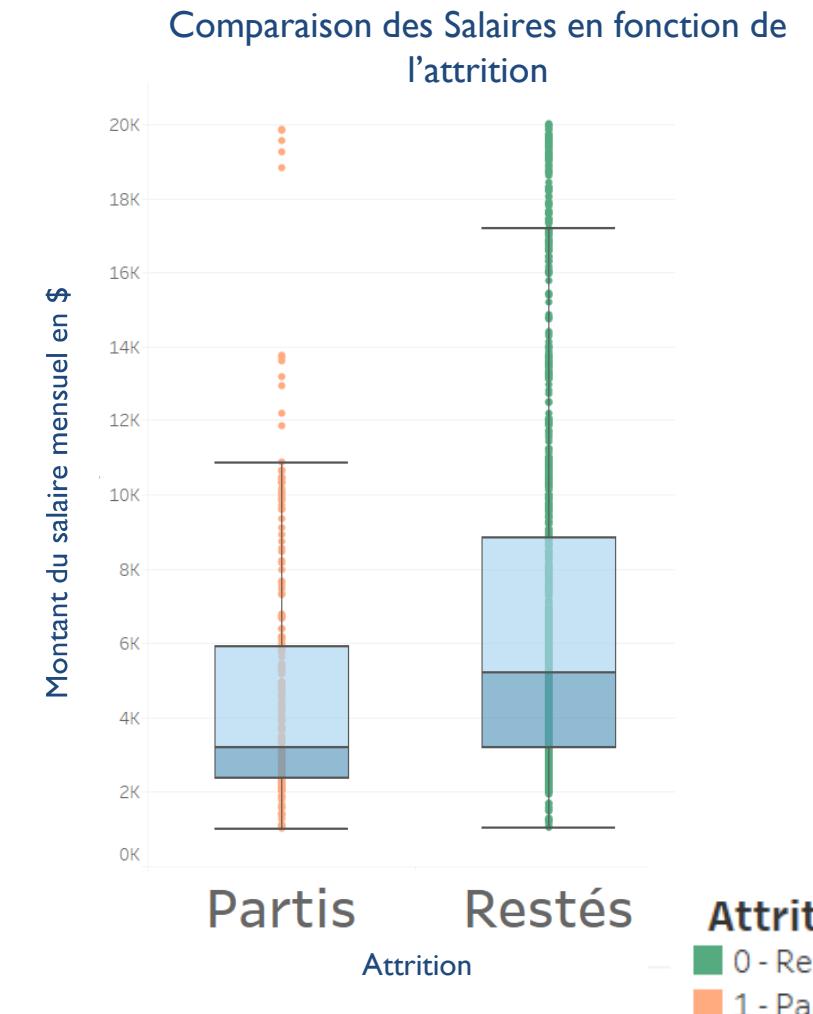
CONCLUSION /

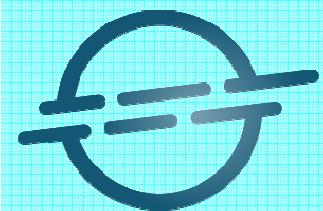
PERSPECTIVES

## Facteurs diminuant l'attrition

### Salaires

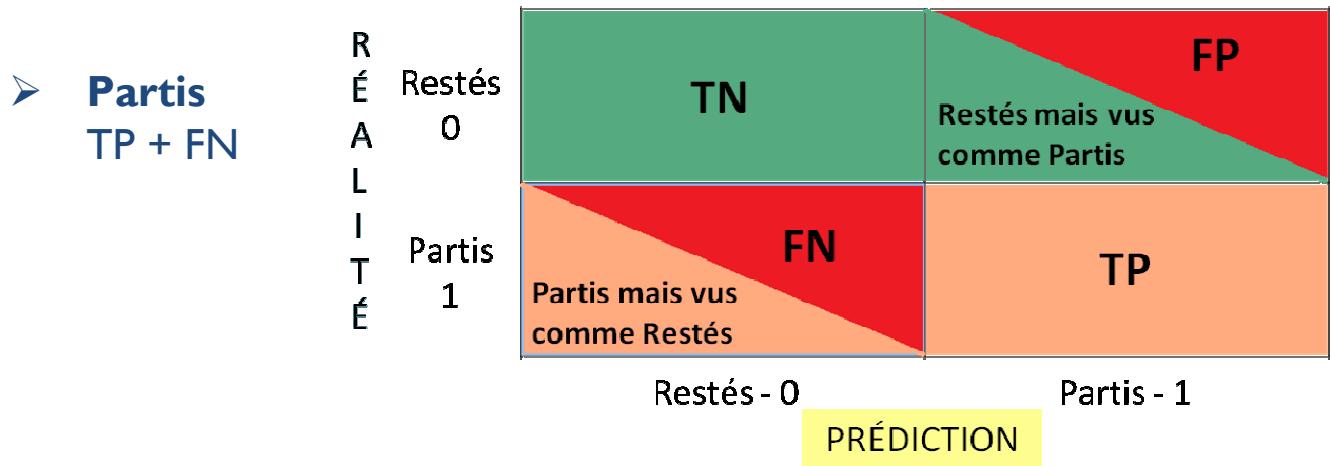
- Salaire médian = 5 000 \$
- Débutants et Juniors moins bien rémunérés => Attrition élevée
- Salaire médian partis < Salaire médian restés
- Peu d'attrition pour le personnel touchant des salaires > 11 K\$

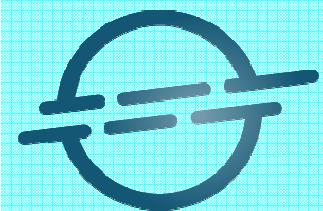




- I. INTRODUCTION
- II. PROBLÉMATIQUE
- III. ANALYSE EXPLORATOIRE
- DES DONNÉES
- IV. MACHINE LEARNING**
- V. RECOMMANDATIONS
- VI. CONCLUSION
- VII. PERSPECTIVES

- **Élimination de variables et d'individus**
  - \* Test statistiques de l'EDA et à nos observations
  - \* Retrait des individus de plus de 55 ans (pas de départ – Retraite)  
=> 1401 lignes (95%) et 20 colonnes
- Optimiser l'Accuracy (Pourcentage de prédictions correctes)
- **Métriques supplémentaires**
  - ROC-AUC (Capacité d'un modèle à discriminer les classes)
  - Recall (Taux de TP sur l'ensemble des Partis)





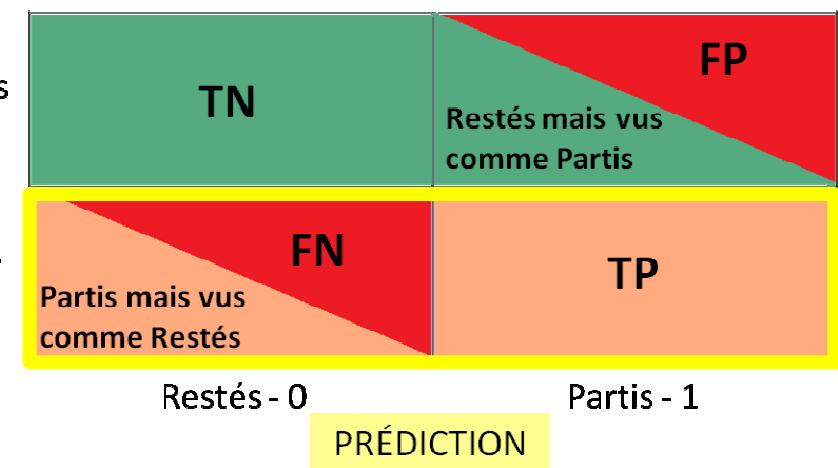
- I. INTRODUCTION
- II. PROBLÉMATIQUE
- III. ANALYSE EXPLORATOIRE
- DES DONNÉES
- IV. MACHINE LEARNING**
- V. RECOMMANDATIONS
- VI. CONCLUSION
- VII. PERSPECTIVES

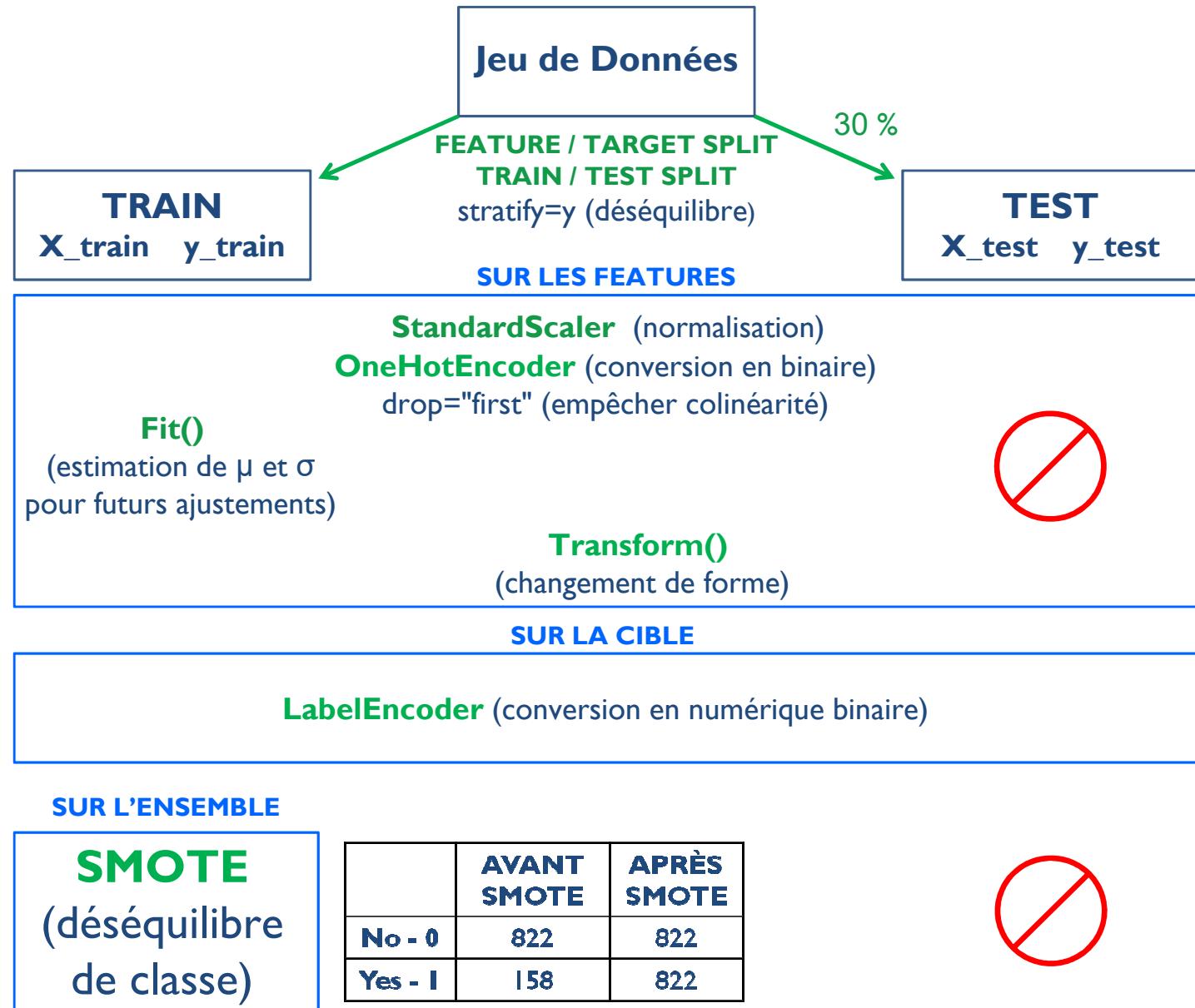
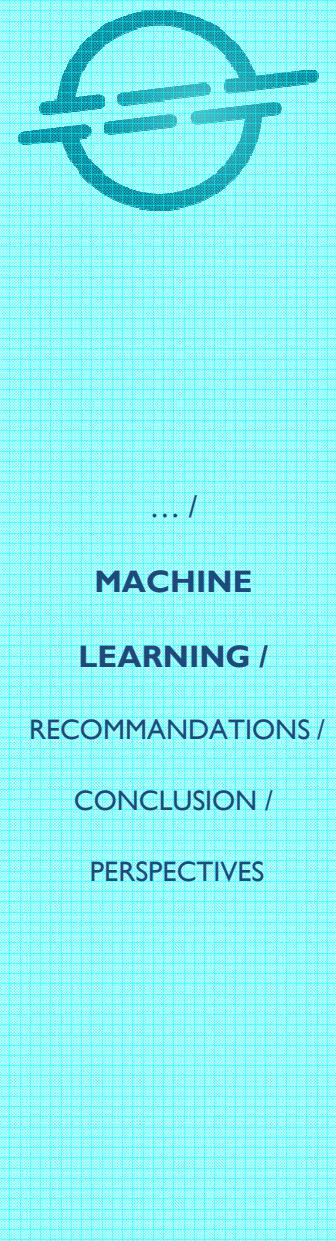
- **Élimination de variables et d'individus**
  - \* Test statistiques de l'EDA et à nos observations
  - \* Retrait des individus de plus de 55 ans (pas de départ – Retraite)  
=> 1401 lignes (95%) et 20 colonnes
- Optimiser l'Accuracy (Pourcentage de prédictions correctes)
- **Métriques supplémentaires**
  - ROC-AUC (Capacité d'un modèle à discriminer les classes)
  - Recall (Taux de TP sur l'ensemble des Partis)

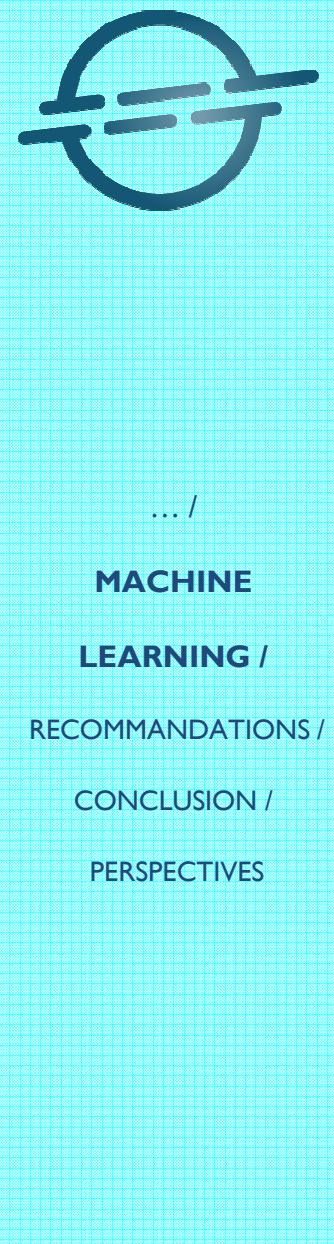
- **Partis**  
TP + FN

R  
É  
A  
L  
I  
T  
É

Restés  
0  
Partis  
1



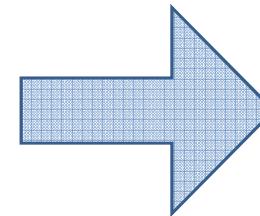




**3 modèles**  
**Logistic Regression**  
**Decision Tree**  
**Random Forest**

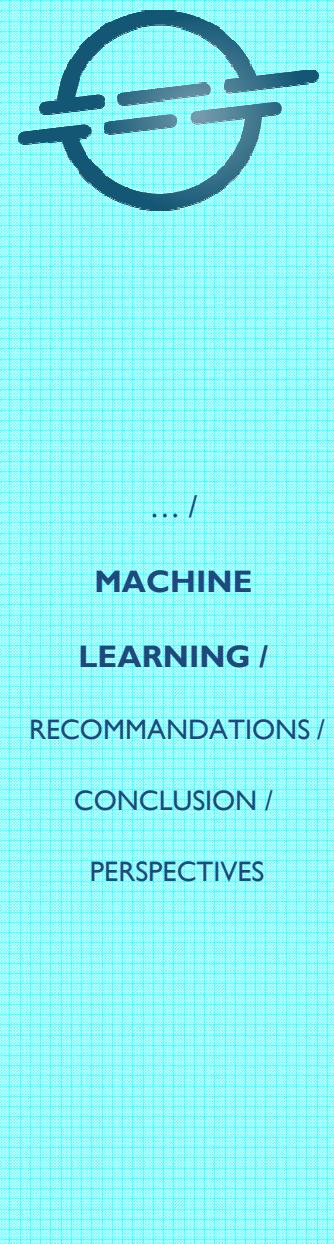


**Hyper  
paramétrages**  
**GridSearchCV**



**Choix**  
**Logistic Regression**  
(Meilleur Recall)

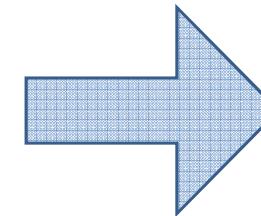
		Logistic Regression AVANT SMOTE	Logistic Regression AVANT HYPER- PARAMETRAGES	Logistic Regression AVANT SELECTION VARIABLES	Logistic Regression AVANT SELECTION INDIVIDUS	Logistic Regression	
T	R	Matrice de confusion Train set	[[849 14] [91 75]]	[[681 182] [133 730]]	[[686 133] [109 710]]	[[849 14] [91 75]]	[[678 144] [117 705]]
ACCURACY_Train-set			0,778	0,817	0,844	0,824	0,841
T	E	Matrice de confusion Test set	[[357 13] [47 24]]	[[307 63] [29 42]]	[[296 56] [23 45]]	[[357 13] [47 24]]	[[289 64] [20 48]]
ACCURACY_Test set			0,794	0,791	0,811	0,790	0,800
		RECALL	34%	59%	66%	68%	71%
		AUC	80%	80%	81%	82%	81%
		F1_score	44%	48%	53%	51%	53%
		Precision	65%	40%	45%	41%	43%
		Train-time	0,06	0,02	0,03	0,06	0,04
		Test-time	0,00	0,00	0,00	0,00	0,00



**3 modèles**  
**Logistic Regression**  
**Decision Tree**  
**Random Forest**



**Hyper  
paramétrages**  
**GridSearchCV**

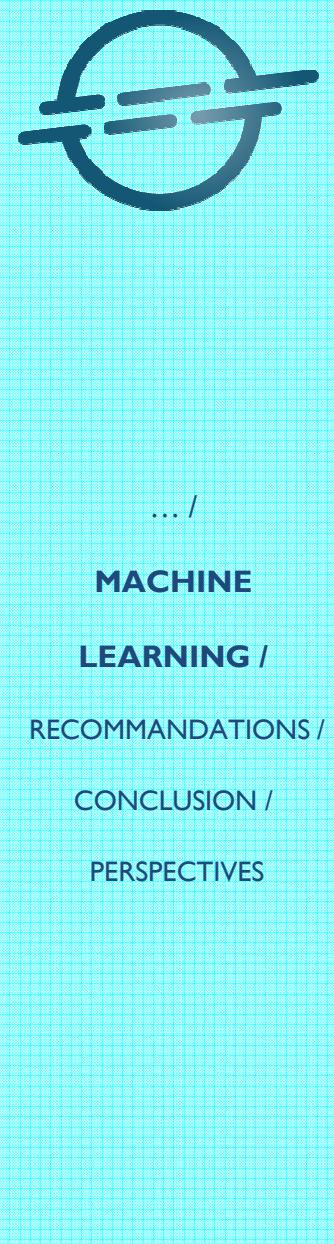


**Choix**

**Logistic Regression**  
(Meilleur Recall)

**VERSION FINALE**

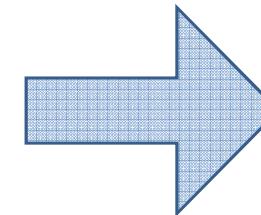
		Logistic Regression AVANT SMOTE	Logistic Regression AVANT HYPER- PARAMETRAGES	Logistic Regression AVANT SELECTION VARIABLES	Logistic Regression AVANT SELECTION INDIVIDUS	Logistic Regression	
T	R	Matrice de confusion Train set	[[849 14] [91 75]]	[[681 182] [133 730]]	[[686 133] [109 710]]	[[849 14] [91 75]]	[[678 144] [117 705]]
ACCURACY_Train-set			0,778	0,817	0,844	0,824	0,841
T	E	Matrice de confusion Test set	[[357 13] [47 24]]	[[307 63] [29 42]]	[[296 56] [23 45]]	[[357 13] [47 24]]	[[289 64] [20 48]]
ACCURACY_Test set			0,794	0,791	0,811	0,790	0,800
		RECALL	34%	59%	66%	68%	71%
		AUC	80%	80%	81%	82%	81%
		F1_score	44%	48%	53%	51%	53%
		Precision	65%	40%	45%	41%	43%
		Train-time	0,06	0,02	0,03	0,06	0,04
		Test-time	0,00	0,00	0,00	0,00	0,00



3 modèles  
Logistic Regression  
Decision Tree  
Random Forest



Hyper  
paramétrages  
**GridSearchCV**

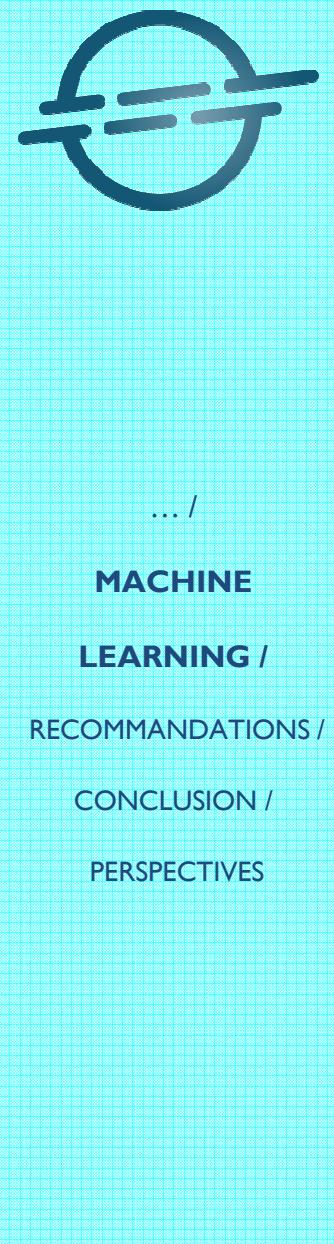


**Choix**  
**Logistic Regression**  
(Meilleur Recall)

**VERSION FINALE**

Équilibrée

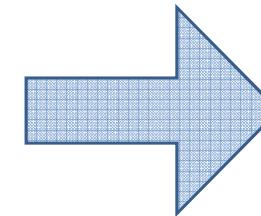
		Logistic Regression AVANT SMOTE	Logistic Regression AVANT HYPER- PARAMETRAGES	Logistic Regression AVANT SELECTION VARIABLES	Logistic Regression AVANT SELECTION INDIVIDUS	Logistic Regression	
T	R	Matrice de confusion Train set	[[849 14] [91 75]]	[[681 182] [133 730]]	[[686 133] [109 710]]	[[849 14] [91 75]]	[[678 144] [117 705]]
ACCURACY_Train-set			0,778	0,817	0,844	0,824	0,841
T	E	Matrice de confusion Test set	[[357 13] [47 24]]	[[307 63] [29 42]]	[[296 56] [23 45]]	[[357 13] [47 24]]	[[289 64] [20 48]]
ACCURACY_Test set			0,794	0,791	0,811	0,790	0,800
		RECALL	34%	59%	66%	68%	71%
		AUC	80%	80%	81%	82%	81%
		F1_score	44%	48%	53%	51%	53%
		Precision	65%	40%	45%	41%	43%
		Train-time	0,06	0,02	0,03	0,06	0,04
		Test-time	0,00	0,00	0,00	0,00	0,00



**3 modèles**  
**Logistic Regression**  
**Decision Tree**  
**Random Forest**



**Hyper  
paramétrages**  
**GridSearchCV**



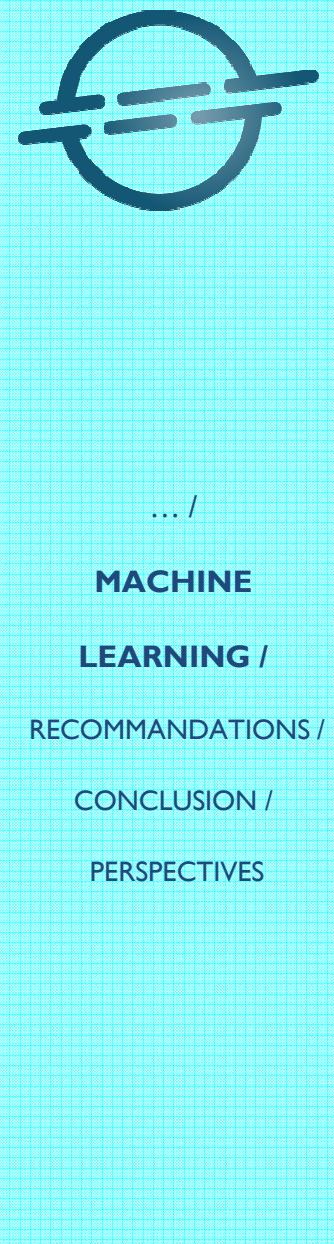
**Choix**

**Logistic Regression**  
(Meilleur Recall)

**VERSION FINALE**

		Logistic Regression AVANT SMOTE	Logistic Regression AVANT HYPER- PARAMETRAGES	Logistic Regression AVANT SELECTION VARIABLES	Logistic Regression AVANT SELECTION INDIVIDUS	Logistic Regression	
T	R	Matrice de confusion Train set	[[849 14] [91 75]]	[[681 182] [133 730]]	[[686 133] [109 710]]	[[849 14] [91 75]]	[[678 144] [117 705]]
ACCURACY_Train-set			0,778	0,817	0,844	0,824	0,841
T	E	Matrice de confusion Test set	[[357 13] [47 24]]	[[307 63] [29 42]]	[[296 56] [23 45]]	[[357 13] [47 24]]	[[289 64] [20 48]]
ACCURACY_Test set			0,794	0,791	0,811	0,790	0,800
		RECALL	34%	59%	66%	68%	71%
		AUC	80%	80%	81%	82%	81%
		F1_score	44%	48%	53%	51%	53%
		Precision	65%	40%	45%	41%	43%
		Train-time	0,06	0,02	0,03	0,06	0,04
		Test-time	0,00	0,00	0,00	0,00	0,00

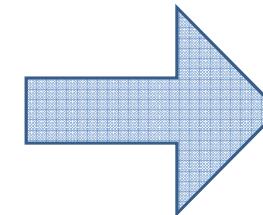
Beaucoup  
d'erreurs



**3 modèles**  
**Logistic Regression**  
**Decision Tree**  
**Random Forest**



**Hyper  
paramétrages**  
**GridSearchCV**



**Choix**

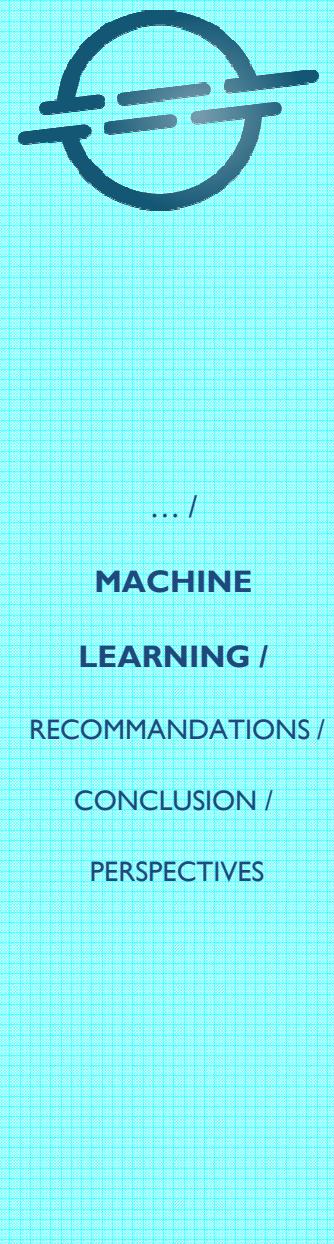
**Logistic Regression**  
(Meilleur Recall)

**VERSION FINALE**

		Logistic Regression AVANT SMOTE	Logistic Regression AVANT HYPER- PARAMETRAGES	Logistic Regression AVANT SELECTION VARIABLES	Logistic Regression AVANT SELECTION INDIVIDUS	Logistic Regression	
T	R	Matrice de confusion Train set	[[849 14] [91 75]]	[[681 182] [133 730]]	[[686 133] [109 710]]	[[849 14] [91 75]]	[[678 144] [117 705]]
ACCURACY_Train-set			0,778	0,817	0,844	0,824	0,841
T	E	Matrice de confusion Test set	[[357 13] [47 24]]	[[307 63] [29 42]]	[[296 56] [23 45]]	[[357 13] [47 24]]	[[289 64] [20 48]]
ACCURACY_Test set			0,794	0,791	0,811	0,790	0,800
		RECALL	34%	59%	66%	68%	71%
		AUC	80%	80%	81%	82%	81%
		F1_score	44%	48%	53%	51%	53%
		Precision	65%	40%	45%	41%	43%
		Train-time	0,06	0,02	0,03	0,06	0,04
		Test-time	0,00	0,00	0,00	0,00	0,00

TP reconnus

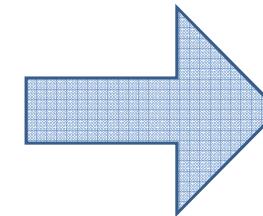
Différence Parti-resté



**3 modèles**  
**Logistic Regression**  
**Decision Tree**  
**Random Forest**



**Hyper  
paramétrages**  
**GridSearchCV**



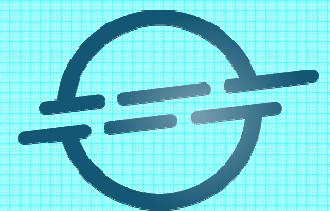
**Choix**  
**Logistic Regression**  
(Meilleur Recall)

**VERSION FINALE**

		Logistic Regression AVANT SMOTE	Logistic Regression AVANT HYPER- PARAMETRAGES	Logistic Regression AVANT SELECTION VARIABLES	Logistic Regression AVANT SELECTION INDIVIDUS	Logistic Regression	
T	R	Matrice de confusion Train set	[[849 14] [91 75]]	[[681 182] [133 730]]	[[686 133] [109 710]]	[[849 14] [91 75]]	[[678 144] [117 705]]
ACCURACY_Train-set			0,778	0,817	0,844	0,824	0,841
T	E	Matrice de confusion Test set	[[357 13] [47 24]]	[[307 63] [29 42]]	[[296 56] [23 45]]	[[357 13] [47 24]]	[[289 64] [20 48]]
ACCURACY_Test set			0,794	0,791	0,811	0,790	0,800
		RECALL	34%	59%	66%	68%	71%
		AUC	80%	80%	81%	82%	81%
		F1_score	44%	48%	53%	51%	53%
		Precision	65%	40%	45%	41%	43%
		Train-time	0,06	0,02	0,03	0,06	0,04
		Test-time	0,00	0,00	0,00	0,00	0,00

**Impacté**  
**Trop de FP**





... /

MACHINE

LEARNING /

RECOMMANDATIONS /

CONCLUSION /

PERSPECTIVES

## IMPORTANCE DES FEATURES

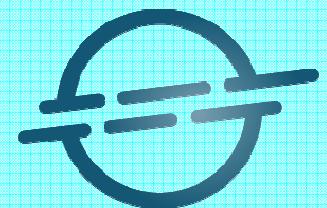
### Calcul des **odds-ratio** et de la **p-value**

Odds-ratios = exponentiation des coefficients de régression  
+ faciles à interpréter

#### **P-value :**

Department  
Age  
YearsAtCompany  
YearsInCurrentRole  
YearsWithCurrManager

**pas suffisamment de preuves** pour affirmer qu'elles sont liées à l'attrition



... /

MACHINE

LEARNING /

RECOMMANDATIONS /

CONCLUSION /

PERSPECTIVES

□ Odds-ratios  $\geq 1$

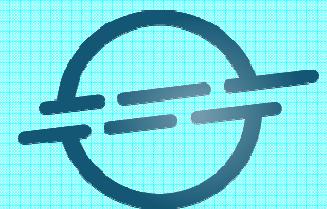


## Features augmentant l'attrition

### ➤ DÉPLACEMENTS PROFESSIONNELS

Si déplacements professionnels FRÉQUENTS  
= 22 fois + de risques de partir que NO TRAVEL

Si déplacements professionnels RARES  
= 7 fois + de chances de partir que NO TRAVEL



... /

MACHINE

LEARNING /

RECOMMANDATIONS /

CONCLUSION /

PERSPECTIVES

□ Odds-ratios  $\geq 1$



## Features augmentant l'attrition

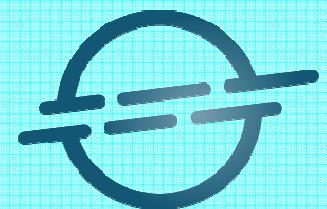
### ➤ DÉPLACEMENTS PROFESSIONNELS

Si déplacements professionnels FRÉQUENTS  
= 22 fois + de risques de partir que NO TRAVEL

Si déplacements professionnels RARES  
= 7 fois + de chances de partir que NO TRAVEL

### ➤ HEURES SUPPLÉMENTAIRES

Si heures supplémentaires effectuées  
= 13,1 fois + de risques de partir



... /

MACHINE

LEARNING /

RECOMMANDATIONS /

CONCLUSION /

PERSPECTIVES

□ Odds-ratios  $\geq 1$



## Features augmentant l'attrition

### ➤ DÉPLACEMENTS PROFESSIONNELS

Si déplacements professionnels FRÉQUENTS  
= 22 fois + de risques de partir que NO TRAVEL

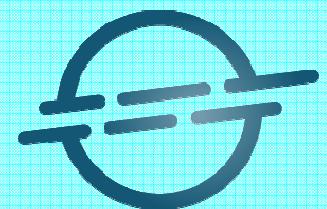
Si déplacements professionnels RARES  
= 7 fois + de chances de partir que NO TRAVEL

### ➤ HEURES SUPPLÉMENTAIRES

Si heures supplémentaires effectuées  
= 13,1 fois + de risques de partir

### ➤ DISTANCE TRAVAIL / DOMICILE

- + la distance est grande
- + le risque de départ est élevé



... /

MACHINE

LEARNING /

RECOMMANDATIONS /

CONCLUSION /

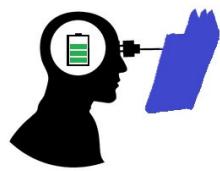
PERSPECTIVES

□ Odds-ratios < 1

Working Years



Training Times



Monthly



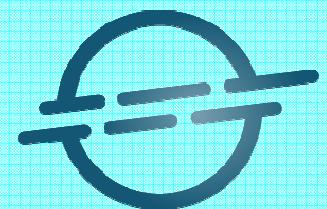
## Features diminuant l'attrition

- **NBRE D'ANNÉES DE TRAVAIL**
  - NBRE DE FORMATIONS SUIVIES**
  - SALAIRE**
  - NBRE DE STOCK OPTIONS**
- ↘ l'attrition à mesure que leurs valeurs ↗



Stock





... /

MACHINE

LEARNING /

RECOMMANDATIONS /

CONCLUSION /

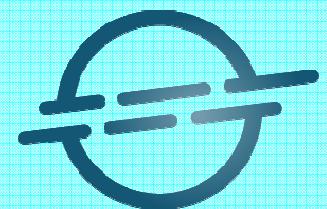
PERSPECTIVES

□ Odds-ratios < 1



## Features diminuant l'attrition

- **NBRE D'ANNÉES DE TRAVAIL**
- **NBRE DE FORMATIONS SUIVIES**
- **SALAIRE**
- **NBRE DE STOCK OPTIONS**
  - ↘ l'attrition à mesure que leurs valeurs ↗
- **SATISFACTION**
- **IMPLICATION**
- **ÉQUILIBRE VIE PRO / VIE PERSO**



... /

MACHINE

LEARNING /

RECOMMANDATIONS /

CONCLUSION /

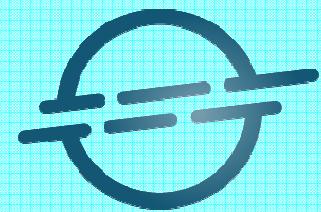
PERSPECTIVES

□ Odds-ratios < 1



## Features diminuant l'attrition

- **NBRE D'ANNÉES DE TRAVAIL**  
**NBRE DE FORMATIONS SUIVIES**  
**SALAIRE**  
**NBRE DE STOCK OPTIONS**  
↘ l'attrition à mesure que leurs valeurs ↗
- **SATISFACTION**
- **IMPLICATION**
- **ÉQUILIBRE VIE PRO / VIE PERSO**
- **DOMAINE D'ÉTUDE**  
Les salariés diplômés en Sciences et en Médical ont – de chances de quitter l'entreprise que les diplômés en Ressources Humaines



- I. INTRODUCTION
- II. PROBLÉMATIQUE
- III. ANALYSE EXPLORATOIRE  
DES DONNÉES
- IV. MACHINE LEARNING
- V.
- RECOMMANDATIONS**
- VI. CONCLUSION
- VII. PERSPECTIVES

*Les recommandations sont exposées dans l'ordre où il est préférable de mener les actions.*

✓ **ACCÈS AUX FORMATIONS**

Proposer des formations



**Bénéfices sur salarié :**  
Perspective d'avancement

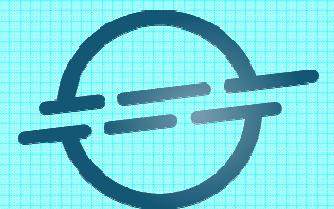
✓ **POLITIQUE DE RÉMUNÉRATION**

Garantir l'équité de la rémunération

Proposition d'avantages



**Bénéfices sur salarié :**  
Facteurs de motivation et de satisfaction



... /  
RECOMMENDATIONS /  
CONCLUSION /  
PERSPECTIVES

## ✓ CONTRAINTES PROFESSIONNELLES

### \* Déplacements professionnels et Heures supplémentaires

Avantage financier (épuisement professionnel)  
Les limiter  
Mieux les répartir sur l'ensemble des salariés



### \* Distance Travail-Maison

Prime au kilométrage  
Télétravail

#### Bénéfices sur salarié :

- ↗ équilibre vie pro/vie perso
- ↗ satisfaction

#### Bénéfices sur salarié :

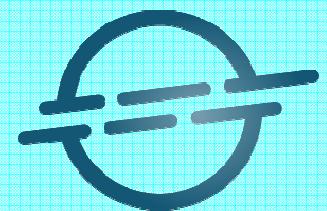
- ↘ temps
- ↘ coûts de trajet

## ✓ EMBAUCHE

Mieux cibler les candidats

Recruter des candidats dont le domaine d'étude est le MÉDICAL ou les SCIENCES DE LA VIE





- I. INTRODUCTION
- II. PROBLÉMATIQUE
- III. ANALYSE EXPLORATOIRE  
DES DONNÉES
- IV. MACHINE LEARNING
- V. RECOMMANDATIONS
- VI. CONCLUSION**
- VII. PERSPECTIVES

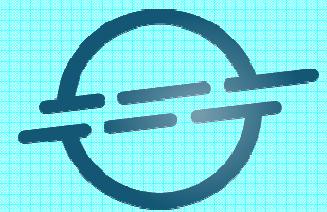
## NOS OBJECTIFS DE DÉPART

**Déterminer les facteurs qui jouent un rôle sur la décision de départ de salariés**

**Proposer des solutions pour diminuer l'attrition :**

Recommandations pour diminuer l'attrition volontaire

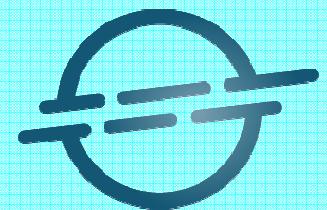
Construction d'un modèle de Machine Learning pour prédire le départ d'un salarié



- I. INTRODUCTION
- II. PROBLÉMATIQUE
- III. ANALYSE EXPLORATOIRE
  - DES DONNÉES
- IV. MACHINE LEARNING
- V. RECOMMANDATIONS
- VI. CONCLUSION**
- VII. PERSPECTIVES

## NOS OBJECTIFS DE DÉPART

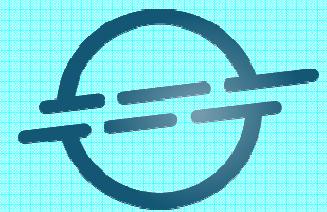
- ✓ **Déterminer les facteurs qui jouent un rôle sur la décision de départ de salariés**  
=> 12 identifiés parmi les 34 facteurs qu'on nous a remis
- Proposer des solutions pour diminuer l'attrition :**
  - Recommandations pour diminuer l'attrition volontaire
  - Construction d'un modèle de Machine Learning pour prédire le départ d'un salarié



- I. INTRODUCTION
- II. PROBLÉMATIQUE
- III. ANALYSE EXPLORATOIRE  
DES DONNÉES
- IV. MACHINE LEARNING
- V. RECOMMANDATIONS
- VI. CONCLUSION**
- VII. PERSPECTIVES

## NOS OBJECTIFS DE DÉPART

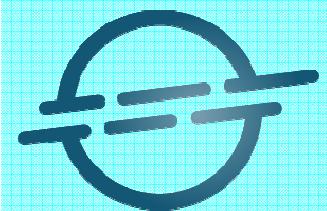
- ✓ **Déterminer les facteurs qui jouent un rôle sur la décision de départ de salariés**  
=> 12 identifiés parmi les 34 facteurs qu'on nous a remis
- ✓ **Proposer des solutions pour diminuer l'attrition :**
  - Recommandations pour diminuer l'attrition volontaire
  - Construction d'un modèle de Machine Learning pour prédire le départ d'un salarié



- I. INTRODUCTION
- II. PROBLÉMATIQUE
- III. ANALYSE EXPLORATOIRE  
DES DONNÉES
- IV. MACHINE LEARNING
- V. RECOMMANDATIONS
- VI. CONCLUSION**
- VII. PERSPECTIVES

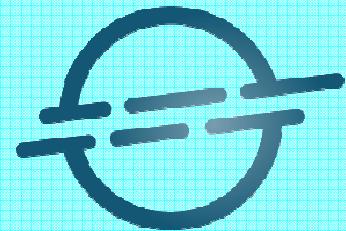
## NOS OBJECTIFS DE DÉPART

- ✓ **Déterminer les facteurs qui jouent un rôle sur la décision de départ de salariés**  
=> 12 identifiés parmi les 34 facteurs qu'on nous a remis
- ✓ **Proposer des solutions pour diminuer l'attrition :**
  - ✓ Recommandations pour diminuer l'attrition volontaire
  - ✓ Construction d'un modèle de Machine Learning pour prédire le départ d'un salarié
    - => Modèle améliorable
    - => **Bonnes perspectives de résultats**
    - => **Possible de le transférer à d'autres secteurs d'activité**

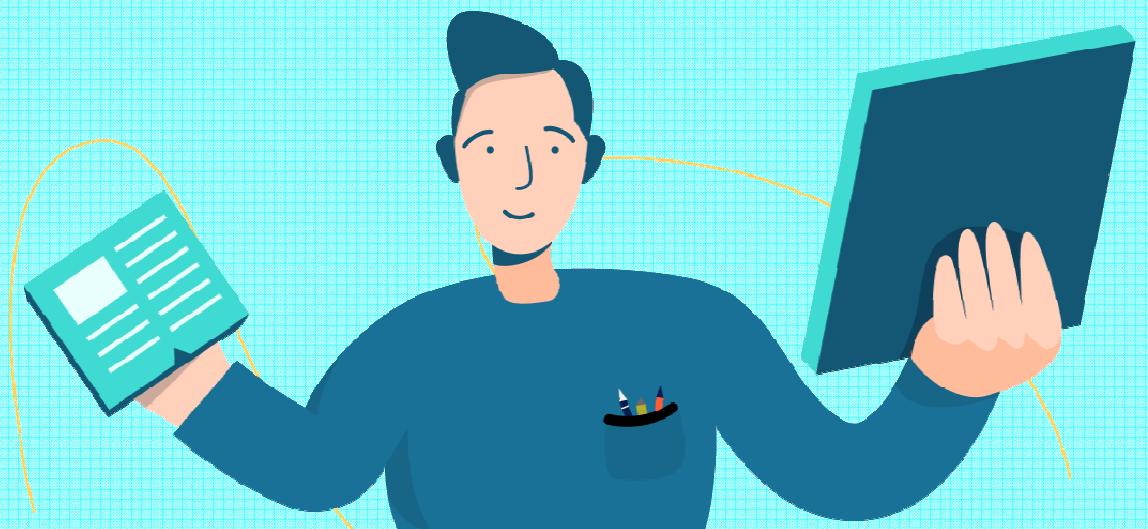


- I. INTRODUCTION
- II. PROBLÉMATIQUE
- III. ANALYSE EXPLORATOIRE  
DES DONNÉES
- IV. MACHINE LEARNING
- V. RECOMMANDATIONS
- VI. CONCLUSION
- VII. PERSPECTIVES**

- **Essayer :**
  - D'autres techniques de sur-échantillonnage (déséquilibre)
  - D'autres outils d'ajustement des paramètres de modèle
  - D'autres modèles de ML
- **Utiliser** Random Forest + adapté au sujet
- **Travailler** avec le département des Ressources Humaines
  - Mieux comprendre les variables
  - Établir un questionnaire + précis (facteurs + pertinents)
- **Catégoriser** la variable Attrition  
(volontaire, involontaire, démographique, retraite)
- **Enrichir** le dataset avec les données collectées les années précédentes et ajouter les nouvelles variables (questionnaire)



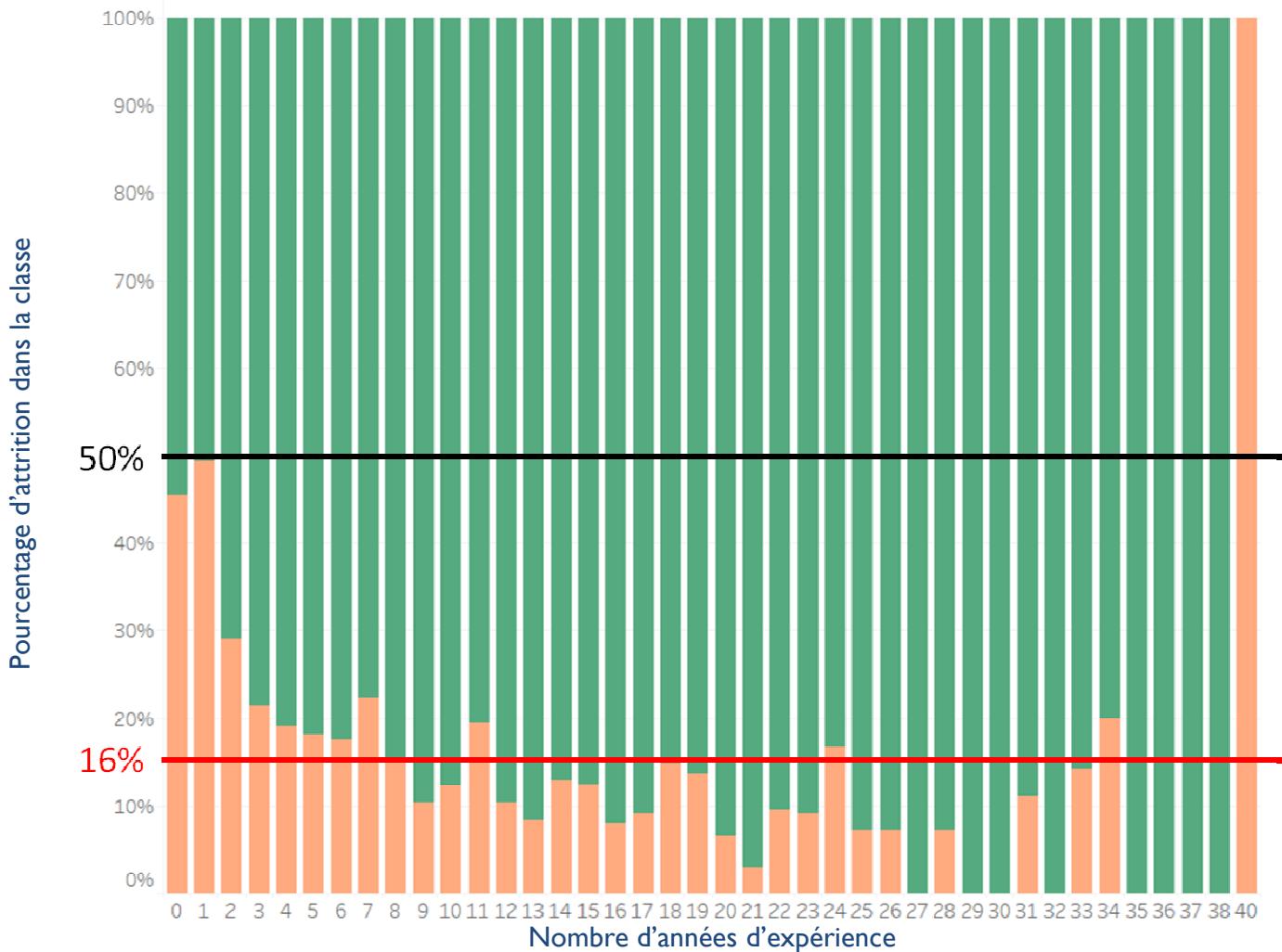
*Merci ...*







## Observations générales



### Expérience professionnelle

- 40 ans d'expérience professionnelle  
= Retraite
- Expérience de 0 à 7 ans  
= Attrition est plus élevée
- Expérience de 0 à 2 ans  
= presque 50 % d'attrition

Attrition1  
0-Restés  
1-Partis

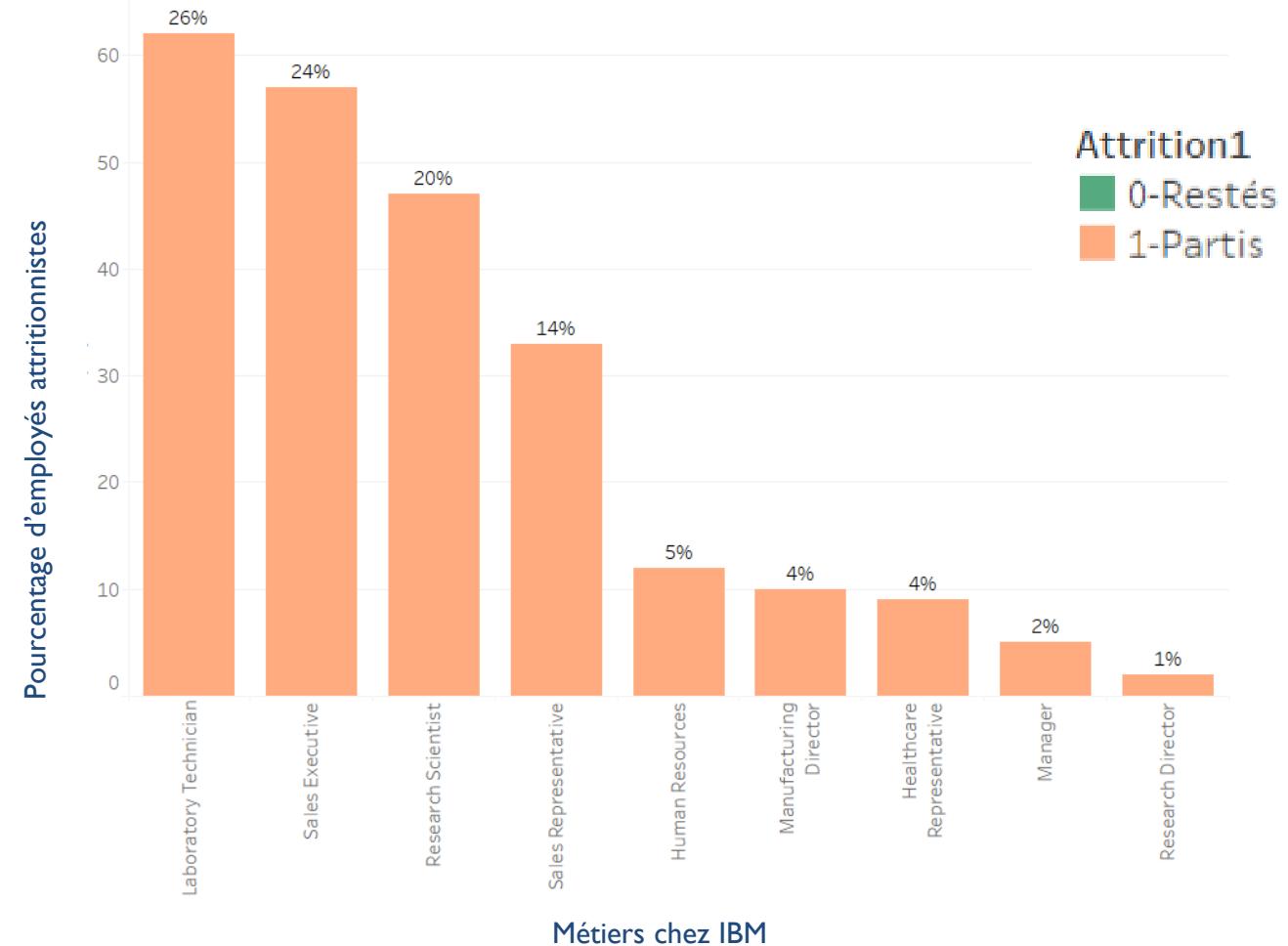


## Observations générales

### Répartition de l'attrition selon les métiers

Attrition plus élevée chez les :

- Techniciens de laboratoire 26 %
- Directeurs de vente 24 %
- Chercheurs 20 %

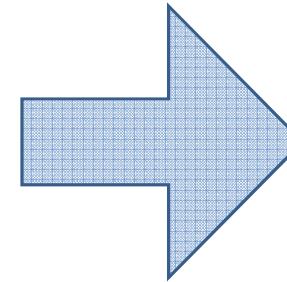




3 modèles  
Logistic Regression  
Decision Tree  
Random Forest



Hyper  
paramétrages  
**GridSearchCV**



Choix  
**Logistic Regression**  
(+ adapté au déséquilibre et Meilleur Recall)

**C = 20**

Contrôle la quantité de  
régularisation  
↓ risque d'overfitting

**class\_weight = 'balanced'**  
Compense le déséquilibre de classes

**penalty = 'l1'**  
Sélectionne les variables les +  
importantes  
↳ risque d'overfitting

```
LogisticRegression(C=20, class_weight='balanced', penalty='l1', random_state=42,  
                   solver='liblinear')
```

**solver = 'liblinear'**

Optimise la fonction de coût basée sur la  
descente de gradient  
Bien pour des LR avec des petits à moyens jeux  
de données

**random\_state = 42**  
Garantit la reproductibilité  
des résultats



		Logistic Regression AVANT SMOTE	Logistic Regression AVANT HYPER- PARAMETRAGES	Logistic Regression AVANT SELECTION VARIABLES	Logistic Regression AVANT SELECTION INDIVIDUS	Logistic Regression
T R A I N	Matrice de confusion Train set	[[849 14] [91 75]]	[[681 182] [133 730]]	[[686 133] [109 710]]	[[849 14] [91 75]]	[[678 144] [117 705]]
	ACCURACY_Train-set	0,778	0,817	0,844	0,824	<b>0,841</b>
T E S T	Matrice de confusion Test set	[[357 13] [47 24]]	[[307 63] [29 42]]	[[296 56] [23 45]]	[[357 13] [47 24]]	[[289 64] [20 48]]
	ACCURACY_Test set	0,794	0,791	0,811	0,790	<b>0,800</b>
		<b>RECALL</b>	34%	59%	66%	68%
		<b>AUC</b>	80%	80%	81%	82%
		<b>F1_score</b>	44%	48%	53%	51%
		Precision	65%	40%	45%	41%
		Train-time	0,06	0,02	0,03	0,06
		Test-time	0,00	0,00	0,00	<b>0,00</b>



		Logistic Regression AVANT SMOTE	Logistic Regression AVANT HYPER- PARAMETRAGES	Logistic Regression AVANT SELECTION VARIABLES	Logistic Regression AVANT SELECTION INDIVIDUS	Logistic Regression
T R A I N	Matrice de confusion Train set	[[849 14] [91 75]]	[[681 182] [133 730]]	[[686 133] [109 710]]	[[849 14] [91 75]]	[[678 144] [117 705]]
	ACCURACY_Train-set	0,778	0,817	0,844	0,824	0,841
T E S T	Matrice de confusion Test set	[[357 13] [47 24]]	[[307 63] [29 42]]	[[296 56] [23 45]]	[[357 13] [47 24]]	[[289 64] [20 48]]
	ACCURACY_Test set	0,794	0,791	0,811	0,790	0,800
		RECALL	34%	59%	66%	68%
		AUC	80%	80%	81%	82%
		F1_score	44%	48%	53%	51%
		Precision	65%	40%	45%	41%
		Train-time	0,06	0,02	0,03	0,06
		Test-time	0,00	0,00	0,00	0,00



		Logistic Regression AVANT SMOTE	Logistic Regression AVANT HYPER- PARAMETRAGES	Logistic Regression AVANT SELECTION VARIABLES	Logistic Regression AVANT SELECTION INDIVIDUS	Logistic Regression
T R A I N	Matrice de confusion Train set	[[849 14] [91 75]]	[[681 182] [133 730]]	[[686 133] [109 710]]	[[849 14] [91 75]]	[[678 144] [117 705]]
	ACCURACY_Train-set	0,778	0,817	0,844	0,824	0,841
T E S T	Matrice de confusion Test set	[[357 13] [47 24]]	[[307 63] [29 42]]	[[296 56] [23 45]]	[[357 13] [47 24]]	[[289 64] [20 48]]
	ACCURACY_Test set	0,794	0,791	0,811	0,790	0,800
		RECALL	34%	59%	66%	68% 71%
		AUC	80%	80%	81%	82% 81%
		F1_score	44%	48%	53%	51% 53%
		Precision	65%	40%	45%	41% 43%
		Train-time	0,06	0,02	0,03	0,06 0,04
		Test-time	0,00	0,00	0,00	0,00



		Logistic Regression AVANT SMOTE	Logistic Regression AVANT HYPER- PARAMETRAGES	Logistic Regression AVANT SELECTION VARIABLES	Logistic Regression AVANT SELECTION INDIVIDUS	Logistic Regression
T R A I N	Matrice de confusion Train set	[[849 14] [91 75]]	[[681 182] [133 730]]	[[686 133] [109 710]]	[[849 14] [91 75]]	[[678 144] [117 705]]
	ACCURACY_Train-set	0,778	0,817	0,844	0,824	0,841
T E S T	Matrice de confusion Test set	[[357 13] [47 24]]	[[307 63] [29 42]]	[[296 56] [23 45]]	[[357 13] [47 24]]	[[289 64] [20 48]]
	ACCURACY_Test set	0,794	0,791	0,811	0,790	0,800
		RECALL	34%	59%	66%	68%
		AUC	80%	80%	81%	82%
		F1_score	44%	48%	53%	51%
		Precision	65%	40%	45%	41%
		Train-time	0,06	0,02	0,03	0,06
		Test-time	0,00	0,00	0,00	0,00



		Logistic Regression AVANT SMOTE	Logistic Regression AVANT HYPER- PARAMETRAGES	Logistic Regression AVANT SELECTION VARIABLES	Logistic Regression AVANT SELECTION INDIVIDUS	Logistic Regression
T R A I N	Matrice de confusion Train set	[[849 14] [91 75]]	[[681 182] [133 730]]	[[686 133] [109 710]]	[[849 14] [91 75]]	[[678 144] [117 705]]
	ACCURACY_Train-set	0,778	0,817	0,844	0,824	<b>0,841</b>
T E S T	Matrice de confusion Test set	[[357 13] [47 24]]	[[307 63] [29 42]]	[[296 56] [23 45]]	[[357 13] [47 24]]	[[289 64] [20 48]]
	ACCURACY_Test set	0,794	0,791	0,811	0,790	<b>0,800</b>
		RECALL	34%	59%	66%	68% <b>71%</b>
		AUC	80%	80%	81%	82% <b>81%</b>
		F1_score	44%	48%	53%	51% <b>53%</b>
		Precision	65%	40%	45%	41% <b>43%</b>
		Train-time	0,06	0,02	0,03	0,06 <b>0,04</b>
		Test-time	0,00	0,00	0,00	<b>0,00</b>