Master Big Data et Intelligence Artificial

Big Data 2 : DATA VISUALIZATION

Progress notebook

RÉALISÉ PAR :
ATANANE Chaima
Hajar EL AAZZOUZI

Encadré par :
Pr . M. El HAJJI &
Tarek AÏT BAHA

1) Introduction

Le projet A vise à analyser les données sur les ressources humaines (HR) d'une entreprise afin de comprendre les facteurs liés à l'attrition des employés. Ce rapport détaille les différentes étapes du projet, depuis l'importation des données jusqu'à la création d'un tableau de bord interactif sur Tableau.

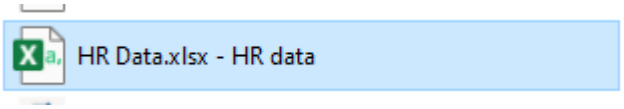
2) Objectif (Fin novembre) :

L'objectif du Projet A est de comprendre les facteurs liés à l'attrition des employés au sein d'une entreprise en utilisant une approche d'analyse de données et de modélisation prédictive.

3) Collecte et Exploration de Données sur les Ressources Humaines : Un Regard Approfondi sur les Facteurs d'Attrition des Employés

La collecte et l'exploration de données sur les ressources humaines représentent une phase cruciale de notre projet, visant à dévoiler les mécanismes sous-jacents de l'attrition des employés au sein de notre organisation. Cette étape fondamentale permettra de rassembler des informations significatives à partir de diverses sources de données, offrant ainsi une perspective holistique sur les facteurs qui influent sur la décision des employés de rester ou de quitter l'entreprise. La compréhension approfondie de ces données sert de fondement essentiel à notre analyse exploratoire et à la construction ultérieure de modèles prédictifs. Ce processus de collecte de données constitue le socle sur lequel reposera notre capacité à fournir des recommandations éclairées pour améliorer la rétention des employés et à élaborer des stratégies adaptées aux besoins spécifiques de notre personnel.

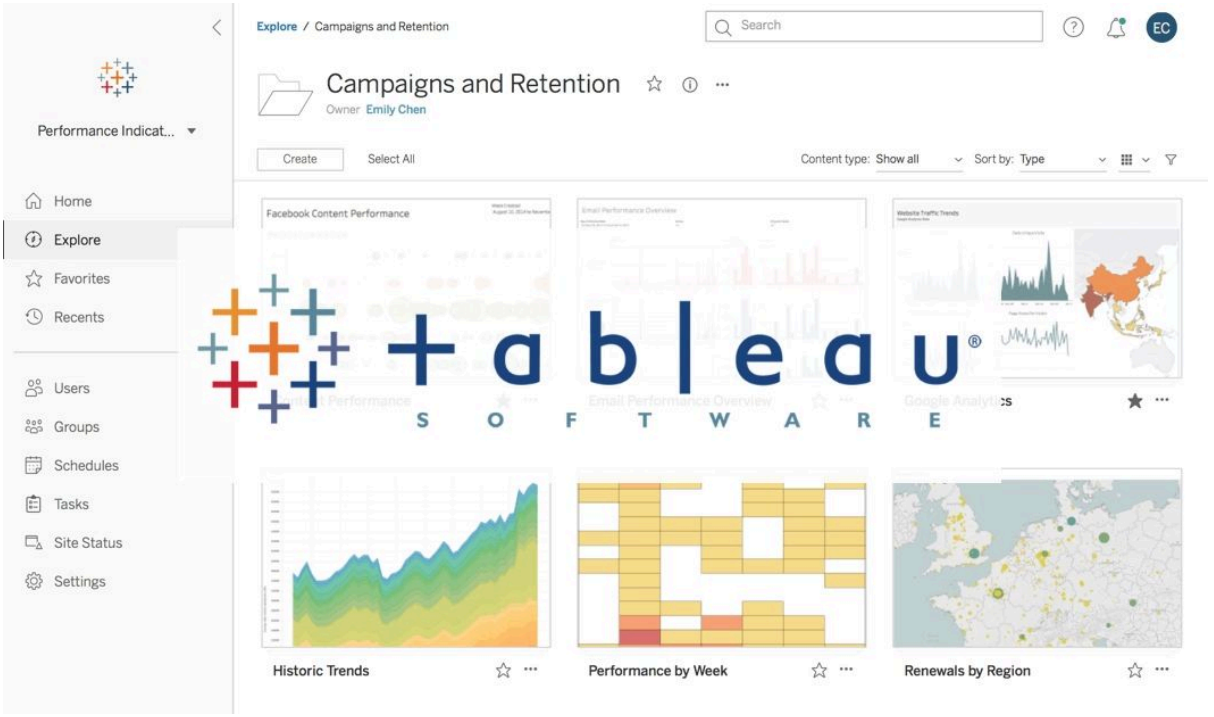
A	B	C	D	E	F	G	H	I	J	K	L	M	N
Attrition	Business Travel	CF_age band	CF_attrition label	Department	Education Field	emp no	Employee Number	Gender	Job Role	Marital Status	Over Time	Over18	Training Times L
Yes	Travel_Rarely	35 - 44	Ex-Employees	Sales	Life Sciences	STAFF-1	1	Female	Sales Executive	Single	Yes	Y	
No	Travel_Frequently	45 - 54	Current Employees	R&D	Life Sciences	STAFF-2	2	Male	Research Scientist	Married	No	Y	
Yes	Travel_Rarely	35 - 44	Ex-Employees	R&D	Other	STAFF-4	4	Male	Laboratory Technician	Single	Yes	Y	
No	Travel_Frequently	25 - 34	Current Employees	R&D	Life Sciences	STAFF-5	5	Female	Research Scientist	Married	Yes	Y	
No	Travel_Rarely	25 - 34	Current Employees	R&D	Medical	STAFF-7	7	Male	Laboratory Technician	Married	No	Y	
No	Travel_Frequently	25 - 34	Current Employees	R&D	Life Sciences	STAFF-8	8	Male	Laboratory Technician	Single	No	Y	
No	Travel_Rarely	Over 55	Current Employees	R&D	Medical	STAFF-10	10	Female	Laboratory Technician	Married	Yes	Y	
No	Travel_Rarely	25 - 34	Current Employees	R&D	Life Sciences	STAFF-11	11	Male	Laboratory Technician	Divorced	No	Y	
No	Travel_Frequently	35 - 44	Current Employees	R&D	Life Sciences	STAFF-12	12	Male	Manufacturing Director	Single	No	Y	
No	Travel_Rarely	35 - 44	Current Employees	R&D	Medical	STAFF-13	13	Male	Healthcare Representative	Married	No	Y	
No	Travel_Rarely	35 - 44	Current Employees	R&D	Medical	STAFF-14	14	Male	Laboratory Technician	Married	No	Y	
No	Travel_Rarely	25 - 34	Current Employees	R&D	Life Sciences	STAFF-15	15	Female	Laboratory Technician	Single	Yes	Y	
No	Travel_Rarely	25 - 34	Current Employees	R&D	Life Sciences	STAFF-16	16	Male	Research Scientist	Divorced	No	Y	
No	Travel_Rarely	25 - 34	Current Employees	R&D	Medical	STAFF-18	18	Male	Laboratory Technician	Divorced	No	Y	
Yes	Travel_Rarely	25 - 34	Ex-Employees	R&D	Life Sciences	STAFF-19	19	Male	Laboratory Technician	Single	Yes	Y	
No	Travel_Rarely	25 - 34	Current Employees	R&D	Life Sciences	STAFF-20	20	Female	Manufacturing Director	Divorced	No	Y	
No	Travel_Rarely	25 - 34	Current Employees	R&D	Life Sciences	STAFF-21	21	Male	Research Scientist	Divorced	Yes	Y	
No	Non-Travel	Under 25	Current Employees	R&D	Medical	STAFF-22	22	Male	Laboratory Technician	Divorced	Yes	Y	
No	Travel_Rarely	45 - 54	Current Employees	Sales	Life Sciences	STAFF-23	23	Female	Manager	Married	No	Y	
No	Travel_Rarely	35 - 44	Current Employees	R&D	Life Sciences	STAFF-24	24	Male	Research Scientist	Single	Yes	Y	
No	Non-Travel	Under 25	Current Employees	R&D	Other	STAFF-26	26	Female	Manufacturing Director	Divorced	No	Y	
Yes	Travel_Rarely	35 - 44	Ex-Employees	Sales	Life Sciences	STAFF-27	27	Male	Sales Representative	Single	No	Y	
No	Travel_Rarely	25 - 34	Current Employees	R&D	Life Sciences	STAFF-28	28	Female	Research Director	Single	No	Y	
No	Travel_Rarely	Under 25	Current Employees	R&D	Life Sciences	STAFF-30	30	Male	Research Scientist	Single	No	Y	
Yes	Travel_Rarely	25 - 34	Ex-Employees	R&D	Medical	STAFF-31	31	Male	Research Scientist	Single	No	Y	



4) Infrastructure Technique :

La mise en œuvre technique de notre projet repose sur l'utilisation de bibliothèques Python spécialisées et l'intégration du logiciel Tableau pour mener une analyse approfondie des données relatives à l'attrition des employés. Le langage de programmation Python, renforcé par des bibliothèques telles que pandas pour la manipulation des données et matplotlib pour la visualisation, constitue le moteur principal de notre approche analytique. Cette combinaison d'outils offre une flexibilité et une puissance considérables pour extraire des insights significatifs à partir des ensembles de données complexes sur les ressources humaines.

En parallèle, l'utilisation du logiciel Tableau ajoute une dimension interactive à notre analyse. Il facilite la création de tableaux de bord visuellement attrayants, permettant une présentation claire et compréhensible des résultats. L'intégration fluide entre Python et Tableau crée une synergie puissante, où la robustesse des bibliothèques Python rencontre l'expressivité visuelle de Tableau, offrant ainsi une plateforme complète pour explorer, analyser et communiquer les tendances relatives à l'attrition des employés.



5) Préparation Initiale - Importation, Lecture et Exploration des Données

Cette première étape marque le point de départ de notre parcours analytique. Elle englobe un triptyque d'actions fondamentales. Tout d'abord, l'importation des bibliothèques nécessaires assure que les outils indispensables à la manipulation et à la visualisation de données sont à notre disposition, établissant ainsi les fondations de notre analyse. Ensuite, la phase de Lecture des données s'engage, où les informations cruciales sur les ressources humaines sont extraites d'un fichier CSV pour être soigneusement organisées dans un DataFrame. Enfin, l'étape d' Exploration des données prend son envol, donnant un premier aperçu du paysage de nos données. Les premières lignes, les informations sur les types de données et les statistiques descriptives viennent éclairer les contours de nos données, amorçant ainsi notre quête d'insights et de compréhension approfondie. C'est dans cette phase initiale que les données se dévoilent, prêtes à dévoiler les histoires qu'elles renferment.

```
  Attrition    Business Travel CF_age band CF_attrition label Departmen
t  \
0       Yes      Travel_Rarely     35 - 44       Ex-Employees      Sale
s
1        No  Travel_Frequently     45 - 54  Current Employees        R&
D
2       Yes      Travel_Rarely     35 - 44       Ex-Employees        R&
D
3        No  Travel_Frequently     25 - 34  Current Employees        R&
D
4        No      Travel_Rarely     25 - 34  Current Employees        R&
D

  Education Field   emp no  Employee Number  Gender                Job R
ole  \
0   Life Sciences  STAFF-1                1  Female        Sales Execut
ive
1   Life Sciences  STAFF-2                2    Male     Research Scient
ist
```

6) Forge des Données - Traitement et Visualization :

L'étape de la Forge des Données marque le passage du brut à l'affiné. Ici, deux volets s'entrelacent harmonieusement. Tout d'abord, le traitement des données se manifeste par la sélection minutieuse des colonnes pertinentes pour notre analyse. C'est le moment où les données brutes sont sculptées pour révéler les aspects cruciaux, créant ainsi un ensemble de données épuré et ciblé. En parallèle, la phase de visualisation des données se déploie avec l'utilisation judicieuse de différents graphiques tels que scatter plots, histogrammes, et diagrammes à barres.

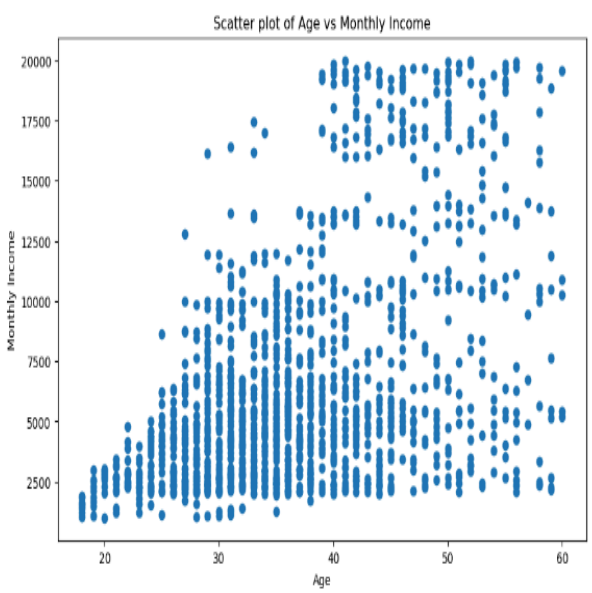
Ces visualisations deviennent des fenêtres sur les caractéristiques des employés, offrant une compréhension visuelle profonde de la distribution des données et des tendances émergentes. Ensemble, ces deux processus incarnent la phase où les données prennent forme et se transforment en un canevas riche en détails, prêt à être interprété et exploré plus avant.

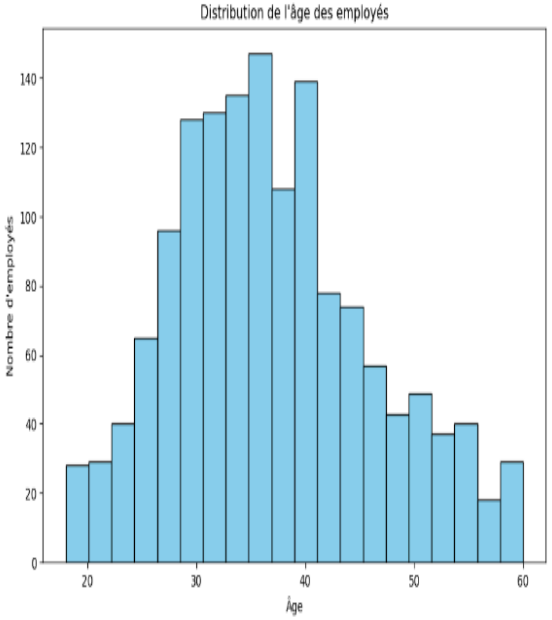
Exploration Visuelle des Traits Caractéristiques

- Le Scatter Plot Age vs Monthly Income offre une vue panoramique des relations potentielles entre l'âge des employés et leurs revenus mensuels. Les points dispersés sur ce graphique révèlent des schémas, mettant en lumière toute corrélation ou tendance notable.

- Le deuxième , l'Histogramme de l'Âge ajoute une couche d'analyse en offrant une perspective de la distribution de l'âge au sein de notre effectif. La hauteur des barres

sur ce graphique crée une représentation visuelle de la concentration d'employés dans différentes tranches d'âge, dévoilant ainsi des insights sur la démographie de l'entreprise.

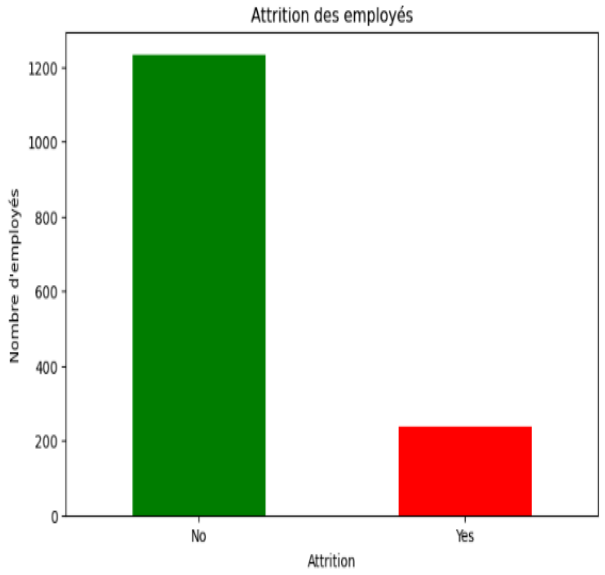


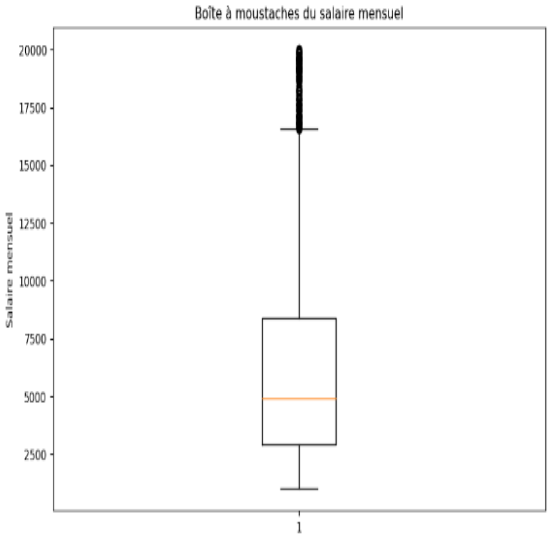


<u>Radiographie de l'Attrition et Économie de Salaire</u>

- Cette section s'ouvre avec le Diagramme à Barres pour l'Attrition qui offre un aperçu immédiat du flux des employés. Les deux teintes, vert et rouge, désignent respectivement ceux qui ont choisi de rester et ceux qui ont pris le chemin du départ. Ce visuel, simple mais puissant, permet une évaluation rapide de la dynamique de l'attrition au sein de l'entreprise.

- En continuant notre exploration, la Boîte à Moustaches pour le Salaire Mensuel entre en scène, délivrant des informations plus détaillées sur la répartition des salaires. Cette boîte à moustaches, par sa structure, dévoile la médiane, les quartiles, et identifie d'éventuelles valeurs aberrantes, offrant ainsi une radiographie visuelle des émoluments mensuels des employés.

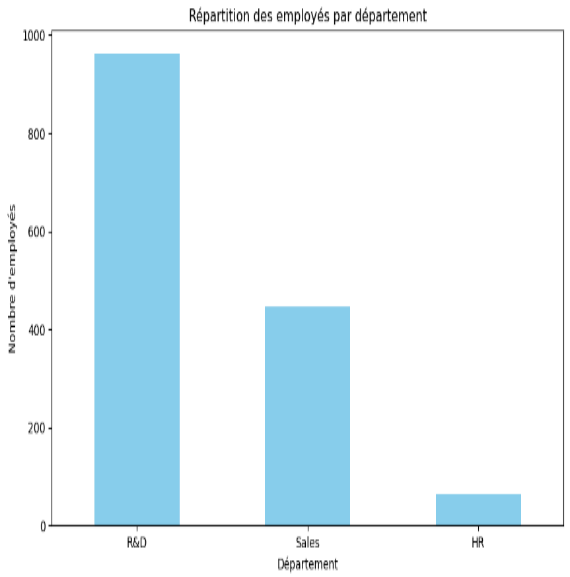


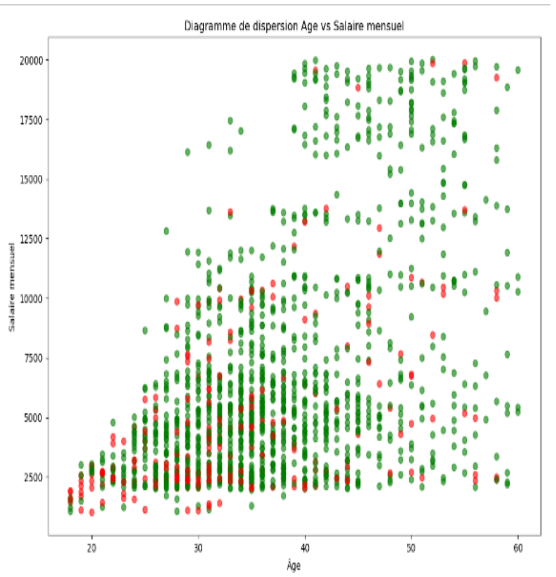


Parcours Départemental et Signature d'Attrition

- Le Diagramme à Barres pour le Département sert comme une cartographie visuelle de la distribution des employés dans différents secteurs de l'entreprise. Les barres bleu ciel se dressent, chaque hauteur représentant le nombre d'employés par département. C'est une exploration visuelle qui offre un aperçu immédiat de la structure organisationnelle et de la répartition des effectifs.

- Le Diagramme de Dispersion Age vs Salaire Mensuel entre en scène avec des nuances de couleurs révélatrices. Ce graphique, à travers la dispersion des points, illustre la relation potentielle entre l'âge des employés et leur salaire mensuel. Les points rouges signalent ceux qui ont choisi de quitter l'entreprise, créant une superposition visuelle de l'attrition sur cette dynamique.





7) Modélisation de l'Attrition des Employés : Anticipation des Départs avec la Régression Logistique

Préparation des Données pour la Modélisation

L'étape initiale englobe l'importation des bibliothèques Python essentielles pour la manipulation de données, la division des ensembles d'entraînement et de test, le codage des étiquettes, ainsi que la création et l'évaluation d'un modèle de Régression Logistique. Cette phase instaure le fondement technique nécessaire à la modélisation. Ensuite, les données issues d'un fichier CSV sont chargées dans un DataFrame de pandas, formant ainsi une structure de données appropriée à l'analyse. Par la suite, le prétraitement des données intervient avec l'utilisation d'un encodeur d'étiquettes, convertissant les valeurs de la colonne 'Attrition' en nombres (0 pour 'No' et 1 pour 'Yes'). Cette étape revêt une importance cruciale, facilitant la compréhension et le traitement des données par le modèle. Ce processus global constitue la première étape essentielle dans la préparation des données pour la modélisation, jetant ainsi les bases d'une analyse approfondie.

Modélisation - Sélection des caractéristiques et Division des données

Dans cette étape cruciale, les caractéristiques pertinentes, telles que l'âge, le salaire mensuel, la distance du domicile, etc., sont sélectionnées comme variables indépendantes, tandis que la variable cible 'Attrition' est définie. Ces caractéristiques choisies forment la base sur laquelle le modèle va être construit. Subséquemment, les données sont fractionnées en ensembles d'entraînement et de test à l'aide de la fonction train test split, avec **80%** des données utilisées pour l'entraînement et **20%** réservées pour les tests. Cette division est cruciale pour évaluer la performance du modèle sur des données non vues. En parallèle, un modèle de régression logistique est

instancié, amorçant ainsi le processus de création du modèle qui sera entraîné sur les données d'entraînement pour anticiper l'attrition des employés en fonction des caractéristiques spécifiées. Cette étape sert de fondation pour le développement du modèle de prédiction.

Modélisation - Entraînement, Prédictions et Évaluation du Modèle :

À cette étape cruciale, le modèle de régression logistique est entraîné sur l'ensemble d'entraînement à l'aide de la méthode fit. Cependant, il est essentiel de noter qu'un avertissement de convergence peut apparaître, indiquant que le modèle n'a pas convergé après un certain nombre d'itérations. Dans ce cas, des ajustements, tels que l'augmentation du nombre d'itérations ou la mise à l'échelle des données, peuvent être nécessaires pour garantir la convergence. Une fois le modèle entraîné, il est utilisé pour faire des prédictions sur l'ensemble de test, anticipant l'attrition des employés en fonction des caractéristiques sélectionnées. Enfin, l'exactitude du modèle, la matrice de confusion et le rapport de classification sont calculés et affichés. Ces métriques évaluent la performance du modèle, fournissant des informations sur sa capacité à faire des prédictions précises et à identifier les employés qui quittent réellement l'entreprise.

```
Entrée [20]: model.fit(X_train, y_train)

C:\Users\hp\AppData\Roaming\Python\Python39\site-packages\sklearn\linear_
model\_logistic.py:460: ConvergenceWarning: lbfgs failed to converge (sta
tus=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown i
n:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-re
gression
  n_iter_i = _check_optimize_result(

Out[20]: LogisticRegression
         LogisticRegression()
```

8) Affichage et Interprétation des Résultats

Les performances du modèle de régression logistique sont présentées à travers des métriques clés. La précision du modèle atteint 85.37%, indiquant le pourcentage d'instances correctement prédites parmi toutes les prédictions. Cependant, une analyse plus approfondie à travers la matrice de confusion révèle des aspects importants. Parmi les 253 employés qui ne quitteront pas l'entreprise (Vrais négatifs), le modèle a

correctement prédit 251 d'entre eux. En revanche, le modèle a fait des erreurs en prédisant à tort le départ de 2 employés qui ne quitteront pas réellement l'entreprise (Faux positifs). De plus, le modèle n'a pas réussi à prédire correctement le départ d'aucun employé qui quittera réellement l'entreprise (Vrais positifs), laissant 41 employés non détectés (Faux négatifs).

Le rapport de classification fournit des métriques spécifiques pour chaque classe (0 et 1). Pour la classe 0 (employés ne quittant pas l'entreprise), la précision élevée (86%) indique une capacité du modèle à identifier correctement ceux qui resteront. Cependant, le recall encore plus élevé (99%) suggère qu'il manque ceux qui quitteront réellement. Pour la classe 1 (employés quittant l'entreprise), la précision de 0% signifie que le modèle n'a correctement prédit aucun départ, tandis que le recall de 0% montre son incapacité à capturer ceux qui ont réellement quitté.

Ces résultats soulignent la nécessité d'une analyse plus approfondie et d'éventuelles améliorations du modèle pour mieux anticiper l'attrition des employés.

Affichage des résultats

Entrée [23]:
```
print(f"Précision du modèle : {accuracy:.2%}\n")
print("Matrice de confusion :\n", conf_matrix)
print("\nRapport de classification :\n", classification_report_str)
```

```
Précision du modèle : 85.37%

Matrice de confusion :
 [[251   2]
 [ 41   0]]

Rapport de classification :
              precision    recall  f1-score   support

           0       0.86      0.99      0.92       253
           1       0.00      0.00      0.00        41

    accuracy                           0.85       294
   macro avg       0.43      0.50      0.46       294
weighted avg       0.74      0.85      0.79       294
```

9) Création de Tableaux et Visualisation Finale

Après avoir réalisé les étapes cruciales de manipulation, traitement, et modélisation des données, la création de tableaux récapitulatifs s'avère nécessaire pour synthétiser les résultats obtenus. Ces tableaux peuvent inclure des statistiques descriptives, les performances du modèle, et d'autres métriques pertinentes. En parallèle, pour une compréhension holistique du projet, une visualisation finale est réalisée. Cette dernière rassemble les insights tirés de l'analyse des données, du traitement, et de la

modélisation, offrant ainsi une représentation graphique claire des tendances, des relations, et des prédictions du modèle. Cette étape finale permet de présenter de manière concise les résultats du projet, facilitant la communication des conclusions aux parties intéressées.