

Communauté Économique et Monétaire de l'Afrique Centrale

(C.E.M.A.C)



Institut Sous-régional de Statistique et d'Economie Appliquée

(ISSEA)

Organisation Internationale

B.P. 294 Yaoundé, (Cameroun)

Tel : (237) 222 22 01 34 Fax : (237) 222 22 95 21 E-mail: isseacemac@yahoo.fr

Projet Machine Learning

Thème :

**PREDICTION DES FINS DE CONTRAT DE
TRAVAIL EN ENTREPRISE : UNE APPROCHE
PAR MACHINE LEARNING**

Proposé par le groupe 14 :

AWADJOU DEUFACK

Rodrigue Pavel,

KENGNE Bienvenu Landry,

NGONO MVOGO Franck,

NJIFON MOUNCHINGAM

Seidou Ayatou

Élèves Ingénieurs Statisticiens

Économistes en 2^{ème} année

Novembre 2024

✠ Résumé ✠

Cette étude visait à développer un modèle de prédiction de fin de contrat de travail en utilisant l'apprentissage automatique de machine Learning afin de voir dans quelle mesure fidéliser les employés à l'entreprise "StrategieConsulting". Il est intéressant de noter que les résultats qu'on mentionnera plus bas seront d'utilité capitale pour l'entreprise dans la mesure où ils serviront de levier à ladite structure d'éviter non seulement la perte de la main d'œuvre qualifiée (départ de l'employé) et aussi d'éviter les coûts liés à la formation des nouveaux employés en cas de démission volontaire ou involontaire d'un travailleur. Ainsi l'étude a été réalisée sur un ensemble de données qui comprend des informations sur 1470 employés actuels et sur ceux qui ont déjà quitté leur emploi, avec près de 35 éléments d'informations précieux sur l'histoire et le profil des employés de ladite entreprise américaine. Cette dernière combine de nombreux facteurs : sociaux, culturels, financiers, professionnels et relationnels. Trois algorithmes d'apprentissage automatique différents ont été utilisés associés à une interface simple et pratique pour un usage de l'entreprise afin de prédire l'attrition d'un employé en entrant juste ses informations. Les résultats de nos analyses montrent que 83,9% des employés sont encore dans l'entreprise tandis que 16,1% ont quitté l'entreprise. La majorité des employés de cette entreprise américaine sont des hommes et voyagent rarement. Plus de la moitié de ces employés travaillent dans le département de recherche et développement. La moitié de ces employés est âgée de 35 ans. De plus, l'analyse des résultats expérimentaux révèle que le Random Forest démontre les meilleures capacités pour prédire l'attrition des employés quel que soit la méthode de resampling utilisée. La meilleure précision de prédiction était de 85% selon la méthode utilisée, ce qui est considéré comme une bonne précision. Les caractéristiques qui déterminent le plus l'attrition sont entre autres : l'âge, la distance entre la maison et le lieu de travail, le revenu mensuel, les voyages, l'évaluation de performance, l'ancienneté, les honoraires...

✠ INTRODUCTION ✠

Contexte

Les entreprises opèrent aujourd’hui dans un environnement économique hyper-concurrentiel, marqué par la mondialisation, l’innovation rapide, la fragmentation des marchés et une complexité croissante des facteurs politiques, sociaux et environnementaux. Ces transformations engendrent à la fois des opportunités et des défis majeurs, accentuant l’incertitude et la nécessité de réactivité. L’exigence de solutions personnalisées pour les clients et les employés, couplée à une concurrence accrue pour les talents, impose aux organisations une adaptation constante.

Sur le marché du travail, le recrutement et la fidélisation de ressources humaines hautement qualifiées sont devenus des enjeux cruciaux. La réduction de l’absentéisme et des démissions, ainsi que la motivation des employés à œuvrer pour les objectifs de l’entreprise, sont des facteurs clés de performance. Dans ce contexte, la gestion des ressources humaines (GRH) se positionne comme un levier stratégique, permettant d’aligner les objectifs organisationnels avec les aspirations individuelles des employés.

La ressource humaine représente la véritable richesse d’une organisation. L’efficacité d’une entreprise repose avant tout sur le potentiel de ses employés : leur enthousiasme, leurs compétences et leur créativité. Face à la « guerre des talents », l’optimisation des processus RH est devenue une nécessité. Ce projet propose d’explorer comment les techniques de *machine learning* peuvent contribuer à améliorer la prise de décision en GRH, en apportant des solutions prédictives et personnalisées pour optimiser le recrutement, la fidélisation et la gestion des performances. L’objectif est de développer des modèles capables d’anticiper les besoins en ressources humaines, d’identifier les profils les plus performants et de proposer des actions de fidélisation ciblées, contribuant ainsi à la croissance et à la pérennité de l’entreprise.

Problématique

Face au désir des employeurs de fidéliser leurs employés, ils se posent les questions suivantes : qu’est qui peut expliquer le départ d’un employé ? Peut-on savoir lorsqu’on employé à l’intention

de démissionner ? Comment peut-on empêcher la démission d'un employé qu'on trouve important ?

Objectifs spécifiques

Cette étude a pour objectifs de prédire la démission d'un employé.

De façon spécifiques, il s'agira de :

- . De déterminer les caractéristiques d'un employé qui a l'intention de démissionner
- . Déterminer les éléments pouvant expliquer la démission d'un employé.

Intérêt de l'étude

Cette étude permettra aux employeurs de mettre sur pieds des mesures pour éviter la démission de clients, de mettre sur pied un environnement de travail favorable, d'éviter de dépenser chaque fois pour les nouveaux recrutements.

0.1 Contexte théorique et revue de littérature

Selon Herzberg (2003), la satisfaction est le résultat de facteurs motivants intrinsèques tels que la reconnaissance, les opportunités de croissance professionnelle et un sentiment positif à l'égard de l'organisation (Herzberg, 2003). Les facteurs extrinsèques contribuant à éviter l'insatisfaction comprennent une direction et un encadrement efficaces, la satisfaction salariale et les avantages sociaux, ainsi que de bonnes relations avec les collègues. Selon la théorie des deux facteurs, en satisfaisant les facteurs extrinsèques, les employés peuvent se sentir neutres, mais pas extra satisfaits (Herzberg, 1959). Si les besoins des facteurs extrinsèques sont satisfaits, les employés peuvent être motivés et, par conséquent, satisfaits par des facteurs intrinsèques. De plus, les raisons du taux de rotation du personnel (attrition) sont principalement liées à leur motivation au travail et aux mesures de satisfaction (Pratt, 2021). Les employés satisfaits auront moins tendance à décider de quitter l'entreprise (Barriball et Farrell., 2007)(Coomber et Barriball, 2007. Les mesures de satisfaction sont également liées à la performance. Par conséquent, Les employés plus satisfaits montrent de meilleures performances (Cameron, 2011).

Dans des études précédentes, la prédiction de la fidélisation du personnel a été réalisée en utilisant différents algorithmes. Les recommandations incluent l'Arbre de décision, les Classificateurs et la Régression logistique, les Forêts aléatoire. La raison de ces différentes recommandations peut être liée à l'ensemble de données utilisé, aux spécifications des objectifs de recherche et au volume de

données disponibles. Pour la présente recherche, nous avons choisi de tester les performances de deux algorithmes : Forêt aléatoire, le KNN et SVM (Support Vector Machine).

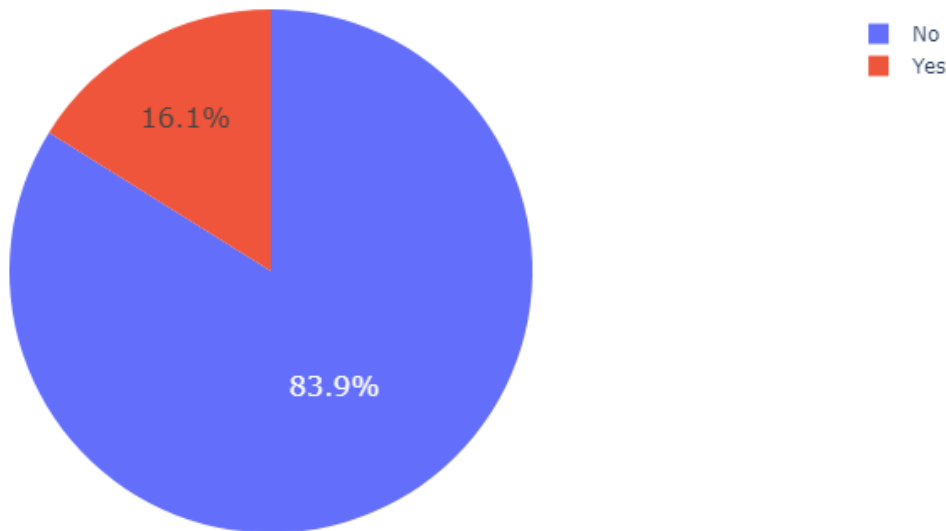
0.2 Description des employés de l'entreprise

Notre étude porte sur un ensemble de données concernant 1470 employés. Après vérification, nous avons constaté l'absence de données dupliquées. Seules quelques informations manquaient concernant la distance domicile-travail pour une petite partie des employés. Nous avons comblé ces manques en faisant une imputation par la médiane du fait d'une distribution étallée à droite des données sur la distance au lieu de travail.

Sur les 1470 employés interrogés, 83,9 % sont encore dans l'entreprise tandis que 16,1 % ont quitté l'entreprise, 71 % des employés de cette entreprise voyagent rarement, 18,8 % voyagent fréquemment et 10,2 % très rarement. En outre, plus de la moitié des employés travaillent dans le département de la recherche et du développement (soit 65,4 %), 31 % des employés ont fait des études en médecine, 41,2 % ont des études en science de la vie ; en ce qui concerne le sexe, les hommes représentent 60 % et les femmes 40% ; un peu moins de la moitié de cette population est mariée, 32% sont célibataires et 22% sont divorcés. En outre, Seulement 28,3% de notre population ont l'habitude de faire des heures supplémentaires. Par ailleurs, 31,2% des employés ont un niveau de satisfaction par rapport à leur emploi très élevé, pour 19,7% est faible. Il en ressort également que 50% des employés sont âgés de moins de 36 ans et il n'y a pas d'employés qui a moins de 18 ans. En outre, 50% des employés sont à moins de 7 unités de distance (puisque nous ne disposons pas d'information) et l'employé qui vit le plus loin de l'entreprise est à 29 unités de distance de là ; 50% des employés ont un revenu inférieur à 4919 unité monétaire, le salaire le plus bas est de 1009 et le plus élevées est de 19999 Unités monétaires.

Pendant l'analyse exploratoire de nos données, nous avons visualisé notre variables cible Attrition et nous remarquons que sur les 1470 employés, seulement 237 c'est-à-dire 16,12% des employés ont eu a quittés l'entreprise.

FIGURE 1 – Répartition des employés suivant l'Attrition



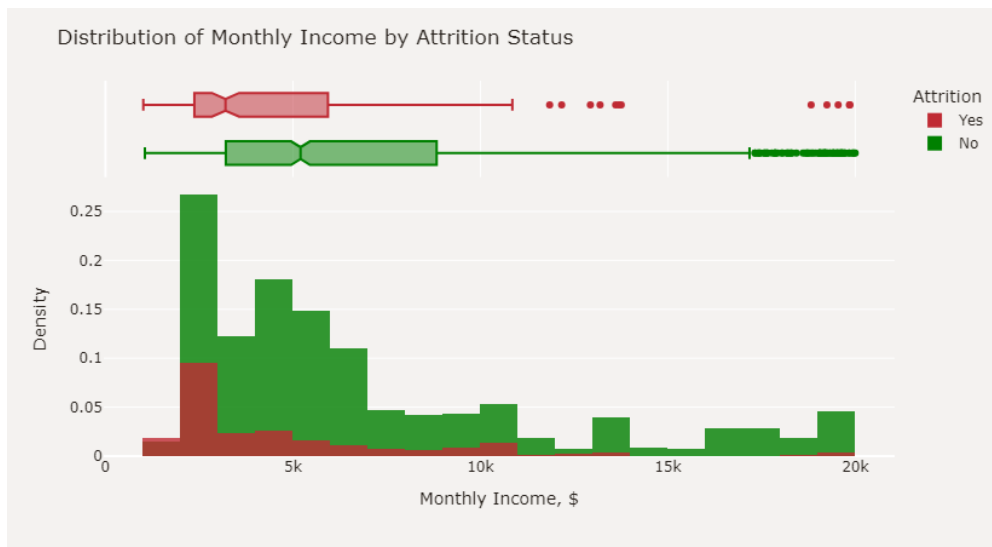
Source: Auteur, Sortie python

0.3 Relation entre les variables et la variable d'intérêt (Attrition)

L'analyse exploratoire révèle plusieurs facteurs potentiellement liés à un taux de départ plus élevé. Premièrement, les employés du département de recherche et développement semblent sur-représentés parmi les départs, de même que les hommes. De plus, une faible satisfaction professionnelle et une faible satisfaction concernant l'environnement de travail sont corrélées à un taux de départ plus important; paradoxalement, une forte satisfaction relationnelle semble également être un facteur. En outre, les techniciens de laboratoire et les chercheurs scientifiques affichent des taux de départ plus élevés, tout comme les employés ayant une formation en sciences de la vie. Il est intéressant de noter que les employés effectuant rarement des voyages d'affaires et ceux ayant une excellente évaluation de performance présentent aussi un taux de départ plus important. Egalement, on note une influence du revenu mensuel dans le comportement des employés. Ces observations suggèrent des liens complexes entre différents facteurs et la décision de quitter l'entreprise, nécessitant une analyse plus approfondie par des techniques de modélisation prédictive

pour confirmer ces relations et identifier les facteurs les plus influents.

FIGURE 2 – Distribution des revenus par attrition



Source: Auteur, Sortie python

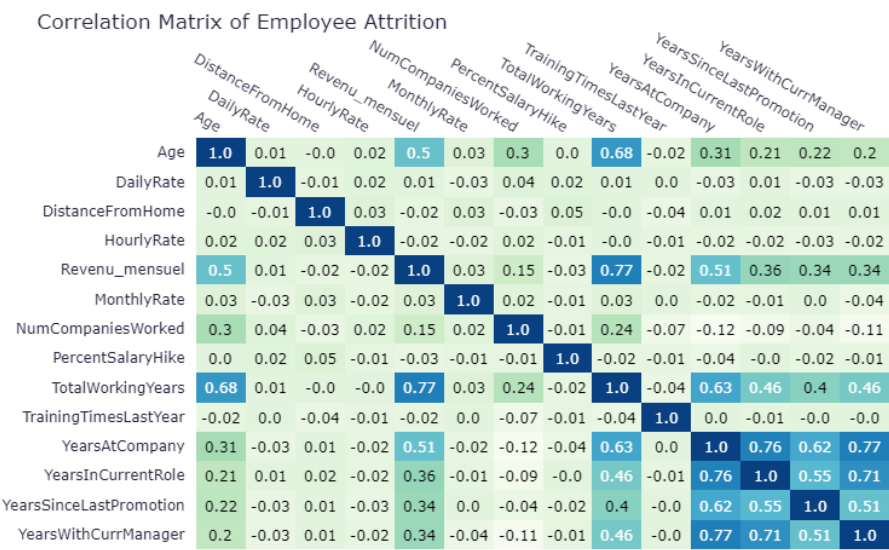
0.4 Préparation des données : avant modélisation

L'analyse exploratoire identifie plusieurs facteurs associés à un taux de départ plus élevé. Premièrement, le département d'affectation joue un rôle significatif, avec un taux de départ plus important observé au sein du département de recherche et développement. De plus, le genre influence les départs, les hommes étant plus susceptibles de quitter l'entreprise. Ensuite, la satisfaction au travail, qu'elle soit faible ou élevée, semble corrélée aux départs. Enfin, d'autres facteurs comme la fréquence des voyages d'affaires, le poste occupé (techniciens de laboratoire et chercheurs scientifiques), la formation en sciences de la vie et les évaluations de performance semblent également contribuer aux départs. En résumé, ces observations suggèrent la nécessité d'une analyse plus approfondie pour mieux comprendre l'interaction de ces facteurs et prédire les départs.

Nous avons supposé que la liaison est forte lorsque les coefficients de corrélation est supérieure à 0,7 et d'après la matrice de corrélation, on remarque que : Revenu-mensuel TotalWorkingYears, YearsAtCompany et YearsInCurrentRole, YearsAtCompany et YearsWithCurrManager et YearsInCurrentRole et YearsWithCurrManager sont fortement corrélés.

Grâce à des tests d'Anova et de Khi-deux plusieurs variables ont été supprimées dans notre base.

FIGURE 3 – Répartition des employés suivant l’Attrition



Source: Auteur, Sortie python

En effet, le test de khi-2 est venu confirmer ou pas certains de nos soupçons. On peut dire que la variable attrition est faiblement liée aux variables suivantes : Voyage-affaires, Department, Environnement de Satisfaction, Implication-dans-emploi, Satisfaction-travail, variable WorkLifeBalance, variable EducationField et enfin État-civil. Tout ceci a été fait dans le but de supprimer certaines variables.

Après tout ceci vient le traitement des valeurs aberrantes. Ensuite, la variable catégorielle attrition a été convertie en champs binaires 0 pour No et 1 pour Yes. En outre, les variables catégorielles hors attrition ont été pour dummeriser. Notre base est donc prête pour le modèle.

0.5 Modélisation et prédiction de la fidélité des employés

Les entreprises peuvent bénéficier de l'utilisation de modèles de prédiction et obtenir une grande précision et prédictions basées sur leurs données. Le modèle peut aider à mettre au jour les facteurs qui contribuent le plus aux résultats, ceux qui ont plus d'effet sur l'ensemble de données. Dans le cadre de cette étude, le modèle de prédiction est à la recherche des facteurs qui influent sur l'attrition et le comportement des employés à l'aide d'explications de prédictions. Le

système comporte deux phases : la phase d'entraînement et la phase de test. Et plusieurs connaissances peuvent être utilisées pour le développement des données d'entraînements.

Les relations entre les caractéristiques sont définies en fonction des données historiques. Ensuite, les algorithmes d'apprentissage automatique traitent les données. Au cours de cette phase, nous utilisons 80 % de l'ensemble de données. Dans la phase d'application, les 20 % restants des données sont utilisés pour vérifier la précision de l'algorithme. Un modèle prédictif est appliqué pour tester si ces employés quitteraient leur organisation. En utilisant ce modèle, les facteurs à fort impact peuvent être identifiés pour aider les organisations à concentrer leurs stratégies et leurs décisions sur les problèmes les plus pertinents

Avant d'appliquer un modèle de prédiction sur les données, il est crucial de rééquilibrer les classes de la variable cible à cause du fort déséquilibre entre les attritions positives et les attritions négatives. Ce qui nous permettra d'éviter le biais dans le modèle qui pourrait réduire sa performance. Pour ce faire il existe plusieurs méthodes parmi lesquelles :

- Random Oversampling : Cette méthode consiste à dupliquer aléatoirement des exemples de la classe minoritaire jusqu'à ce que les classes soient équilibrées. Comme avantage, cette méthode est simple à implémenter, mais l'inconvénient est qu'elle peut conduire à un surapprentissage si les données sont peu nombreux.
- Random undersampling : Elle consistant à réduire la taille de la classe majoritaire, offre des avantages en termes de réduction du temps de calcul, de simplification du modèle, et de diminution du risque de surapprentissage sur la classe majoritaire, conduisant potentiellement à une meilleure généralisation. Cependant, il présente l'inconvénient majeur d'une perte d'information potentiellement importante, pouvant nuire à la performance du modèle et introduire un biais si mal effectué. Le choix du sous-échantillonnage doit donc être soigneusement considéré en fonction du jeu de données et des risques liés à la perte d'information.
- Les méthodes hybrides : Elles combinent sous-échantillonnage et sur-échantillonnage pour obtenir un meilleur équilibre des classes, réduire le temps de calcul et améliorer la performance du modèle en atténuant les inconvénients de chaque approche seule.

Dans ce devoir nous allons utiliser la méthode de "Combinaison de sous-échantillonnage aléatoire et de SMOTE" qui est une méthode hybride : Elle consiste à réduire la taille de la classe majoritaire avec un sous-échantillonnage aléatoire, puis appliquer SMOTE pour augmenter la taille

de la classe minoritaire. Cela permet de réduire le temps de calcul et d’éviter un surapprentissage excessif.

0.6 Présentation des Résultats

0.6.1. Modélisation sur les données de base

Ici il s’agit de lancer quelques modèles de prédiction sur les données sans au préalable rééquilibrer la base. Pour ce faire les deux modèles utilisés sont ici, le KNN et le RandomForest. Les résultats sont les suivants :

Nous présentons ci-dessous une analyse comparative des performances de trois modèles d’apprentissage automatique : Random Forest, KNN et SVM, appliqués à un jeu de données initialement déséquilibré, puis rééquilibré à l’aide de la technique SMOTE. L’analyse porte sur quatre métriques : Accuracy, Recall, Précision et F1-score. Nous examinons également les limitations des modèles et les stratégies d’amélioration mises en œuvre.

TABLEAU 1 – Comparaison des performances des modèles RandomForest et KNN avec données de base

| Modèle | Accuracy | Recall | Precision | F1-score |
|--------------|----------|--------|-----------|----------|
| RandomForest | 0,82 | 0,82 | 0,83 | 0,82 |
| KNN | 0,73 | 0,73 | 0,78 | 0,75 |

Comme le montre le Tableau 1, avec les données déséquilibrées, le modèle Random Forest surpasse légèrement le modèle KNN sur toutes les métriques. Cependant, même pour le meilleur modèle, les performances restent modestes.

Après rééquilibrage des données à l’aide de SMOTE (Tableau 2)et normalisation des variables quantives, nous observons une amélioration significative des performances de tous les modèles. L’Accuracy, la Précision, le Recall et le F1-score sont sensiblement plus élevés que dans le cas des données déséquilibrées. Cela démontre l’importance de traiter le déséquilibre des classes pour obtenir des résultats plus fiables et plus précis.

TABLEAU 2 – Comparaison des performances des modèles KNN, Random Forest et SVM, avec données de base rééquilibrées

| Modèle | Accuracy | Précision | Recall | F1-score |
|---------------|----------|-----------|----------|----------|
| KNN | 0.907618 | 0.850704 | 0.986928 | 0.913767 |
| Random Forest | 0.918963 | 0.947552 | 0.885621 | 0.915541 |
| SVM | 0.917342 | 0.973978 | 0.856209 | 0.911304 |

Sur cette base, le modèle le plus approprié pour faire notre prédiction semble être le **Random-Forest**. Pour faciliter l'usage, il sera implémenté sur un interface de prédiction en entrant juste les informations de l'employé.

Nous obtenons là des valeurs encourageantes qui montre que cette méthode rend plus efficace Random Forest et le KNN.

Discussions

En se basant sur ce qui précède, on constate que le Random forest est l'algorithme le plus efficace dans les deux cas. Ainsi, Nous l'avons utilisé pour détecter les caractéristiques les plus importantes responsables de l'attrition des employés (voir note book). En accord avec notre travail, les éléments qui influent le plus sont : pour la première méthode nous avons comme 5 variables importantes : stockoptionLevel-1, Revenu-mensuel, YearAtCompany, Age et Etat-civil-Married et pour la deuxième méthode nous avons voyage d'affaires, age, Daillyrate, Totalworkingyears et Évaluation-performance.

En comparaison avec les recherches précédentes menées par les auteurs sur un échantillon plus faible de répondants, ces résultats confirment non seulement la littérature disponible en la matière en ce qui concerne la performance d'un employé, mais également ils révèlent d'autres facteurs susceptibles d'influencer le statut de l'employé au sein de l'entreprise il s'agit entre autres de sa rémunération par unité de temps, du nombre de voyages qui lui sont accordés, et du nombre total de jours au cours desquels il est contraint de travailler dans l'année. La taille plus élevée de notre échantillon a augmenté la précision de l'algorithme, et la gamme plus large de caractéristiques incluses a apporté davantage de perspectives sur le problème de l'attrition. En utilisant trois algorithmes et un ensemble de données plus important, le modèle proposé s'est avéré à plusieurs reprises le meilleur ajustement pour calculer l'attrition des employés.

✠ CONCLUSION ✠

En conclusion de ce devoir, nous pouvons donc dire que la fidélisation des salariés a aujourd'hui pris une dimension beaucoup plus large que celle de la simple volonté de réduire le turnover. En effet, si le fait de limiter les départs volontaires dans les entreprises constitue toujours un des objectifs de l'utilisation des outils de fidélisation, ils ont également une toute autre portée. Cette autre portée, les entreprises comme les salariés l'ont bien comprise, et chacun sait en jouer. Dans cette autre dimension, les outils de fidélisation sont en effet perçus comme de véritables leviers pouvant agir directement sur les comportements et les sentiments des salariés dans le milieu professionnel. Par comportements et sentiments, nous entendons motivation, implication, satisfaction et fidélité des salariés envers leur entreprise. Avoir des employés fidèles et durables est essentiel pour la plupart des organisations. Dans ce devoir, nous avons présenté l'expérience visant à trouver l'algorithme le plus adapté pour appréhender les comportements des employés dans le contexte donné. L'étude est basée sur les recherches précédentes des auteurs et est améliorée grâce à une plus grande quantité de données. Elle complète les recherches et les conclusions précédentes des auteurs selon lesquelles le random forest est l'algorithme le plus approprié pour les ensembles de données décrivant la satisfaction et l'attrition des employés, avec une précision de 85% ou 88% selon la méthode de suréchantillonnage. Après l'entraînement des données et l'utilisation des algorithmes, les résultats se sont améliorés et plusieurs caractéristiques différentes ont été mises en évidence. Les caractéristiques les plus importantes sont voyage d'affaires, age, Daily-rate, Totalworkingyears et Évaluation-performance. Nous suggérons à l'entreprise STRATEGICS CONSULTING d'accorder davantage de voyages aux employés, d'augmenter leur taux horaire entre autres.

✠ BIBLIOGRAPHIE ✠

- Alao, D. A. (s.d.). Analyzing employee attrition using decision tree algorithms, Computing, Information Systems, Development Informatics and Allied Research Journal. 2013.
- Alao, D. A. (s.d.). Analyzing employee attrition using decision tree algorithms, Computing, Information Systems, Development Informatics and Allied Research Journal,. 2013.
- Barriball, C. e., Farrell., R. e. (2007). Impact of job satisfaction components on intent to leave and turnover for hospital-based nurses : A review of the research literature. International Journal of Nursing Studies :s. <https://doi.org/10.1016/j.ijnurstu.2006.02.004>.
- Cameron, W. e. (2011). Developing management skills. 8. Ed, Pearson, Prantice Hall.
- Herzberg. (1959). The motivation to work. . New York : John Wiley Sons.
- Herzberg, F. (2003). One More Time : How Do You Motivate Employees ? Harvard Business Review.
- Pratt, M. C. (2021). Motivation in a Business Company Using Technology-Based Communication. In Artificial Intelligence in Industry 4.0. Studies in Computational Intelligence . DOI :10.1007/978-3-030-61045-6₂.
- Gaston Yves MVENG MINKOULOU (2006). Fidélisation du personnel et performance de l'entreprise : une application au personnel d'encadrement de Guinness Cameroun s.a.