01/01/2020

Analyse et prédiction des attritions

PROJET MACHINE LEARNING

Sabrine rahhou
HITEMA

Introduction

Chaque année, de nombreuses entreprises recrutent plusieurs employés. Les entreprises investissent du temps et d'argent dans la formation de ces employés, pas seulement dans le cadre de cette initiative, mais il existe également des programmes de formation dans les entreprises pour leurs employés actuels. L'objectif de ces programmes est d'accroître l'efficacité de leurs employés.

Dans cette étude de cas nous nous intéressant à L'analyse des ressources humaines qui est un domaine de l'analyse qui consiste à appliquer des processus d'analyse au service des ressources humaines d'une organisation dans le but d'améliorer la performance des employés et d'obtenir ainsi un meilleur retour sur investissement.
L'analyse des ressources humaines ne se limite pas à la collecte de données sur l'efficacité des employés. Au lieu de cela, il vise à fournir un aperçu de chaque processus en rassemblant des données, puis en les utilisant pour prendre des décisions pertinentes sur la manière d'améliorer ces processus.

Nous nous intéressant essentiellement à l'attrition des ressources humaines qui fait référence à la perte progressive d'employés au fil du temps. En général, le taux d'attrition relativement élevé est problématique pour les entreprises. Les professionnels des ressources humaines assument souvent un rôle de premier plan dans la conception des programmes de rémunération de l'entreprise, de la culture de travail et des systèmes de motivation permettant à l'organisation de retenir les meilleurs employés.

Nous allons essayer d'étudier les facteurs qui conduisent à l'attrition des employés.

I. Préparation de la base de données

A. Chargement de la base

Tout d'abord nous allons d'abord créer un nouveau dossier que nous allons appeler « projet python » ou nous allons placer notre base de donnée.
Sur jupyter , nous commençons notre premier pas avec python. Il faut charger les différentes librairies et, éventuellement, les configurer selon nos attentes.

Après cette étape, il faut spécifier Le chemin d'accès au fichier puis importer notre fichier csv. Par ailleurs j'ai renommé le fichier principale csv en « attrition » pour plus de faciliter lors de l'insertion.

B. Les propriétés de notre base

Le type DataFrame est bien reconnu. Voyons maintenant l'architecture de la structure de notre base.
Via la fonction shape, nous pouvons voir que le nombre de lignes est de1470 ligne, 35 variables (colonnes)

```
#Nom des colonnes
Nom = attrition.columns.values
print(Nom)
```

Cette option nous donne le nom des colonnes donc le nom des variables qui sont :

```
[Age , Attrition , BusinessTravel, DailyRate , Department,

 DistanceFromHome, Education , EducationField , EmployeeCount

, EmployeeNumber , EnvironmentSatisfaction, Gender HourlyRate

 JobInvolvement, JobLevel, JobRole , JobSatisfaction

MaritalStatus , MonthlyIncome , MonthlyRate NumCompaniesWorked

, Over18 , OverTime , PercentSalaryHike , PerformanceRating

RelationshipSatisfaction , StandardHours , StockOptionLevel,

TotalWorkingYears , TrainingTimesLastYear , WorkLifeBalance

YearsAtCompany , YearsInCurrentRole ,

YearsSinceLastPromotion , YearsWithCurrManager]
```

Explication des variables

Dans notre base nous avons des variables déjà codé en variables dichotomique, c'est variables sont :

Education : 1= niveau inférieur au collège ,2= collège ,3 =baccalauréat, 4 =Master ,5 = Doctorat

Satisfaction dans l'environnement : 1 =Faible, 2 = Moyen, 3 = élevé, 4 = très élevé

Engagement dans le travail : 1 =Faible, 2 = Moyen, 3 = élevé, 4 = très élevé

Satisfaction dans le travail : 1 =Faible, 2 = Moyen, 3 = élevé, 4 = très élevé

Performance 1 = faible, 2= bien, 3= Excellent, 4 exceptionnel

Satisfaction des relations dans l'entreprise : 1 =Faible, 2 = Moyen, 3 = élevé, 4 = très élevé

Balance vie/travail 1 =Faible, 2 = Moyen, 3 = élevé, 4 = très élevé

Type de chaque colonne

Nos variables sont soit sous forme de texte ou numérique, pour savoir le type de chacune, nous utilisons la fonction suivante

```
#type de chaque colonne
print(attrition.dtypes)
```

```
Age                          int64
Attrition                   object
BusinessTravel              object
DailyRate                    int64
Department                  object
                              ...
WorkLifeBalance              int64
YearsAtCompany               int64
YearsInCurrentRole           int64
YearsSinceLastPromotion      int64
YearsWithCurrManager         int64
Length: 35, dtype: object
```

Avec :

Pandas dtype	Python type	Usage
object	Str	Text
int64	int	Integer numbers
float64	float	Floating point numbers
bool	bool	True/False values

En tout, nous avons 34 variables comprenant à la fois les caractéristiques catégorielles et numériques. La variable qui nous intéresse est "attrition" de l'employé qui peut être un YES ou un NO.

Supprimer des variables

La base contient des variables que nous n'allons pas utiliser par la suite, nous allons alors les supprimer. Tout d'abord, nous avons utilisé la commande . colonnes pour avoir les noms exacts de chaque variable. Ensuite, nous utilisons la fonction. drop() pour les enlever.
A la fin nous représentons les résultats obtenus dans une liste, pour être sûr que l'opération de suppression des champs a été effectuée.

Information sur les données

En utilisent la fonction ''attriction.info()'' nous aurons le type de chaque variable et le nombre de colonnes non nulles.

```
Age                         0
Attrition                   0
BusinessTravel              0
DailyRate                   0
Department                  0
                           ..
WorkLifeBalance             0
YearsAtCompany              0
YearsInCurrentRole          0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
Length: 35, dtype: int64
```

Nous pouvons voir sur le Nootebook que nous n'avons aucune variable nulle ni manquante.
Notre base de données est donc prête à être utilisé.

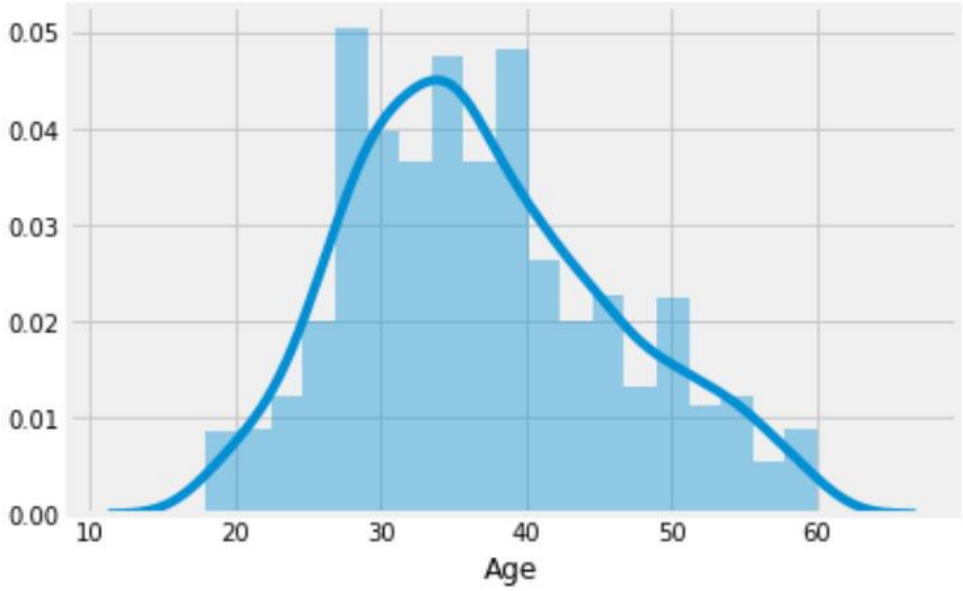
II. Statistique descriptive

C. Analyse univarié

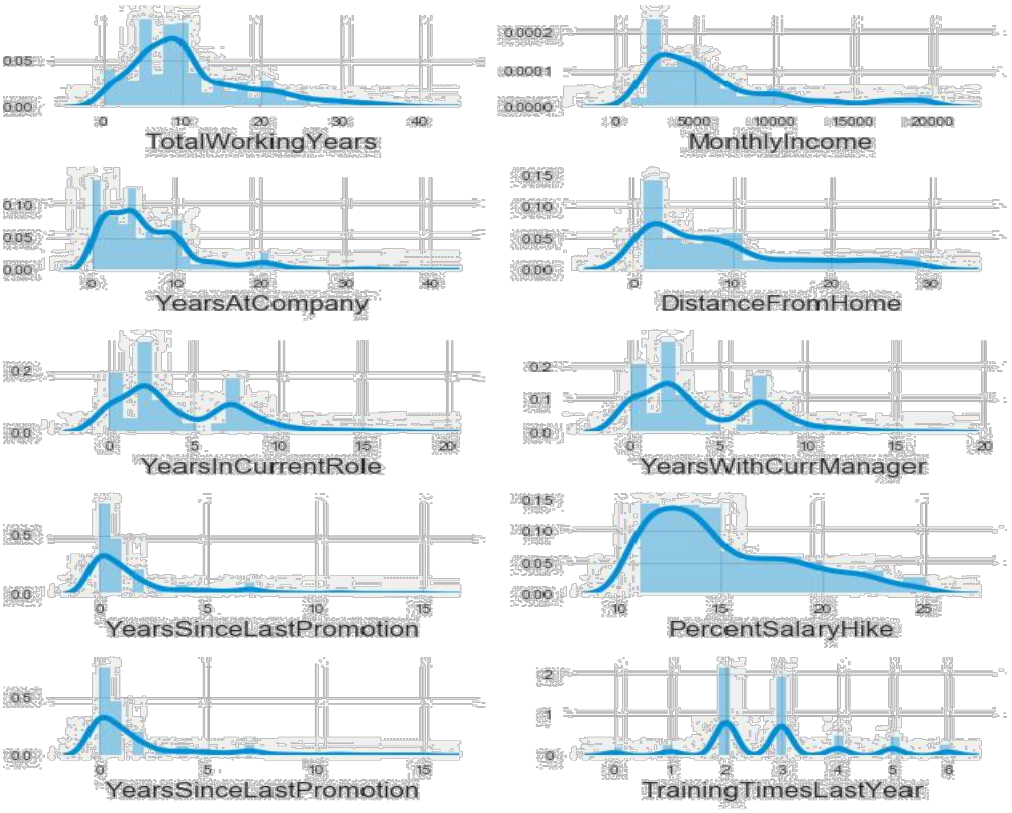
Dans un premier temps nous allons présenter les statistiques descriptives de notre base de données de façon généraliste.

Variables numériques

Analysons d'abord les différentes variables numériques. Pour ce faire, nous pouvons réellement tracer une boîte à moustaches montrant toutes les caractéristiques numériques mais vu que variables ont des échelles assez différentes, tracer une boîte à moustaches n'est pas une bonne idée. Nous pouvons tracer un kdeplot montrant la distribution de la variable. Ci-dessous, nous avons un kdeplot pour la variable 'Age'.



L'âge de l'employé est normalement réparti, la majorité se situe entre 20 et 50 ans, ce qui ne se contrarie pas avec la vie réelle, la moyenne est d'environ 35 ans.

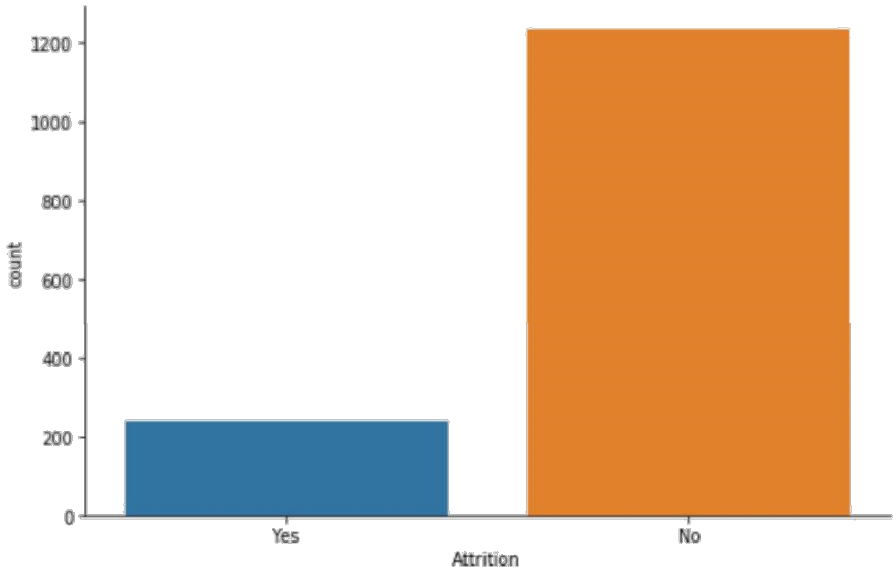


Les individus présents sur notre base présente des employés d'IBM, nous pouvons voir que La majorité des gens gagne moins de 10000 par mois, la distribution est décalée à gauche.
L'expérience professionnelle de la population est très dense jusqu'à 15 ans dans une entreprise, puis diminue rapidement, la distribution est déviée à gauche. De même pour les années au sein de l'entreprise qui montre que la majorité des employés en passé moins que 10 ans dans l'entreprise.

Variables caractéristiques

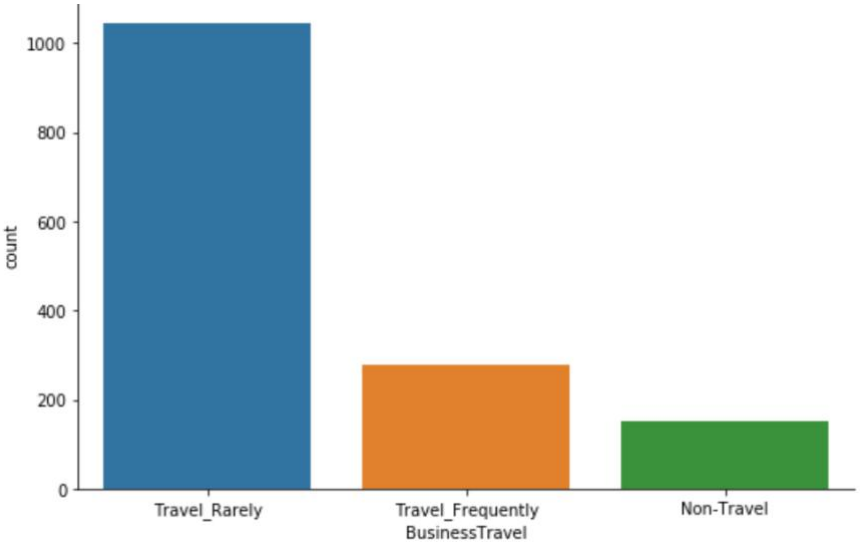
Analysons maintenant les différentes variables caractéristiques. Nous allons utiliser un diagramme de comptage pour montrer le nombre relatif d'observations de différentes catégories.

➢ Pour la variable attrition :



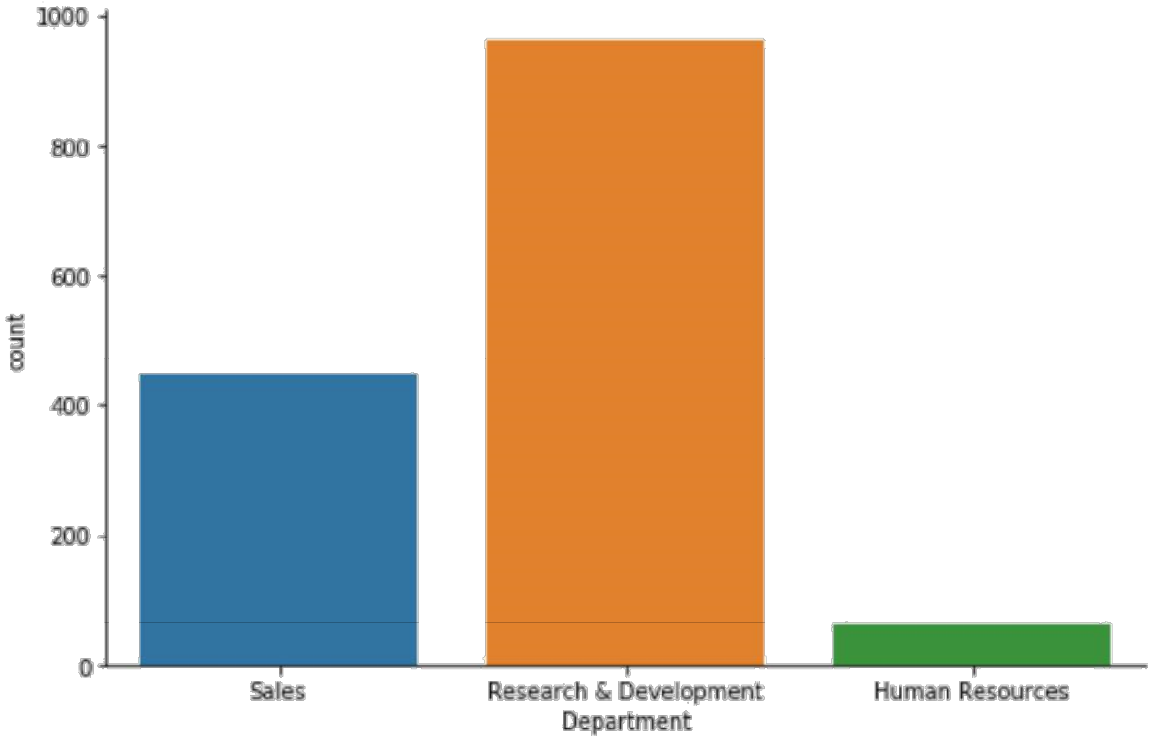
nous pouvons déjà voir que le nombre d'observations appartenant à la catégorie «Non» (1233) est bien supérieur à celui appartenant à la catégorie «Oui» (237), notre base contient plus d'employée qui n'ont pas quitter leur entreprise.
Analysons de la même manière d'autres variables catégorielle.

➢ Pour la fréquence de voyage pour le travail



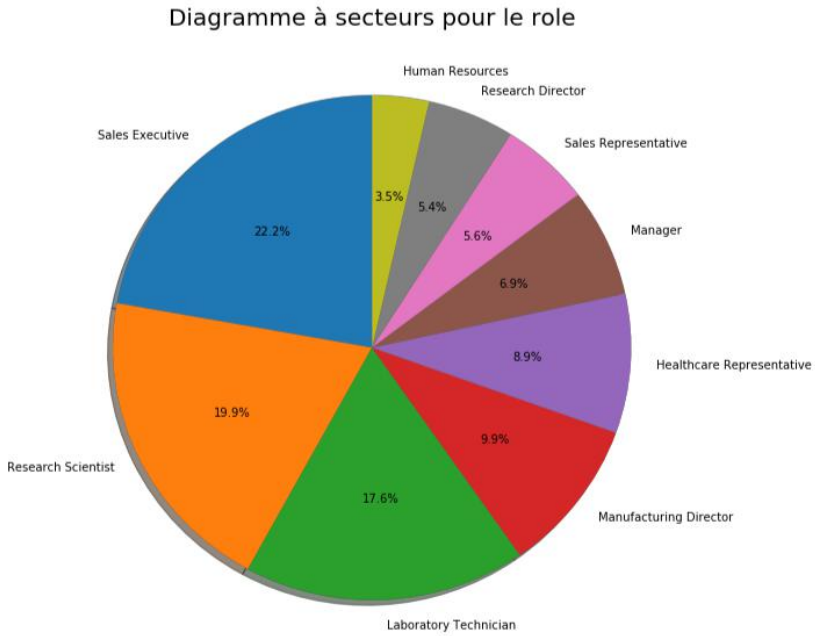
La grande majorité des employés voyage rarement, ceux sont suivis par les employés qui voyagent fréquemment. La minorité est représentée par les employés qui ne voyagent jamais.

➢ Pour le département du travail



Les grandes majorités des employés occupent des postes dans le département de recherche et développement, par la suite, les employés ont des postes dans le services des vente, la minorité occupe des postes dans le département des ressources humaines.

➢ Pour le rôle dans l'entreprise

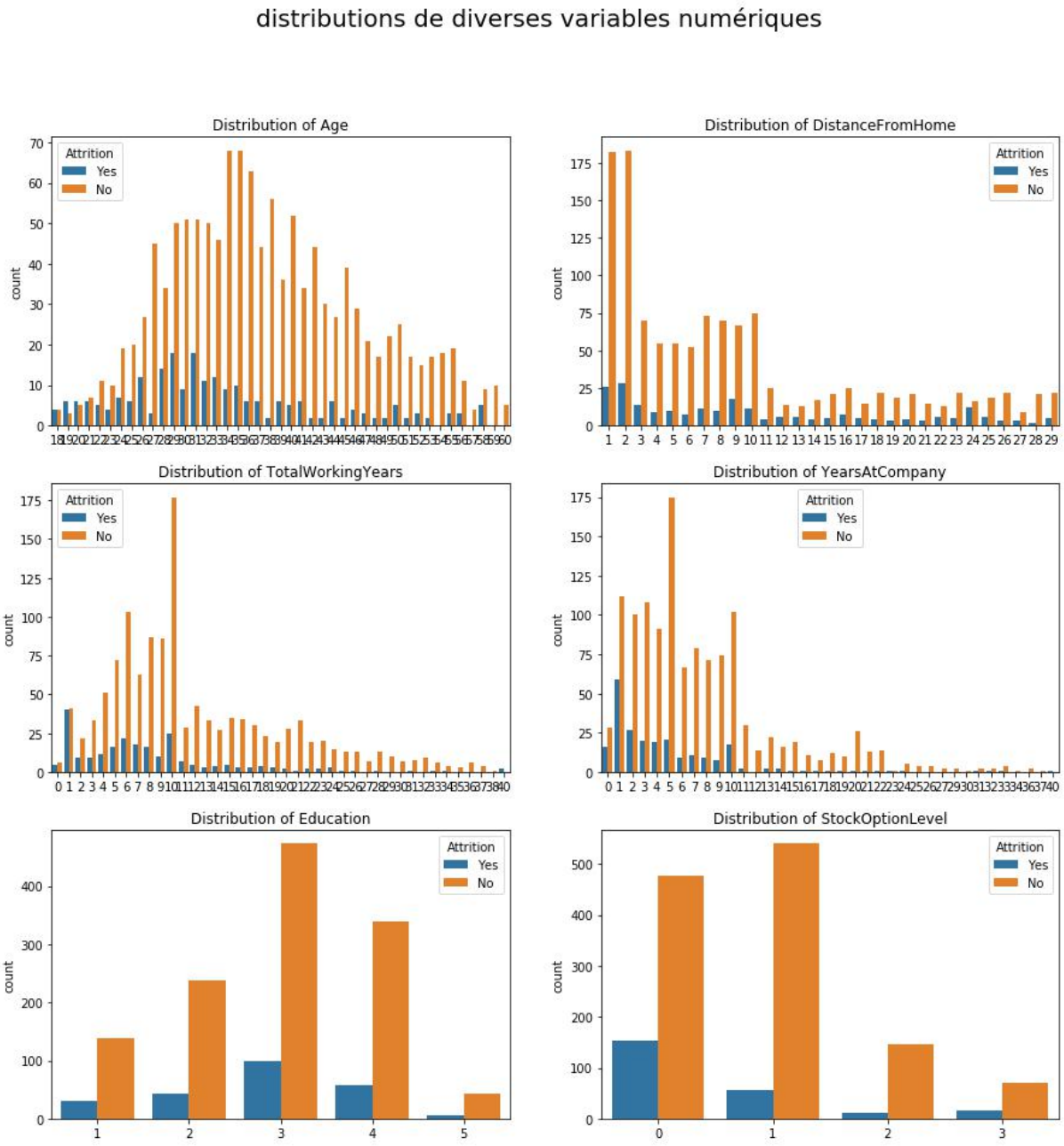


Les emplois occupés par les employés sont au majoritairement dans les postes de responsable des ventes, puis de recherche et développement, puis de technicien de laboratoire. La minorité est représentée par les employés des ressources humaines et les directeurs de recherche.

D. Statistique bivarié

Dans cette partie, nous allons croiser les variables intéressantes avec notre variable d'intérêt qui est l'attrition, afin d'avoir une première idée sur la relation entre les variables.

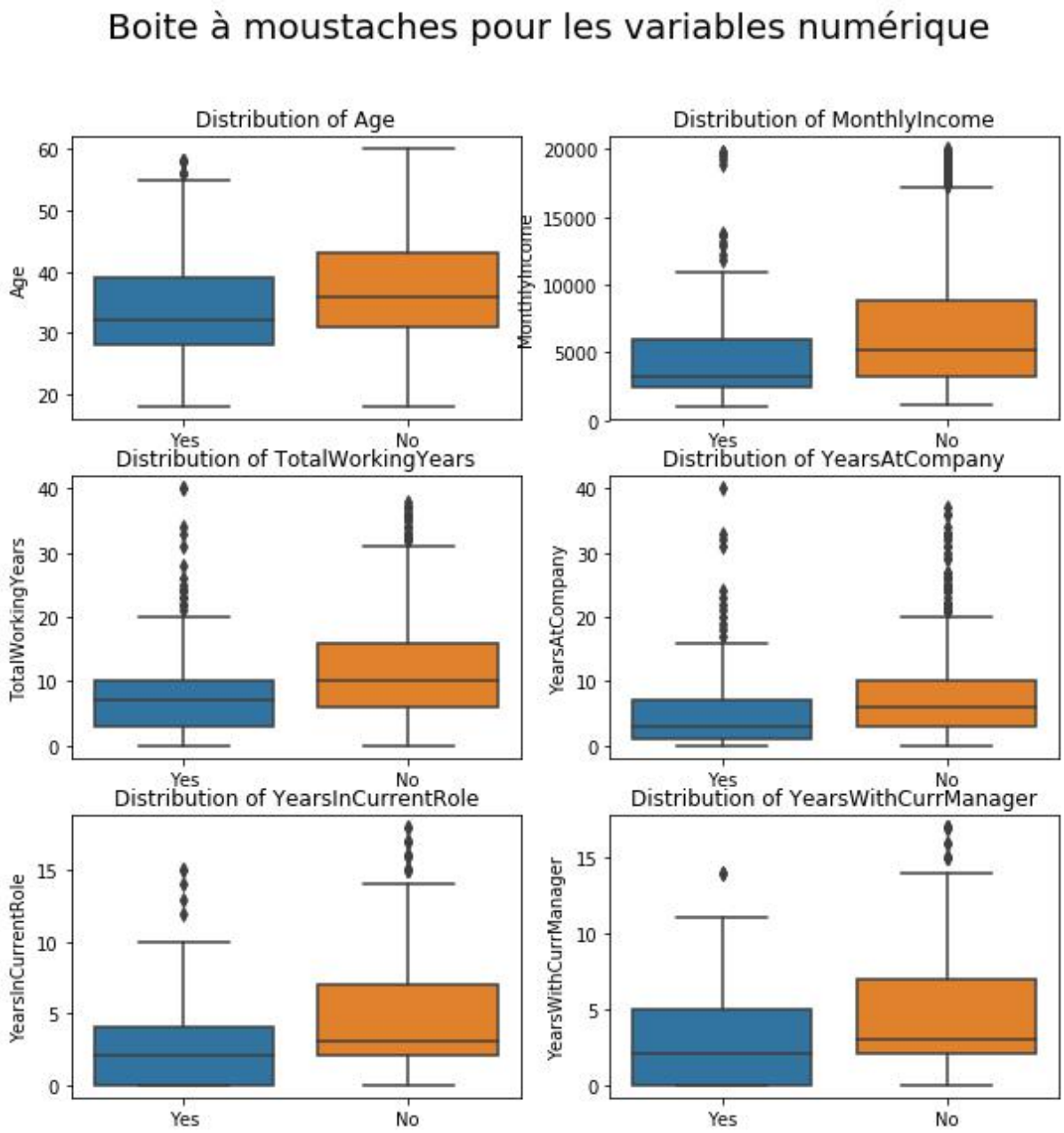
Distribution des variables numériques



Nous pouvons voir que L'attrition est élevée à l'âge de 28,29 & 30 ans ce qui est logique vu qu'un vieux n'as pas intérêt à changer de travaille par rapport à un jeune, Les employés ayant un an d'expérience sont eux qui abandonnent le plus, plus les années d'expérience augmente plus la probabilité de quitter sont travail baisse. Mais arrivé a 10 an d'expérience,

cette probabilité augmente une autre fois puis continu de baisser.

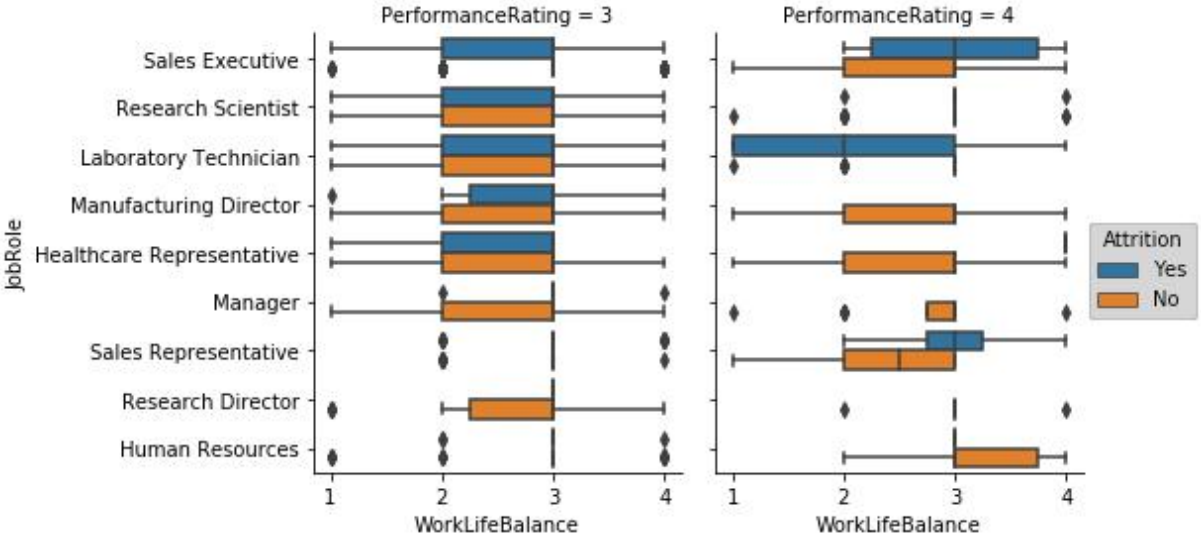
Les employés ayant un baccalauréat ou un master sont les plus à avoir quitté leur travail.
Pour la distance entre le travail et la maison, nous pouvons voir que les personnes résidents proche de leur milieu de travail ont une faible probabilité de quitter leur entreprise par rapport aux autres.
Nous pouvons faire des boites a moustaches pour les autres variables numérique



Comme nous avons déjà vu, cette distribution montre que l'attrition est plus élevée chez les employés qui ont entre les 30 à 40 ans.
L'attrition est plus élevée pour les employés ayant les faibles revenus mensuels
Les gens sont plus attirés quand ils ont moins d'expérience de travail.

Croisement de plusieurs variables

Nous allons croiser la balance ou l'équilibre travail/vie avec le rôle dans entreprise, et la performance. Le résultat se présente comme suit :



Nous pouvons voir que les directeurs des ventes (Sales Executives) , même avec une excellente performance, ont tendance à quitter leur travail et aussi les Techniciens de laboratoire avec des performances exceptionnelles mais avec un faible équilibre travail-vie sont ceux qui démissionnent le plus.

Etude de moyenne

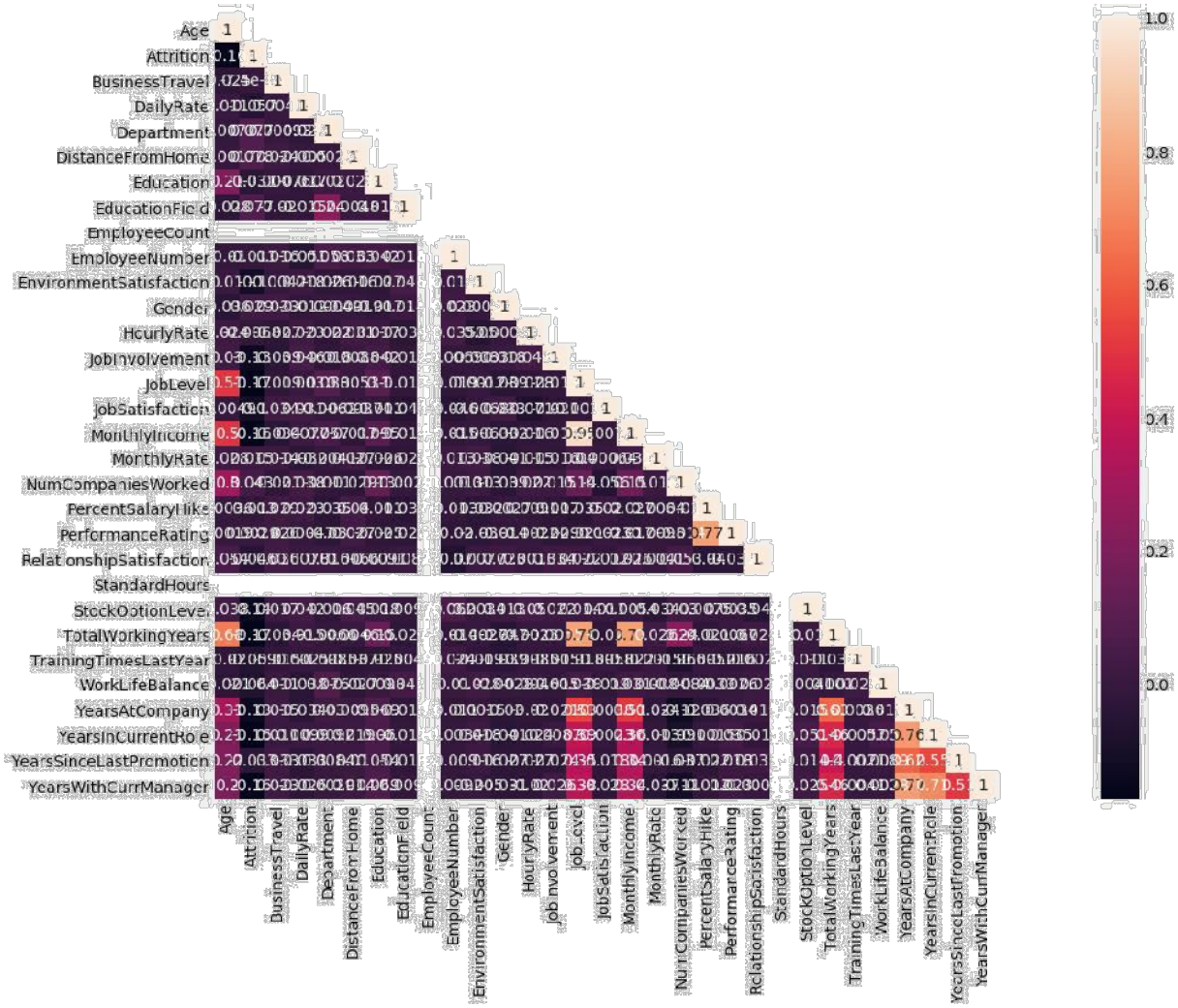
Afin de pourvoir voir comment les valeurs moyennes diffèrent pour différentes variables, nous allons groupez les données par Attrition en utilisent la fonction :

```
attriction.groupby('Attrition').mean()
```

Cette dernière montre que le niveau de satisfaction moyen des employés qui restent dans l'entreprise est supérieur à celui des employés qui partent. Les employés non satisfaits sont ceux qui quittent le plus.
Les personnes ayant le même rôle pendant plus longtemps sont restées plus longtemps dans l'entreprise, La distance de la maison a également contribué à L'attrition, la moyenne étant plus élevée pour les personnes ayant quitté.

Etude des corrélations

Nous réalisons une matrice de corrélation afin de savoir les variables les plus corréler avec notre variable d'intérêt qui est l'attrition et aussi pour voir si les variables sont corréler entre eux.



Le niveau d'emploi est étroitement lié à l'âge, MonthlyIncome (revenu mensuel) est très fortement lié à joblevel (poste),La performance est étroitement liée à la hausse du salaire en pourcentage ,le nombre total d'années de travail est étroitement lié au niveau d'emploi , Années en entreprise est liée aux années dans le rôle actuel (logique).

III. Modélisation

Pour prédire l'attrition des employés nous avons lancer 3 types de modélisation : la régression logistique, l'arbre de décision, et les réseaux de neurones
Dans un premier temps nous avons fait une partition aléatoire de nos données en :

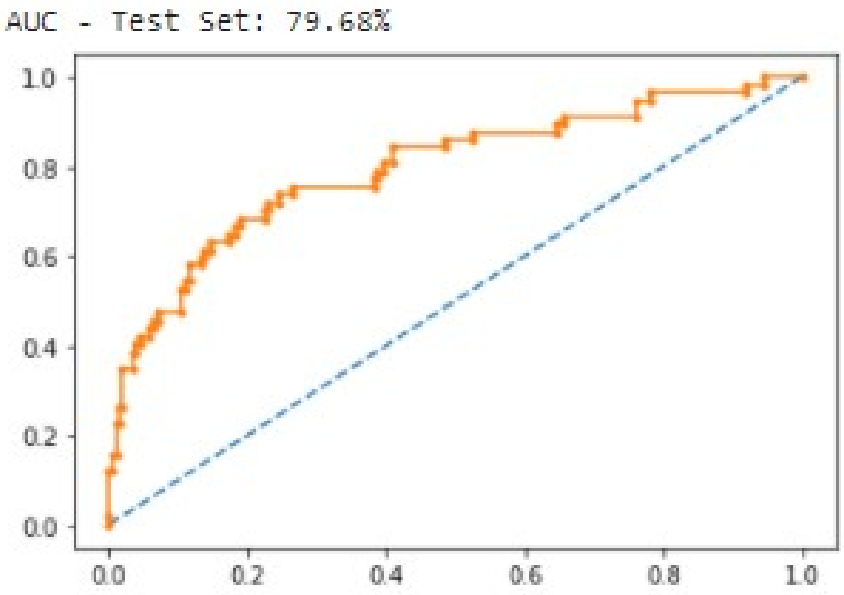
Echantillon d'apprentissage (75%)
Echantillon de test (25%).

Régression logistique

Nous avons par la suite lancé le modèle de la régression logistique, les résultats de ce dernier sont les suivant

	Attrition observer	Non-attrition observer
Attrition prédit	305 (VP)	6 (FP)
Non-attrition prédit	40 (FN)	17 (VN)

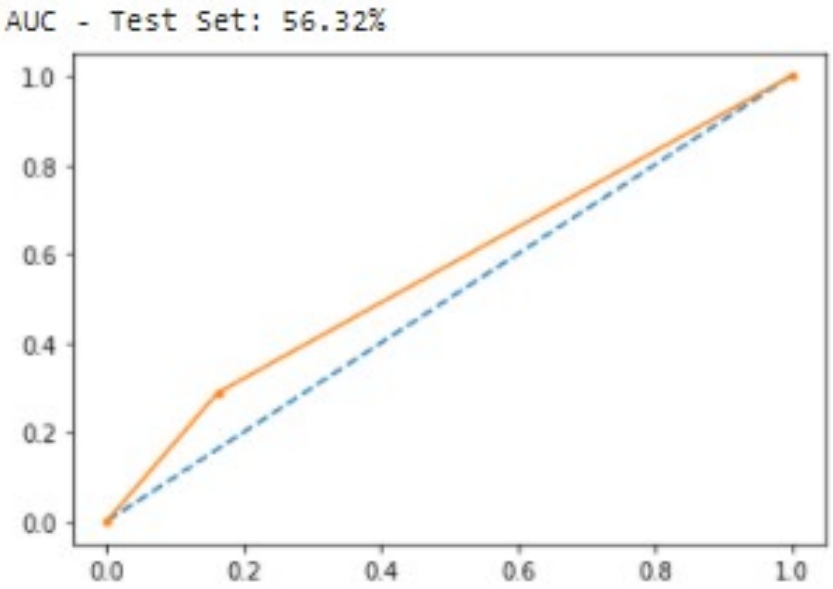
Courbe de ROC



Arbre de décision

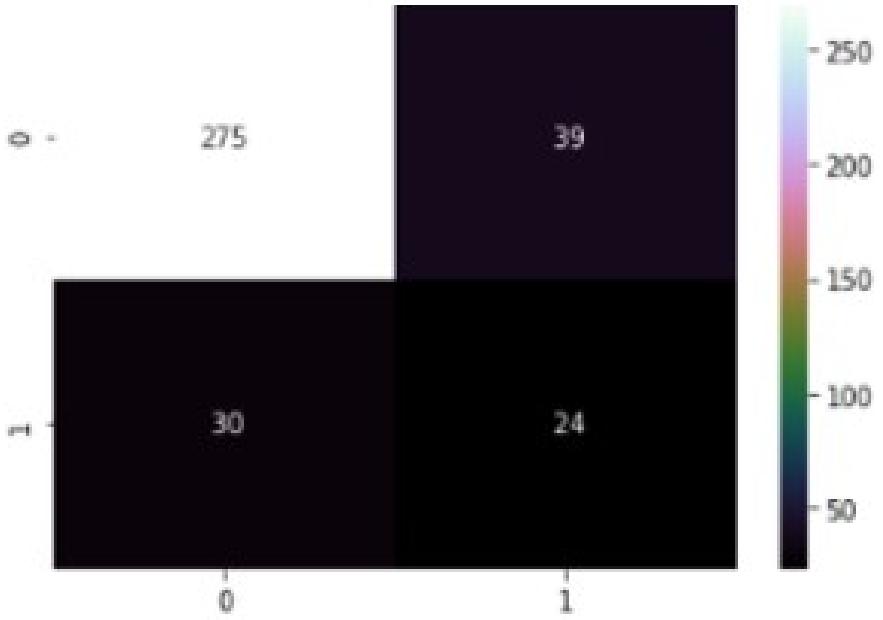
	Attrition observer	Non-attrition observer
Attrition prédit	257 (VP)	57 (FP)
Non-attrition prédit	44 (FN)	10 (VN)

Courbe de ROC

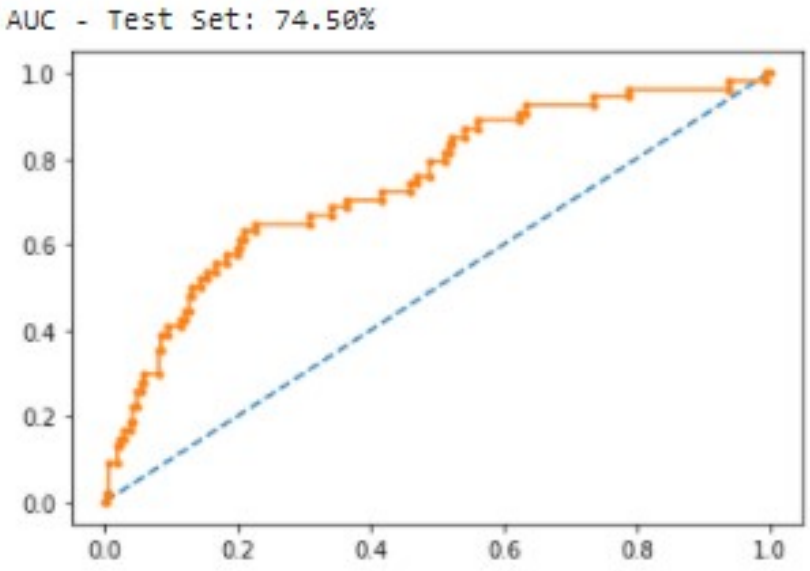


Réseaux de neurones

- Précision : 81%



Courbe de ROC



Conclusion

Durant cette étude, Nous avons étudié les facteurs qui conduisent à l'attrition des employés et donc qui posent un problème aux entreprises qui perd leur main d'œuvre au fil de temps.

L'idée que la plupart ont est le fait que la rémunération est le facteur essentiel, c'est vrai ! Mais ils existent d'autres facteurs qui poussent l'employé à quitter son travail et d'autres qui l'encourage à rester.

Grace à cette analyse, nous pouvons dire que l'expérience professionnel joue un rôle important dans l'attrition, moins l'employé à d'expériences plus il changera sont travail, de même pour l'Age, plus on est âgé moins nous avons envie de quitter notre travail.

L'emplacement de l'entreprise est aussi important, un employé qui résidents près de son travail aura moins envie de changé par rapport à un autre qui habite loin.

Après avoir réalisé les différentes statistiques descriptives et tester des différents modèles de prédiction nous pouvons conclure que le meilleur modèles qui nous permets de prédire l'attrition au sein d'une entreprise et la régression logistique avec une précision de 87%.