Université Nice Sophia Antipolis

Membre de UNIVERSITÉ CÔTE D'AZUR

Projet d'Analyse de Données
Sujet : Prédiction de l'Attrition d'Employés

- Présenté par :

  Ayoko Claudia AYIKA

  Kokou Laris EDJINEDJA

- Chargé de cours : Nicolas PASQUIER

- Année scolaire : 2020-2021

STRUCTURE DU PROJET

I. Etude et compréhension du sujet

1.1. Définition :

Attrition = réduction d'effectifs due à des départs comme la retraite ou la démission. En règle générale, attrition = perte.

1.2. Objectif : Construire un modèle de prédiction de l'attrition afin d'assurer la pérennité des compétences et savoir-faire dans l'entreprise.

Ensemble de données : Contient des informations sur :

- Employés ayant résilié leur contrat
- Employés n'ayant pas résilié leur contrat

On dispose de 1470 instances (=lignes) et chaque instance représente un employé décrit par 34 variables.

Variable à prédire : Attrition.

Variables non utilisées : EmployeeCount, Over18, StandardHours.

On dispose de deux fichiers de données :

- Data_Projet_1.csv → 1470 instances dont la classe réelle est connue.
- Data_Projet_1_New.csv → 150 instances à prédire.

Quels sont donc les points à retenir pour ce projet ?

- Générer plusieurs classifieurs et les tester (en utilisant les paramétrages afin d'optimiser le résultat).
- Minimiser le risque de ne pas prévoir l'attrition d'un employé.
- Définir un ou des critères en fonction des critères du classifieur décrit précédemment.
- Comparer les résultats des classifieurs générés selon ces critères afin d'en identifier le plus performant.
- Appliquer le classifieur sélectionné à l'ensemble de données à prédire afin de savoir si un employé est susceptible de résilier ou non son contrat.

II. Préparation des données

La fonction Read.csv nous a permis de lire les fichiers qui étaient en format csv et pour vérifier que toute les variables était présentes pour notre étude, nous avons utilisé les commandes view et str disponibles sur R .

III. Exploration des données

3.1. Vue de l'ensemble des données

A l'aide de la fonction summary, voici les statistiques générales de base du jeu de données étudié.

```
      Age        Attrition           BusinessTravel   DailyRate
 Min.   :18.00   No :1233   Non-Travel       : 150   Min.   : 102.0
 1st Qu.:30.00   Yes: 237   Travel_Frequently: 277   1st Qu.: 465.0
 Median :36.00              Travel_Rarely    :1043   Median : 802.0
 Mean   :36.92                                       Mean   : 802.5
 3rd Qu.:43.00                                       3rd Qu.:1157.0
 Max.   :60.00                                       Max.   :1499.0

                 Department  DistanceFromHome   Education
 Human Resources       : 63   Min.   : 1.000   Min.   :1.000
 Research & Development:961   1st Qu.: 2.000   1st Qu.:2.000
 Sales                 :446   Median : 7.000   Median :3.000
                              Mean   : 9.193   Mean   :2.913
                              3rd Qu.:14.000   3rd Qu.:4.000
                              Max.   :29.000   Max.   :5.000

         EducationField EmployeeCount EmployeeNumber   EnvironmentSatisfaction
 Human Resources : 27    Min.   :1     Min.   :   1.0   Min.   :1.000
 Life Sciences   :606    1st Qu.:1     1st Qu.: 491.2   1st Qu.:2.000
 Marketing       :159    Median :1     Median :1020.5   Median :3.000
 Medical         :464    Mean   :1     Mean   :1024.9   Mean   :2.722
 Other           : 82    3rd Qu.:1     3rd Qu.:1555.8   3rd Qu.:4.000
 Technical Degree:132    Max.   :1     Max.   :2068.0   Max.   :4.000

    Gender      HourlyRate     JobInvolvement    JobLevel
 Female:588   Min.   : 30.00   Min.   :1.00   Min.   :1.000
 Male  :882   1st Qu.: 48.00   1st Qu.:2.00   1st Qu.:1.000
              Median : 66.00   Median :3.00   Median :2.000
              Mean   : 65.89   Mean   :2.73   Mean   :2.064
              3rd Qu.: 83.75   3rd Qu.:3.00   3rd Qu.:3.000
              Max.   :100.00   Max.   :4.00   Max.   :5.000

                      JobRole    JobSatisfaction  MaritalStatus MonthlyIncome
 Sales Executive          :326   Min.   :1.000   Divorced:327   Min.   : 1009
 Research Scientist       :292   1st Qu.:2.000   Married :673   1st Qu.: 2911
 Laboratory Technician    :259   Median :3.000   Single  :470   Median : 4919
 Manufacturing Director   :145   Mean   :2.729                  Mean   : 6503
 Healthcare Representative:131   3rd Qu.:4.000                  3rd Qu.: 8379
 Manager                  :102   Max.   :4.000                  Max.   :19999
 (Other)                  :215
  MonthlyRate    NumCompaniesWorked Over18   OverTime   PercentSalaryHike
 Min.   : 2094   Min.   :0.000      Y:1470   No :1054   Min.   :11.00
 1st Qu.: 8047   1st Qu.:1.000               Yes: 416   1st Qu.:12.00
 Median :14236   Median :2.000                          Median :14.00
 Mean   :14313   Mean   :2.693                          Mean   :15.21
 3rd Qu.:20462   3rd Qu.:4.000                          3rd Qu.:18.00
 Max.   :26999   Max.   :9.000                          Max.   :25.00

 PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel
 Min.   :3.000     Min.   :1.000            Min.   :80    Min.   :0.0000
 1st Qu.:3.000     1st Qu.:2.000            1st Qu.:80    1st Qu.:0.0000
 Median :3.000     Median :3.000            Median :80    Median :1.0000
 Mean   :3.154     Mean   :2.712            Mean   :80    Mean   :0.7939
 3rd Qu.:3.000     3rd Qu.:4.000            3rd Qu.:80    3rd Qu.:1.0000
 Max.   :4.000     Max.   :4.000            Max.   :80    Max.   :3.0000

 TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
 Min.   : 0.00     Min.   :0.000         Min.   :1.000   Min.   : 0.000
 1st Qu.: 6.00     1st Qu.:2.000         1st Qu.:2.000   1st Qu.: 3.000
 Median :10.00     Median :3.000         Median :3.000   Median : 5.000
 Mean   :11.28     Mean   :2.799         Mean   :2.761   Mean   : 7.008
 3rd Qu.:15.00     3rd Qu.:3.000         3rd Qu.:3.000   3rd Qu.: 9.000
 Max.   :40.00     Max.   :6.000         Max.   :4.000   Max.   :40.000


 YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
 Min.   : 0.000     Min.   : 0.000          Min.   : 0.000
 1st Qu.: 2.000     1st Qu.: 0.000          1st Qu.: 2.000
 Median : 3.000     Median : 1.000          Median : 3.000
 Mean   : 4.229     Mean   : 2.188          Mean   : 4.123
 3rd Qu.: 7.000     3rd Qu.: 3.000          3rd Qu.: 7.000
 Max.   :18.000     Max.   :15.000          Max.   :17.000
```

On constate l'absence des Valeurs manquante NA.

Nous avons deux types de variables : les variables quantitatives et les variables qualitatives.

Comme variables qualitatives, nous distinguons les variables catégorielles à l'instar de notre classe, Gender ; et les variables nominales.

Pour mener à bien notre étude nous avons transformé ces variables en facteurs. Par conséquent, elles ne sont pas décrites sous R par des quartiles, mais par les effectifs de chaque modalité,

En revanche, les variables quantitatives qui sont ici discrètes et continues sont décrites par la moyenne, le minimum la médiane et les quartiles.

3.2. Analyse des variables continues par rapports à la classe attrition

Avant de commencer la modélisation, nous avons examiné l'ensemble des variables par rapport à l'ensemble à prédire.

3.2.1. Relation entre Age ,Monthly_Rate des employés ayant résiliés ou non leurs contrats et Attrition

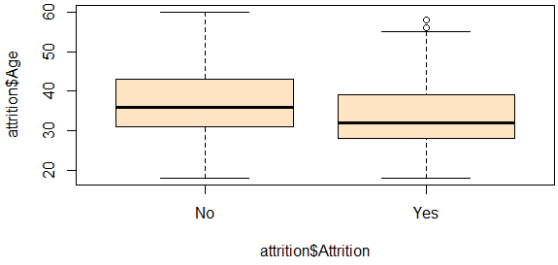

Fig1

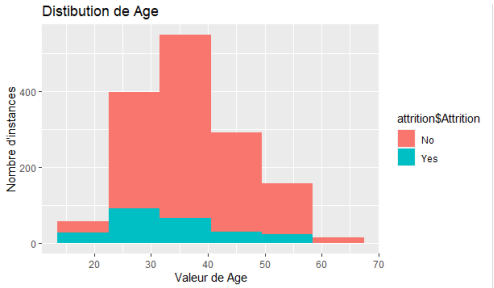

Fig2

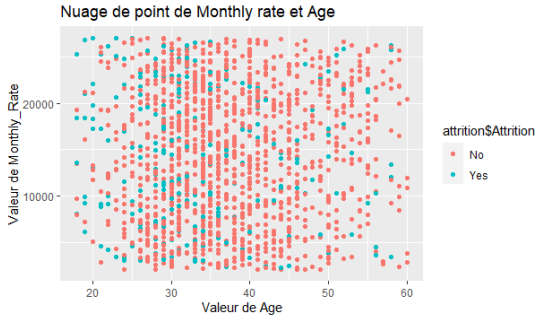

Fig3

La boite à moustache nous révèle qu'il y a plus de sujets qui gardent leurs contrats que de personnes qui partent entre l'Age minimal et l'Age médian. (fig1)

Cette hypothèse est confirmée sur la figure donnée par l'histogramme d'effectifs. (fig2)

Ensuite, en analysant le Nuage de points de Monthly Rate par rapport à l'Age et à notre classe, on voit qu'il y'a une forte corrélation entre l'âge et attrition, par contre pour la distribution de monthly rate on a une répartition plus ou moins constante.

3.2.2. Relation entre Employeecount-Attrition,StandarHours–Attrition et over18 –Attrition

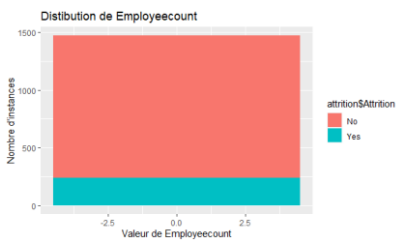

Fig4

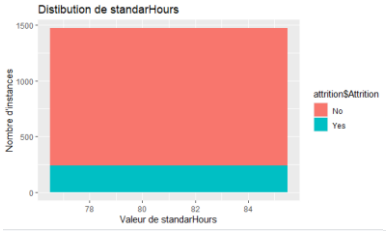

Fig5

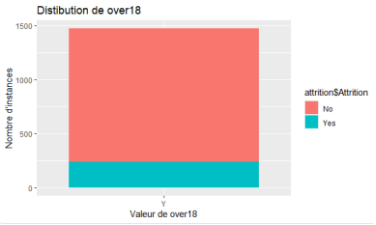

Fig6

Les diagrammes en bâtons générés par Employeecount-Attrition,StanderHours-Attrition et Over18-Attrition nous montre que la relation entre ces trois variables et attrition est statistiquement non significative.

Il n'y a pas de répartition pouvant nous aider à trouver une relation entre ces variables et l'attrition des employés. Nous pouvons les considérer comme des variables statistiquement non significatives.

Cela nous donne une possibilité de les supprimer de nos jeux de donnés pour notre étude.

3.3 . Analyse des variables continues, discrète, catégorielle et nominale par rapports à la clase attritions

3.3.1. Relation entre HourlyRate,JobRolle des employés ayant résiliés ou non leurs contrats et Attrition

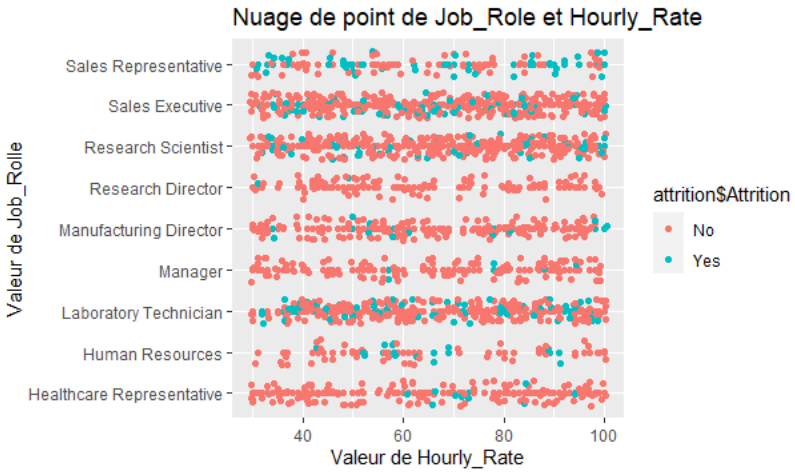

Fig7

Nous constatons une forte concentration d'attrition=non pour les employés qui ont leurs Job compris entre Research Scientist et Manager et qui ont une revenus journaliers comprise entre 40 et 100.

Cela nous montre une dépendance de ces variables et Attrition=oui ou non des employés. Donc ces données sont importantes pour notre prédiction.

3.3.2. Relation entre MonthlyRate,MaritaleStatus des employés ayant résiliés ou non leurs contrats et Attrition

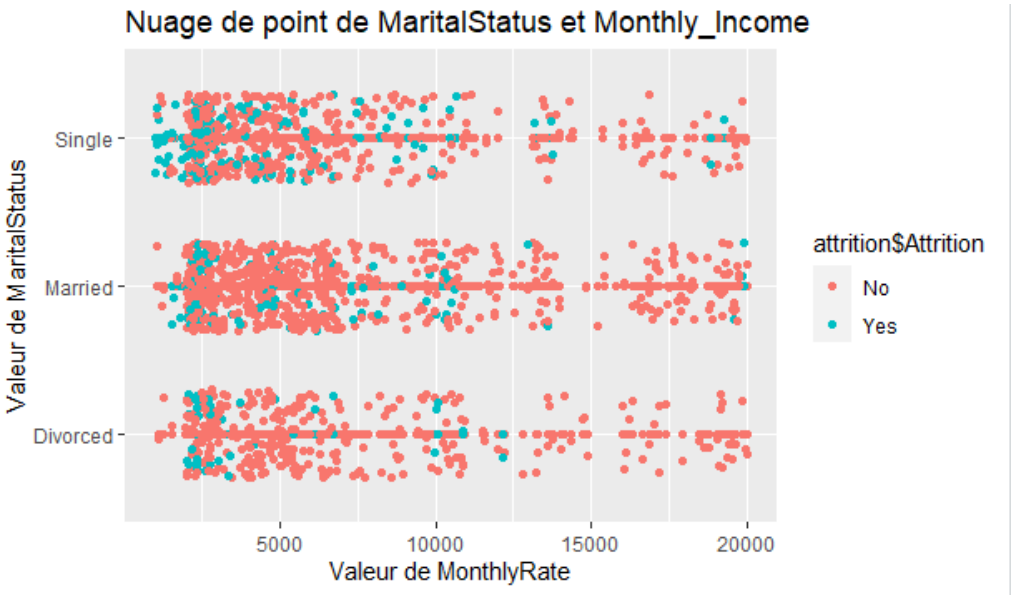

Fig8

On constate qu'ici il y a plus de départ des employés pour les revenus inférieurs à 5000, et par rapport à la situation maritale, il y a plus de départ des sujets qui sont célibataires. Nous déduisons qu'il y a une corrélation entre ces variables et Attrition oui ou non des employés.

3.3.3. Relation entre DistanceFromHome,JobSatifaction des employés ayant résiliés ou non leurs contrats et Attrition

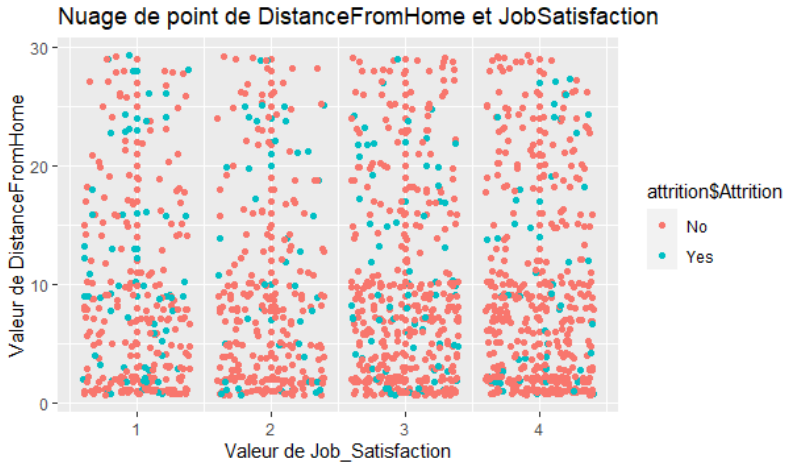

Fig9

La concentration Attrition=no est plus importante pour une distance comprise entre 0 et 10.

Si nous prenons une distance entre 0 et 10 et jobSatifaction medium on voit qu'il y a une présence plus forte d'Attrition No que de yes. Face à cette inégale répartition, on peut dire qu'il existe une corrélation avec la classe donc ces variables peuvent nous aider à faire une bonne prédiction.

IV. Définition de la méthode d'évaluation des classifieurs

Nous évaluerons les classifieurs à l'aide des critères suivants :

- **Matrice de confusion** : notre but étant de minimiser le risque de ne pas prévoir l'attrition d'un employé, la matrice de confusion nous permettra de mettre en évidence le taux de Faux Négatifs le plus faible.

- **Taux de succès global** : grâce à ce critère, nous pourront connaître le taux global de succès de chaque classifieur pour un choix plus optimal.

- **Calcul du Taux de Vrais Négatifs** : pour savoir si l'on peut se fier au classifieur pour les prédictions négatives.

- **Courbe ROC et indice AUC** : pour établir une relation entre le taux de vrais positifs et le taux de faux positifs.

V. Description de la méthode de création des données d'apprentissage et de test

L'ensemble d'apprentissage, que nous nommerons ici **attrition_EA** sera constitué des 2/3 des instances de l'ensemble de données que nous nommerons **attrition**. L'ensemble de test quant à lui sera nommé **attrition_ET** et sera constitué du dernier tiers de l'ensemble attrition.

« **attrition_EA** » sera donc constitué des 980 premières instances de **attrition** et « **attrition_ET** » sera constitué des 490 dernières.

VI. Description des configurations des classifieurs générés

6.1. Choix des classifieurs

Nous utiliserons les classifieurs suivants pour notre étude :

- Rpart()
- Random_Forest()
- Svm()
- Neural_network()
- Naive_Bayes()

6.2. Définition des paramétrages pour les classifieurs

- Pour rpart(), nous utiliserons les paramétrages : **split** = 'gini' ou 'information' et **minbucket** = 10 ou 5.

- Pour random_forest(), nous utiliserons **ntree** = 500 ou 300 et **mtry** = 3 ou 5.

- Pour nnet(), nous travaillerons avec **size** = 25 ou 50, **decay** = 0.01 ou 0.001 et **maxit** = 100 ou 300 .

- Pour svm(), on utilisera **kernel** = 'linear' , « polynomial', 'radial' ou 'sigmoid'.

- Pour naive_bayes(), ce sera **laplace** = 0 ou 20 et **usekernel** = FALSE ou TRUE .

Pour chacun de ces classifieurs, nous avons écrit une fonction afin de pouvoir tester plusieurs paramétrages et ainsi dégager le paramétrage nous donnant le meilleur résultat. Ensuite, nous avons comparé les classifieurs retenus dans chaque cas et nous avons retenu « le » plus performant suivant la méthode d'évaluation.

VII. Description du classifieur sélectionné

Après avoir testé tous ces paramétrages nous avons donc sélectionné pour chaque classifieur celui qui donnait les résultats les plus pertinents. Pour ce faire, nous avons utilisé les critères susmentionnés.

Il s'agit donc de:

- rpart() avec les paramétrages **split = 'gini' et minbucket = 10** .

- random_Forest() avec les paramétrages **ntree = 300** et **mtry = 5**.

- nnet() avec les paramétrages **size = 25, decay = 0. 01** et **maxit = 300.**

- svm() avec le paramétrage **kernel = linear.**

- naive_Bayes avec les paramétrages **laplace = 0**, **usekernel = FALSE**.

Les classifieurs ainsi retenus, nous avons donc pu passer à leur comparaison, toujours en fonction de nos critères de sélection et, deux classifieurs se sont démarqués. Il s'agit de :

- naive_Bayes()
- svm()

Les résultats sont les suivants :

Naive Bayes	**Support vector machines**
test_nb(0, FALSE, FALSE, "red")	test_svm("linear", FALSE, "red")
nb_class No Yes No 325 91 Yes 26 48	svm_class No Yes No 403 13 Yes 37 37
FN = 26	**FN = 37**
succes5 = 0.7612245	succes3= 0.8979592
TVN_nb = 0.9259259	**TVN_svm = 0.9159091**
AUC = 0.774525727650727	AUC = 0.81876949064449

En effet, selon les critères **Matrice de confusion** (pour mettre en évidence le taux le plus faible de Faux Négatifs) et **Calcul du taux de Vrais Négatifs** (pour savoir si l'on peut se fier au classifieur pour les prédictions négatives), le classifieur qui l'emporte est naive_Bayes(). En revanche, selon les deux autres critères, à savoir : **Taux de succès global** et **Courbe ROC et indice AUC**, svm() est plus performant.

Mais, étant donné que nous cherchons à « minimiser le risque de ne pas prévoir l'attrition d'un employé », c'est-à-dire minimiser le taux de Faux Négatifs, il nous a semblé plus judicieux de retenir le classifieur naive_Bayes() .

A l'issue des tests réalisés, le classifieur retenu et que nous appliquerons à notre ensemble à prédire est donc **naive_Bayes** avec les paramétrages **laplace = 0, usekernel = FALSE.**

VIII. Résumé des résultats de l'application du classifieur sélectionné à l'ensemble de données à prédire (distribution des classes prédites, probabilités minimales, maximales et moyennes associées à chacune des classes)

Notre ensemble à prédire, que nous avons nommé **attrition_new** est composé de 150 instances. Après avoir appliqué le classifieur à cet ensemble, nous avons obtenu les prédictions suivantes :

- 118 employés sont susceptibles de ne pas résilier leur contrat.
- 32 employés sont susceptibles de résilier leur contrat.

En ce qui concerne les probabilités, nous avons les résultats suivants :

- **Attrition = Non**
  - Probabilité minimale : 0.0106026
  - Probabilité maximale : 1.0000000
  - Moyenne : 0.7297545

- **<u>Attrition = Oui</u>**
    - Probabilité minimale : 0.0000000
    - Probabilité maximale : 0.9893974
    - Moyenne : 0.2702455

IX. <u>Conclusion résumant les autres observations sur cette application et les résultats, les difficultés rencontrées</u>

A l'issue de ce projet, nous avons pu prédire l'attrition des employés grâce au classifieur Naive_Bayes qui nous a donné des résultats convaincants en fonction des critères de sélection que nous avons choisis. Notre classifieur nous a permis de mener à bien notre étude car nous voulions minimiser le risque de ne pas prévoir l'attrition d'un employé. Cependant, force est de constater que l'algorithme SVM a obtenu globalement le meilleur taux de succès pour notre étude. En choisissant SVM comme classifieur, nous n'aurions ni la valeur minimale de Faux Négatifs, ni le Taux de Vrais Négatifs le plus élevé alors que ces critères sont assez déterminants dans notre étude. C'est la raison pour laquelle, le classifieur Naive_Bayes nous a semblé plus convaincant. Nous avons donc réussi à prédire l'attrition de 32 employés sur les 118 présentés par notre ensemble à prédire. En tenant compte de l'étude globale que nous avons réalisée, nous déduisons des relations entre des variables comme l'âge des employés, leur revenu, leur secteur d'activité et leur attrition. Grâce aux prédictions réalisées, l'entreprise pourra s'organiser en vue de limiter ces attritions ou pour restructurer ses équipes.