**PROJET IA**

**DOCUMENT ETHIQUE**

**JUSTIFICATIF DU CHOIX DE DONNEES**

Djamel Hemch

Shoukry Badani

Karim Kanoun

Said Amokrane

Abdeljalil Hadjhabib

Table des matières

[1-*Données retenues:* 3](#_Toc98160465)

[*2-Données retenues:* 5](#_Toc98160466)

[3-*les models testés:* 6](#_Toc98160467)

[A. Régression Logistique: 6](#_Toc98160468)

[B. Random Forest Classifier: 6](#_Toc98160469)

[C. Stochastic Gradient Decent Classifier: 6](#_Toc98160470)

[D. C-Support Vector Classification (SVC): 6](#_Toc98160471)

[E. Perceptron: 6](#_Toc98160472)

[F. Multi Layered Perceptron (MLP): 6](#_Toc98160473)

[4-*les models retenus:* 6](#_Toc98160474)

[*A.* *Multi Layered Perceptron (MLP):* 6](#_Toc98160475)

[ Validation croisée 7](#_Toc98160476)

[ GridSearchCV 7](#_Toc98160477)

[ Paramètres optimaux 7](#_Toc98160478)

[ Validation croisée 7](#_Toc98160479)

[ Matrice de confusion 7](#_Toc98160480)

[ Rappel, Precision avec threshold 7](#_Toc98160481)

[ Courbe ROC 7](#_Toc98160482)

[*B.* *Stochastic Gradient Decent Classifier:* 7](#_Toc98160483)

[ Validation croisée 7](#_Toc98160484)

[ GridSearchCV 7](#_Toc98160485)

[ Paramètres optimaux 7](#_Toc98160486)

[ Validation croisée 7](#_Toc98160487)

[ Matrice de confusion 7](#_Toc98160488)

[ Rappel, Precision avec threshold 7](#_Toc98160489)

[ Courbe ROC 7](#_Toc98160490)

[*C.* *Random Forest Classifier:* 7](#_Toc98160491)

[ Validation croisée 7](#_Toc98160492)

[ GridSearchCV 7](#_Toc98160493)

[ Paramètres optimaux 7](#_Toc98160494)

[ Validation croisée 8](#_Toc98160495)

[ Matrice de confusion 8](#_Toc98160496)

[ Rappel, Precision avec threshold 8](#_Toc98160497)

[ Courbe ROC 8](#_Toc98160498)

# **1-Données retenues:**

* Age:

L’Age peut influer sur la forme de l’employé et donc de son retrait ou non de la société

* Info poste actuel (joblevel / jobrole / monthlyincome / standardhours / stockoptionlevel / yearsatcompany / yearssincelastpromotion / employeecount / precentsalaryhike / trainingtimeslastyear):

Toutes ses informations font écho à la situation actuelle de l’employé au sein de l’entreprise ce qui nous permet de conclure quel facteurs ou données sont les plus impactant dans le taux de turn-over de cette liste.

Exemple : un employé qui travaille depuis 10 ans au sein de l’entreprise sans recevoir de promotion pour autant (YearsAtCompany/PercentSalaryHike)

* Sexe de l’employé (Gender / maritalstatus) :

Pour estimer quel est le genre des employés et s’il existe une certaine discrimination entre le sexe et le statut marital de la personne, et si le sexe est un facteur dans le turn-over ou non chose qui est fortement envisageable.

Exemple : Le taux de turn-over est plus élevé chez les femmes mariées que chez les hommes mariés.

* Nombre d'entreprises pour lequel le salarié a travaillé avant de rejoindre HUMANFORYOU (numcompaniesworked) :

Ce paramètre peut nous indiquer si l’employé est du type frivole ou plus à rester stable dans une entreprise.

* Distance entre Maison et travail :

Si la personne vit loin de là où elle travaille il serait fort probable qu’a un certain temps l’employé veuille quitter son job pour une entreprise proche de son domicile.

* **Notes managériales : (*yearswithcurrentmanager + Données MANAGER) :***

Peut indiquer si le manager influe sur la performance de l’employé en bien comme en mal.

* Données de l’enquête employés :

Nous indique la qualité de vie au sein de l’entreprise.

* **Heure d'entrée et de sortie :**

Peut-être jointe à un nombre d’autres données de la liste.

***Exemple :***

* si l’employé vit loin et arrive souvent en retard
* Si l’employé sort souvent tard alors que dans son contrat il n’est pas stipulé qu’il doit faire des heures supplémentaires etc…

# **2-Données écartées:**

* Niveau éducatif et spécialité:

N’influe en aucun cas sur le turn-over, car du moment où l’employé est engagé il est estimé apte à fournir le travail proposé et donc son niveau éducatif ainsi que sa spécialité ont déjà été évalués et n’influeront plus sur les décisions.

* **Emplolyeeid :**

Ne concerne que l’identité de l’employé et n’influe pas sur le taux de turn-over.

* **Over18 :**

Les employés mineurs ne sont tout simplement pas engagés dans l’entreprise.

* **Voyage professionnel :**

Nous estimons que des voyages d’affaires n’auront aucun impact sur la volonté de quitter l’entreprise.

# **3-les models testés:**

# **Régression Logistique:**

La régression logistique est un modèle statistique permettant d’étudier les relations entre un ensemble de variables qualitatives Xi et une variable qualitative Y. Il s’agit d’un modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien.

Un modèle de régression logistique permet aussi de prédire la probabilité qu’un événement arrive (valeur de 1) ou non (valeur de 0) à partir de l’optimisation des coefficients de régression. Ce résultat varie toujours entre 0 et 1. Lorsque la valeur prédite est supérieure à un seuil, l’événement est susceptible de se produire, alors que lorsque cette valeur est inférieure au même seuil, il ne l’est pas.

# **Random Forest Classifier:**

Random Forest est un classificateur qui contient un certain nombre d'arbres de décision sur divers sous-ensembles de l'ensemble de données donné et prend la moyenne pour améliorer la précision prédictive de cet ensemble de données. Au lieu de s'appuyer sur un arbre de décision, la forêt aléatoire prend la prédiction de chaque arbre et sur la base des votes majoritaires des prédictions, et elle prédit la sortie finale.

# **Stochastic Gradient Decent Classifier:**

La descente de gradient stochastique (SGD) est une approche simple mais très efficace pour ajuster les classificateurs linéaires et les régresseurs sous des fonctions de perte convexes telles que les machines à vecteurs de support (linéaires) et la régression logistique. Même si SGD existe depuis longtemps dans la communauté de l'apprentissage automatique, il a récemment reçu une attention considérable dans le contexte de l'apprentissage à grande échelle.

# **C-Support Vector Classification (SVC):**

Les machines à vecteurs de support (SVM) sont un ensemble de méthodes d'apprentissage supervisé utilisées pour la classification, la régression et la détection des valeurs aberrantes.

Les avantages des machines à vecteurs de support sont :

* Efficace dans les espaces de grande dimension.
* Toujours efficace dans les cas où le nombre de dimensions est supérieur au nombre d'échantillons.
* Utilise un sous-ensemble de points d'apprentissage dans la fonction de décision (appelés vecteurs de support), il est donc également efficace en mémoire.
* Polyvalent : différentes fonctions du noyau peuvent être spécifiées pour la fonction de décision. Des noyaux communs sont fournis, mais il est également possible de spécifier des noyaux personnalisés.

# **Perceptron:**

Un réseau de neurones artificiels simple est composé d'entrées et d'une seule sortie. Le cheminement des informations se fait par un circuit de liens qui relient entre eux des perceptrons ou algorithmes traitant les données. Le perceptron simple ou monocouche fonctionne sur un modèle basique de type classifieur linéaire (qui sépare deux classes).

# **Multi Layered Perceptron (MLP):**

Un perceptron multicouche est un modèle de réseaux artificiels, qui comprend plusieurs couches cachées, et permet de produire un séparateur non linéaire. Il est constitué de plusieurs entrées et sorties.

Il s'agit d'un réseau à propagation directe (feedforward) disposant d'un nombre de neurones artificiels variable, qui compose plusieurs couches du système neuronal. On dépasse alors le machine learning pour entrer dans l'ère du deep learning, avec des systèmes d'apprentissage pouvant traiter des données en profondeur.

# **4-les models retenus:**

Définitions des techniques utilisées :

**Coté optimisation :**

1. Validation croisée :

Cela consiste à découper le dataset en K sous-ensemble (ou K folds) puis prendre un des K sous-ensemble comme dataset de validation (validation set) et les K-1 restants comme dataset d’entrainement (training set). On répète l’opération sur toutes les combinaisons possibles. On obtient K mesures de performance dont la moyenne représente la performance de l’algorithme.

1. Optimisation du model avec gridsearch cv :

Un hyperparamètre de modèle est une caractéristique d'un modèle qui est externe au modèle et dont la valeur ne peut pas être estimée à partir des données. La valeur de l'hyperparamètre doit être définie avant le début du processus d'apprentissage. Par exemple, c dans Support Vector Machines, k dans k-Nearest Neighbors, le nombre de couches cachées dans les réseaux de neurones.

En revanche, un paramètre est une caractéristique interne du modèle et sa valeur peut être estimée à partir des données. Exemple, coefficients bêta de régression linéaire/logistique ou vecteurs de support dans les machines à vecteurs de support.

**Vérification de la qualité du model :**

1. La matrice de confusion :

La matrice de confusion est les données brutes pour tout ce que nous allons montrer ici: c'est un tableau qui contient le nombre de vrais positifs (TP), de vrais négatifs (TN), de faux positifs (FP) et de faux négatifs (FN). L'avantage d'utiliser les courbes ROC et PR est qu'elles résument les informations de performance pertinentes pour les problèmes de classification binaire. Il est plus facile de lire et d'interpréter que d'avoir plusieurs matrices de confusion - une pour chaque seuil - et ensuite de calculer certains ratios pour avoir un aperçu de ce que signifient ces nombres (TP, TN, FP, FN). Et voilá , nous venons de définir ce que représentent les courbes ROC et PR - des rapports de nombres que nous obtenons à partir de plusieurs matrices de confusion. La matrice de confusion prend en charge plusieurs classes mais dans cet article, nous allons nous concentrer sur le problème de classification binaire.

1. courbe ROC :

 ( **Receiver Operating Feature Curve** ) est un graphique montrant les performances d'un modèle de classification à tous les seuils de classification. Cette courbe trace deux paramètres :

* Taux de vrais positifs
* Taux de faux positifs

**Le taux de vrais positifs** ( **TPR** ) est synonyme de rappel et est donc défini comme suit :

JPR=JPJP+FN

**Le taux de faux positifs** ( **FPR** ) est défini comme suit :

FPR=FPFP+JN

Une courbe ROC trace le TPR par rapport au FPR à différents seuils de classification. Abaisser le seuil de classification classe plus d'éléments comme positifs, augmentant ainsi à la fois les faux positifs et les vrais positifs. La figure suivante montre une courbe ROC typique.

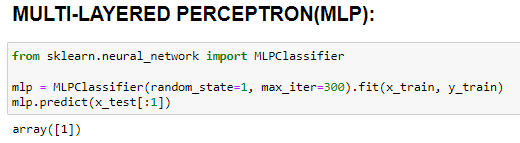
1. Précision et rappel :

Dans la reconnaissance de formes, la récupération d'informations et la classification (apprentissage automatique), la précision (également appelée valeur prédictive positive) est la fraction des instances pertinentes parmi les instances récupérées, tandis que le rappel (également appelé sensibilité) est la fraction des instances pertinentes qui ont été récupérées. La précision et le rappel sont donc tous deux basés sur la pertinence.

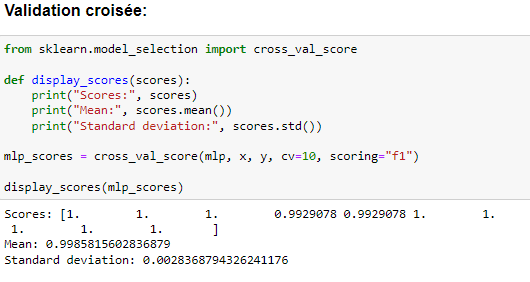
# **Multi Layered Perceptron (MLP):**

Avant optimisation :

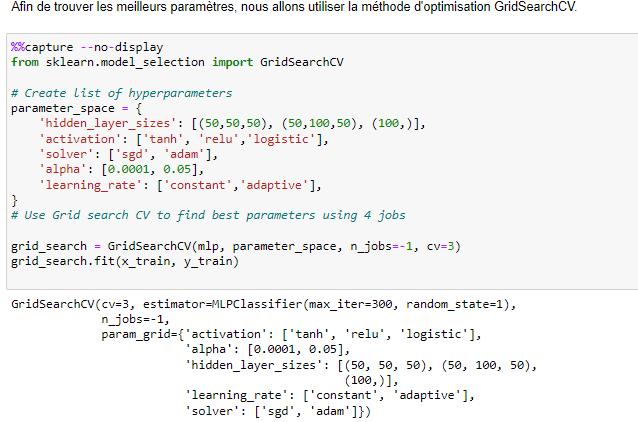
## Entrainement du model



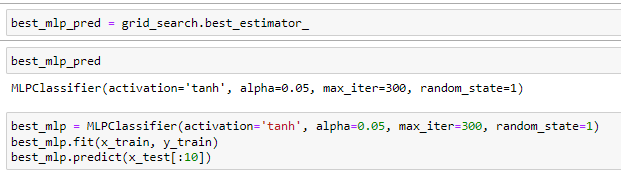
## Validation croisée



## GridSearchCV

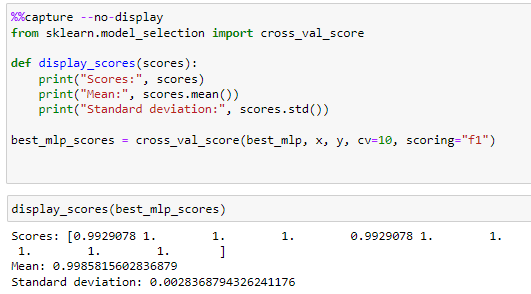


## Re-entrainement avec les nouveaux paramètres

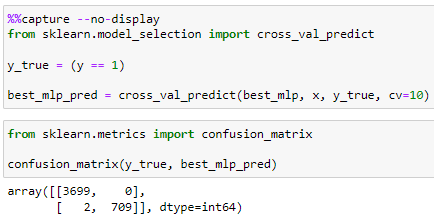


Apres optimisation :

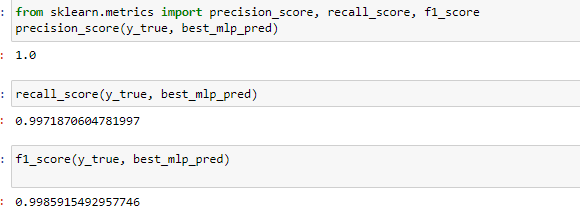
## Validation croisée



## Matrice de confusion



## Rappel, Precision avec threshold

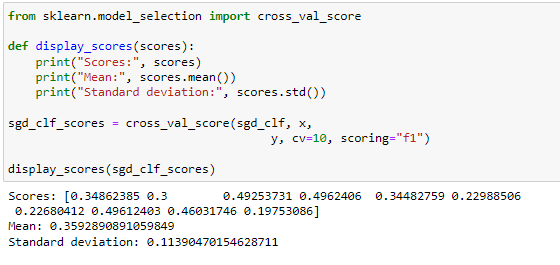


## Courbe ROC

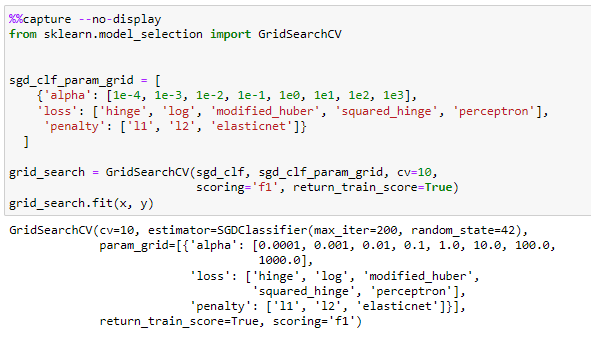
# **Stochastic Gradient Decent Classifier:**

Avant optimisation :

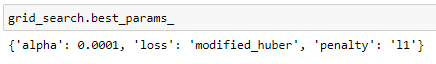
## Validation croisée



## GridSearchCV

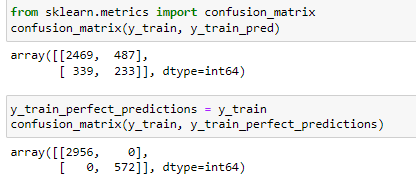


## Paramètres optimaux

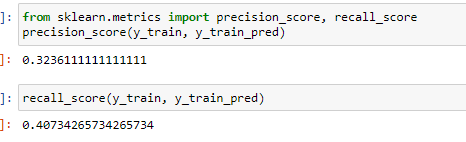


Apres optimisation :

## Matrice de confusion



## Rappel, Precision avec threshold

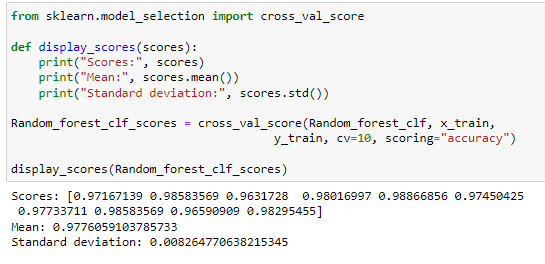


## Courbe ROC

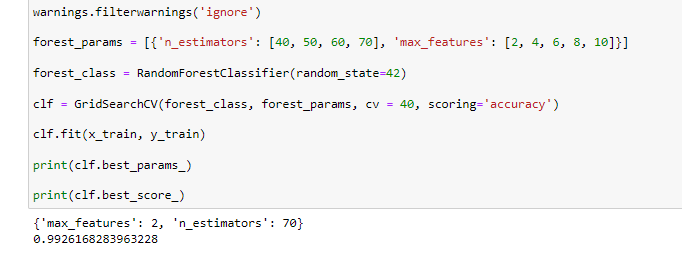
# **Random Forest Classifier:**

Avant optimisation :

## Validation croisée

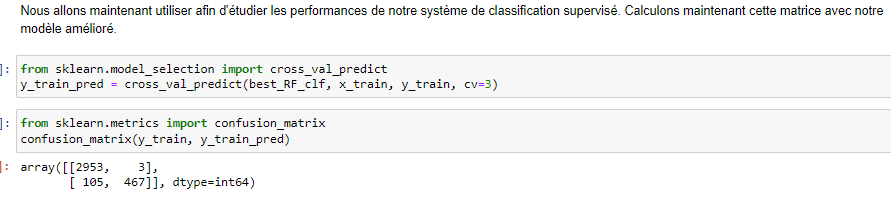


## GridSearchCV

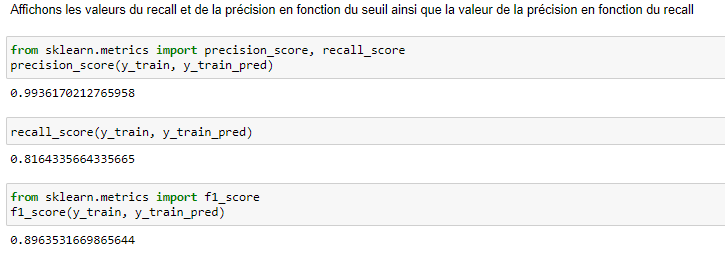


Apres optimisation :

## Matrice de confusion



## Rappel, Precision avec threshold



## Courbe ROC

1. ***- Proposition de pistes d'amélioration:***
2. Offrir une rémunération compétitive.  
   Même si la rémunération n’est pas le premier facteur de motivation pour tous les salariés, offrir un salaire légèrement supérieur (10% par exemple) à celui du marché montrera que l’entreprise prête attention au bien-être de ses salariés. Il est également important de penser à attribuer des primes pour que les salariés sentent leurs efforts récompensés.
3. Offrir des avantages en nature ou offrir d’autres possibilités.  
   En effet, fournir une voiture de fonction, des tickets restaurants, des outils de travail comme un ordinateur ou un Smartphone montrera que vous voulez que le travail de vos salariés se fasse dans les meilleures conditions possibles.  
   De plus, les salariés recherchent plus de flexibilité afin de trouver un équilibre entre vie professionnelle et vie personnelle, ainsi offrir la possibilité du télétravail peut être une solution.
4. Offrir des perspectives de promotion  
   Offrir des perspectives de promotion en interne ou permettre la mobilité géographique pour valoriser le travail des salariés. Leur donner plus de responsabilités ou leur confier une équipe est un moyen de les faire sentir plus importants pour l’entreprise. De cette manière, ils s’impliqueront davantage.
5. Travailler avec un cabinet de recrutement  
   Pour offrir à tout moment des perspectives d’évolution intéressantes à vos salariés  ou éviter des erreurs de recrutement. En effet, l’entreprise qui recrute en urgence du fait d’un départ imprévu n’aura pas pu planifier le recrutement. En se précipitant, l’entreprise risque de faire une erreur de casting entraînant le départ anticipé d’une personne. Il ne restera plus qu’à recommencer encore et encore. Recourir à un cabinet de recrutement permet d’avoir des candidatures de qualités et rapidement pour éviter ses erreurs et ne pas aggraver le turnover en entreprises.

**Selon le résultat des données les plus pertinentes qui causent un turnover en entreprise**:

L’âge et le temps de travail sont les deux facteurs qui influent le plus sur le retrait des salariés de leur entreprise.

Contre ces deux facteurs nous allons proposer quelques idées pour éviter cela:

**Pour l’âge** :

Afin de pousser les plus jeunes à rester en entreprise nous devons proposer plus d’activités capables de donner envie aux jeunes de rester en entreprise comme organiser des fêtes, des promotions ou encore des événements interentreprises. Aussi proposer aux plus jeunes une prise en charge par les plus vieux afin de les intégrer au mieux aux salariés de l’entreprise.

Pour les plus vieux quant à eux pour éviter une sortie rapide ou avant l’âge de la retraite l’entreprise peut proposer une retraite plus conséquente aux personnes qui restent plus longtemps en entreprise Proposer aussi aux personnes les plus âgées / plus anciens de l’entreprise un travail moins affaiblissants et moins difficiles à réaliser.

Aussi nous pourrions proposer une couverture médicale aux plus nécessiteux

**Le temps de travail moyen :**

Proposer des récompenses aux personnes qui travaillent le plus et qui font des heures supplémentaires.

Proposer à ces derniers aussi des jours de congé supplémentaire en fonction de leur temps de travail et de leur et des tâches réalisées.