[Exploration de données statistiques Pandas et Seaborn (moncoachdata.com)](https://moncoachdata.com/blog/exploration-de-donnees-statistiques/)

Dans le cadre de mon master 1 en **Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales** qui est une formation en alternance. N’ayant pas trouvé une alternance, je dois réaliser un projet data science pour pouvoir pratiquer les notions apprises au cours du premier semestre.

C’est dans ce contexte que j’ai décidé de travailler sur un projet data science qui porte sur la prédiction de la probabilité qu’une personne prenne le vaccin contre la grippe h1n1, ou contre la grippe saisonnière. Le sujet consiste à prédire si les gens ont reçu les vaccins contre la grippe H1N1 et la grippe saisonnière en utilisant les informations qu'ils ont partagées sur leurs antécédents, leurs opinions et leurs comportements en matière de santé.

Comme dans tout projet data science je me suis plongé tout d’abord dans le jeu de données : étudier ses propriétés, ses distributions. Ainsi, j’ai commencé par Analyse exploratoire des données, visualisation données, statistiques descriptives et la prédiction en dernière partie.

Présentation du sujet

Le projet que je réalise est une compétition en cours sur la plateforme Drivendata.

Pour ce projet, j’ai à ma disposition quatre jeux de données. Ces données proviennent d’une 'enquête nationale 2009 sur la grippe H1N1 (NHFS) parrainée par le National Center for Immunization and Respiratory Diseases (NCIRD) et menée conjointement par le NCIRD et le National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC). La NHFS était une enquête téléphonique à composition aléatoire assistée par liste auprès des ménages, conçue pour surveiller la couverture vaccinale contre la grippe au cours de la saison 2009-2010. La population cible de la NHFS était toutes les personnes de 6 mois ou plus vivant aux États-Unis au moment de l'entretien. Les données de la NHFS ont été utilisées pour produire des estimations en temps opportun des taux de couverture vaccinale pour les vaccins monovalents contre la grippe pH1N1 et trivalents contre la grippe saisonnière.

Pour ce projet , j’ai à ma disposition, quatre jeux de données, training\_set\_features.csv , qui le dataset des variables explicatives, c’est ma base de données d'entraînement. Ce dataset possède trente-six (36) colonnes et 26707 entières! Donc c’est plus de 26700 individus.

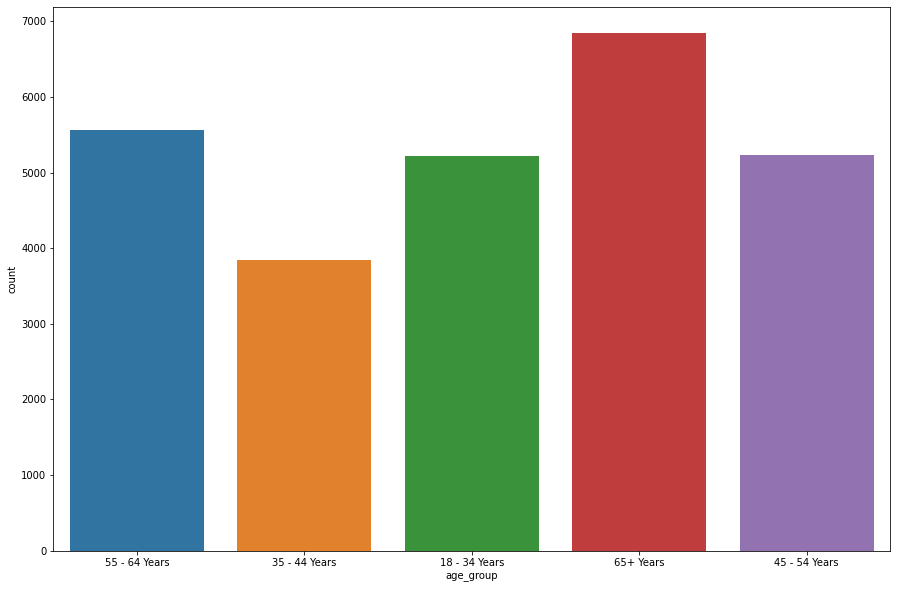
Le dataset train\_set\_labels.csv c’est la dataset d'entraînement des variables à prédire à savoir h1n1 vaccine et seasonal\_vaccine. Ce dataset comporte 26707 lignes et deux colonnes . Et enfin le dataset test\_set\_features.csv pour test mes modèles à la fin. Parmi les 36 variables, seuls deux variables sont qualitatives, les 34 restantes sont des variables qualitatives majoritairement binaires. Chaque ligne de l’ensemble de données représente une personne qui a répondu à l’Enquête nationale sur la grippe H1N1 de 2009. Les deux variables cibles h1n1\_vaccine et seasonal\_vaccine sont binaires : 0 = Non et 1 = Oui. Certains répondants (individu) n’ont reçu aucun vaccin, d’autres n’en ont reçu qu’un seul et certains ont reçu les deux donc c’est un problème multi label.

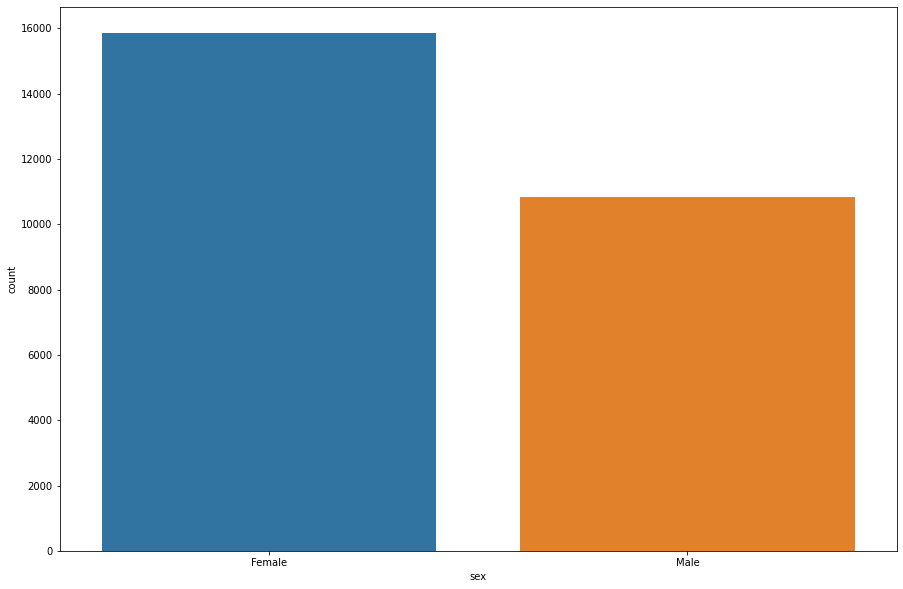
### 2 - Exploration et visualisation des données

Tout d’abord, avant d’essayer de recueillir les informations, j’'aimerais avoir une meilleure connaissance du jeu de données dans lequel je dois travailler. Combien de données sont manquantes? Combien y a-t-il de colonnes différentes, quels sont les différents types de variables? Ce sont là les questions que j'ai commencé par me poser avant de commencer l’exploration des données.

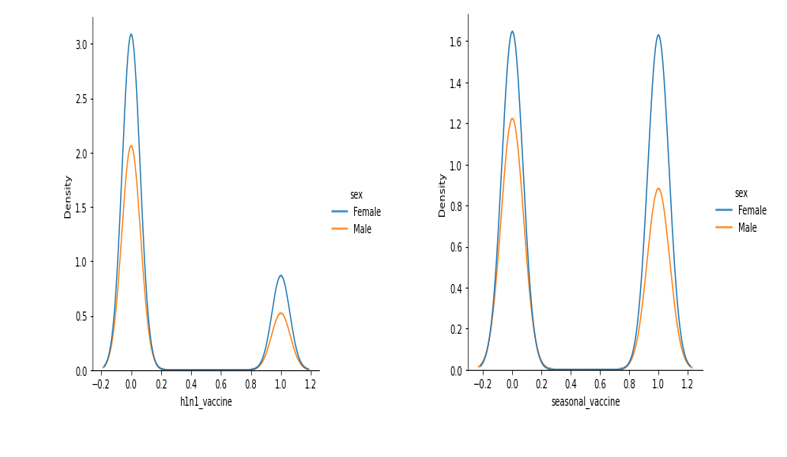
Afin de faire une analyse pertinente des données sous différents angles, de manière à en extraire des relations, des associations et patterns qui vont me permettre d’obtenir de précieuses informations du dataset. J’ai utilisé pour cette analyse un notebook Jupiter qui est un outil parfait pour explorer rapidement un jeu de données. Pour la suite, j’ajouterai des extraits d’image pertinents . Pour effectuer cette analyse exploratoire, j’ai utilisé la bibliothèque [Pandas](https://moncoachdata.com/blog/analyse-exploratoire-de-donnees/) pour l’analyse de données et la bibliothèque [Seaborn](https://seaborn.pydata.org/tutorial.html) pour la visualisation de données. j’ai préféré Seaborn car c’est un formidable outil de visualisation d’un point de vue esthétique..

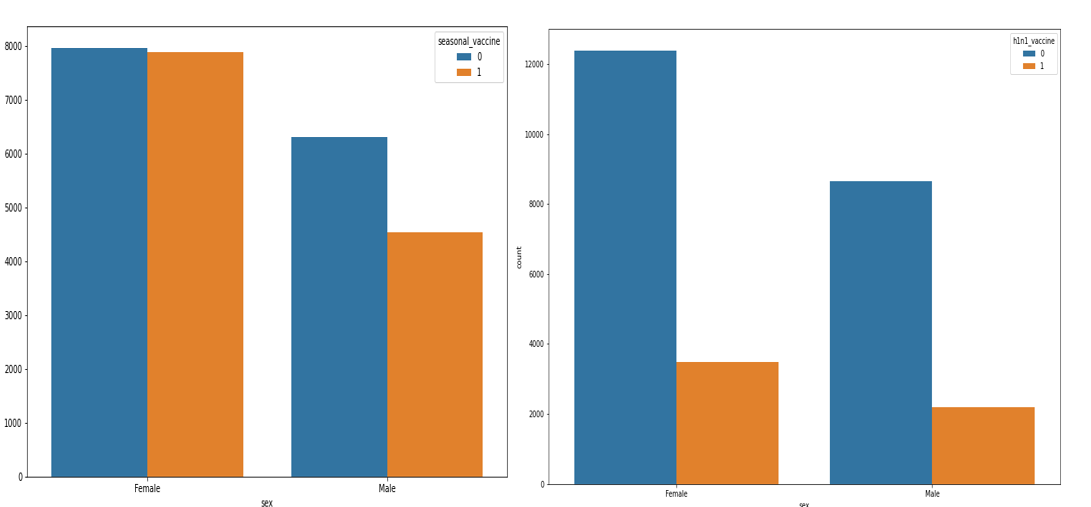
Voici quelques visualisations réalisées, l’image ci-dessous montre la répartition des individus en fonction de leur classe d’âge. On constate que c’est la classe de 65 ans et plus qui est la classe la plus représentée avec environ 26% des individus, suivis par la classe des 55 à 64 ans avec 20%. Les classes de 45 à 54 ans et 18 à 34 ans ont toutes 20% et enfin la classe de 35 à 44 ans avec 14% des individus.



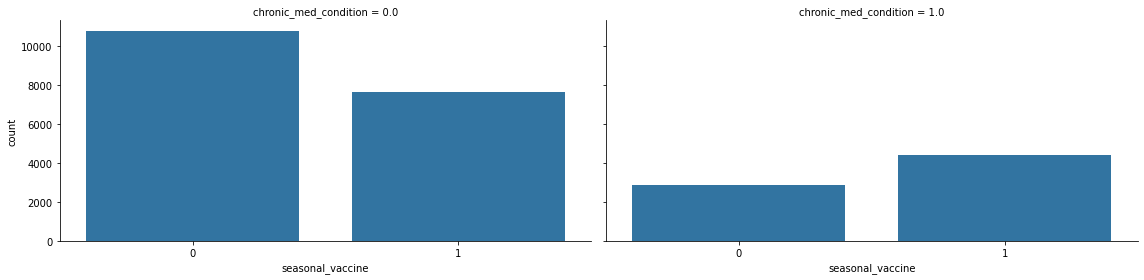


Notre échantillon est constitué de 60% de sexe feminin et 40% du sexe masculin.

La distribution des deux variables cibles en fonction du sexe grâce à Kernel density estimation (KDE) qui est une méthode non paramétrique pour estimer la fonction de densité de probabilité d'une variable aléatoire. On constate que ça sur le vaccin h1n1\_vaccine ou seasonal\_vaccine les valeurs d’espression du sexe feminin sont plus élevées que celle du sexe maculin.



Le graphe ci-dessus montre la distribution de l’ensemble de la population étudiée autour des deux variables cibles. En analysant attentivement de graphe on constate qu’il y a plus d’individu qui ont accepté le vaccin contre la grippe saisonnière que le vaccin contre la grippe h1n1 et c’est les mêmes tendance au niveau du sexe. **Je peux émettre une hypothèse en disant il semble que le le vaccin contre la grippe saisonnière est plus accepté que celui de la grippe h1n1**

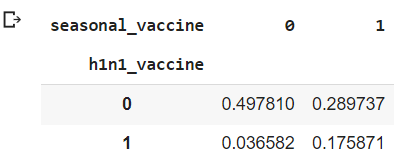


Ce graphe nous montre distribution de la variables seasonal vaccine en fonction de la variable chronic med condition qui est une variable binaire, sa valeur 0 si l’individu n’a aucune des conditions médicale chronique et 1 si l’individu possède une des conditions médicales chroniques suivantes: asthme ou une autre affection pulmonaire, diabète, une maladie cardiaque, une maladie rénale, une anémie falciforme ou une autre anémie, une affection neurologique ou neuromusculaire, une affection hépatique ou un système immunitaire affaibli causée par une maladie chronique ou par des médicaments pris pour une maladie chronique. On constate que les individus acceptent le vaccin contre la grippe saisonnière sans qu’ils aient une maladie chronique ou pas. Contrairement à ce qu’on pouvait croire.

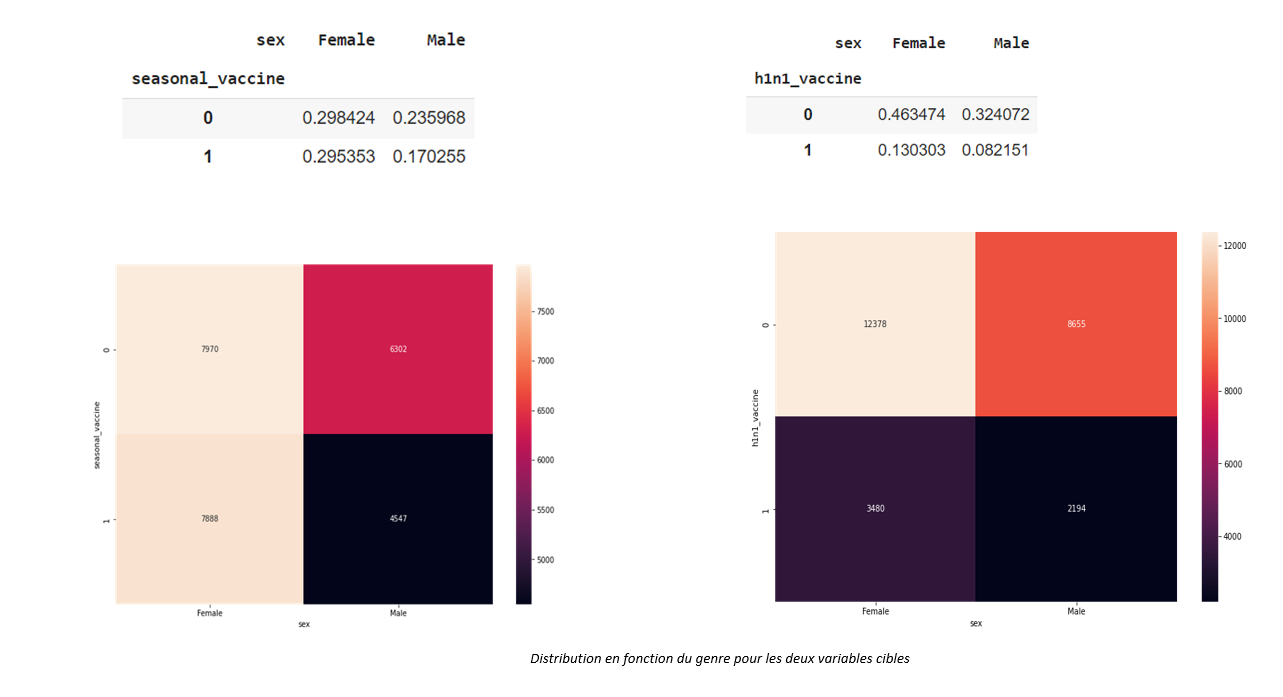
### statistiques descriptives

Sans doute la première question qui m’est venue est combien d'individus de l’ensemble de la population étudiée on accepté les deux vaccins? Le graphe ci-dessous m’a permis d’avoir une idée sur cette question.

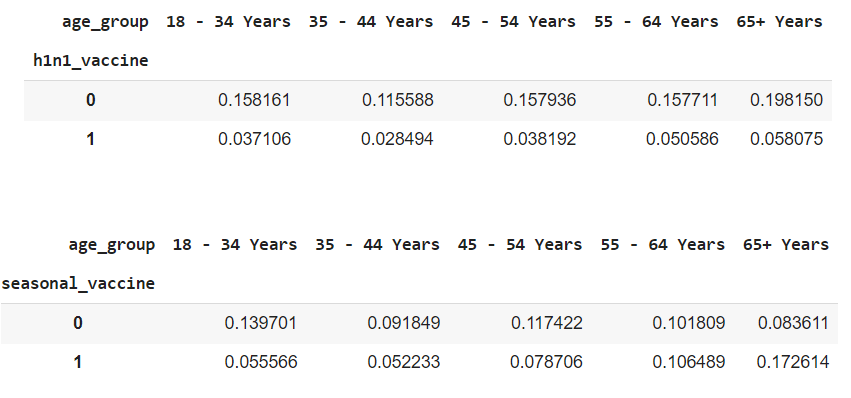




Après cette visualisation on peut dire qu' environ 50% des individus ont refusé. Le vaccin h1n1 ont également refusé le vaccin contre la grippe saisonnière environ 29% des individus qui ont refusé le vaccin contre la grippe h1n1 ont accepté le vaccin contre la grippe saisonnière seasonal\_vaccine seulement 3.6% ont refusé le vaccin contre la grippe saisonnière seasonal\_vaccine et ont fait le vaccin contre la grippe h1n1\_vaccin en fin 17.5% des individus ont fait à la fois les deux vaccins.



Ces deux tableaux croisés nous montre les distributions des deux variables cibbles seasonal\_Vaccine et h1n1\_vaccine en fonction du sexe. On constate que environ 30% des individus du sexe feminin on accepté le vaccin contre la grippe saisonnière et seulement 17 % des individus de sexe masculin ont accepté le vaccin contre la grippe saisonnière. Alors que seulement 13% des individus du sexe feminin on accepté le vaccin contre la grippe h1n1 et seulement 8% des individus de sexe masculin l’on accepté. Donc, on peut émettre une hypothèse en disant qu’il semblerait que le vaccin contre la grippe saisonnière est plus accepté que celui de la grippe h1n1.



Avec le tableau ci-dessous on peut dire que c’est la classe des individus de 65 et plus qui accepte le plus de faire le vaccin contre la grippe car on constate que pour la grippe h1n1 c'est la classe la plus représentée dans au niveau des deux variables cibles. On peut formuler une hypothèse : Les individus qui ont plus de 65 ans et plus acceptent le vaccin contre la grippe que les individus qui ont moins de 65 ans.

Sans surprise, on constate que la distribution des deux variables cibles n'est pas du tout uniforme, on constate, supériorité considérable des femmes sur les deux variables. Cependant, notons que 60% de la population étudiée est du sexe feminin et 40% des individus sont du sexe masculin ce qui semble expliquer cette superiorité du sexe feminin au niveau des deux variable. Pour en savoir plus, je vais approfondir mon analyse avec des méthodes statistiques plus avancées afin de confirmer ou infirmer les hypothèses émises au cours de l'exploration et la visualisation des données.

### Réduction de dimension: ACM

Pour mieux comprendre les données et mettre en pratique les notions apprises pour ce premier semestre en analyse multidimensionnelle , j’ai appliqué une Analyse des Composantes Multiples sur mes données. Les graphes et tableaux réalisés à cet effet sont dans l’annexe de mon rapport.

### Traitement et Préparation des données

La base de données d'entraînement est composée de 26707 lignes (individus) et de 36 colonnes (variables) .

Traitement des valeurs manquantes: En science de données on sait que les valeurs manquantes sont courantes et peuvent parfois avoir un effet significatif sur notre analyse et jouer sur la performance des différents modèles de prédiction. Ainsi, après les premières visualisations de mon dataframe j’ai constaté la présence des valeurs manquantes sur un certain nombre de variables. Ensuite j’ai essayé de comprendre si ces valeurs peuvent être considérées comme des valeurs manquant si ce n’ pas des espaces ou caractères spéciaux avant de penser à leur remplacement ou suppression.

Ainsi, j’ai constaté qu’il y a trois variables à savoir employment\_occupation , employment\_industry et health\_insurance qui signifient respectivement Type d’industrie dans laquelle le répondant est employé. Les valeurs sont représentées sous forme de courtes chaînes de caractères aléatoires, type de profession du répondant. Les valeurs sont représentées sous forme de courtes chaînes de caractères aléatoires et A une assurance maladie ont des valeurs manquant les plus élevées environ 50% de valeurs manquantes.

Compte tenu de ses valeurs manquantes et du contexte de l'étude, j'ai décidé de les retirer de la base de données.

Pour les variables qui ont des valeurs manquantes inférieur à 50%, j'ai fait un traitement cas par cas. Pour les variables qualitatives , j’ai remplacé les valeurs manquantes par le mode de la variable concernée. Pour les variables quantitatives, j’ai remplacé les valeurs manquantes par la médiane de la variable concernée car la moyenne est souvent influencée par les valeurs aberrantes. Pour confirmer mon choix, j’ai comparé l'écart type avant le remplacement et après le remplacement des valeurs manquantes il y a une variation pas assez importante de la valeur de l'écart type ce qui veut dire que ces remplacements de valeurs manquantes n'influencent pas trop la variance de ces variables raison pour laquelle j’ai opté pour ces techniques de remplacement.

Pour ce qui est des valeurs aberrantes, mon jeu de données étant composé de 98% de variables qualitatives, il n’y a pas de valeur aberrante au niveau des différentes variables et même sur les deux variables quantitatives il n’y a pas de valeurs aberrantes.

### Apprentissage, choix des modèles, entrainement et ajustement des modèles

Après le nettoyage des données, j’ai procédé à l'implémentation des différents modèles qui sont appropriés à la nature de mon travail. Les variables à prédire étant des variables qualitatives je dois faire une classification et comme je connais les différentes étiquettes de mes labels alors elle sera supervisée. Ainsi , en premier lieu j’ai essayé le classifieur de RandomForest pour classer en les individus deux groupes.

### 

### Bibliographie

La régression logistique Par Sonia NEJI et Anne-Hélène JIGOREL

## 

Les organismes de recherche en santé, les laboratoires qu’ils soient publics privés mènent de très nombreux travaux de recherche clinique dans tous les domaines de la santé en impliquant les populations plus particulièrement les patients ou les profils de risque.

La science de données permet aujourd’hui de concevoir des études et des évaluations, d’effectuer des analyses de données complexes à partir des résultats de différentes analyses. Ces différents travaux permettent de mettre en place des politiques de santé les mieux adaptées pour une couverture sanitaire rassurante.

Dans le cas de ce projet, on peut dire après nos analyses on peut dire avec une certaine certitude les personnes susceptibles de faire le vaccin contre la grippe h1n1 ou la grippe saisonnière. Grâce à mes analyses, je peux dire que les personnes susceptibles d’accepter le vaccin antigrippal sont celles dont le médecin à recommander de le faire, les personnes âgées de 65 ans et plus etc.. Il s’avère nécessaire de retenir que parmi les trois hypothèses émises lors de l’exploration des données une seule a été confirmée après la conception des modèles. Donc on peut dire que parfois ce qu’on pense exact ne l’ai pas, et on ne peut se fier aux visualisations et observations des données pour prendre des décision, il faut toujours l’avis des experts métier pour pouvoir mieux comprendre et confirmer nos résultats. En matière de compétences ce projet m'a permis de voir mes limites dans les techniques d'apprentissage supervisé mais m’a également permis de mettre en pratique les notions apprises au cours de ce semestre. Il m’a permis de comprendre plus le fonctionnement des algorithmes de classification. Les différents obstacles et barrières que j’ai rencontrés ont été de bonnes leçons pour moi. Comme perspective je continuerai d'améliorer mon modèles pour une participation à cette compétition.

Etudiant en master 1 MIASH,