Rapport de Projet: Statistiques en Grande Dimension

Sommaire:

Table des annexes :

1- Contexte de l'étude

Ces données sont extraites d'une étude portant sur l'admission au département des urgences de patients adultes présentant divers symptômes aux Etats-Unis. Il en résulte soit l'admission du patient, soit son refus de prise en charge, qui est la variable d'intérêt du problème.

Un total de 972 variables explicatives sont présentes dans la base de données de départ. Elles se décomposent en plusieurs types (voir liste des variables en Annexe):

- **Démographiques (9)** tel que l'âge, le sexe ou l'ethnicité.
- **Variables de triage (13)** tel que le mode d'arrivée (ambulance), le département d'arrivée à l'hôpital (A, B ou C), le mois/jour d'arrivée et les heures d'arrivée réparties en intervalles de 4 heures (23-02, 03-06, 7-10, etc).
- **Historique sur le patient par l'hôpital** tel que son admission lors de son dernier passage, le nombre de visites à l'hôpital la dernière année ou encore le nombre d'opérations l'année dernière.
- **Variables dites de plainte (200)** telles que fatigue, mal de tête ou encore problème lié à l'alcool.
- **Antécédents médicaux (200)** tels qu'une hernie abdominale, insuffisance rénale..
- **Médecine ambulatoire réalisée sur le patient (48)** tels que le nombre d'anesthésiants, antibiotiques… administrés au patient.
- **Imagerie de diagnostic réalisé sur le patient (9)** tels que le nombre d'écho cardiogrammes réalisés ou imagerie par résonance magnétique.
- **Mesures vitales (28)** telles que la mesure du poul ou la tension lors de la dernière année ou de la visite en cours.
- **Mesures numériques du laboratoire (352)** telles que la teneur en créatinine, sodium ou potassium mesurés par les différents laboratoires (dernière valeur, minimum, maximum et médiane).
- **Mesures catégoriques du laboratoire (27)** qui sont soit binaires soit entières pour les différents laboratoires (dernière valeur, nombre de résultats positifs, et nombre de tests).

Nous remarquons que sur les 972 variables explicatives ci-dessus, 461 variables sont des données quantitatives, tandis que 511 sont des variables catégoriques ou binaires. De nombreuses données sont manquantes étant donné que lors du passage d'un potentiel patient à l'hôpital, l'ensemble des tests ne lui sont pas effectués en fonction de ses propres syndromes. Pour cette raison, certaines variables ne contiennent que très peu de données.

2 - Objectif

L'enjeu du problème est de parvenir à prédire au mieux la variable d'intérêt qui est l'admission du patient (représentée par la classe 1) ou le refus de prise en charge (représenté par la classe 0) aux départements des urgences, en fonction des variables explicatives à disposition. Étant donné le grand nombre de données sur les patients arrivant à l'hôpital, une partie considérable du travail reposera sur le processus de sélection des variables.

3 - Prétraitement des données

La première étape était de traiter les données au préalable les données avant de pouvoir passer aux autres étapes. Les données de base contiennent un grand nombre de lignes (560.486 patients). Le nombre de patients a été réduit pour faciliter l'exécution des différents algorithmes. Nous n'avons gardé que 2500 patients.

Après importation des données, et puisque nous avons pris un nombre réduit de patients, il se peut que certaines colonnes ne contiennent que des zéros ou que des NaN. Nous avons donc décidé d'éliminer les colonnes qui ne contiennent que des zéros, et celles qui ne contiennent que des NaN. Nous nous sommes retrouvés avec un nombre de colonnes égal à 895. Ensuite nous avons scindé notre base de données en deux bases de données : une base de données pour les variables numériques et une autre pour les variables binaires et catégoriques. Par la suite, pour pouvoir utiliser les variables catégoriques et binaires, nous avons utilisé la fonction *get_dummies* de la librairie *pandas* sur *python* pour transformer ces dernières en colonnes binaires, qui sont égales à 1 si l'attribut appartient à la classe et 0 sinon (par exemple si la colonne Y_j peut prendre trois valeurs {a, b, c}, Y_j est transformée en 3 colonnes binaires $\{Y_a, Y_b, Y_c\}$ où :

$$\forall (x_{ij}, x_{ia}) \in Y_j \times Y_a,\ x_{ia} = \begin{cases} 1 & \text{si} \quad x_{ij} = a \\ 0 & \text{sinon} \end{cases}$$

Les variables numériques contenant toujours des valeurs NaN, nous ne pouvons pas nous contenter de supprimer les lignes contenant les valeurs manquantes car notre base de données serait réduite à la base de données vide. Plusieurs méthodes permettent de pallier ce problème de données manquantes consistant à remplacer les cellules vides par plusieurs valeurs (notamment: la moyenne, le maximum, le minimum, zéro). Nous allons par la suite essayer les différentes méthodes pour voir laquelle est la plus efficace pour nos données numériques. Nos variables se résument alors à 392 variables numériques et 534 variables catégoriques. Nous avons choisi 4 modèles différents pour en étudier les performances sur notre problème de classification binaire.

4 - Variables numériques

Avant de commencer, et pour choisir quelle méthode nous allons utiliser pour remplacer les valeurs vides. Cette hypothèse que nous prendrons sera forte puisqu'elle impactera tout notre jeu de données. Ainsi, nous avons effectué une validation croisée à 10 passes avec tous les modèles de notre étude sur l'ensemble de nos variables numériques afin de sélectionner la méthode de remplacement la plus efficace.

Algorithmes	Max		Min		Moyenne		Zero	
	Précision	Rappel	Précision	Rappel	Précision	Rappel	Précision	Rappel
Régression logistique	73%	20%	70%	27%	75%	21%	78%	37%
KNN	54%	33%	60%	34%	60%	38%	57%	37%
Arbre de décision	64%	61%	67%	61%	66%	61%	64%	59%
Random Forest	81%	65%	75%	59%	76%	62%	80%	60%
Moyenne	**68%**	**45%**	**68%**	**45%**	**69%**	**46%**	**70%**	**48%**

On peut voir que la méthode qui a la meilleure moyenne de précision et de rappel sur les 4 algorithmes, est celle du remplissage par zéro. Dans la suite de notre étude, nous décomposerons de manière aléatoire notre jeu de données en deux parties, un jeu de données d'entraînement sur lequel nous optimiserons l'erreur empirique et un jeu de données test pour pouvoir faire une validation a la fin de notre étude nous permettant de voir si notre modèle se généralise bien sur de nouvelles données.

4.1 - PCA

Après l'exécution d'une PCA avec 10 composantes sur toutes les variables, nous avons pu voir que seulement 2 composantes expliquent 82% de la variance. Nous avons alors effectué une PCA avec 2 composantes, et nous allons essayer de tester les différents algorithmes avec les 2 composantes. La **Figure 1** nous montre que le PCA ne nous permet pas d'avoir une frontière entre les deux classes. Ainsi, l'utilisation de ces composantes pour la prédiction des classes, ne pourra pas donner de bons résultats.

Nous allons calculer le f1-score (`F1 = 2 * (précision * rappel) / (précision + rappel)`) en utilisant les 2 composantes PCA pour les différents algorithmes, et comparer avec le f1-score en utilisant toutes les variables.

Algorithmes	f1-score PCA	f1-score toutes les variables
Régression logistique	50%	50%
KNN	45%	45%
Arbre de décision	64%	63%
Random Forest	70%	69%

En utilisant les 2 composantes PCA, on peut avoir à peu près la même performance qu'avec toutes les variables. Nous allons donc essayer d'autres méthodes.

4.2 - PLS Discriminant Analysis

Nous allons utiliser un PLS avec différents nombres de composantes, et tester la performance de notre modèle pour les différents nombres de composantes (**Figure 2**). Avec 5 composantes, on remarque que la précision et le rappel se stabilisent. Ce modèle nous donne une précision de 71% et un rappel de 45%.

4.3 - ANOVA

Pour le choix de variables, nous allons tout d'abord essayer de faire un choix en utilisant un test ANOVA (Analyse de la variance) et choisir les variables avec les k meilleures F-Values. Nous allons tester nos différents algorithmes avec différentes valeurs k. Nous essaierons de trouver le nombre de variables où la précision et le rappel se stabilisent. On remarque sur la **Figure 3** qu'en prenant les 30 premières variables avec Random Forest et Decision Tree, nous arrivons à un niveau stable de f1-score, tandis qu'on a besoin de plus de variables pour se stabiliser pour les autres algorithmes (Nous avons zoomé sur les 100 premières variables).

Vu que le f1 score arrive un maximum pour le Random Forest, nous allons refaire le test en calculant la précision et rappel de cet algorithme, en zoomant sur les 30 premières variables. On observe sur la **Figure 4** que la précision et rappel arrivent à leurs maximum et se stabilise en prenant les k = 14. On remarque qu'avec un nombre réduit de variables, nous avons pu avoir à peu près les mêmes précisions et rappels pour les données d'apprentissage et de test, en utilisant l'algorithme de Random Forest. En utilisant toutes les variables, la précision pour les données d'apprentissage était de 80%, et est passée à 81% en utilisant 14 variables. Par rapport au rappel, il était de 77% en utilisant toutes les variables numériques, et est resté au même niveau en utilisant 14 variables. Concernant les données de test, on peut voir qu'en utilisant 14 variables, on peut avoir presque les mêmes rappels et précisions. (voir **Figure 5**

et **Figure 6**). Parmi les variables que le test ANOVA retourne en premier, on retrouve "meds_cardiovascular", qui est le nombre de médicaments cardiovasculaires prescrits pour le patient. Si ce nombre est élevé, le patient a plus de chance d'être admis aux urgences. Une 2eme variable est: "triage_vital_dbp" qui représente les pulsations cardiaques lors de l'arrivée du patient. Si les pulsations sont élevées, le patient a encore plus de chance d'être admis. D'autres variables sont la température du patient, la pression artérielle, la fréquence respiratoire, ainsi que d'autres métriques relevées à l'arrivée du patient, et d'autres variables liées à des maladies chroniques. Une autre variable qui est classée parmi les 14 premières variables par le test ANOVA, est le nombre d'admissions pour l'année précédente, qui quand il est élevé, le patient a plus de chance d'être admis. On peut conclure que les patients qui sont dans un état grave, reflété par les métriques mesurées à l'arrivée, ou les patients ayant des maladies chroniques et ceux qui ont déjà été admis plusieurs fois aux urgences, ont plus de chance d'être admis.

Le test ANOVA nous a donc permis de réduire le nombre de paramètres numériques utilisés, et donc d'avoir une meilleure interprétation des résultats, et un temps d'exécution beaucoup plus rapide. En même temps, nous avons une précision supérieure à 80% et un rappel supérieur à 60%, avec une bonne généralisation sur les données de validation. Par la suite, nous allons faire une étude de variables catégoriques.

5 - Variables Catégoriques

5.1 - Chi2

Nous avons étudié les variables catégoriques et numériques à part étant donné les problématiques liées aux données manquantes. Nos 511 variables catégoriques se décomposent en 13 variables multi-classes et 498 variables binaires. Nous avons effectué une transformation des variables multi-classes en plusieurs variables binaires puis les avons regroupé avec les variables binaires. Nous obtenons 534 variables binaires à l'issue de cette opération. Nous effectuons alors un test de Chi2 avec comme hypothèse nulle H0: "Les variables X_j et y sont indépendantes". Nous obtenons les p-values classées par ordre décroissant indiquées sur la **Figure 7**.

En commençant avec la variables avec la plus petites p-value, pour k paramètres on considère un vecteur augmenté des k-1 paramètres avec des p-values plus faible et on effectue une régression logistique avec validation croisée à 10 passes afin d'observer les performance du modèle à l'aide des trois métriques suivantes (précision, rappel et f1-score). On observe les résultats sur la **Figure 8** et on remarque que le score associé aux trois métriques diminue quand le nombre de paramètres s'agrandit (k>20). En "zoomant" sur les 20 premiers

paramètres avec la Régression Logistique, on obtient un score optimal pour les trois métriques pour k=16, soit les paramètres suivants: ['esi_4', 'previousdispo_Admit', 'employstatus_Retired', 'esi_2', 'arrivalmode_ambulance', 'insurance_status_Medicare', 'cc_shortnessofbreath', 'dep_name_C', 'o2_device_max', 'copd', 'employstatus_Full Time', 'htn', 'coronathero', 'maritalstatus_Widowed', 'chrkidneydisease', 'maritalstatus_Single'].

Parmi ces différents paramètres, on remarque que la variable esi revient deux fois (esi_4 et esi_2) qui est l'Emergency Severity Index, indiquant l'indice de gravité de l'urgence, ce qui semble cohérent avec l'admission du patient à l'hôpital. On remarque également que si le patient a été déjà admis à l'hôpital par le passé, il a de forte chance d'être encore admis lors de sa nouvelle visite. Notons également que si le patient a été conduit à l'hôpital par voie ambulancière, il a également plus de chance d'être pris en charge. De plus, si le patient a une faiblesse respiratoire, une hyperoxie (teneur anormalement élevée en dioxygène), de l'hypertension, une maladie cardiaque coronarienne, une maladie respiratoire obstructive chronique ou encore une maladie rénale chronique, celui-ci a de grandes chances d'être pris en charge. Enfin il est intéressant de remarquer que le statut marital ou celui de l'employé influence également sa prise en charge. Ainsi si le patient est veuf ou célibataire, retraité ou employé à temps-plein, il a plus de chance d'être admis à l'hôpital, on suppose que cela peut refléter son état de santé et sa capacité à se déplacer à l'hôpital (si le patient est tout seul, cela a plus de chance de signifier que son état est critique).

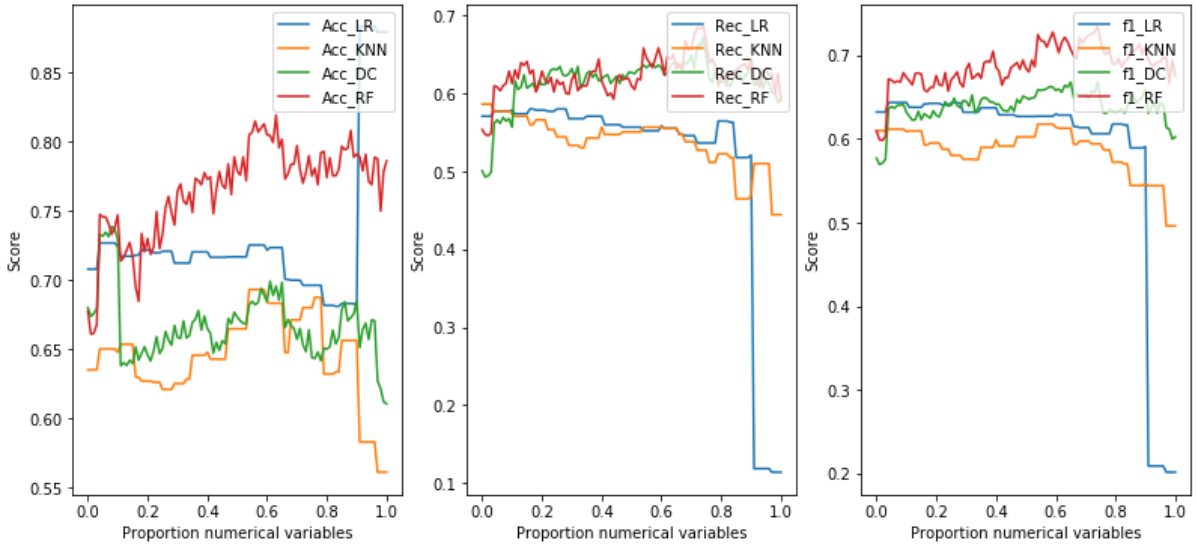
5.2 - Multi-Component-Analysis

Nous effectuons maintenant une Multi-Component-Analysis (MCA) afin de voir si nous parvenons à obtenir une meilleure performance de notre modèle de régression logistique. La MCA est comparable à la Principal Component Analysis (PCA) mais s'effectue sur des variables catégoriques. Un de ses inconvénients toutefois est le manque de visibilité des paramètres retenus (qui sont une combinaison linéaire des paramètres). La projection des variables sur les deux premières composantes du MCA est indiquée sur la **Figure 9**. En augmentant progressivement le nombre de composantes retenues, on obtient le graphique de la **Figure 10**. On observe de meilleurs résultats pour un nombre de facteurs égal à 100 qu'avec notre approche initiale, cependant nous recherchions un nombre plus faible de facteurs permettant une analyse des paramètres combinaison de ces derniers. Ce nombre élevé ne nous permet pas de différencier quels paramètres pourrait être à l'origine d'une prise en charge ou non du patient à l'hôpital et de pouvoir expliquer nos résultats.Nous décidons donc de garder nos 16 paramètres retenus à l'issu du test du Chi2 et de les ajouter aux paramètres numériques retenus dans la première partie et d'effectuer une comparaison des métriques pour plusieurs algorithmes différents.

6 - Résultats et conclusion :

Nous retenons donc les paramètres numériques et catégoriques des deux parties. Nous avons au total 16 paramètres catégoriques et 14 numériques. Afin de déterminer quelle proportion de paramètres numériques (resp. catégoriques) conserver pour nos prédictions finales, nous faisons varier la proportion de nos paramètres numériques (soit p la proportion de nos paramètres numériques alors la proportion de nos paramètres catégoriques est de 1-p). Nous évaluons ensuite la performance de nos différents algorithmes sur les métriques précision/rappel/f1-score comme le montre la figure ci-dessous:

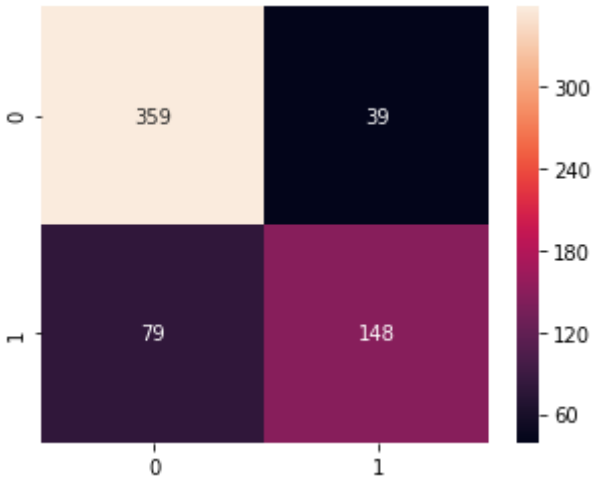
Figure 12: Performance du corpus d'algorithmes sur la répartition entre variables numériques et catégoriques.



On observe que la valeur p = 0.6 pour nos paramètres numériques est optimale, soit 6 paramètres catégoriques ainsi que 8 paramètres numériques sont finalement retenus. On remarque également que l'algorithme de Random Forest est le plus performant en termes de précision et f1-score et sensiblement identique en termes de rappel avec l'algorithme de Decision Tree. Notons également que le KNN sous-performe par rapport aux autres algorithmes sur l'ensemble des métriques. La Régression Logistique fournit des résultats honorables en moyenne mais n'est plus utilisable dès que le nombre de paramètres catégoriques devient trop petit.

Effectuons maintenant la validation finale de notre modèle retenu (régression logistique) avec nos paramètres finaux retenus (8 catégoriques et 7 numériques).

Figure 13 : Matrice de confusion de notre modèle de Random Forest sur notre jeu de données test.



Cette étude nous a permis de sélectionner les variables expliquant au mieux notre variable d'intérêt du problème et de sélectionner l'algorithme le plus performant de l'éventail choisi (Régression Logistique, KNN, Decision Tree et Random Forest). On retient donc finalement les 17 paramètres suivants :

Variables Catégoriques	Variables numériques
'esi_4'	'meds_cardiovascular'
'previousdispo_Admit'	'triage_vital_dbp'
'employstatus_Retired'	'triage_vital_temp'
'arrivalmode_ambulance'	'triage_vital_sbp'
'insurance_status_Medicare'	'meds_diuretics'
'esi_2'	'triage_vital_hr'
	'triage_vital_rr'
	'meds_gastrointestinal'

Annexe 1: Approche PCA/PLS sur les variables numériques

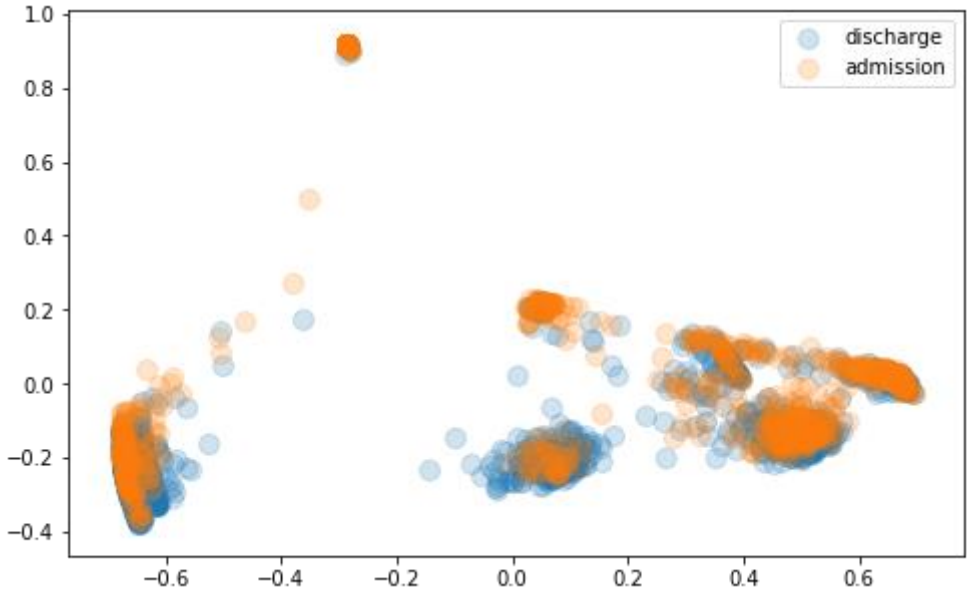


Figure 1. Matrices de confusion avec PCA

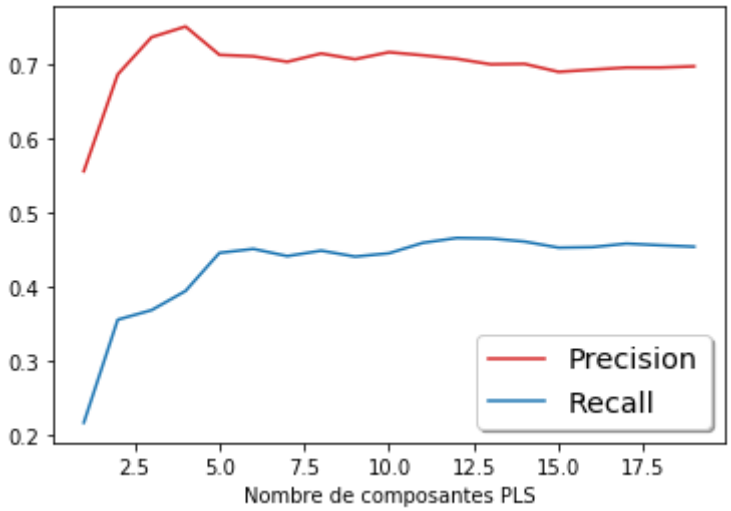


Figure 2. Précision et rappel du PLS en fonction du nombre de composantes principales

Annexe 2: Approche ANOVA sur les variables numériques

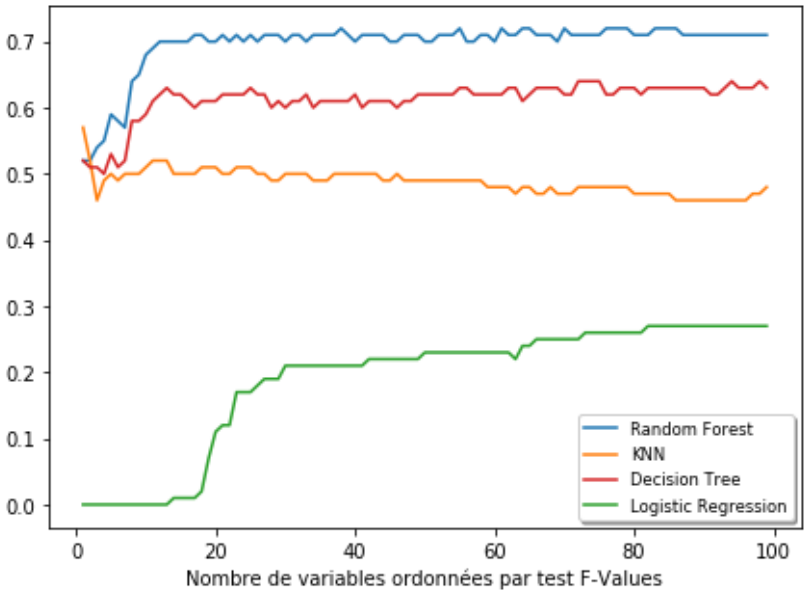


Figure 3: F1- score en fonction des nombres de variables

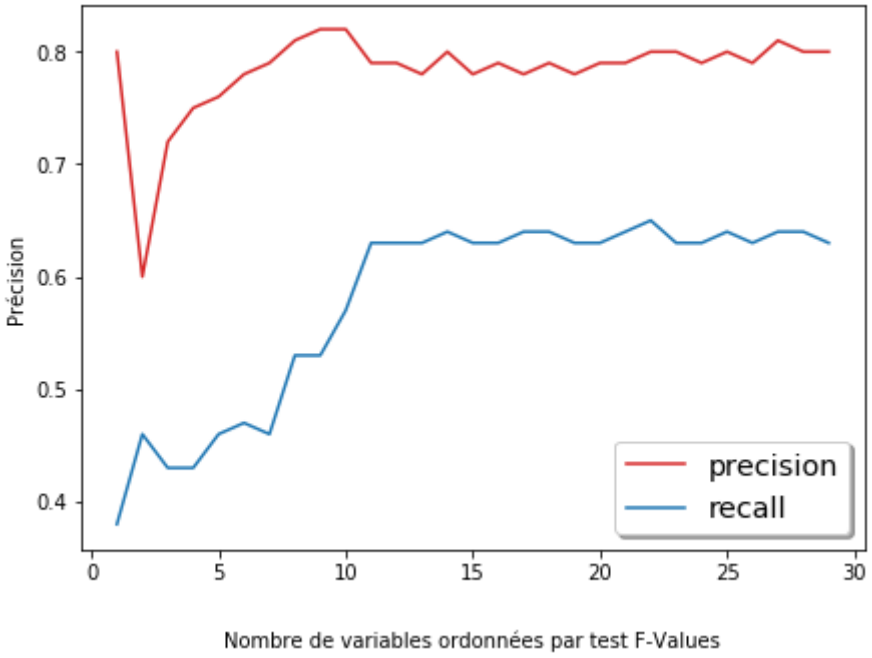


Figure 4: Précision et rappel score en fonction des nombres de valeurs

Annexe 3: Approche ANOVA sur les variables numériques

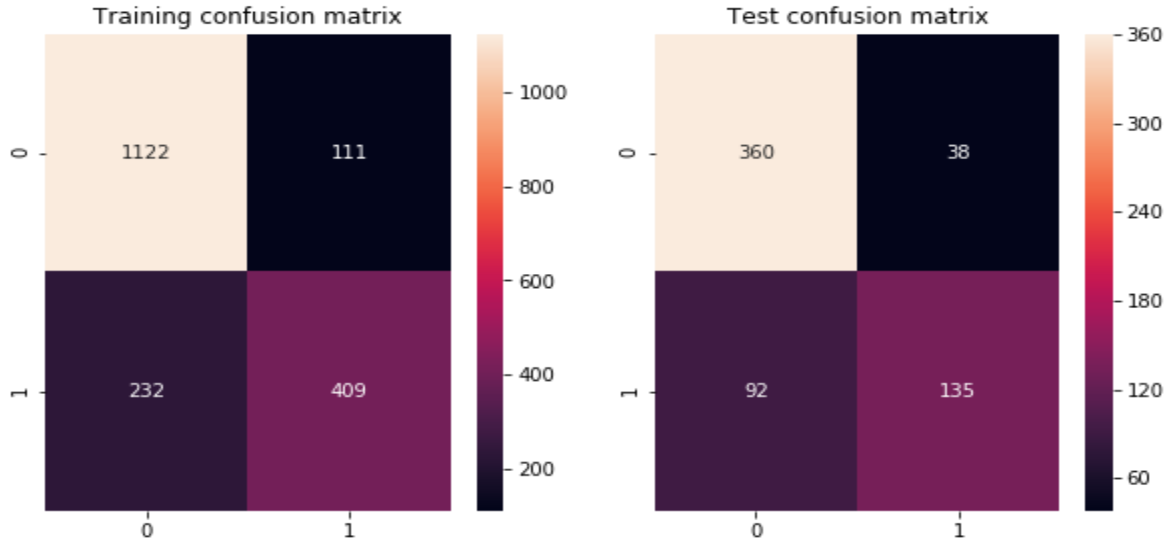


Figure 5: Matrices de confusion avec 14 variables numériques

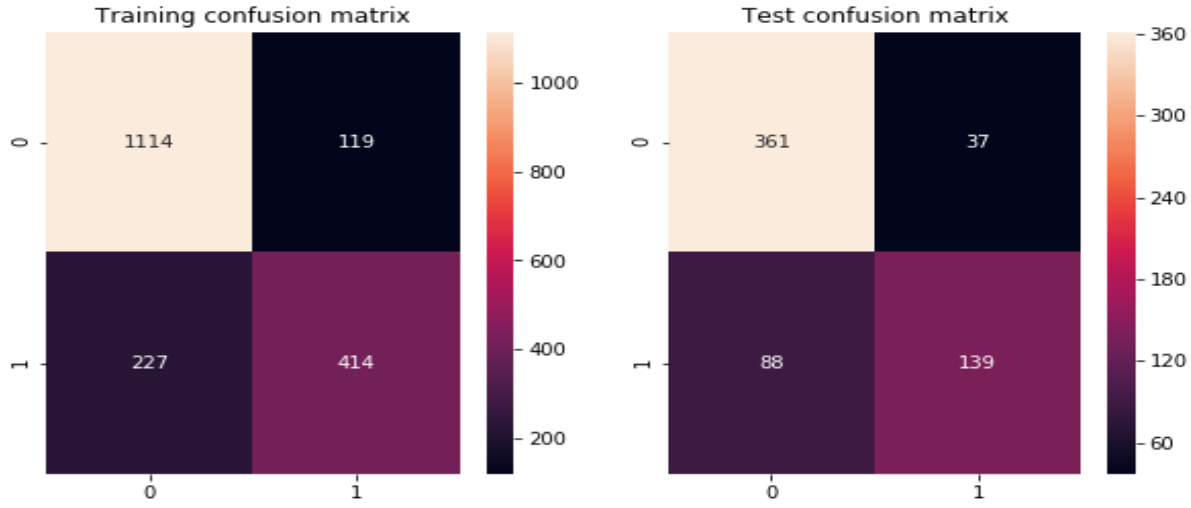


Figure 6: Matrices de confusion avec toutes les variables numériques

Annexe 3: Approche Chi2 sur les variables catégoriques

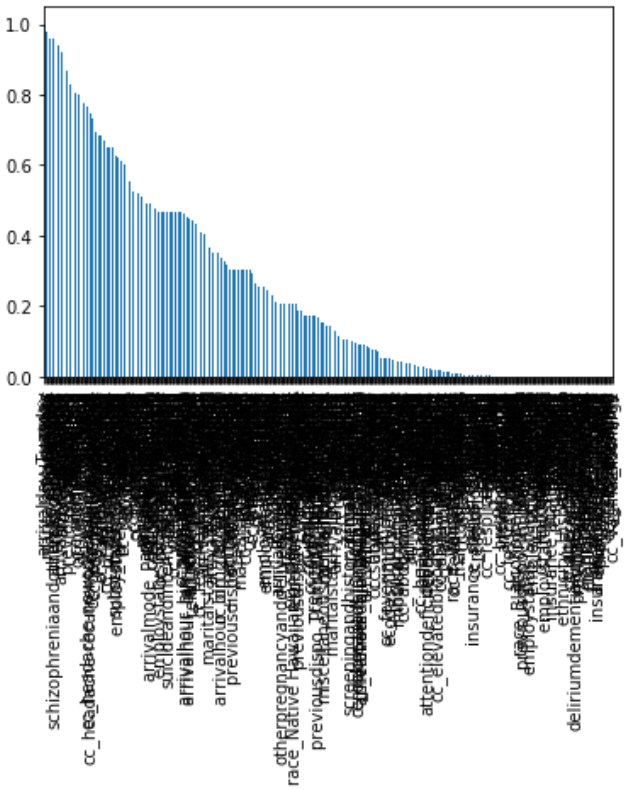


Figure 7: P-values obtenues par paramètre catégorique

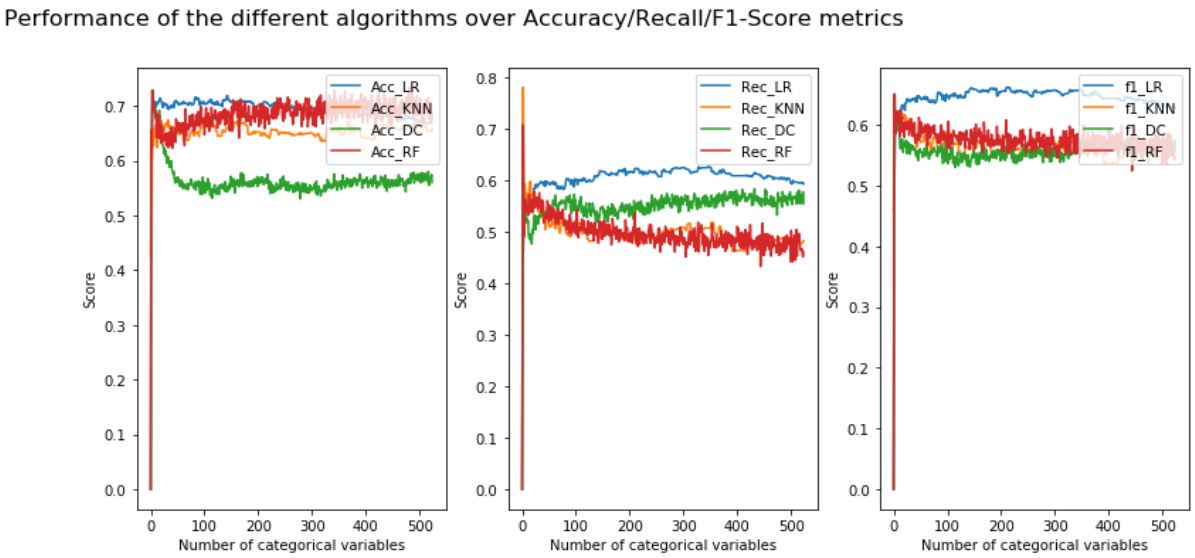


Figure 8: Précision, Rappel et F1-score pour un nombre de paramètres croissant

Annexe 4: Approche MCA sur les variables catégoriques

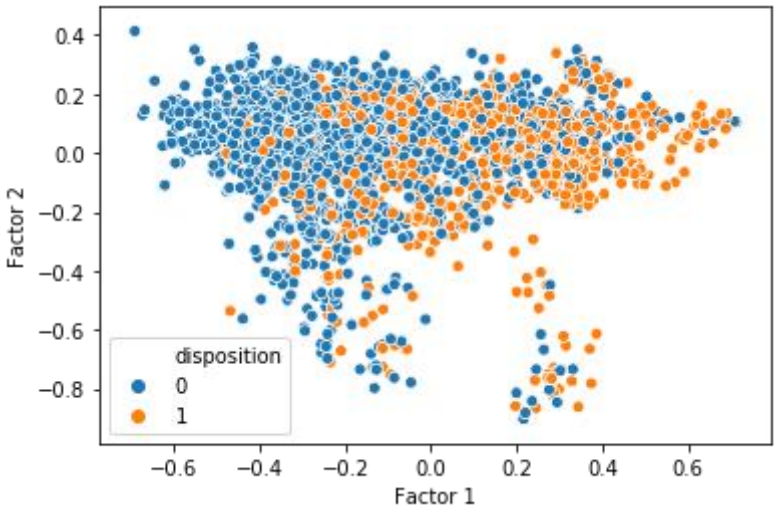


Figure 9: Représentation des paramètres projetés sur deux composantes du MCA

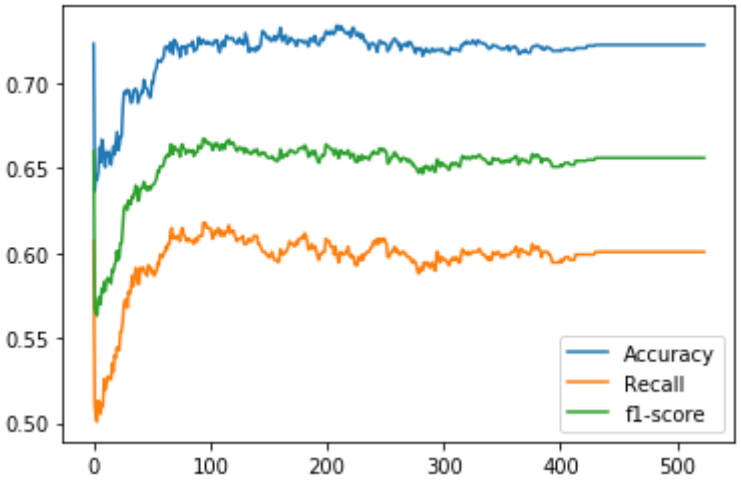


Figure 10 : Précision, Rappel et F1-score pour un nombre de facteurs croissant du MCA