Journal of Economics and Applied Statistics
Volume 17 number 3, December 2020
ISSN : 1112-234X EISSN : 2600-664
Pages : 84-100

STATISTICAL LEARNING FOR KNOWLEDGE DISCOVERY FROM HEALTH DATA: ESSAY ON COLORECTAL CANCER DATASET

APPRENTISSAGE STATISTIQUE POUR L'EXTRACTION DE CONNAISSANCES A PARTIR DE DONNEES SANITAIRES : ESSAI SUR DONNEES DU CANCER COLORECTAL

* **Dalia ATIF**
Université de Tipasa- institut d'économie, commerce et gestion
atif.dalia@cu-tipaza.dz

Rachid BENAMIROUCHE
Ecole Nationale Supérieure de Statistique et d'Economie Appliquée ENSSEA
rbena2002@hotmail.com

Reçu le: 2020/02/26 **Accepté le :** 2020/07/05 **Publication en ligne le**: 2020/12/31

ABSTRACT: Hospital information systems store increasingly large and heterogeneous data volumes, the one of them is the clinical information system of colorectal cancer (CRC), which is focused on the patient's folder, we were interested in this work to the knowledge discovery from this informational environment, in the form of prognostic factors acting on the recurrence of the disease. We worked for that on a sample of patients with histologically proven CRC. Several steps were then necessary for the KDD process, according to the biomedical data characteristics: the treatment of missing values, the discretization of continuous variables, the preselecting of variables and the rebalancing of classes. The constructed model exhibited excellent validation performance, with superior sensitivity to specificity.

Keywords: data, knowledge, CRC, recurrence, IS.

JEL Classification: C51, C52, I10.

RESUME : Les systèmes d'information hospitaliers (SIH) stockent des volumes de données de plus en plus importants et hétérogènes, parmi ces derniers figure le SI clinique du cancer colorectal (CCR) qui est centré sur le dossier du malade ; nous nous sommes intéressés dans ce travail à l'extraction de connaissances (ECD) de cet environnement informationnel, sous forme de facteurs pronostiques agissant réellement sur la récidive de la maladie. Nous avons travaillé pour cela sur un échantillon de patients avec un CCR histologiquement prouvé. Plusieurs étapes furent alors nécessaires au processus ECD, selon les caractéristiques spécifiques aux données biomédicales à savoir : le traitement des valeurs manquantes, la discrétisation des variables continues, la présélection des variables et le rééquilibrage des classes. Le modèle construit a présenté d'excellentes performances en validation avec une sensibilité supérieure à la spécificité.

Mots clés : donnée, connaissance, CCR, récidive, SI.

* Auteur Correspondant

1. INTRODUCTION :

Nous nous intéressons dans ce travail au processus d'extraction de connaissances dans le domaine sanitaire, cette discipline apparue que très récemment suite au développement des capacités de stockage et de calcul, a vu son champ d'application s'étendre à plusieurs domaines d'applications telles que : la gestion de la relation client, la détection de fraudes, le contrôle de production…etc. Partant de ce principe tous les secteurs ont intérêt à valoriser leurs données dans une stratégie d'aide à la prise de décision. Parmi tous ces secteurs, le domaine sanitaire constitue un environnement informationnel riche en données hétérogènes, qui nécessitent des outils de traitement sophistiqués ; afin de pouvoir les transformer en connaissances utiles.

Le but étant de présenter une classification pronostique des malades avec un CCR histologiquement prouvé, classés stade III ou IV selon la classification TNM, et qui guiderait le clinicien dans l'optimisation de la prise en charge thérapeutique du malade, dans une perspective de rationalisation des dépenses de santé. Pour cela nous avons dégagé les étapes essentielles au processus ECD par rapport aux propriétés des données biomédicales. En outre, l'imputation des valeurs manquantes et l'apprentissage en présences de classes déséquilibrées constituent les principales démarches à entreprendre pour l'établissement d'un modèle prédictif

Beaucoup de travaux ont été menés sur les différents problèmes liés aux données biomédicales mais séparément, on peut citer (Buuren, 2007), (Cottrell and al, 2009), (Kouadjo and al, 2013), (Lee and Carlin, 2010) pour l'imputation des valeurs manquantes, et (Lee, 2000), (Menardi and Torelli, 2012) pour le rééquilibrage et la subdivision optimale de l'échantillon en présence de classes déséquilibrées. Nous proposons dans ce papier de traiter ces deux problèmes simultanément, afin d'aboutir à des connaissances utiles et de qualité.

La suite de l'article est organisée comme suit : la partie 2 évoque le cadre théorique associé aux données manquantes, la partie 3 présente la méthodologie de l'imputation multiple, la partie 4 le rééquilibrage des classes, la partie 5 illustre l'application empirique entreprise dans ce travail et enfin les parties 6 et 7 synthétisent les principaux résultats issus de ce travail.

2. LE CADRE THEORIQUE:

La typologie des mécanismes qui gouvernent les valeurs manquantes revient à (Little and Rubin, 2002) avant d'énoncer cette classification quelques notations sont nécessaires :

On note le jeu de données incomplet par $X=(X^{obs},X^{mis})$ avec une partie observée X^{obs} et une partie manquante X^{mis}, où $X=(Y,Z_1,Z_2,...,Z_k)$ est constitué par un vecteur de variables complètes Y et $(Z_1,Z_2,...,Z_k)$ un vecteur de variables incomplètes, et on note R le mécanisme de réponse, ce mécanisme est représentée par une matrice d'indicatrices $(R_{i,j} / j=1,..,p)$, avec $R_{ij}=1$ si la valeur de l'individu i $(i=1,..,n)$ est manquante, et 0 sinon. Ce mécanisme est supposé avoir une distribution dépendante de X et d'un vecteur de paramètres ϕ, caractérisé par une distribution conditionnelle $p(R/X,\phi)$

2.1. Le mécanisme MCAR (missing completly at random) :

$p(R/X,\phi)=p(R/\phi)$, dans ce cas les données manquent d'une manière totalement aléatoire, on rencontre cette situation dans le cas de la perte d'une partie de l'échantillon ou bien que ce dernier soit inexploitable pour l'étude envisagée.

2.2. Le mécanisme MNAR (missing not at random) :

$P(R/X,\phi)$, signifie que le mécanisme peut dépendre des valeurs observées et non observées (données manquantes et variables non recueillies) et donc d'une information non disponible.

2.3. Le mécanisme MAR (missing at random) :

$p(R/X,\phi)=p(R/X^{obs},\phi)$, ici le mécanisme dépend des valeurs observées (les valeurs des variables complètes plus les valeurs observées des variables incomplètes), c'est-à-dire de l'information disponible. Il est impossible de vérifier si la condition MAR est satisfaite et la raison de manière intuitive est claire. Comme nous ne connaissons pas les valeurs des données manquantes, nous ne pouvons pas faire de comparaison entre le fait que le mécanisme dépende des valeurs manquantes ou bien celles qui sont observées (MAR vs MNAR).

2.3.1. Le mécanisme ignorable :

L'ignorabilité signifie que le processus d'estimation peut *ignorer* le mécanisme de réponse, ce dernier est considéré comme étant ignorable si:

(a) Il est de type MAR ;

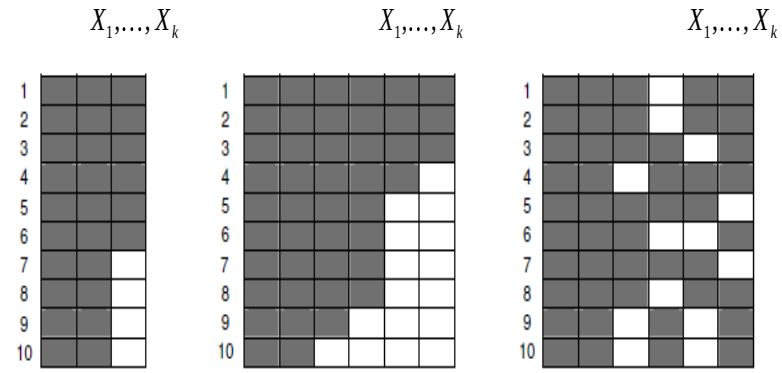
(b) Les paramètres qui régissent le mécanisme de réponse ϕ sont indépendants des paramètres d'intérêt θ. (Allison, 2001).

La vraisemblance des données observées s'écrit dans ce cas

$$L(\theta,\phi / X^{obs},R)=\prod_{i=1}^{n} p(R_i / X_i^{obs},\phi)\times p(X_i^{obs} / \theta)$$

Ce principe indique qu'il n'est pas nécessaire de modéliser le mécanisme de réponse comme une partie du processus d'estimation. Selon (Allison, 2001), la condition (b) est toujours satisfaite dans la réalité, ce qui rend les deux notions de (mécanisme MAR et mécanisme Ignorable) comme étant équivalentes. Lorsque le mécanisme des données manquantes s'avère Ignorable avec un arrangement des données arbitraire (non-monotone), le recours aux méthodes itératives dites d'imputation multiple reste essentiel.

Figure N°1: **la structure des données manquantes.**



Source : (Buuren, 2007).

3. L'IMPUTATION MULTIPLE :

L'idée de base de cette méthodologie est d'imputer chaque donnée manquante par M valeurs différentes, celle-ci maintient l'incertitude liée à l'information indisponible, en considérant les paramètres et les données comme des variables aléatoires (Nakache, 2005); ainsi l'algorithme est divisé en deux étapes : l'imputation et l'analyse.

3.1. L'étape d'imputation :

L'étape d'imputation se déroule en deux étapes :

- Les données manquantes sont imputées par tirages dans les distributions les distributions conditionnelles prédictives $p(X^{mis} / X^{obs}, \psi)$;
- Tirage d'un vecteur de paramètres du modèle d'imputation dans la distribution à postériori $p(\psi / X^{obs})$.

Soit, on considère le schéma itératif suivant sous l'hypothèse d'un mécanisme de réponse ignorable. Etant donné la valeur courante des paramètres ψ^t au temps t :

(I) : on tire aléatoirement des valeurs dans les distributions conditionnelles prédictives $X^{mis,(t+1)} \sim p(X^{mis} / X^{obs}, \psi^{(t)})$.

(P) : on tire aléatoirement la valeur des paramètres dans la distribution à posteriori $\psi^{(t+1)} \sim p(\psi / X^{obs}, X^{mis,(t+1)})$, et on itère le schéma (I-P), à partir d'une valeur ψ^0 de départ pour aboutir à une chaine de Markov $(X^{mis,(t)}, \psi^{(t)}) / t = 1, 2, \ldots$ qui converge après un certain nombre d'itérations vers la distribution prédictive à postériori des données $p(X^{mis}, \psi / X^{obs})$ (Shafer, 2000). On obtient par la suite M jeux de données complets en construisant soit une chaine simple ou une chaine multiple.

Cependant, les méthodes itératives se distinguent selon deux approches : la première selon la spécification d'une loi jointe (JM, Joint Modelling), et la seconde selon la spécification de distributions conditionnelles appelée aussi imputation par régression séquentielle (FCS, Fully Conditionnal Spécification). Nous nous restreindrons par la suite de détailler la première approche, car même si elle présente des bases théoriques plus solides, le problème se pose lorsque la loi jointe n'est pas valide, aussi nous renvoyons le lecteur intéressé à l'ouvrage de (Shafer, 2000) pour plus de détails. La seconde approche est plus flexible, car elle se restreint de faire des suppositions sur la loi jointe, nommée aussi méthode MICE (Multivariate Imputation by Chained Equations), elle permet de générer des imputations variable par variable par équations chainées, en spécifiant pour chaque variable incomplète une distribution conditionnellement aux autres variables $Z_j^t \sim p(Z_j / Y, Z_{-j}^{t-1}, \psi_j^t)$ et met donc en œuvre l'algorithme de Gibbs, qui converge vers la distribution jointe que l'on n'a pas eu à spécifier, et permet donc de ramener un problème à k dimensions à k problèmes univariés.

Soit l'itération n° t :

$Z_1^t \sim p(Z_1 / Y, Z_2^{t-1}, Z_3^{t-1}, ..., Z_k^{t-1}, \psi_1^t)$

$Z_2^t \sim p(Z_2 / Y, Z_1^t, Z_3^{t-1}, ..., Z_k^{t-1}, \psi_2^t)$

.

.

.

$Z_k^t \sim p(Z_k / Y, Z_1^t, Z_2^t, ..., Z_{k-1}^t, \psi_k^t)$

3.1.1. Predictive Mean Matching (PMM) **:**

Est une autre manière d'opérer l'imputation multiple, en produisant des valeurs réalistes selon l'hypothèse suivante : la valeur manquante appartient à un ensemble de candidats issus de la même distribution. L'approche suivie est ainsi la même que celle qui vient d'être décrite, la différence apparait une fois que l'on a tiré aléatoirement, la valeur des paramètres dans la distribution à postériori ; on génère ensuite des valeurs pour l'ensemble des observations sur la variable incomplète Z_j (valeurs observées et manquantes), on constitue après un groupe de candidats qui ont des valeurs proches pour cette valeur manquante, un tirage aléatoire est alors effectué parmi ces candidats, et la valeur observée de ce candidat est utilisée pour imputer la valeur manquante. Cette méthode est le plus souvent utilisée pour l'imputation des variables quantitatives anormalement distribuées, car elle génère des valeurs qui ressemblent aux valeurs réelles, ainsi si la variable est asymétrique, les valeurs imputées le seront également.

3.2. L'étape d'analyse :

Une fois l'étape d'imputation terminée, on procède à des calculs simples prescrits au départ par (Little and Rubin, 2002) et qui permettent d'obtenir des paramètres combinés. Soit $\hat{\theta}$ l'estimateur de θ (moyenne, coefficients de régression…etc) et $\hat{U}$ l'estimateur de sa variance, pour M imputations on obtient :

$$\hat{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m$$

Quand à la variance de l'estimateur, elle est décomposée en deux parties : la variance intra-imputation et inter-imputation :

$$V_{\text{intra}} = \frac{1}{M}\sum_{m=1}^{M}\hat{U}_m$$

Et

$$V_{\text{inter}} = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta})(\hat{\theta}_m - \hat{\theta})'$$

La variance totale est alors :

$$\hat{U} = V_{\text{intra}} + (1+\frac{1}{M})V_{\text{inter}}$$

De cette manière, on peut obtenir des intervalles de confiance pour le paramètre θ $IC = \hat{\theta} \pm t_\nu \sqrt{\hat{U}}$: où t_ν est un quantile de la distribution de Student à υ degré de liberté :

$$\nu = (M-1)(1+\frac{M.V_{\text{intra}}}{(M+1).V_{\text{inter}}})^2$$

3.3. Le nombre d'itérations nécessaires :

Le nombre d'itérations nécessaires à la convergence de l'algorithme MICE, est un nombre réduit (entre 5 et 10) et est de ce fait beaucoup moins important que celui qui est nécessaire dans la majorité des techniques MCMC (milliers d'itérations). Cette convergence rapide est due à la procédure univariée de l'imputation qui génère des imputations statistiquement indépendantes, pour une valeur donnée des paramètres de régression. (Buuren, 2011)

3.4. Le nombre de jeux de données nécessaires:

La littérature disponible s'accorde sur un nombre M de 3 à 5, et cela en s'appuyant sur un critère d'efficacité de l'estimation de la quantité θ qui est basée sur M imputations, nommé à la base par Rubin par efficacité relative, elle est égale à $RE = (1+\frac{\lambda}{M})^{-1}$, où le paramètre $\lambda = \frac{(1+M^{-1}).V_{inter}}{V_{intra}}$ est la proportion de la variance totale de θ qui est due aux valeurs manquantes. Ultérieurement Schafer démontre dans des travaux empiriques (Shafer, 2000) que pour une valeur de λ=0.1 un nombre d'imputation $M = 5$ aboutit à une efficacité =98% et si l'on double le nombre d'imputations à 10, l'efficacité est alors de 99%, on obtient donc un gain qui est dérisoire par rapport à l'augmentation du nombre d'imputations. Cependant, il faut noter que le paramètre λ est un bon indicateur de la qualité d'estimation de θ, en fonction qu'il soit inférieure à la proportion de valeurs manquantes de chaque variable incomplète (Brand, 1999).

4. LE REEQUILIBRAGE DES CLASSES :

Ce cas de figure se présente très fréquemment dans le cas des données biomédicales, où la proportion des deux classes (positive et négative) n'est pas équilibrée ; on parle alors de données déséquilibrées (unbalanced data). On s'intéresse plus précisément à ce problème car il altère les performances de la classification; pour résoudre ce problème plusieurs approches existent dans la littérature : la première consiste à sur-échantillonner la classe minoritaire par un tirage aléatoire avec remise (over-sampling), son principal désavantage est la duplication des observations (création de doublons) et donc l'absence d'information supplémentaire; la seconde stratégie consiste à sous-échantillonner la classe majoritaire induisant à une perte d'information potentiellement utile (profils rares) ; la dernière stratégie est la plus intelligente, au sens où au lieu de dupliquer des observations de la classe minoritaire, on génère des données synthétiques des deux classes Y_c selon le principe du smooth bootstrap. Nous tirons pour cela de l'ensemble d'apprentissage une observation appartenant à l'une des deux classes (choisies en donnant la même probabilité aux deux classes) et on génère une nouvelle observation dans son voisinage, un voisinage, qui est déterminé par les contours du noyau et sa largeur est régie par le paramètre de lissage. La mise en œuvre de cette stratégie se déroule selon les étapes suivantes :

- Choix aléatoire uniforme de la classe ;
- Choix aléatoire uniforme de l'observation de la classe en question ;
- Choix d'un point proche de l'observation.

Répéter l'opération d fois permet la création d'un nouvel ensemble d'apprentissage synthétique (Lunardon and al, 2014) d'une taille d où approximativement le même nombre d'observations appartient aux deux classes. La taille d peut être définie sur la taille n du jeu d'apprentissage original ou choisie arbitrairement. Cette démarche s'apparente à la modélisation par noyau de la densité, qui consiste à choisir une fonction noyau et un paramètre de lissage ; qui se doit d'être suffisamment faible pour éviter la perte de détails. Cette technique appelée ROSE (Random Over Sampling Examples) aide ainsi le classifieur à estimer une règle de classification plus précise, car la même attention sera portée aux deux classes. La procédure qui vient d'être décrite est fournie par la fonction rose.real() du package ROSE sur R ; où le noyau utilisé est gaussien, et avec une adaptation naïve au cas nominal avec conservation des valeurs.

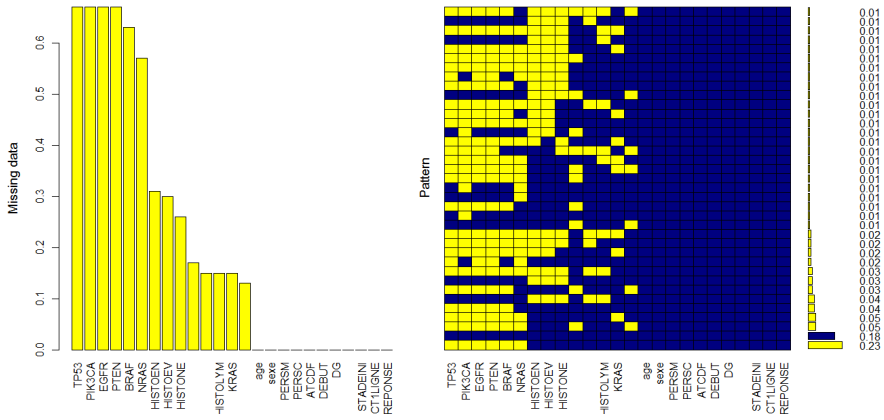
5. METHODES ET MATERIELS :

L'étude empirique a porté sur 100 dossiers médicaux s'étendant entre Novembre 2011 jusqu'à Aout 2018. Il s'agit donc d'une étude rétrospective dans la région de l'Algérois, chez des patients atteints de CCR localement avancés et/ ou métastatiques opérés ou non, classés stade III ou IV selon la classification TNM ; englobant des données cliniques (sexe, âge, antécédents personnels médicaux de CCR ou d'adénome ou de maladie inflammatoire chronique de l'intestin **PERSM**, antécédents personnels chirurgicaux de néoplasie **PERSC** , antécédents familiaux de néoplasie **ATCDF**, siège de la tumeur **DG**, délai entre l'apparition des premiers symptômes et la date de diagnostic **DEBUT**, stade initial **STADEINI**), des données anatomopathologiques (grade histologique **DIAGNOSTIC**, type histologique **HISTOTYPE**, présence d'emboles vasculaires **HISTOEV**, présence d'engainements périnerveux **HISTOEN**, présence de mitose des cellules malignes **HISTOMITO**, présence de prolifération lymphocytaire **HISTOLYM,** présence de nécrose **HISTONE**), et données biologiques qui portent sur le dosage de l'antigène carcino-embryonnaire (**ACE**) et de l'antigène tumoral (**CA-199**), ainsi que les mutations des gènes intervenant dans les deux voies de signalisation (**KRAS**, **NRAS**, **BRAF**, **PIK3CA**, **TP53, PTEN, EGFR**).

5.1. Imputation des valeurs manquantes :

Sachant que le jeu de données présente des valeurs manquantes, il nous faut spécifier en premier lieu la structure et la typologie du mécanisme qui en est responsable. Une simple visualisation de la figure 2 nous permet d'affirmer une structure arbitraire. Ensuite, pour ce qui est de la typologie du mécanisme, on associe à chaque variable incomplète une indicatrice R_i (qui vaut 1 si la donnée manque et 0 sinon), nous devons ensuite croiser cette indicatrice avec chacune des autres variables, le lien statistique recherché est un test du khi deux significatif (p-value<0.05).

Figure N° 2: l'arrangement des données manquantes.



source : output de R fait par les auteurs.

Nous avons plusieurs croisements qui sont significatifs, le mécanisme n'est plus totalement aléatoire car il dépend d'une autre variable du jeu de données. Le constat est que la structure des données manquantes est arbitraire avec plusieurs variables incomplètes et selon un mécanisme aléatoire; l'apport des méthodes d'imputation multiple intervient donc dans cette situation. Il nous faut préciser que rien ne nous permet de réfuter un mécanisme non ignorable, même si l'hypothèse MAR n'est pas très réaliste, elle nous permet des simplifications numériques importantes.

Afin de rendre cette hypothèse encore plus réaliste, nous avons introduit toutes les variables lors du processus d'imputation sans tenir compte de l'analyse qui sera faite ultérieurement et cela selon le principe de (Buuren, 1999). L'arrangement des données s'avérant non-monotone, nous avons eu recours aux méthodes itératives et plus précisément à l'algorithme FCS, vu la nature des données biomédicales (présence de variables qualitatives), qui est fourni par la fonction mice () du package MICE sur R. Nous avons utilisé cette procédure automatisée pour générer 5 jeux de données complets, chacun après 50 itérations, afin de nous assurer de la stabilité des résultats, sachant que même si la procédure est automatisée, nous décidons d'utiliser la méthode PMM pour imputer les deux variables ACE et CA199 vu l'anormalité de leurs distributions. Quant au vecteur de paramètres du modèle d'imputation, le tirage aléatoire s'est fait dans la loi asymptotique de l'estimateur. Ensuite nous avons subdivisé chaque jeu de données imputé, à hauteur de

70%-30% pour « l'apprentissage-validation », selon une stratégie de stratification proportionnelle.

5.2. Rééquilibrage des classes :

Tous les patients ont été suivis avec un examen clinique, un scanner thoraco-abdomino-pelvien (ou une radiographie de thorax et une échographie abdominale), le dosage de l'ACE et des coloscopies de contrôle. La récidive de la maladie était évaluée tous les 3 mois les 2 premières années, puis tous les 6 mois les 3 années suivantes. Durant ce suivi, la présence ou l'absence de récidive était noté. La figure 3 illustre l'effectif de chaque classe et démontre que nous sommes en présence d'un apprentissage supervisé déséquilibré.

Figure N° 3 : la distribution des classes.

Source : fait par les auteurs.

Notre objectif est donc d'identifier les facteurs pronostiques influant sur le risque de récidive, Afin d'établir un classifieur avec précision, il nous faut d'abord rééquilibrer les classes, car le déséquilibre mène souvent aux problèmes de sur-apprentissage, avec des modèles qui collent à l'échantillon avec lequel ils ont été construits et donc à des modèles avec une faible capacité de généralisation. La solution adoptée dans notre travail est la génération de nouvelles observations synthétiques ; nous utilisons pour cela la fonction rose.real () seulement sur les 5 échantillons d'apprentissage (les 5 échantillons de validation ne sont pas rééquilibrés). À noter qu'une racine (seed) différente fut utilisée pour chaque échantillon d'apprentissage, afin que les classifieurs construits soient décorrélés entre eux.

5.3. Discrétisation des variables continues :

Nous avons discrétisé toutes les variables continues et cela pour plusieurs raisons :

- Rendre homogène les données ;
- Contourner le problème des odds ratios (OR) des variables continues, car ces derniers ne prennent pas en compte la non linéarité des liaisons ;

- Diminuer l'effet exagéré de certaines valeurs influentes.

La stratégie de discrétisation suivie pour les quatre variables continues fut non supervisée : les connaissances à priori d'un cancérologue nous ont permis de découper ces variables comme suit :

- ACE>5 et ACE<5, ce seuil n'est pas anodin car il est considéré par la communauté médicale comme étant un seuil critique ;
- CA199>34 et CA199<34, ce seuil fut choisi pour les mêmes raisons évoquées précédemment ;
- AGE>50 et AGE<50, DEBUTS>12 et DEBUTS<12.

Tout en évitant d'avoir un nombre de classes différent d'une variable à une autre.

5.4. Présélection des variables :

Cette étape consiste à estimer l'association entre la présence de récidive et chaque variable X_j. Les OR ainsi calculés sont dits « bruts » ou « non ajustés », cependant, il est impossible de combiner les résultats selon les règles de Rubin pour les OR, sans une transformation adéquate qui est dans ce cas de type logarithmique, le sommaire s'est donc construit par l'ensemble des résultats sur les 5 échantillons d'apprentissage. Nous avons retenu pour l'analyse multivariée, toutes les variables dont le degré de significativité est inférieur à 0.20, afin de prendre en compte des variables qui seraient des facteurs de confusion ou des facteurs d'interaction.

6. LES RESULTATS D'ANALYSE MULTIVARIEE COMBINEE:

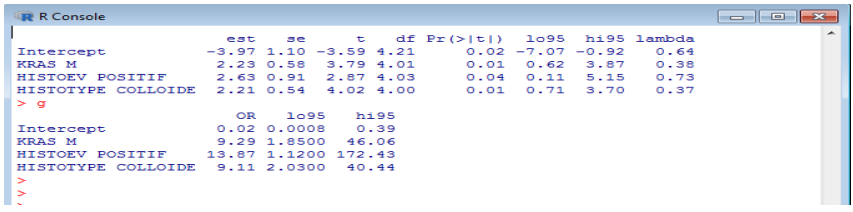
Nous avons construit une régression logistique sur chacun des 5 échantillons d'apprentissage, selon une stratégie forward basée sur le critère AIC pour éviter les effets de la colinéarité. Enfin, pour obtenir le modèle combiné, nous avons utilisé les règles de Rubin énoncés précédemment, soit le logit pour l'individu ω :

$$\log it(\omega) = \ln\left[\frac{\varphi(\omega)}{1-\varphi(\omega)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

Avec : $\varphi(\omega) = P(Y(\omega) = 1 / X(\omega))$ la probabilité à postériori pour un individu ω d'être positif.

Il s'agit alors d'estimer les coefficients $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)_{j=0,..,p}$ et d'estimer leurs variances $\hat{\sigma}_{\hat{\beta}_0}^{\ 2}, \hat{\sigma}_{\hat{\beta}_1}^{\ 2}, ..., \hat{\sigma}_{\hat{\beta}_p}^{\ 2}$ sur les 5 échantillons d'apprentissage. Ensuite, pour tester la significativité des coefficients de régression, le test de Wald est approximé par le test de Student; même si la littérature recommande un nombre restreint de jeux de données pour l'imputation multiple, nous avons tout de même remarqué un impact sur la largeur des intervalles de confiance, lorsque $M = 5$. Étant donné que dans notre cas, la variabilité des estimateurs est dominée par la variance inter-imputation, la précision peut être améliorée selon (Rubin, 2004) en augmentant le nombre de M. Nous avons donc repris les arguments d'efficacité relative de Rubin, nous remarquons une amélioration de l'efficacité lorsqu'on augmente le nombre de jeux de données imputées, ceci dit au-delà de $M = 20$ l'amélioration devient dérisoire, nous nous en tiendrons donc à ce seuil ; sachant que la perte d'efficacité se traduit par une augmentation de la variance de l'estimation et une perte de puissance statistique.

Figure N° 4: la régression logistique combinée pour $M = 20$.



Source : output de R fait par les auteurs.

Soit :

$$\text{logit}(\omega) = -3.97 + 2.23\ (\text{KRAS M}) + 2.63.(\text{HISTOEV POSITIF}) + 2.19.(\text{HISTOTYPE COLLOIDE})$$

6.1. La validation du classifieur :

Nous avons testé les performances de ce classifieur sur les 5 échantillons de validation, et les résultats sont dressés sur le tableau 1.

Tableau N°1 : performances du classifieur par régression logistique.

N° du jeu de données.	1	2	3	4	5
Précision	0.81	0.81	0.83	0.80	0.81
Sensibilité	0. 8293	0. 8293	0.8537	0.8171	0.8293
Spécificité	0.7222	0.7222	0.7222	0.7222	0.7222
AUC	0.776	0.776	0.788	0.770	0.776

Source: fait par les auteurs.

Un classifieur construit à des fins pronostiques et un autre créé pour poser un diagnostique, ne poursuivront pas les mêmes objectifs. Le premier devra limiter le nombre de faux négatifs afin de ne pas négliger une personne à risque. Alors que le second devra limiter le nombre de faux positifs. (Bertrand, 2010) Dans ce contexte, notre modèle est intéressant au sens où il présente de bonnes performances en sensibilité.

7. DISCUSSION :

Les résultats de la régression logistique combinée pour $M = 5$, sont jugés mauvais sur leurs écart-types, et donc une augmentation de la largeur des IC, une conséquence directe est l'absence de relation significative pour la présence d'emboles vasculaires. En décidant d'augmenter le nombre d'imputations, nous avons constaté une amélioration dérisoire de l'efficacité relative au-delà de $M = 20$, nous nous sommes donc arrêtés à ce nombre ; et les résultats combinés pour ce seuil démontrent une baisse de la variabilité des estimateurs, ce qui se traduit par une réduction de la largeur des IC, on parvient aussi à mettre en évidence une relation significative entre la présence d'emboles vasculaires et la probabilité de récidive, l'augmentation de l'efficacité a donc comme conséquence la diminution de la variabilité, car cette dernière est dominée par la variance inter-imputation.

Toutefois, on remarque que le paramètre lambda reste tout de même très élevé pour la variable HISTOEV, ce qui augmente sensiblement sa variance, en dépit de l'augmentation du nombre d'imputations, et cela à cause de la proportion élevé de valeurs manquantes pour cette variable et du processus d'imputation bâti sur l'hypothèse d'un mécanisme ignorable, pas très réaliste en présence de données rétrospectives ; où le recueil de données reste rudimentaire et où le mécanisme peut dépendre de variables non recueillies. Le

classifieur présente de meilleures performances en matière de sensibilité, une qualité très recherchée lorsqu'on souhaite poser un pronostique. En parallèle, les OR ajustés significativement reliés à la probabilité de récidive, ont été approuvés par un cancérologue et par la littérature :

- La présence d'emboles vasculaires est souvent associée à la survenue de métastases à distance. (Frosst and al, 1995), (Mitry and Rachet, 2006) ;
- La présence des mutations KRAS est associée à un risque plus élevé de récidive. (Andreyev and al, 2001), (Esteller and al, 2001), malheureusement aucune thérapie ciblée n'a pu être mise en place pour enrayer leurs actions ;
- Le type d'adénocarcinome colloïde (mucineux) est reconnu comme un histo-pronostique. (El Housse and al, 2015).

8. CONCLUSION :

L'hétérogénéité des populations étudiées impose la stratification des malades grâce à une classification pronostique, afin de cibler les patients éligibles à un traitement précis, car l'augmentation des dépenses en santé justifie la rationalisation des ressources, soit à maitriser les couts par la qualité des soins. Le classifieur obtenu à partir de données rétrospectives, est un modèle qui fut construit à partir de variables cliniques, anatomopathologiques et biologiques permettant de prédire la probabilité d'une récidive. Cependant certaines réserves sont à prendre en considération :

- Plus l'échantillon d'apprentissage sera représentatif de la population, plus le modèle appris sera précis ; ce qui nécessite et implique la mise en place d'un SIS intégré, en développant un plan de coordination et de standardisation des sources de données sur l'ensemble du territoire ; afin de pouvoir coupler les données des différentes bases (issues des différents établissements de santé) selon des procédures simples et rapides.
- Un classifieur comme tout modèle mathématique a une durée de vie et doit être réévalué en fonction de l'actualisation des données, ceci sera aussi possible grâce à un SIS intégré permettant d'offrir un bilan quotidien des différents paramètres souhaités.

BIBLIOGRAPHIE :

1. Allison PD.: missing data, Ed Sage Publications, 2001.

2. Andreyev HJ., Norman AR., Cunningham D., Oates J., DixBR., Iacopetta BJ., et al. : "Kirsten ras mutations in patients with colorectal cancer : the 'RASCAL II'study", Br J Cancer ; 85, 2001, pp.692-6.

3. Bertrand D., Fluss J., Billard C., and Ziegler JC. : "Efficacité, sensibilité, spécificité: comparaison de différents tests de lectures", L'ANNEE PSYCHOLOGIQUE, 2010/2(Vol.110), pp.299-320.

4. Brand JPL. : Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets, Rotterdam: Erasmus University; 212, 1999.

5. Buuren SV., Boshuizen HC., and Knook DL. : "Multiple imputation of missing blood pressure covariates in survival analysis*"*, Stat Med, 18(6), 1999, pp.681-94.

6. Buuren SV. : "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specifiation", Statistical Methods in Medical Research, 16(3), 2007, pp.219-242.

7. Buuren SV., and Groothuis-Oudshoorn K. : "mice: Multivariate Imputation by Chained Equations in R*"*; Journal of Statistical Software, Volume 45, Issue 3(2011).

8. Cottrell G., Cot M., et Mary JY. : « L'imputation multiple des données manquantes aléatoirement concepts généraux et présentation d'une méthode Monte-Carlo » ; Revue d'Epidémiologie et de Santé Publique, 57, 2009, pp.361–372.

9. El Housse H., and al. : « Profils épidémiologique et anatomoclinique d'une population marocaine atteinte de cancer colorectal », Journal Africain du Cancer, Volume 7, Issue 2 (2015), pp.95–99.

10. Esteller M., Gonzalez S., Risques RA., Marcuello E., Mangues R., Germa JR., et al. : « K-ras and p16 aberrations confer poor prognosis in human colorectal cancer", J Clin Oncol; 19, 2001, pp.299-304.

11. Frosst P., Blom HJ., Mitos R., Goyette P., Sheppard CA., Matthews RG., and al. : « A candidate genetic risk factor for vascular disease : a common mutation in methylenetetrahydrofolate reductase », Nat Genet; 10(1), 1995, pp.111–3.

12. Kouadjo JM., Kouakou JA., et Kanga KD. : « Méthodologie d'obtention d'une base de données imputées » ; The African Statistical Journal, Volume 16, May 2013, pp.61-81.

13. Lee KJ. , and Carlin JB. : "Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation*"*; Am J Epidemiol, 171, 2010, pp.624–632.

14. Lee S.: "Noisy replication in skewed binary classification", Computational Statistics and Data Analysis, 34, 2000, pp.165-191.

15. Little RJ., and Rubin, DB. : Statistical Analysis with Missing Data, John Wiley & Sons, 2002.

16. Lunardon N., Menardi G., and Torelli N. : "ROSE: A Package for binary imbalanced learning", The R Journal, Vol. 6/1, June 2014, pp.79-89.

17. Menardi G., and Torelli N.: "Training and assessing classification rules with unbalanced data", Data Mining and Knowledge Discovery, 28 (1), 2012, pp.92-122.

18. Mitry E., and Rachet B. : « Pronostic des cancers colorectaux et inegalites socio-economiques », Gastroenterol clin biol; 30(4), 2006, pp.598–603.

19. Nakache JP., and Gueguen A. : « Analyse multidimensionnelle des données incomplètes », Rev. Statistique Appliquée, LIII (3), 2005, pp.35-62.

20. Rubin DB. : "Multiple Imputation for Nonresponse in Surveys", New York: John Wiley and Sons; 2004.

21. Shafer JL. : "Analysis of incomplete multivariate data". Chapman & Hall/CRC, 2000.