ECONOMETRIE

Justine CHARLEY – Mohamed HAIDARA – Manon MOTTIER

Université Panthéon Sorbonne

Table des matières

I - Introduction

Peut-on prédire une élection ? c'est le casse tête auquel se sont confrontés des étudiants de l'école Télécome Paris Tech pour les élections présidentielles de 2017. Basé sur une analyse de sentiment de Twitter ainsi que sur des données sociodémographiques, ces étudiants ont tenté de prédire qu'elle personnalité politique monterait sur le trône présidentiel.

Le verdict ? Un second tour annoncé entre Fillon et Le pen, nous ne sommes pas passés loin de la montée du Front National au gouvernement.

Leur prédiction était majoritairement basée sur une analyse des tweets ainsi que des recherches sur google. Ils n'avaient pas anticipé le fait que certains candidats avaient un électorat extrêmement actif sur les réseaux par rapport à d'autres et n'avaient pas pondéré le poids de ces nouvelles variables entraînant un biais important au point de promulguer Mr Fillon au 2^{e} tour à la place de Mr Macron.

L'utilisation de données issues de l'open data est une méthode encore très fragile et à manipuler avec précaution. La plus part des instituts de sondages se basent sur des enquêtes afin d'ajuster au mieux leurs algorithmes (bien qu'ils intègrent de plus en plus l'open data).

La plupart des instituts de sondage effectuent leur analyse statistique sur un ensemble d'informations collectées auprès d'une population cible. Cet échantillon de la population doit être constitué avec une grande attention pour coller au mieux à la population globale. Cette échantillon stratifié peut être réalisé conditionnellement à un ensemble de variables de contexte (Sexe, Etude, Geographique …). Bien souvent, le recours à l'echantillonage et à sa méthodologie (grappe, stratifié, quotas…) est la résultante d'une problèmatique de cout.

Une autre approche à considérer est le recours aux réseaux bayesiens afin d'intégrer la dépendance et l'indépendance conditionnelle entre les variables. Ils permettent très facilement de modéliser les dépendances causales entre les variables et donc de mieux capter leurs effets les unes sur les autres. Ainsi des variables qui paraissaient comme indépendante de la variable cible peuvent grâce aux réseaux bayesiens voir leurs interactions avec d'autres variables intégrées au modèle et améliorer de façon significatives son pouvoir prédictif.

Ce projet sera découpé en 2 grandes parties :

- **L'analyse exploratoire** : cette partie nous permettra de traiter les données et d'effectuer des regroupements entre les variables
- **La modélisation** : nous effectuerons dans un premier temps une régression logistique sur des variables d'opinion et de contexte afin de sélectionner les variables pertinentes puis nous modéliserons un réseau bayesien afin de prédire la probabilité de vote pour DSK des électeurs.

II – Analyse exploratoire

L'analyse exploratoire est une étape cruciale dans le pré-traitement des données. En effet, cette étape de découvrir et de connaître les variables, de mieux appréhender leurs regroupements, et surtout rester le plus objectif possible en évitant les idées reçues.

Initialement nous avions à notre disposition une base de données regroupant un peu plus de 230 variables issues d'une étude soutenue par le Ministère de l'Intérieur et de l'Aménagement du territoire ; « Le Baromètre Politique Français 2006 – 2007 ». Ce baromètre se décomposait en 4 vagues d'enquêtes réalisées à partir d'un échantillon représentatif des électeurs français inscrits sur des listes électorales. Nous avions à notre disposition la première vague de cette enquête.

Ces données rassemblaient des informations personnelles (sexe, métiers, âge etc) ainsi que sur des indicateurs de politisation (personnalité politique soutenue ou encore la perception des enjeux et des thèmes politiques) sur les électeurs. L'idée principale de ce baromètre était de prévoir pour quelle personnalité politique les électeurs allaient voter au prochain tour des élections présidentielles.

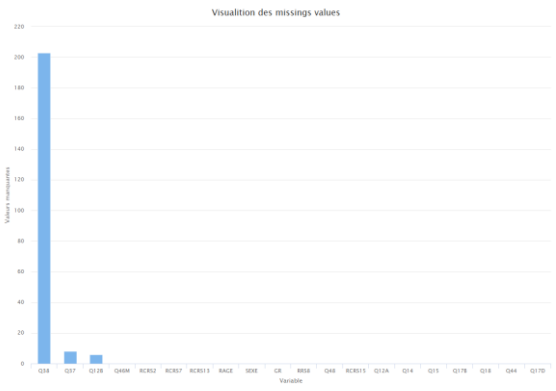
Dans le cas de notre projet nous avons choisi de nous intéresser à la probabilité de vote des électeurs envers le candidat Dominique Strauss-Kahn. Pour cela nous avions à notre disposition 2 types de variables

- Des variables de contexte comme le niveau du diplôme, les catégories sociaux professionnelles ou encore la tranche d'âge.
- Des variables d'opinion analysant les réponses des électeurs à des questions politiquement orientées comme connaître leur position sur l'homosexualité, les journeaux tv etc.

a – Les valeurs manquantes

Avant d'effectuer des regroupements ou commencer la modélisation il a été nécessaire d'analyser le pourcentage de valeurs manquantes par variables. On a constaté que seulement 3 variables étaient incomplètes :

- Q12B : « Problème le plus important pour le France en second »
- Q37 : « Informations TV : nombre de jours regardées »
- Q38 : « Journal TV régulièrement regardé »



Graphique du pourcentage de valeurs manquantes par variables

Ces variables présentaient respectivement 6, 8 et 203 valeurs manquantes, ce qui représente environ 3,8% de notre base de données, ce qui est très marginale. Nous aurions pu les supprimer mais afin de conserver au maximum l'information, nous avons mis en place une stratégie d'imputation en remplaçant les valeurs manquantes via la modalité la plus fréquente car la répartition des observations pour chacune de ses variables présentaient un fossé énorme entre la modalité la plus fréquente et la deuxième modalité la plus fréquente.

Par exemple pour la variable Q38 qui représente la réponse au journal TV le plus régulièrement regardé, on a constaté que 1 825 électeurs ont votés pour le journal de 20h sur TF1 et 992 électeurs pour le journal de 20h sur France 2. L'écart étant tellement important, affecter la modalité du journal de 20h sur TF1 à ces valeurs manquantes ne modifiera pas la distribution de la variable ni sa significativité. Nous aurions également pu avoir recours à une méthode d'imputation multiple comme Mice.

Même constat pour les variables Q37 et Q17B qui représentent respectivement le nombre de jours pendant lesquel un électeur regarde la télé ainsi que leur avis sur l'homosexualité ; la part de valeurs manquantes parmi les données étant très faible, nous les avons également affectées aux modalités les plus représentées.

b – Les regroupements

Une fois les données pré-traitées et complètes, nous avons analysé la répartition des classes par variables. Un nombre trop important de classe parmi les variables peut-être source de biais dans l'estimation. Nous avons donc choisi d'effectuer certains regroupements selon plusieurs critères et en distinguant 2 types de variables :

- Traitement des NSP
- Traitement des autres modalités

- **<u>Les sans réponses / Ne souhaite pas répondre</u>**

Nous avons commencé les regroupements par la recodificiation des individus ayant répondu NSP. Bien que ces modalités représentent une infime minorité, ils doivent être intégrés à une autre modalité.

Cette catégorie NSP « Ne se prononce pas » permet aux personnes non concernées de tout de même répondre à la question sans être tenté de répondre au hasard et donc de créer un biais de questionnaire.

Nous sommes partis sur l'hypothèse que la personne n'ose peut-être pas répondre ou exprimer son point de vue de peur d'être jugée et donc nous avons choisi de les regrouper avec la réponse la moins « avouable » / la plus « polémique » à l'exception des réponses neutres.

	Variable	*Thématique*	*Catégorie intiale*	*Catégorie après regroupement*
Variable de contexte	RCRS13	Ascendance Famille	Nsp	A un parent immigré
	RRS8	Activité profes.	Nsp	Autre
	Q48	Statut marital	Nsp	Divorcé
	RCRS15	Religion	Nsp	Autre religion
Variable d'opinion	Q12B	Problème en France (en second)	Nsp	Sécurité / immigration
	Q14	Chomage	Nsp	Stable
	Q15	Delinquance	Nsp	Stable
	Q17B	Homosexualité	Nsp	Non
	Q37	Nbr jours tv regardé	Non réponse	Rarement
	Q38	Journal tv regardé	Nsp / Non réponse	Autre
	Q44	Mondialisation	Nsp	Un danger
	Q17D	Peine de mort	Nsp	Oui

Prenons l'exemple de la religion (RCRS15), on peut supposer que les personnes affiliées à des groupes religieux de minorité n'auront pas la volonté de faire connaître leur appartenance religieuse au public et vont donc préférer – au lieu de mentir – de ne pas exprimer leurs positions sur ce sujet.

C'est d'autant plus vrai lorsque l'on entame un débat sur la légalisation de la peine de mort ou de l'homosexualité, ces sujets sont très polémiques et très sensibles pour le grand public. Ce sont des sujets encore tabous dans notre société d'autant plus durant la période pendant laquelle a été effectué cette enquête.

- **<u>Les regroupements purs</u>**

Concernant les autres variables, nous les avons examinés individuellement afin de déterminer la meilleure façon de faire les regroupements. Pour cela nous avons eu recours à plusieurs critères

- Distribution du Y par rapport aux modalités de la variable considérée : Les histogrammes vont mettre l'accent sur les fréquences de chaque modalité et également nous permettre de confirmer ou non, nos intuitions quant aux regroupements des modalités entre elles.
- Proximité des modalités via l'ACM
- Corrélation entre les modalités par rapport à la variable cible : afin d'effectuer des regroupements intelligents, deux modalités fortement corrélées peuvent être regroupées en une seule classe.
- Article de recherche & connaissances personnelles du sujet.

Nous allons vous présenter certains exemples de regroupements. Vous trouverez la liste exhaustive des regroupements effectués pour chaque modalité dans l'onglet « variable » du dictionnaire des variables.

	Initial			Transformation
	Modalité	Label	Nbr	Modalité
Q38 (Journal TV régulièrement regardé)	1	20h TF1	2028	à droite
	2	13h TF1	735	
	3	20h France 2	992	à gauche
	4	13h France 2	245	
	5	19-20 France 3	581	
	6	Soir 3 France 3	155	
	7	12-14 France 3	85	
	8	Canal +	81	
	9	Arte Info	100	
	10	66 minutes M6	322	à droite
	11	12h50 M6	55	
	12	LCI	147	
	13	I-télé	73	
	14	Autre	17	Autres
	15	Aucun	23	
	16	Nsp	11	

La variable Q38 concernant le journal TV regardé était initialement composée de 17 modalités. Nos recherches concernant l'orientations politiques par média, nous ont conduits à un regroupement en 3 classes : « Droite » / « Gauche » / « Autres ». Plusieurs études ont démontré que les électeurs de certains mouvements politiques favorisaient le visionnage de certaines chaînes/programme tv. Il est de notoriété commune que presque l'intégralité des médias tous une couleur politique[1]. Ci-contre les regroupements effectués pour cette variable.

Certaines variables d'opinions étaient quant à elles décomposées en 5 modalités. Prenons l'exemple de la question Q17D qui interrogeait les électeurs quant à leur avis sur le rétablissement de la peine de mort.

	Initial			Transformation
	Modalité	Label	Nbr	Modalité
Q17D (rétablir la peine de mort)	1	Tout à fait d'accord	894	Oui
	2	Plutôt d'accord	983	
	3	Plutôt pas d'accord	786	Non
	4	Pas du tout d'accord	2971	
	5	Nsp	16	Oui

Nous avons choisi de regrouper les modalités selon l'avis général qu'elles exprimaient ; d'un côté les électeurs ayant un avis positif (« Tout à fait d'accord », « Plutôt d'accord », « Nsp ») et de l'autre côté les réponses caractérisant un avis négatif (« Plutôt par d'accord », « Pas du tout d'accord »).

Concernant le statut marital du couple (Q48). Il est notable que l'homogamie politique règne au sein d'un couple et qu'il y règne une certaine entente politique. « *70% des individus connaissant l'orientation politique de leur conjoint se situent exactement sur le même positionnement* »[2].

	Initial			Transformation
	Modalité	Label	Nbr	Modalité
Q48	1	Célibataire	1157	Célibataire
	2	Marié(e)	2822	Couple / concubinage
	3	Vivant en couple sans être marié(e)	710	Vivant en couple sans être marié(e)
	4	Pacsé(e)	43	Couple / concubinage
	5	Divorcé(e) ne vivant pas en couple	310	Divorcé + nsp
	6	Divorcé(e) vivant en couple	77	
	7	Veuf ou veuve ne vivant pas en couple	484	VEUF
	8	Veuf ou veuve vivant en couple	45	
	9	Nsp	2	Divorcé + nsp

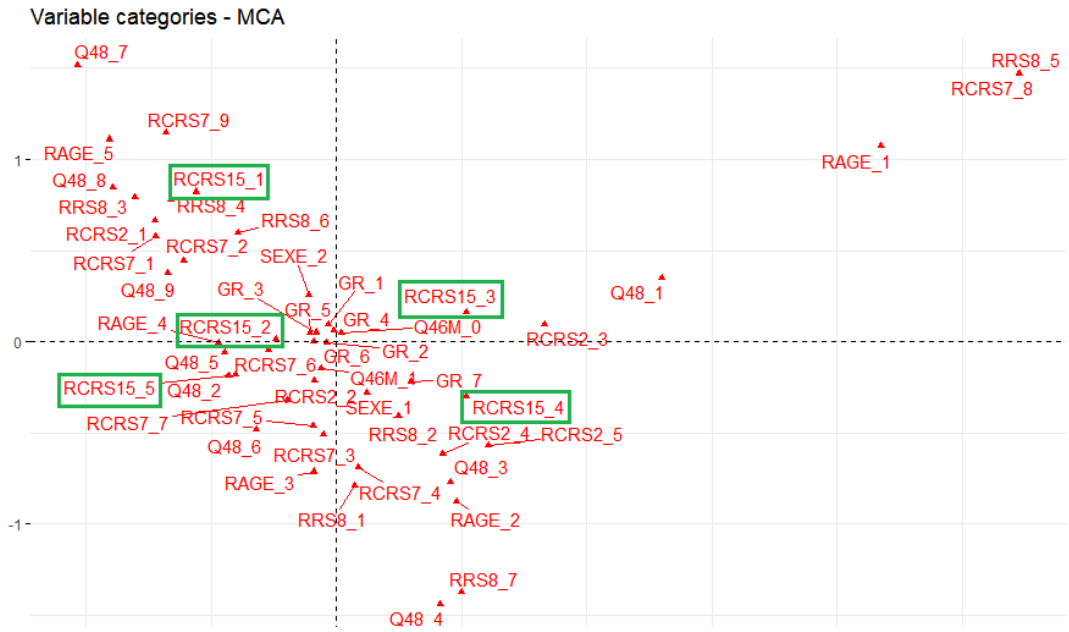
C'est pour cela que nous avons choisi de regrouper les personnes mariées avec celles qui sont pacsées, de laisser les célibataires, les divorcés et les personnes en couple dans une catégorie à part. les

[1] https://www.planet.fr/dossiers-de-la-redaction-dis-moi-qui-tinforme-je-te-dirai-pour-qui-tu-votes.193652.1466.html

[2] https://www.pleinevie.fr/vie-quotidienne/societe/elections-2017-dans-un-couple-on-vote-de-la-meme-facon-18149

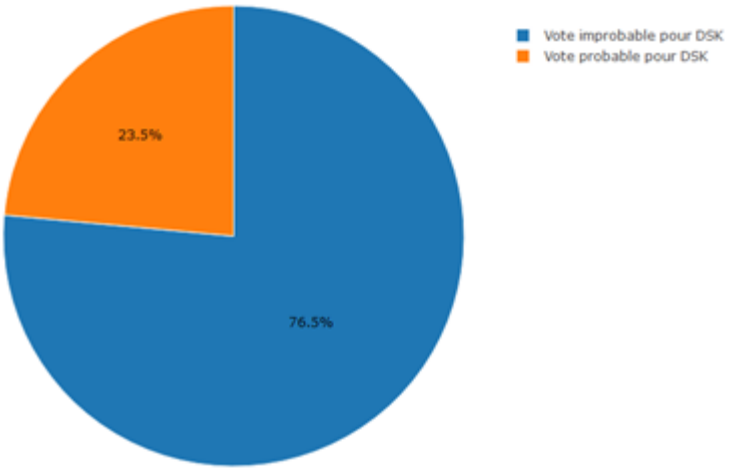
personnes vivant en couple ne sont pas aussi engagées à notre sens que des personnes mariées ou pacsées c'est pour cela que nous avons choisi de ne pas les regrouper.

Prenons en dernier exemple le statut religieux de l'électeur. Au départ nous avions 5 modalités différentes ; nous avons choisi de regrouper la catégorie NSP avec la classe « autre religion ». En analysant l'ACM des variables de contexte on constate qu'après les avoirs projetés dans un nouvel espace, les modalités ne sont pas proches les unes des autres mais semblent aux contraires éloignées. Nous aurions pu regrouper la catégorie NSP avec la classe « catholique non pratiquant » d'après l'ACM mais nous avons choisi d'utiliser le critère du bon sens pour cette catégorie et l'ACM pour les autres modalités. D'après l'ACM nous n'effectuerons pas d'autres regroupements sur cette variable.



c – Premières analyses

Sur nos 5650 individus ayant répondu au questionnaire, 4320 n'ont pas d'intention de vote porté sur Dominique Strauss-Kahn. Ce déséquilibre entre les classes est notamment expliqué par la multitude de candidats possibles. (Les électeurs avaient le choix entre pas moins de 16 candidats, allant de Bayrou à Villepin en passant par Le Pen) ;



d – Analyse des corrélations et dépendances statistiques

Après avoir traité les valeurs manquantes des données il a été nécessaire d'analyser les possibles corrélations entre les différentes variables et notre variable cible. L'intégralité de nos variables étant des variables catégorielles, nous n'avons pu calculer les coefficients de corrélations. Pour ce faire nous avons eu recours à d'autres tests tels que le test du Khi2, le Kappa de Cohen, le V de cramer, les coefficients de contingence ou encore le T de Tschuprow. Le tableau ci-dessous synthétise tous les résultats obtenus. Nous les commenterons dans la partie suivante.

Variable	Khi2	Result_khi2	Phi	Result_Phi	Cramer	Result_Cramer	Contingence	Result_Contingence	T_Tschuprow	Result_Tschuprow
GR	0,0002255	Dépendance	0,0678	Très faible	0,068	Très faible	0,068	Très faible	0,0433	Très faible
Q12A	0,0000000	Dépendance	0,1178	Faible	0,118	Faible	0,117	Faible	0,0647	Très faible
Q12B	0,0000000	Dépendance	0,1136	Faible	0,114	Faible	0,113	Faible	0,0610	Très faible
Q14	0,0173586	Dépendance	0,0493	Très faible	0,049	Très faible	0,049	Très faible	0,0330	Très faible
Q15	0,0028257	Dépendance	0,0566	Très faible	0,057	Très faible	0,057	Très faible	0,0378	Très faible
Q17B	0,0004770	Dépendance	0,0596	Très faible	0,060	Très faible	0,060	Très faible	0,0422	Très faible
Q17E	0,0000000	Dépendance	0,1434	Faible	0,143	Faible	0,142	Faible	0,1014	Faible
Q18	0,0045110	Dépendance	0,0481	Très faible	0,048	Très faible	0,048	Très faible	0,0365	Très faible
Q37	0,6289983	Indépendance	0,0305	Très faible	0,030	Très faible	0,030	Très faible	0,0187	Très faible
Q38	0,0000000	Dépendance	0,1747	Faible	0,175	Faible	0,172	Faible	0,0888	Très faible
Q44	0,0000270	Dépendance	0,0650	Très faible	0,065	Très faible	0,065	Très faible	0,0494	Très faible
Q48	0,4595715	Indépendance	0,0370	Très faible	0,037	Très faible	0,037	Très faible	0,0220	Très faible
RAGE	0,0000003	Dépendance	0,0795	Très faible	0,080	Très faible	0,079	Très faible	0,0562	Très faible
RCRS13	0,0019498	Dépendance	0,0513	Très faible	0,051	Très faible	0,051	Très faible	0,0390	Très faible
RCRS15	0,0006799	Dépendance	0,0585	Très faible	0,058	Très faible	0,058	Très faible	0,0413	Très faible
RCRS2	0,0000000	Dépendance	0,0983	Très faible	0,098	Très faible	0,098	Très faible	0,0695	Très faible
RCRS7	0,0000000	Dépendance	0,1157	Faible	0,116	Faible	0,115	Faible	0,0688	Très faible
RRS8	0,0006514	Dépendance	0,0645	Très faible	0,064	Très faible	0,064	Très faible	0,0412	Très faible
SEXE	0,0000261	Dépendance	0,0564	Très faible	0,056	Très faible	0,056	Très faible	0,0564	Très faible

- **Test du Khi2 de Pearson**

Ce test statistique est utilisé pour tester l'indépendance entre deux variables aléatoires qualitatives et détecter la présence éventuelle de liaison.

H_0: $Les\ distributions\ comparées\ sont\ identiques\ (ie\ les\ deux\ variables\ sont\ indépendantes)$
H_1: $Les\ variables\ ne\ sont\ pas\ indépendantes$

En regardant les résultats obtenus, on constate que l'intégralité des variables semble avoir un lien statistique avec notre variable cible à l'exception de Q48 (Situation familiale) et Q37 (Nbr de jour de TV regardés). Intuitivement nous serions tentés d'exclure ces variables à cette étape car nous ne sommes pas capables de confirmer ou d'infirmer l'existence d'un lien statistique significatif entre Q48 et Q37 et notre variable cible. Néanmoins nous allons conserver ces variables afin de les utiliser en tant que variables instrumentales ; il est possible que l'interaction de Q48 avec une autre variable ait un effet significatif sur notre probabilité de vote. Si ces potentielles interactions ne sont pas significatives, elles seront éliminées durant la phase de sélections des interactions.

- **Test du Phi**

Basé sur le Khi2 de Pearson, ce test permet de s'affranchir des effets de taille du test du Khi2 en prenant la racine du Khi2 divisé par les effectifs totaux.

Les variables Q48 et Q37 sont celles possédant le plus petit coefficient Phi.

- **Test du V de Cramer**

Basé sur le test du Phi, ce test permet de comparer l'intensité du lien (s'il existe) entre deux variables qualitatives. Il est généralement à interpréter en parallèle du test du Khi2 et beaucoup plus facile à lire que le test du Phi. Un résultat proche de 0 mettra en évidence des variables complètement indépendantes tandis qu'un résultat proche de 1 caractérisera des variables dépendantes. Ce test est considéré comme plus fiable que celui du Khi2 de Pearson dans la mesure où il permet de s'affranchir de la taille des données.

On remarque que 5 variables ont une intensité faible et toutes les autres très faibles. Ces résultats peuvent paraîtres très surprenants dans la mesure où certaines variables ont des liens confirmés par plusieurs études statistiques sur l'intention de vote comme le niveau de diplôme et pourtant cette variable possède le plus petit V de Cramer.

Il a été maintes et maintes fois prouvé dans la littérature qu'avec la mondialisation, le niveau de diplôme est devenu un élément influençant énormément les intentions de votes. Récemment il a été prouvé qu'aux dernières élections présidentielles les classes moyennes ont eu une propension forte à voter pour les candidats Macron et Fillon[3]. Bien qu'il s'agisse de résultats aux élections de 2017, ces clivages existaient bien avant et deviennent juste de plus en plus forts au fil des années.

[3] http://www.slate.fr/story/138311/candidat-premiers-classe

« Le vote de classe structure toujours la présidentielle », [4]titre d'un article écrit par les échos en 2017 qui explique que le paysage politique est déstructuré, refaçonné par le poid des classes sociales de plus en plus fortes et creusé par de nouveaux clivages économiques et sociaux.

- **Test des coefficients de contingence (C de Pearson)**

Afin de décrire la relation entre deux variables catégorielles, il est possible d'utiliser les coefficients de contingence. Ils permettent de mesurer l'intensité d'association entre notre variable cible et nos variables catégorielle. En d'autres termes il permet de faire une inférence statistique sur la relation éventuelle de notre y et des autres variables. Ce coefficient varie entre 0 et 1, 0 caractérisant une indépendance parfaite entre les deux variables testées.

Dans le cas de nos données, les conclusions sont similaires à celles faites par l'analyse du V de Cramer. A ce stade de l'analyse les résultats paraissent étonnants ; aucune variable ne semble forte liée avec notre variable cible.

- **Test de Tschuprow**

Le test de Tschuprow permet de mesurer l'association entre 2 variables qualitatives mais surtout il permet tout comme le V de Cramer de s'affranchir de la taille de l'échantillon. Les résultats obtenus sont très surprenants. D'après ce test, toutes les variables semblent très faiblement liées à notre y. les variables étant identifiées comme indépendantes de notre y sont celles possédant les plus petits T de Tschuprow.

Etant donné que tous ces indicateurs sont des mesures dérivées du Khi2, leurs résultats étaient prévisibles. Globalement toutes les variables semblent indépendamment les unes des autres posséder quelques liens statistiques très faibles avec y.

L'utilisation d'interaction entre les variables pourrait nous permettre d'obtenir des variables corrélées et ayant des dépendances fortes avec notre y.

III – La modélisation

a – La sélection des variables

Dans cette partie nous allons adopter une stratégie de sélection de variables en 2 temps : dans un premier temps nous sélectionnerons d'abords les effets d'interactions pertinents puis nous effectuerons une sélection de variables finales parmi las variables et les interactions au préalable choisies.

[4] https://www.lesechos.fr/20/04/2017/lesechos.fr/0211992880832_le-vote-de-classe-structure-toujours-la-presidentielle.htm

Phase 1 : sélection des effets d'interactions

L'ajout d'une interaction au modèle augmente la dimension explicative du modèle en faisant le produit du nombre de dimensions des variables une à une.

La construction d'un modèle avec interaction se fait normalement par l'ajout d'un (« ^2 ») dans la spécification du modèle. Cependant, le nombre de dimensions et le nombre de variables importants, nous ont empêché de construire une telle modélisation par conséquent il n'est pas possible d'estimer le modèle saturé.

Dans ce souci de grandeur, nous avions deux choix possibles à notre disposition :

- Choisir des interactions « intuitives »
- Choisir des interactions parmi un ensemble groupé

Nous avons opté pour la seconde mais de manière méthodique. En effet, la stratégie a été de tester toutes les interactions possibles pour une variable donnée et de conserver en mémoire toutes les interactions qui ont été significatives.

Une fois ce schéma répété pour toutes les variables, nous avons écrit le modèle de la régression logistique avec toutes les variables additionnées aux différentes interactions retenu afin d'effectuer la sélection de variable globale cette fois-ci.

		Variable croisée								
		RCRS2	RCRS7	RCRS13	RAGE	SEXE	GR	RRS8	Q48	RCRS15
Variable de référence	RCRS2		N	N	N	N	N	N	O	N
	RCRS7	N		N	N	N	N	N	N	N
	RCRS13	N	N		N	O	N	N	N	N
	RAGE	N	N	N		O	N	N	N	N
	SEXE	N	N	O	O		O	N	N	N
	GR	N	N	N	N	O		N	N	N
	RRS8	N	N	N	N	O	N		N	N
	Q48	O	N	N	N	N	N	N		N
	RCRS15	N	N	N	N	N	N	N	N	

		Variable croisée										
		Q12A	Q12B	Q14	Q15	Q17B	Q18	Q37	Q38	Q44	Q17D	Q17E
Variable de référence	Q12A		N	N	N	N	N	N	N	N	N	N
	Q12B	N		N	N	N	N	N	N	N	N	N
	Q14	N	N		N	N	O	O	N	N	O	O
	Q15	N	N	N		N	N	N	N	N	N	N
	Q17B	N	N	N	N		N	N	N	O	N	N
	Q18	N	N	O	N	N		N	N	N	N	N
	Q37	N	N	O	N	N	N		N	N	N	N
	Q38	N	N	N	N	N	N	N		N	N	N
	Q44	N	N	N	N	O	N	N	N		N	N
	Q17D	N	N	O	N	N	N	N	N	N		N
	Q17E	N	N	O	N	N	N	N	N	N	N	

Ci-dessus deux tableaux récapitulant les résultats obtenus à la suite de cette sélection des termes d'interactions. Le premier tableau concerne les variables de contexte et le deuxième, les variables d'opinions.

Les variables en colonnes représentent les variables qui ont été croisée et celles en colonne celles qui ont servi de référence pour l'analyse. Un « O » nous indique l'interaction significative à conserver. Par

exemple la première ligne du premier tableau nous indique que la première régression effectuée a été faite par l'intégralité des variables de contexte par rapport à la variable RCRS2 :

RCRS2 * RCRS7 + RCRS2 * RCRS13 + RCRS2 * RAGE + RCRS2 * SEXE + RCRS2 * GR + RCRS2 * RRS8 + RCRS2 * Q48 + RCRS2 * RCRS15

La variable RCRS2 a été croisée avec l'intégralité des autres variables et la sélection backward nous a indiqué de conserver uniquement l'interaction RCRS2*Q48 parmi toutes les interactions.

Toutes ces régressions nous ont permis de présélectionner un certain nombre d'interactions qui semblent pertinentes pour notre modélisation pour chaque type de modèle (contexte et opinion).

Nous avons donc retenu les interactions suivantes :

- Contexte : RCRS2*Q48 + RCRS13*SEXE + RAGE*SEXE + SEXE*GR + RRS8*SEXE
- Opinion : Q14*Q18 + Q14*Q37 + Q14*Q17D + Q14*Q17E + 17B*Q44

Phase 2 : sélection finale

Dans le cadre de ce projet, la régression logistique a été utilisée dans le but de faire une sélection de variable. La première régression concernera les variables de contexte, puis une seconde les variables d'opinions.

Nous avons choisi de tester plusieurs méthodes : backward, forward et stepwise afin de comparer leurs résultats. Dans les trois cas, nous sommes sur une méthode itérative de sélection de variable minimisant l'AIC et un second modèle minimisant le BIC (Le meilleur modèle est obtenu lorsque toutes les variables ont un impact significatif sur notre régression) basé sur les variables initiales ainsi que sur les interactions choisies dans la partie précédente.

- **Backward** : Cette méthode consiste à partir du modèle complet intégrant toutes les variables puis à chaque étape, d'enlever la variable qui a la plus grand p-value non significative jusqu'à ce que toutes les variables non significatives (au seuil de 10% par défaut) du modèle aient été enlevées.
- **Stepwise** : Le modèle est estimé avec aucune variable à l'exception de la constante et il ajoute les variables une par une (celle ayant la plus grande statistique significative) jusqu'à ce qu'il n'y'ai plus de variables avec une statistique significative à ajouter. En revanche, à chaque étape, la méthode réexamine les variables sélectionnées à l'étape précédente dans le modèle et choisi de les conserver si elles ont toujours une p-value significative sinon il les enlève. Ce processus continu jusqu'à ce qu'on ne puisse plus ajouter de variable significative au modèle ou bien si celle à ajouter est celle qui vient d'être supprimée.
- **Forward** : Cette méthode est très largement inspirée de la stepwise à l'exception du fait qu'une fois qu'une variable est sélectionnée pour le modèle, elle ne peut plus être retiré même si sa p-value devient non significative.

Comme énoncé, nous avons utilisé l'AIC (distance Kallbach-Leibesh) et le BIC (distance Bayesienne) comme mesure du pouvoir explicatif au détriment d'une distance quadratique (le R^2 ou le C(p) de Mallows par exemple). Une fois que nous aurons sélectionné 2 modèles, nous utiliserons le MSE et le RMSE pour les départager et réussir à sélectionner un seul modèle par catégorie de variable.

Afin de réaliser cette sélection de variable nous avons eu recours à la fonction stepA afin de tester l'effet des variables sur notre probabilité de vote pour DSK. Cette fonction permet de trouver le modèle le plus parcimonieux. Par défaut cette fonction réalise la sélection de variable par le biais de l'AIC (k = 2). Afin d'utiliser également le BIC il a été nécessaire de modifier la valeur du k par le logarithme du nombre de lignes de la base.

Cette sélection de variable est une étape primordiale. En effet, un modèle constitué de variables sans aucun impact significatif sur notre variable cible modifierait les paramètres et conduirait à un modèle médiocre.

a) Variable de contexte

Nous avons effectué une sélection de variable en deux temps ; d'abord d'après le critère AIC puis d'après le critère BIC et comparé les résultats. Peu importe la méthode itérative retenue lorsque nous effectuons la sélection de variable grâce à l'AIC, le modèle obtenu est identique. Résultat similaire avec le BIC. Finalement nous nous retrouvons à comparer 2 modèles :

$$y_{AIC} = \beta_0 + \beta_1 RCRS2 + \beta_3 RCRS7 + \beta_4 RCRS13 + \beta_5 RAGE + \beta_6 SEXE + \beta_7 GR + \beta_8 RCRS15 + \beta_9 RCRS13 * SEXE + \beta_{10} RAGE * SEXE + \beta_{11} SEXE * GR$$

$$y_{BIC} = \beta_0 + \beta_1 RCRS2 + \beta_3 RAGE + \beta_4 RCRS15 + \beta_5 SEXE$$

Afin de départager les deux modèles et sélectionner les variables les plus pertinentes nous avons eu recours au MSE et au RMSE.

	AIC	**BIC**
MSE	0.1729864	0.1754833
RMSE	0.4159163	0.4189072

On constate que le modèle regroupant les variables sélectionnées par l'AIC est le modèle minimisant le MSE et le RMSE. Nous choisirons donc ce modèle pour les variables de contexte.

b) Variable d'opinion

La méthodologie reste la même que celle concernant les variables de contexte. Peu importe la méthode itérative utilisée nous obtenons un modèle unique par critère :

$$y_{AIC} = \beta_0 + \beta_1 Q12A + \beta_3 Q12B + \beta_4 Q14 + \beta_5 Q18 + \beta_6 37 + \beta_7 38 + \beta_8 Q44 + \beta_9 Q17D + \beta_{10} Q17E + \beta_{11} Q14 * Q18 + \beta_{12} Q4 * Q37 + \beta_{13} Q14 * Q17D + \beta_{14} Q14 * Q17E$$

$$y_{BIC1} = \beta_0 + \beta_1 Q38 + \beta_3 Q17B + \beta_4 Q17E + \beta_5 Q14 * Q17E$$

$$y_{BIC2} = \beta_0 + \beta_1 Q38 + \beta_3 Q17E + \beta_4 Q17D$$

	AIC	BIC1	BIC2
MSE	0.1697231	0.1723193	0.173
RMSE	0.4119747	0.4151136	0.4159327

On constate que le modèle regroupant les variables sélectionnées par l'AIC est le modèle minimisant le MSE et le RMSE. Nous choisirons donc ce modèle pour les variables d'opinion.

Après avoir effectué dans un premier temps une sélection des termes d'interactions puis des variables dans leur globalité, nous obtenons le modèle final suivant :

$$y = \beta_0 + \beta_1 Q12A + \beta_3 Q12B + \beta_4 Q14 + \beta_5 Q18 + \beta_6 37 + \beta_7 38 + \beta_8 Q44 + \beta_9 Q17D + \beta_{10} Q17E + \beta_{11} Q14 * Q18 + \beta_{12} Q4 * Q37 + \beta_{13} Q14 * Q17D + \beta_{14} Q14 * Q17E + \beta_{15} RCRS2 + \beta_{16} RCRS7 + \beta_{17} RCRS13 + \beta_{18} RAGE + \beta_{18} SEXE + \beta_{20} GR + \beta_{21} RCRS15 + \beta_{22} RCRS13 * SEXE + \beta_{23} RAGE * SEXE + \beta_{24} SEXE * GR$$

C'est à partir de ces variables et des interactions sélectionnées que nous modéliserons un réseau bayésien.

b – Le réseau bayésien

Reposant sur les théories probabilistes et plus précisément sur la formule inversée de Bayes, les réseaux bayésiens permettent de traiter des grosses quantités de données et surtout de les transformer afin qu'elles soient interprétables. Il permet d'obtenir des graphiques représentés par des arcs et des nœuds présentant les différentes dépendances et liens causaux entre les variables sans avoir besoin de se référer à la table de données pour les comprendre.

Les réseaux bayésiens sont souvent assimilés à des systèmes experts car ils permettent d'extraire des raisonnements logiques de l'analyse du graphique tout comme un expert (homme) pourrait formuler, ils permettent entre autres de faire de l'inférence statistiques.

Un exemple simple est la représentation schématique des variables de causalité provoquant un cancer et donc estimer l'ensemble des facteurs à risque. Les liens de causalité déterminent l'influence d'une variable sur une autre dans le cas d'une relation orientée, ou une influence réciproque de deux variables dans le cadre d'une relation non dirigée.

Un autre exemple plus parlant est celui de la médecine et plus précisément de l'action de poser un diagnostic sur une maladie. Il n'y'a pas une liste prédéterminée de symptômes pouvant être à l'origine d'une maladie mais plutôt des millions de combinaisons de différents symptômes et les réseaux bayésiens permettent donc de gérer ces liens de causalité.

Les réseaux bayésiens et plus généralement les réseaux probabilistes permettent de mettre en évidence les corrélations et dépendances conditionnelles entre les variables, effectuer des diagnostics (dans le cas de la médecine de déterminer la cause d'une maladie grâce à l'analyse des liens causaux

des symptômes), effectuer des prédictions (prédire la probabilité que le patient ait cette maladie) ou encore de la classification.

Les réseaux bayésiens sur R sont implémentés à l'aide du package bnlearn, et s'adaptent à tout type de variables : Discrètes (catégorielles) ou continues. Ils nous fournissent un ensemble d'algorithmes combinés à différents tests d'indépendance conditionnelle ce qui en fait toute sa puissance. Si le package bnlearn dispose de nombreuses fonctions permettant de mettre en exergue les réseaux bayésiens, nous avons décidé de retenir deux algorithmes afin de modéliser notre base de données.

Les options suivantes ont été renseignées afin de construire le réseau bayésien :

- **Blacklist**
  - Le vote ne peut influencer ni le contexte ni l'opinion : il s'agit d'un lien logique, ce n'est pas parce que vous allez voter à droite que du jour au lendemain vous aller changer de sexe par exemple.
  - L'opinion ne peut pas influencer le contexte : raisonnement similaire à celui-ci-dessus, que vous changiez d'avis quant à votre position sur la peine de mort, ce n'est pas pour autant que vous allez par exemple passer de la tranche d'âge 35 – 45 ans à la tranche d'âge moins de 18 ans.

Extrait :

1	Q38	RCRS2
2	Q17E	RCRS2
3	Q17D	RCRS2
4	Q12A	RCRS2
5	Q12B	RCRS2
6	Q14	RCRS2
7	Q44	RCRS2
8	Q37	RCRS2
9	Q18	RCRS2

- **Whitelist**
  - Nous avons décidé de représenter les termes d'interactions en whitelist afin de forcer le lien entre ces deux variables qui ont été prouvées comme étant explicative par la sélection des variables dans la partie précédente. Pour rappel voici les interactions sélectionnées :

1	SEXE	RCRS13
2	SEXE	RAGE
3	SEXE	GR
4	Q14	Q17E
5	Q14	Q17D
6	Q14	Q37
7	Q14	Q18

- **Optimized** = True
  - Initialisation du processus d'apprentissage de chaque nœud.

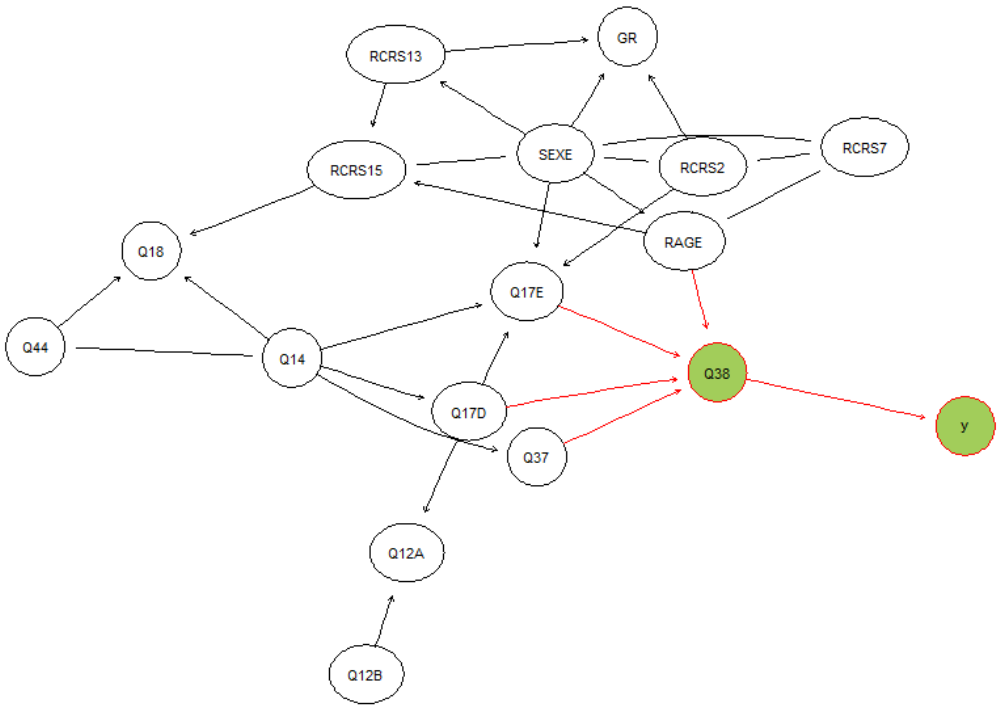
A partir de cette liste de spécification, nous avons décidé de modéliser 2 algorithmes par la fonction GS et HC. La première spécification nous a permis d'intégrer des tests d'indépendances entre les variables, les interactions et notre variable cible et la deuxième d'obtenir des probabilités de vote pour notre candidat DSK.

- **L'algorithme de Grow-Shrink (GS) :** introduit par Margaritis en 2003, cet algorithme travaille en utilisant les tests statistiques pour apprendre la structure des réseaux bayésiens. Plus généralement il a recours aux chaînes de Markov pour identifier la frontière de Markov de chaque nœud afin d'apprendre sur les différents liens existants entre les variables.
- **L'algorithme Hill-Climbing (HC) :** introduit par, Tsamardinos, Brown et Aliferis en 2005, cet algorithme part d'une solution locale qu'il optimise à chaque itération. Il permet entre autres de modéliser les réseaux bayésiens à l'aide des scores de modélisations et ainsi d'obtenir des probabilités afférentes à la variable cible (vote) en utilisant le BIC comme métrique par défaut.

- **<u>Modélisation avec la fonction gs</u>**

La modélisation avec l'algorithme gs nous permet d'explorer la structure du réseau bayésien testant en parallèle l'indépendance des variables avec notre probabilité de vote pour DSK.

- **Test**
  - Parmi l'ensemble des tests d'indépendances existants, nous avons choisi l'option *« smc-x2 »* qui est particulièrement adaptée aux variables discrètes.
  - Ce test correspond exactement à une séquence partielle de Monte-Carlo (suivant asymptotiquement une loi du χ2).



Résultat

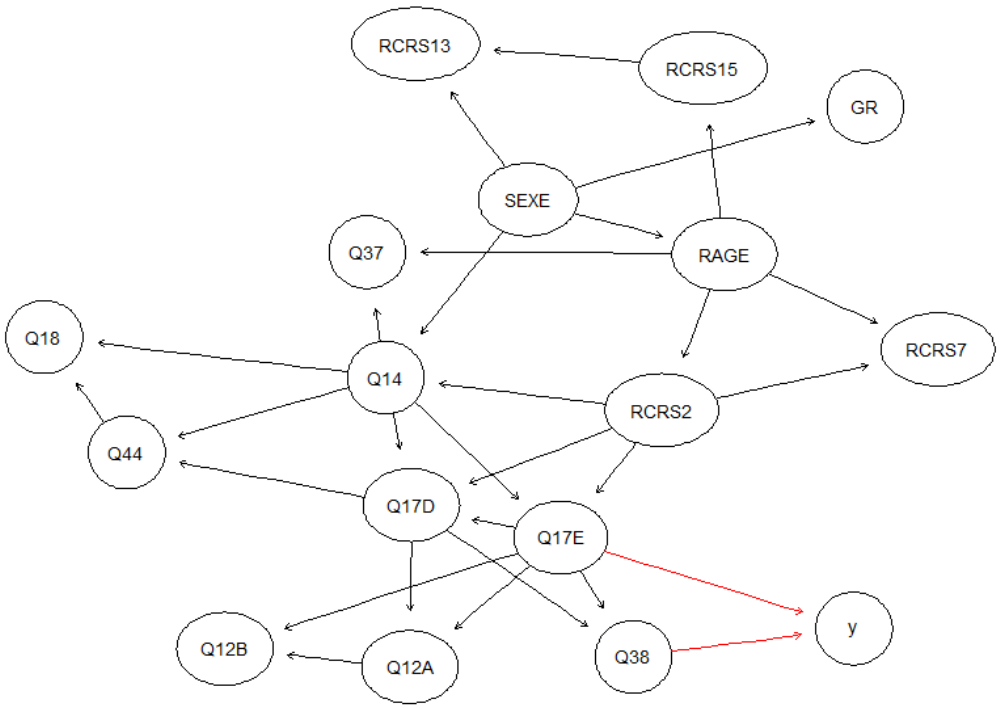
- 7 relations sont non dirigées, c'est-à-dire que les variables se causent mutuellement. Cela se repère lorsque la droite n'est pas munie d'une flèche.
- 23 relations sont dirigées, c'est-à-dire qu'une variable influence une autre sans que la réciproque soit vraie. Cela se repère lorsque la droite unissant deux variables est munie d'une flèche.

    - Les points suivants ressortent de la modélisation avec la fonction GS :

- La variable vote est causée par les programmes télévisuels (Q38)
  La variable programmes télévisuels (Q38) constitue un Backdor entre le vote et l'opinion sur la peine de mort (Q17D) ; la tranche d'âge de l'électeur (RAGE) ; l'opinion sur les immigrés (Q17E).

- **<u>Modélisation avec la fonction hc</u>**

L'intérêt principal que revêt cet algorithme est qu'il permet d'obtenir les probabilités de vote pour un candidat (en l'occurrence DSK) conditionnellement aux variables qui causent le vote. Un score est attribué à chaque réseau candidat reflétant la qualité de l'estimation que l'algorithme maximise.



A l'issue de la modélisation avec HC, nous obtenons un résultat légèrement différent de l'algorithme GS en ce que les mesures utilisées ne sont pas identiques.

La modélisation avec la fonction HC ne prend en compte que les relations dirigées, à savoir qu'une variable influencent l'autre sans que la réciproque soit vraie.

Deux variables causent le vote, elles-mêmes causées par d'autres variables. Cela nous permet de dire que l'Opinion sur les immigrés (Q17E) et les programmes télévisuels (Q38) influencent le vote pour un candidat.

Résultat

Les probabilités sont obtenues via la fonction bn.fit du package bnlearn. Cette fonction permet d'obtenir une pléthore de probabilités correspondant à la structure du réseau. Nous avons décidé de récupérer que celles afférentes à la probabilité de vote.

```
proba_vote
```

```
##    y    Q38 Q17E      Freq
## 1  0 Autres  Non 0.5454545
## 2  1 Autres  Non 0.4545455
## 3  0 DROITE  Non 0.7694962
## 4  1 DROITE  Non 0.2305038
## 5  0 GAUCHE  Non 0.6390753
## 6  1 GAUCHE  Non 0.3609247
## 7  0 Autres  Oui 0.8275862
## 8  1 Autres  Oui 0.1724138
## 9  0 DROITE  Oui 0.8508634
## 10 1 DROITE  Oui 0.1491366
## 11 0 GAUCHE  Oui 0.7639198
## 12 1 GAUCHE  Oui 0.2360802
```

Les probabilités de vote pour le candidat DSK sont obtenues conditionnellement aux variables programmes télévisuels et opinion sur les immigrés.

La forte probabilité de vote pour le candidat étant 0, on a déduit que la probabilité de vote pour le candidat est très fort lorsque les électeurs sont favorables à l'immigration. Les proportions sont partagées pour ceux qui ne reconnaissent ni à droite ni à gauche et qui sont contre l'immigration.

Conclusion

A travers ce projet nous avons tenté de prédire si les électeurs d'un panel allaient voter pour le candidat Dominique Strauss-Kahn. Pour cela nous avions opéré une première sélection de variables sur les termes d'interactions pour déterminer lesquels pourraient améliorer le pouvoir prédictif de notre modèle.

Puis nous les avons intégrés à nos modèles regroupant les variables d'opinion et de contexte afin de sélectionner les variables ayant un réel impact significatif sur notre variable cible. Ces étapes de sélections de variables sont primordiales dans la mesure où elles sont la prémices de la modélisation. Elles sont comme les fondations d'une maison ; si elles sont inutiles ou mauvaises, la maison s'écroulera tout comme nos prédictions.

Nous avons donc modélisé 2 réseaux bayésiens ; un premier afin de déterminer les liens de causalité impactant la probabilité de vote pour DSK et un deuxième afin d'attribuer une probabilité de vote pour DSK aux électeurs.

Les réseaux bayésiens permettent d'interpréter très facilement les liens causaux existants entre les différentes variables et surtout de modéliser rapidement les probabilités conditionnelles. Ils sont un outil très puissant en matière d'apprentissage automatique.

Bbliiographie

- « Réseaux bayésiens et apprentissage ensembliste pour l'étude différentielle de réseaux de régulation génétique » - HT.Nguyen

- « Learning Bayesian Network with the bnlearn R Package » - M.Scutari

- « De l'identification de structure de réseaux bayésiens à reconnaissance de formes à partir d'informations complètes ou incomplètes » - O.François