Fouille de données médicales : Analyse et prédiction de données sur les maladies cardio-vasculaires

Nelly Barret
nelly.barret@etu.univ-lyon1.fr
Université de Lyon
Lyon, France

Valérian Acier
Université de Lyon
Lyon, France

Juliette Reisser
Université de Lyon
Lyon, France

Nour Medjedel
Université de Lyon
Lyon, France

ABSTRACT

La détection des maladies cardio-vasculaires se base sur un grand nombre d'indicateurs, qu'ils soient numériques (e.g. l'âge, le nombre de battements par minute du coeur, ...) ou par palier (e.g. le type de douleur au repos, le type de pente du segment ST, ...). Cette complexité fait émerger des difficultés à comprendre et à prédire les maladies cardio-vasculaires. Dans cet article, nous décrivons une approche pour détecter les indicateurs révélateurs et prédire de nouveaux cas. D'une part, nous avons déterminé les attributs caractéristiques des patients malades via une analyse fine du jeu de données. D'autre part nous avons mis en oeuvre des algorithmes de Data Mining afin de prédire de nouveaux cas.

CCS CONCEPTS

• **Computing methodologies → Supervised learning by regression**; **Model verification and validation**; • **Information systems → Content analysis and feature selection**.

KEYWORDS

maladie cardio-vasculaire, prédiction, corrélation, classification

ACM Reference Format:
Nelly Barret, Valérian Acier, Juliette Reisser, and Nour Medjedel. 2020. Fouille de données médicales : Analyse et prédiction de données sur les maladies cardio-vasculaires. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

L'explosion de la quantité de données (estimée à plus de 40 zettabytes en 2020 [1]) ainsi que l'augmentation du nombre de sources produisant des données (e.g. le nombre d'objets connectés en 2020 est estimé à 50 milliards [3]) posent de nouvelles problématiques. Bien que cette abondance de données ne cesse d'augmenter, les connaissances se font rares : "We are drowing in data, but starving for knowledge" (John Naisbitt, 1982). Avec l'évolution des technologies et dans le but de répondre à ces différents besoins sociétaux, le Data Mining a vu le jour dans les années 1990. Le Data Mining est un processus qui permet d'extraire d'un jeu de données de taille conséquente de la connaissance. Plus précisément, cette connaissance se caractérise par l'exploitation de motifs intéressants (non-triviaux, implicites, inconnus et potentiellement utiles) à partir de cette vaste quantité de données. De nos jours, le Data Mining est devenu une technique d'exploration reconnue qui permet d'ajouter une plus-value aux données existantes. Ces avancées tant au niveau des données que de leur exploitation mettent en lumière plusieurs cas d'applications, e.g. en recherche médicale, et plus particulièrement dans la recherche sur les maladies cardio-vasculaires. Bien que le nombre de malades diminue depuis les dernières décennies, les maladies cardio-vasculaires demeurent la principale cause de mortalité dans les pays de l'OCDE [6]. Près d'un tiers des décès en 2013 étaient dûs à ces maladies. Il semble donc important de comprendre et d'identifier les facteurs influants afin de diagnostiquer (ou prédire le diagnostic) les malades. Notre projet se base sur deux objectifs : prédire le diagnostic des patients et utiliser des algorithmes de Data Mining afin de mieux comprendre les facteurs qui agissent dans le cadre des maladies cardio-vasculaires. Notre approche propose l'implémentation de différents algorithmes de Data Mining pour la prédiction de diagnostics ainsi que des visualisations selon différentes métriques pour la compréhension des indicateurs. Les scripts sont programmés en Python et nous utilisons les librairies Scikit-Learn pour les algorithmes, Matplotlib pour les tracés et Pandas pour le traitement et la manipulation des données. Nous distinguerons quatre sous-objectifs :

(1) Choix d'un jeu de données
(2) Pré-traitement du jeu de données choisi
(3) Choix et utilisation d'algorithmes de Data Mining pertinents par rapport à la problématique énoncée
(4) Visualisation et interprétation des résultats

La Section 2 décrit les méthodes que nous avons mises en place pour la collecte et la préparation des données. En Section 3, nous expliquons les algorithmes que nous avons utilisés et nous montrons les résultats obtenus grâce à ce travail. Nous terminerons en Section 4 par une discussion sur les limitations et les perspectives de ce projet.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 MÉTHODES

2.1 Collecte des données

Nous avons choisi le jeu de données *Heart Disease Dataset* créé par David W. Aha en 1988 et disponible sur le répertoire *UCI Machine Learning* [2]. Le contenu de ce jeu de données représente les facteurs pouvant influer sur le risque de maladies cardiaques. Cette base de données regroupe des informations qui sont récoltées à partir de quatre établissements dans différentes villes :

(1) Hungarian Institute of Cardiology, Budapest: Andras Janosi, M.D.[1] → 294 instances
(2) University Hospital, Zurich, Switzerland: William Steinbrunn, M.D. → 123 instances
(3) University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
(4) V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D. → 200 + 303 instances

Chaque base de données contient 76 attributs comme le montre la Table 5 dont 14 (c.f. Table 1) qui sont souvent utilisés. Cette réduction d'attributs a été suggérée par un expert du domaine, ainsi, il est possible de prendre en compte seulement ces facteurs les plus impactants dans le cadre de la prédiction de diagnostic par exemple. Toutefois, nous avons décidé de garder la totalité des attributs afin d'obtenir des résultats détaillés.

Id	Description
3	âge (en années)
4	sexe ([homme, femme])
9	type de douleur thoracique (4 types)
10	tension artérielle (systolique) au repos (120 mmHg)
12	taux de cholestérol ([1.8 ... 2] g/L)
16	glycémie ([0.5 ... 1.5] g/L)
19	résultats des électrocardiogrammes au repos (3 types)
32	fréquence cardiaque maximale ([60 ... 100] bpm)
38	angine induite par l'effort ([non, oui])
40	dépression ST induite par l'exercice par rapport au repos (TODO)
41	pente du segment ST pendant l'effort (3 types)
44	nombre de gros vaisseaux sanguins ([0 ... 3])
51	thalassémie (3 types)
58	diagnostic final (oui, non) (*)

Table 1: Les 14 indicateurs définis par un expert du domaine

Plusieurs points importants sont à noter. Cette Table contient l'indicateur *diagnostic final*, repérable par (*), qui permet d'évaluer si un patient est malade ou non. C'est l'attribut qui sera utile pour la partie prédiction de notre projet. En effet, grâce à ces labels déjà pré-établis nous pourrons entraîner nos algorithmes de manière supervisée (voir Section 3) à diagnostiquer les maladies cardiaques. Le deuxième point important concerne les domaines de définition des indicateurs. Ils sont en fonction du type de celui-ci. En effet, les indicateurs de type palier, e.g. la douleur pendant l'effort ou le type de pente du segment ST, auront des domaines de définition de type énumération de valeurs. Ces valeurs sont ensuite converties en valeurs numériques afin d'être compatibles avec les indicateurs de type numériques. Par exemple, les résultats des électrocardiogrammes au repos sont définis comme tel : [ordinaire, anomalie de l'onde ST-T, hypertrophie ventriculaire gauche]. Ces indicateurs peuvent ensuite être transcrits comme des entiers : 0 pour ordinaire, 1 pour anomalie de l'onde ST-T et 2 pour hypertrophie ventriculaire gauche.

[1]M.D. : Doctor of Medicine

2.1.1 Analyse du jeu de données. La répartition des données dans le jeu de données est un paramètre qui peut aider dans le choix d'algorithmes pertinents et dans l'interprétation des résultats. Nous avons choisi de représenter les données via Matplotlib selon différentes métriques afin d'observer par exemple des caractéristiques particulières ou des caractéristiques importantes. Dans cette optique, nous avons tout d'abord visualisé la répartition des patients selon leur âge, comme le montre la Figure 1. Nous pouvons observer que le jeu de données contient plus de patients "jeunes" (moins de 50 ans) que de seniors. La moyenne d'âge, représentée par la droite rouge, est de 53 ans. Enfin, nous remarquons qu'il y a peu de données en termes de patients pour les âges plus élevés (au dessus de 65 ans).

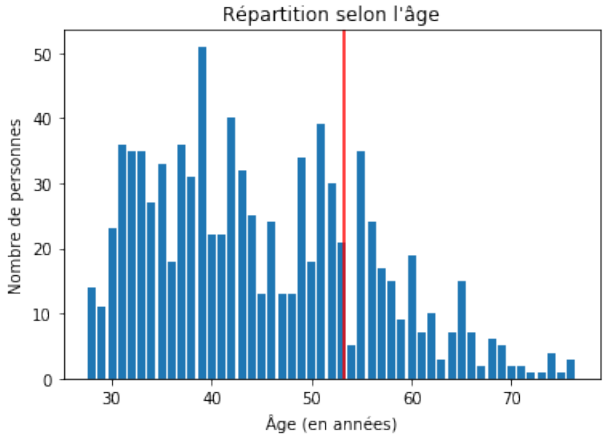


Figure 1: Répartition des patients selon l'âge

Le graphique précédent nous donne une vue d'ensemble de la répartition des données, il semble donc intéressant de visualiser comment les patients se répartissent en terme de diagnostic. La Figure 2 montre un *stacked bar chart* illustrant cette répartition selon les diagnostics. Cinq diagnostics sont possibles (entre 0 et 4) : aucun problème cardio-vasculaire (0), problème détecté (entre 1 et 4). Le jeu de données ne précise pas qu'elle est la signification de cette métrique. Nous pouvons imaginer qu'il s'agit de quatre stades : problème réversible avec peu de contre-indications (1) à problème irréversible avec contre-indications (4). Plusieurs tendances sont à noter dans cette visualisation. Très peu de patients de plus de 50 ans ont un diagnostic négatif (pas de problème cardio-vasculaire). En effet, l'âge semble être un facteur important dans les maladies cardio-vasculaires, ce qui semble pertinent par rapport à la connaissance actuelle dans ce domaine. Inversement, peu de jeunes

patients (moins de 50 ans) ont des diagnostics sévères (3 ou 4), ce qui semble pertinent avec le point ci-dessus.

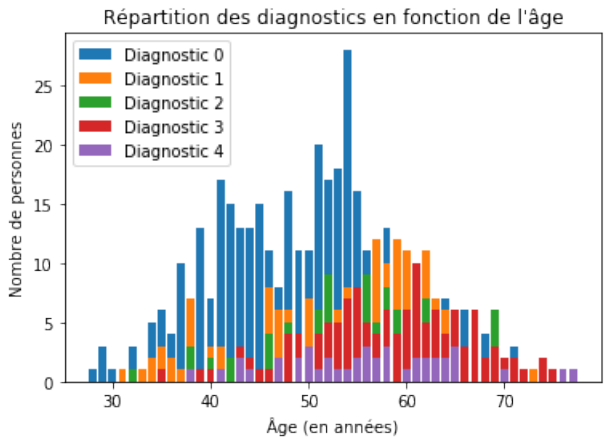


Figure 2: Répartition des diagnostics selon l'âge

Enfin, nous avons voulu comprendre la corrélation des indicateurs du jeu de données. Pour ce faire, nous avons utilisé une représentation via une matrice de corrélation. Soit X et Y, deux indicateurs, chacun respectivement sur l'axe X et l'axe Y. Chaque indice de corrélation calculé entre X et Y est défini sur [-1 ; +1], ce qui représente la force de l'indicateur X par rapport à l'indicateur Y. Quand le coefficient est proche de 1 (rouge foncé), cela indique que X est très corrélé à Y. Inversement quand il est très proche de -1 (bleu foncé), cela indique que Y est très corrélé à X. Quand il est proche de 0 (blanc), cela indique que les variables ne sont pas corrélées (du moins, les données actuelles ne permettent pas de détecter de corrélation). Nous avons réalisé cette matrice sur les 14 indicateurs les plus importants comme le montre la Figure 3, ainsi que sur l'ensemble des indicateurs comme le montre la Figure 7 (en Section 5). Nous pouvons voir que la plupart des indicateurs sont symétriques : quand X est corrélé à Y alors Y est corrélé à X. Notons tous de même que nous avons dû modifier quelque peu les données pour pouvoir les utiliser dans la matrice de corrélation. Comme la matrice utilise un quotient pour calculer la corrélation des variables, il ne faut pas qu'il y ait de dénominateur nul. Les données présentent un certain nombre de données manquantes, que nous avons choisi de remplacer arbitrairement par ∞. Cela mérite plus amples expérimentations afin de déterminer comment traiter ces données manquantes, sujet que traite cet article [7].

Nous avons vérifié les résultats obtenus dans notre matrice de corrélation (générée via la méthode *corr()* de Pandas) via une seconde méthode. Nous avons utilisé la fonction *pearsonr()* de SciPy, qui prend en paramètre les attributs dont on veut calculer la corrélation.

```
# 4 and 5 are highly correlated
0.9958609307383137
# 5 and 6 are highly correlated
0.9886758266578343
```

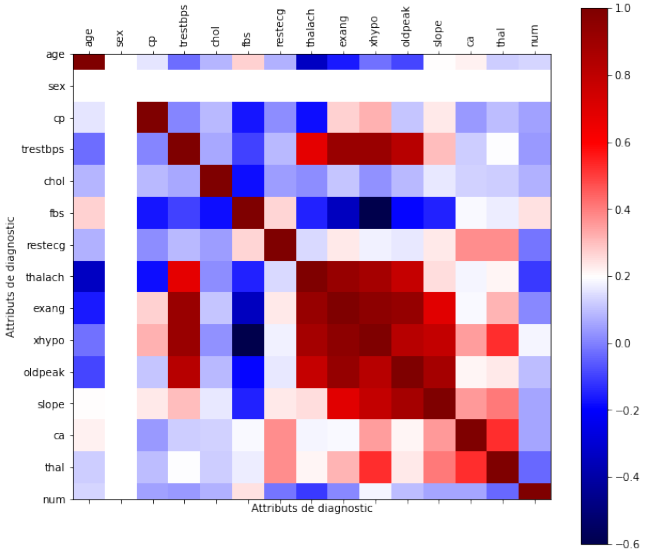


Figure 3: Matrice de corrélation entre les 14 indicateurs importants

Ci-dessus, nous montrons quelques résultats de cette deuxième méthode. Les résultats semblent cohérents avec la Figure 3, ce qui confirme de manière bilatérale les résultats de cette matrice de corrélation.

2.2 Préparation des données

Maintenant que nous avons analysé la répartition des données et que nous avons une vue plus précise des indicateurs et de leur corrélation, nous pouvons appliquer des méthodes de pré-traitement sur les données.

La Figure 4 représente l'importance des 76 indicateurs. Pour ce faire, elle représente l'écart-type via les lignes verticales dans chaque barre. L'écart-type mesure la dispersion des valeurs autour de leur moyenne. Cela implique que plus l'écart-type d'un indicateur est faible, plus les valeur sont homogènes. Nous pouvons observer que l'écart type de chaque indicateur est assez important. Nous pouvons le corréler avec le fait que les données sont hétérogènes au niveau des valeurs. Par exemple, certains indicateurs ont des valeurs par palier et des valeurs inconnues. Cela renforce la dispersion des données et donc augmente l'écart-type. Deuxièmement, nous observons que l'écart-type est de moins en moins élevé quand l'importance des indicateurs diminue.

Nous avons effectué un ensemble de pré-traitements pour supprimer les dimensions trop corrélées ou ayant une valeur unique. Via la matrice de corrélation appliquée sur les 76 features (c.f. Figure 7) nous avons éliminé un ensemble de 9 dimensions extrêmement corrélées (supérieur à 95%).

En utilisant une méthode d'élimination de dimensions récursive (*RFECV*) couplé à un estimateur *Random Forest*, nous sommes parvenus à établir que seulement 9 dimensions sont utiles pour classifier les malades comme le montre la Figure 5.

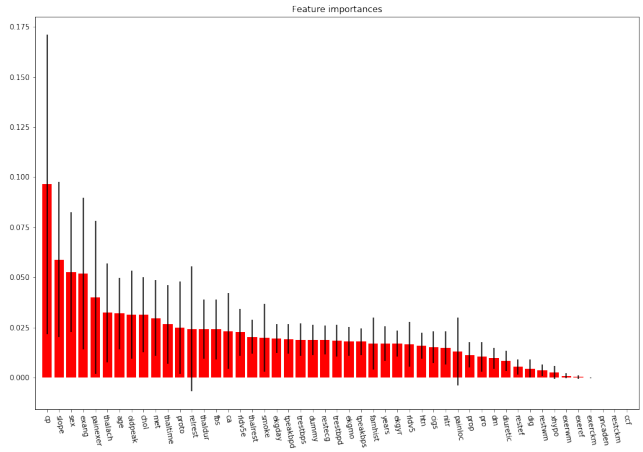


Figure 4: Importance des indicateurs

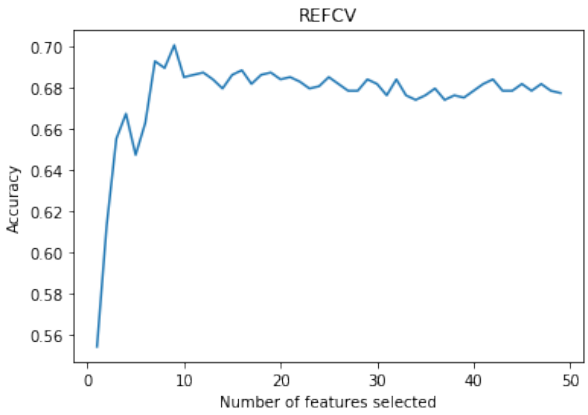


Figure 5: Précision de la méthode d'élimination RFECV en fonction du nombre de dimensions sélectionnés

Nous avons ensuite comparé différents algorithmes de réduction de dimensions avec certains classifieurs sur nos données, comme illustré par la Table 2.

3 RÉSULTATS

Par souci de concision, nous ne détaillerons pas les caractéristiques et l'utilisation des algorithmes utilisés dans cette section et ceux référencés en Table 2.

3.1 Classification

Nous pouvons voir sur la Table 2 que les meilleurs résultats obtenus le sont via l'utilisation d'un *Random Forest* sur un *PCA* avec un taux d'accuracy de 79%. Nous effectuons donc une recherche aléatoire sur les hyperparamètres du *Random Forest* pour ainsi améliorer les résultats nous permettant d'atteindre jusqu'à 81% d'accuracy et 81.7% courbe roc.

	PCA	SVD	ICA	ISOM	TSNE
RF	**79.0**	65.0	71.7	62.9	66.2
KNN	70.7	66.1	68.7	65.8	69.0
Simple Bayes	62.9	60.3	61.2	61.7	62.5
SVC	55.1	53.9	55.1	60.7	69.5
Bagging (SVC)	55.1	53.8	55.7	60.6	69.2
AdaBoost	77.3	64.6	72.6	65.7	67.1
MLP 10x	44.3	44.9	73.3	49.5	69.6
MLP 20x	45.4	44.9	68.1	44.9	70.2

Table 2: Comparatif des résultats de différents algorithmes sur des méthodes de réduction de dimensions (accuracy en %).

Abbréviations : RF : Random Forest; KNN: K-Nearest Neigbhor; SVC : Support Vector Classification; MLP : Multi Layer Perceptron; PCA : Principal Component Analysis; SVD : Singular Value Decomposition; ICA : Independent Component Analysis; ISOM : Isometric Mapping; TSNE : T-distributed Stochastic Neighbor Embedding

	PF	PT	
AF	96	33	0.74
AT	31	55	0.64
	0.75	0.62	0.80

Table 3: Matrice de confusion du *Random Forest* sur un jeu de données de test.
Abbréviations : PF, Predicted False; PT Predicted True; AF, Actual False; AT, Actual True

3.2 Sous-groupes

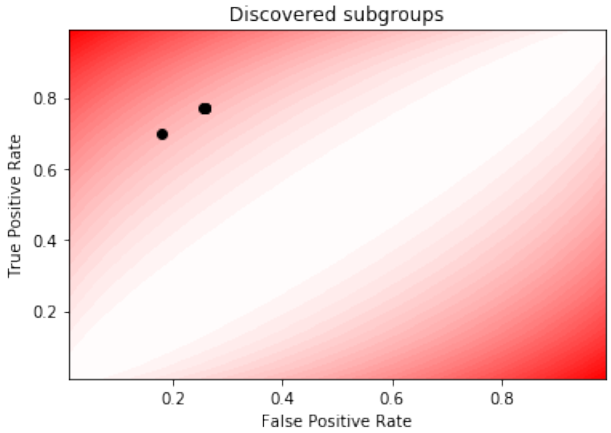
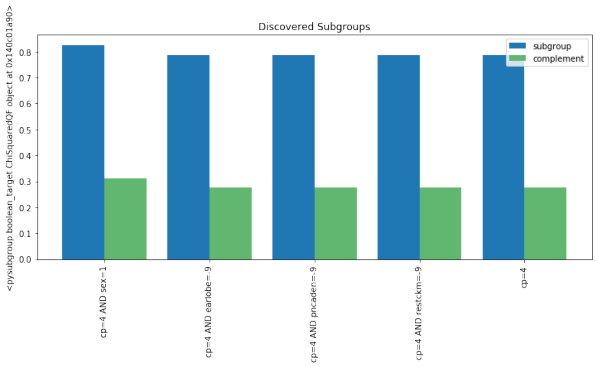


Figure 6: Représentation des sous-groupes dans l'espace ROC.
Représentation issue de la librairie pysubgroup [5]

Grâce à l'analyse de sous-groupes basé sur la librairie Pysubgroup [5] visible dans la Table 4, nous avons pu extraire différentes informations notamment le fait que les cas de problèmes cardiaques



Les barres bleues représentent les sous-groupes découverts, les barres vertes leur complément dans les données. La hauteur des barres montre leur présence dans les données, la largeur représente le nombre d'instances qu'elles couvrent. Représentation issue de la librairie pysubgroup [5].

Quality	Subgroup	Target_share_sg
240.13	cp=4 AND sex=1	82.6
237.16	cp=4 AND earlobe=-9	78.7
234.96	cp=4 AND pncaden=-9	78.5
234.96	cp=4 AND restckm=-9	78.5
234.96	cp=4	78.5

Table 4: Comparatif des différents sous-groupes identifiés.

sont majoritairement sur des hommes ayant une douleur asymptomatique à la poitrine.

4 DISCUSSION

Dans cette dernière section, nous allons présenter les limitations de notre projet, des perspectives en réponse à celles-ci et une conclusion résumant les principaux points de cet article. Premièrement, le jeu de données que nous avons utilisé propose un nombre de diagnostics important, pourtant cela semble peu dans le cadre de la prédiction de maladies cardio-vasculaires. Un plus ample jeu de données pourrait permettre de répondre à d'autres problématiques telles que :

- Amélioration de la prédiction de nouveaux diagnostics
- Prédiction aussi pertinente que ce soit pour un homme ou une femme, un jeune patient ou un senior, une personne ayant des antécédents médicaux ou non, ...
- Détecter des points communs dans les habitudes des malades (alimentaires, environnement ...)

Deuxièmement, les indicateurs décrits dans le jeu de données permettent de décrire des symptômes (e.g. la douleur pendant l'effort), des caractéristiques propres au patient (e.g. l'âge et le sexe) et un état de santé selon des constantes vitales (e.g. la fréquence cardiaque et le taux de cholestérol). Nous notons que 76 indicateurs permettent de décrire de manière plutôt précise l'état de santé actuel du patient. Seulement, cet ensemble d'indicateurs ne permet pas de prendre en compte certains facteurs extérieurs mais qui pourtant influencent la santé du patient, e.g. le diabète héréditaire, un patient plus sujet au cholestérol que la moyenne, un patient présentant un souffle au coeur, ... Ces caractéristiques peuvent être des facteurs influants qu'il semble nécessaire de prendre en compte dans le cadre de la prédiction de diagnostics de maladies cardio-vasculaires. De plus, ce type de données extérieures peut permettre de les corréler aux indicateurs existants pour peut-être trouver de nouvelles explications. Une troisième perspective serait de modéliser ces indicateurs selon une ontologie pour apporter une autre échelle de connaissances. Plusieurs choix sont possibles pour la mise en place de cette hiérarchie. Un premier choix serait de regrouper les indicateurs par catégorie. Un second choix serait de construire une hiérarchie d'indicateurs. Cela permettrait entre autres d'avoir une seconde version de la corrélation entre les indicateurs. En effet, un indicateur qui est fils d'un autre est corrélé à son parent. Cela pourrait augmenter la connaissance sur la corrélation des indicateurs entre eux et donc mieux comprendre les diagnostics, voire même prédire de manière encore plus fiable le diagnostic de nouveaux patients.

Pour conclure, la réduction du nombre d'indicateurs et la recherche de sous-groupes a montré qu'il est possible d'extraire les facteurs importants des maladies cardio-vasculaires, et donc d'aller encore plus loin avec des jeux de données plus conséquents. Nous avons aussi montré que nous pouvons atteindre plus de 80% de détection correctes, ce qui est un score dans la moyenne des détections faites par un cardiologue [4].

REFERENCES

[1] Khalid Adam. 2015. Big Data Analysis and Storage.
[2] Matthias Pfisterer Andras Janosi, William Steinbrunn and Robert Detrano. 1992. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Heart+Disease
[3] Gary Davis. 2018. 2020: Life with 50 billion connected devices. In *2018 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 1–1.
[4] Mark Hlatky, Elias Botvinick, and Bruce Brundage. 1982. Diagnostic accuracy of cardiologists compared with probability calculations using Bayes' rule. *The American Journal of Cardiology* 49, 8 (1982), 1927 – 1931. https://doi.org/10.1016/0002-9149(82)90211-9
[5] Florian Lemmerich and Martin Becker. 2019. pysubgroup: Easy-to-Use Subgroup Discovery in Python. In *Machine Learning and Knowledge Discovery in Databases*, Ulf Brefeld, Edward Curry, Elizabeth Daly, Brian MacNamee, Alice Marascu, Fabio Pinelli, Michele Berlingerio, and Neil Hurley (Eds.). Springer International Publishing, Cham, 658–662.
[6] OCDE. 2015. *Cardiovascular Disease and Diabetes: Policies for Better Health and Quality of Care.* 180 pages. https://doi.org/https://doi.org/10.1787/9789264233010-en
[7] Azizur Rahman and Ajit Majumder. 2013. Effects of Missing Value Estimation Methods in Correlation Matrix- A Case Study of Concrete Compressive Strength Data. *International Journal of Advanced Science and Technology* 52 (04 2013).

5 ANNEXES

Table 5: Liste des 76 indicateurs disponibles dans le jeu de données

Indicateur important	Numéro	Label	Description	Domaine de définition	Domaine de définition numérique
	1	id	identifiant du patient		
	2	ccf	numéro de sécurité sociale		
✓	3	age	âge (en années)	[28, 77]	[28, 77]
✓	4	sex	sexe	[homme, femme]	[1, 0]
	5	painloc	emplacement de la douleur thoracique	[substernal, autre]	[1, 0]
	6	painexer	douleur pendant l'effort	[dûe à l'effort, autre]	[1, 0]
	7	relrest	?	[soulagé après le repos, autre]	[1, 0]
	8	pncaden	sommes des indicateurs 5, 6 et 7	_	[0, ..., 3]
✓	9	cp	type de douleur thoracique	[angine de poitrine normale, angine de poitrine anormale, douleur ne provenant pas de l'angine, asymptomatique]	[1, 2, 3, 4]
✓	10	trestbps	tension artérielle au repos (en mmHg)	[90/60, 130/80]	[90/60, 130/80]
	11	htn	hypertension	[oui, non]	[1, 0]
✓	12	chol	taux de cholestérol sérique (en mmol/l)	environ 4,64	± 4,64
	13	smoke	patient fumeur	[oui, non]	[1, 0]
	14	cigs	nombre de cigarettes par jour		
	15	years	nombre d'années de tabagisme		
✓	16	fbs	glycémie (en mg/dl)	[>120, <= 120]	[1, 0]
	17	dm	diabète héréditaire	[oui, non]	[1, 0]
	18	famhist	antécédants familiaux sur les maladies de l'artère coronaire	[oui, non]	[1, 0]
✓	19	restecg	résultats des ECG au repos	[normal, courbe ST-T anormale, hypertrophie ventriculaire gauche]	[0, 1, 2]
	20	ekgmo	mois de l'ECG	[1, 12]	[1, 12]
	21	ekgday	jour de l'ECG	[1, 31]	[1, 31]
	22	ekgyr	année de l'ECG		
	23	dig	utilisation de digitales (plantes) pendant l'ECG	[oui, non]	[1, 0]
	24	prop	utilisation de bêta-bloquants pendant l'ECG	[oui, non]	[1, 0]
	25	nitr	utilisation de nitrates pendant l'ECG	[oui, non]	[1, 0]
	26	pro	utilisation de bloqueurs de canaux calciques pendant l'ECG	[oui, non]	[1, 0]
	27	diuretic	utilisation de diurétiques pendant l'ECG	[oui, non]	[1, 0]
	28	proto	protocole pour l'exercie	[Bruce, Kottus, McHenry, fast Balke, Balke, Noughton, vélo (150 kPa), vélo (125 kPa), vélo (100 kPa), vélo (75 kPa), vélo (50 kPa), rameur]	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
	29	thaldur	durée de l'exercice (en minutes)		
	30	thaltime	moment où la dépression mesurée par ST est notée		
	31	met	METS (MEtabolic EquivalentS) réussis	[0, 1, 2, ...]	[0, 1, 2, ...]
✓	32	thalach	fréquence cardiaque maximale atteinte (en bpm)		
	33	thalrest	fréquence cardiaque au repos (en bpm)		
	34	tpeakbps	pression artérielle maximale (en Pa)		
	35	tpeakbpd	pression artérielle maximale (en Pa)		

	36	dummy	_	_	_
	37	trestbpd	pression artérielle au repos (en Pa)		
✓	38	exang	effort qui induit une angine de poitrine	[oui, non]	[1, 0]
	39	xhypo	?	[oui, non]	[1, 0]
✓	40	oldpeak	dépression ST induite par l'effort		
✓	41	slope	pente du segment ST lors de l'effort		
	42	rldv5	hauteur du sommet au repos		
	43	rldv5e	hauteur du sommet lors de l'effort		
✓	44	ca	nombre de vaisseaux (colorés par fluoroscopie)	[0, 1, 2, 3]	[0, 1, 2, 3]
	45	restckm	_		
	46	exerckm	_		
	47	restef	radionuclide au repos		
	48	restwm	anomalie de mouvement	[aucune, modérée, sévère, akinésie]	[0, 1, 2, 3]
	49	exeref	?	?	?
	50	exerwm	?	?	?
✓	51	thal	thalassémie	[normal, défaut corrigé, défaut réversible]	[3, 6, 7]
	52	thalsev	_	_	_
	53	thalpul	_	_	_
	54	earlobe	_	_	_
	55	cmo	mois du cathétérisme cardiaque	[1, ..., 12]	
	56	cday	jour du cathétérisme cardiaque	[1, ..., 31]	
	57	cyr	année du cathétérisme cardiaque		
✓	58	num	diagnostic final (oui/non)	[<50%, >= 50%]	[0, 1]
	59	lmt	_	_	_
	60	ladprox	_	_	_
	61	laddist	_	_	_
	62	diag	_	_	_
	63	cxmain	_	_	_
	64	ramus	_	_	_
	65	om1	_	_	_
	66	om2	_	_	_
	67	rcaprox	_	_	_
	68	rcadist	_	_	_
	69	lvx1	_	_	_
	70	lvx2	_	_	_
	71	lvx3	_	_	_
	72	lvx4	_	_	_
	73	lvf	_	_	_
	74	cathef	_	_	_
	75	junk	_	_	_
	76	name	nom du patient	_	_

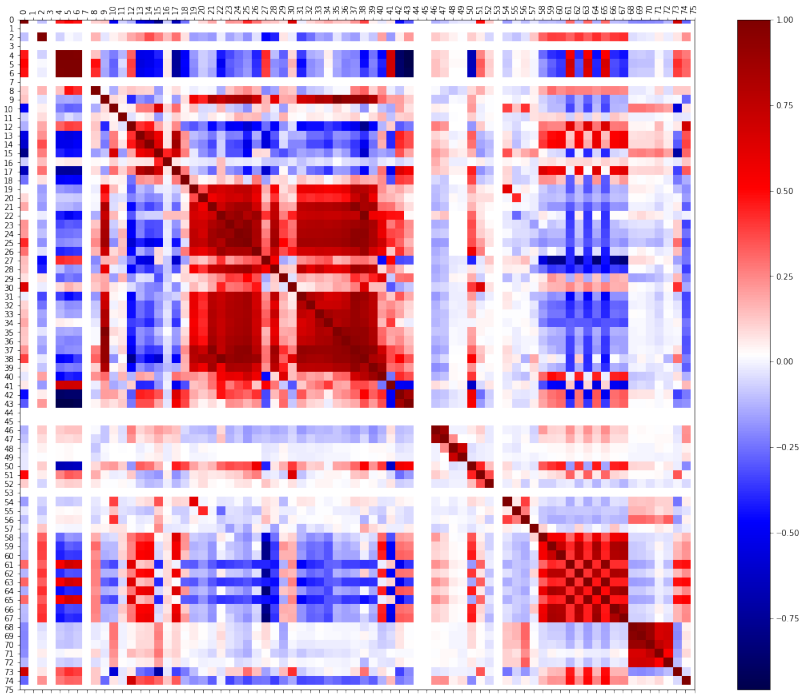


Figure 7: Matrice de corrélation entre les 76 indicateurs
La Figure 7 montre les coefficients de corrélation entre tous les indicateurs sous la forme d'une matrice. Comme indiqué pour la Figure 3, deux indicateurs non corrélés seront blancs, deux indicateurs fortement corrélés seront soit rouge foncé soit bleu foncé en fonction de l'indicateur qui domine. Un dégradé de couleur se réalise pour représenter toutes les valeurs des coefficients de corrélation.