Journal of Economics and Applied Statistics
Volume 17 number 3, December 2020
ISSN : 1112-234X EISSN : 2600-6642
Pages : 137-153

EVALUATION OF RISK FACTORS RELATED TO THE TOPOGRAPHICAL CHARACTERISTICS OF DEEP VEIN THROMBOSIS USING A DATA MINING APPROACH

EVALUATION DES FACTEURS DE RISQUE LIES AUX CARACTERISTIQUES TOPOGRAPHIQUES DE LA THROMBOSE VEINEUSE PROFONDE SELON UNE APPROCHE DE FOUILLE DE DONNEES.

[*]Nora LOUNICI MOSBAH
Ecole Nationale Supérieure de Statistique et d'Economie Appliquée (ENSSEA)
Laboratoire de Statistique Appliquée(LASAP)
noralounici@yahoo.fr

Khadidja SADI
Ecole Nationale Supérieure de Statistique et d'Economie Appliquée (ENSSEA)
Laboratoire de Statistique Appliquée(LASAP)
sadikh00@gmail.com

Asma SENIANI
Ecole Nationale Supérieure de Statistique et d'Economie Appliquée (ENSSEA)
senianialex@gmail.com

Reçu le : 2020/01/18 **Accepté le :** 2020/06/12 Publication **en ligne le** : 2020/12/31

ABSTRACT: Our study focuses on the risk factors of the three forms of venous thrombosis: Proximal, distal and superficial and at its main complication pulmonary embolism. The prospective study took place in a care unit of the hospital in Boumerdes, a city located 45 km east of Algiers. For the occasion, we digitized patient records over a period of 9 years between 2007 and 2017. The sample is composed of 302 individuals aged 16 to 95 years. We seek by this work to highlight the application of data mining techniques in the health field. The most predictive variables were identified by two techniques: *decision trees and multinomial logistic* regression. The results, which fit with the reality of the field, allowed us to confirm that the variables: *postpartum, age, surgery, history of MTEV and the complication of pulmonary embolism* are the most pertinent, they contribute strongly to explain the target variable TVP TOPOGRAPHY.

keywords : venous thromboembolic diseases(MTEV), risk factors, data mining, decision tree, multinomial logistic regression

JEL Classification : C14, C19, C25, D81, D83

RESUME : Notre étude s'intéresse à l'extraction des facteurs de risque des trois forment de la thrombose veineuse : Proximale, distale et superficielle ainsi qu'à sa principale complication, l'embolie pulmonaire(EP). L'étude prospective s'est déroulée dans une unité de soin de l'hôpital de Boumerdes, ville situé à 45 km à l'est d'Alger. Nous avons pour l'occasion numérisée les dossiers des malades sur une période de 9 ans entre 2007 et 2017. L'échantillon est composé de 302 individus âgés de 16 ans à

[*] Auteur Correspondant

95 ans. On cherche par ce travail à mettre en avant l'application des techniques d'exploration de données dans le domaine de la santé. Les variables les plus prédictives ont été identifiées par deux techniques : les *arbres de décision et la régression logistique multinomiale*. . Les résultats qui s'adaptent avec la réalité du terrain, nous ont permis de confirmer que les variables : *post-partum, âge, chirurgie, antécédent de MTEV et la complication d'embolie pulmonaire* sont les plus pertinentes, elles contribuent fortement à expliquer la variable cible TOPOGRAPHIE TVP.

Mots clés : maladies thromboemboliques veineuses (MTEV), facteurs de risque, data mining, arbre de décision, régression logistique multinomiale

1. INTRODUCTION

La thrombose veineuse profonde (TVP) résulte de la formation de caillots de sangs à l'intérieur des veines profondes, généralement dans les jambes et altère la circulation sanguine. Elle peut survenir à la suite d'une intervention chirurgicale, lors d'une grossesse ou en cas d'immobilisation prolongée. La TVP regroupe une affection grave, la thrombose veineuse profonde proximale, pouvant se compliquer en embolie pulmonaire (EP). La TVP distale et la TV Superficielle peuvent se compliquer, mais leur gravité à long terme est discutée[†]. La prévalence de la TVP est de 25 % chez les patients qui se présentent avec une suspicion de thrombose. Ces TVP sont de nature proximale dans 80 % des cas (SS. Anand, PS. Wells 1985) et le taux de mortalité liée à une complication EP est d'environ 25 %, en l'absence de traitement. Une aggravation des TVP distales vers une EP a été évoquée dans 4 à 20 % des cas.

Des travaux ont montré que 20 % des TVP distales se compliquent en proximales dans les 10 premiers jours suivants leur apparition (CI. Lagerstedt CI, & al 1985), (V. Kakkar & al. 1969) (R. Moreno-Cabral R, & al 1976) ont retrouvé des cas d'embolie pulmonaire chez 8-34 % des patients avec une TVP distale. Quant à Lohr et al., ils relatent dans leur article un taux d'extension des TVP distale de 32 % (JM. Lohr & al. 1995). Il existe donc un certain nombre de données dans la littérature qui suggèrent que la TVP distale est potentiellement dangereuse.

Des études néerlandaise et canadienne ont émis l'hypothèse que seules les TVP Proximales étaient redoutables, et que le risque d'EP dans le cas d'une TVP distale pouvait être négligé (A. Cogo & al. 1998), (RA. Kraaijenhagen & al 2002). En revanche, dans les pays comme la France et l'Angleterre, les sujets souffrant de TVP distale sont systématiquement soumis à des examens complémentaires, pour éviter tout risque d'aggravation. Un des facteurs de la TVP est l'âge. L'incidence de la TVP (M. Méan & D. Aujesky 2009) est d'un cas pour 1000 /année avant 50 ans. Après 70 ans, elle est 10 fois supérieure et est affligée d'une surmortalité immédiate et tardive (2 à 3 ans après l'événement).

[†]https://www.has-sante.fr/portail/upload/docs/application/pdf/2010-12/fiche_de_bon_usage_Compression_ medicale_dans_le_traitement _ de_la_maladie_thrombo-embolique_veineuse.pdf

En Algérie, les chiffres du registre des malades atteints d'une TVP manifeste une croissance incontestable de cette pathologie. Cependant, il n'y a pas d'étude sérieuse publiée permettant d'apporter des informations sur l'épidémiologie de la TVP.

C'est dans ce contexte que se situe ce travail qui a comme objectif d'identifier les facteurs de risque liés aux caractéristiques topographies de la TVP de façon épidémiologique et descriptive selon une démarche de data mining (S. Tuffery 2012). Les variables que nous avons collectées concernent le diagnostic, les antécédents familiaux et personnels, ainsi que le bilan de guérison des patients atteints d'une TVP. Pour cerner identifier l'apparition des facteurs de risque de la TVP selon la topographie (proximale, distale et superficielle), nous avons appliqué deux modèles : les arbres de décision et la régression logistique multinomiale. L'extraction des connaissances a été réalisée au moyen de deux logiciels de data mining : *Tanagra et Weka*.

2. METHODOLOGIE

2.1 Description, nettoyage préparation du jeu de données

Les données ont été recueillies à partir d'un seul outil d'investigation : l'analyse et la collecte du contenu des dossiers des patients hospitalisés pour une TVP et sa complication l'EP. Elles proviennent de l'hôpital de Thenia, Boumerdes (service de médecine interne). Nous avons ainsi, collecté et numérisé des dossiers de 302 malades d'âge $\geq 16 ans$ sur une période de 9 ans[‡]. 21 variables représentant les facteurs de risque de la TVP, ont été sélectionnées selon leur fréquence d'apparition. Lors de cette phase cruciale (D. Pyle 1999), nous avons d'abord repérer et éliminer les erreurs, nous avons également procédé au traitement des valeurs manquantes et aberrantes. Certaines variables, présentant un effectif estimé trop faible ont été éliminées, nous avons également fusionné certaines variables (ex : *cancer* et *situation médicale aigue)* afin de constituer un échantillon représentatif. La variable âge a été discrétisée en quatre classes**.** *Après nettoyage et préparation des données, notre BD comprend désormais 302 individus (162 femmes et 140 hommes) et 15 variables qualitatives dont la variable à expliquer Topographie(Prxl (proximale); Dstl (Distale) ; Sprl (superficielle)* (Tableau 1).

Tableau1: caractéristiques des attributs de la base de données

Variables	**Modalités**	**Variables**	**Modalités**
Sexe	F : femme ; H : homme	Complication EP	Oui / Non
Age	[15- 35[; [35-55[; [55- 75[; >=75	Obésité	Oui / Non
Chirurgie	Oui / Non	Hypertension(HTA)	Oui / Non
Grossesse	Oui / Non	Diabète	Oui / Non

‡ Entre 2007 et 2010 un incendie s'est produit au service de médecine interne «service homme», ce qui a conduit à la déperdition de certains dossiers.

Post-partum	Oui / Non	Cardiopathie	Oui / Non
Tabac	Oui / Non		
Antécédent MTEV[§] (ATCD MTEV)	P-EMBL*(personnel)/P-PHLB**(personnel)/F-EMBL (familiale) / F-PHLB(familiale)/Non		
Traumatisme	Imo-pltr(immobilisation plâtré)/Fracture/Autres/Non		
SME***	Handicapé, Cancéreux, VRC(varices), AVC, PSY(pbs psychiatriques)/Autres		
Topographie	**Prxl (proximale)/ Dstl (Distale) ; Sprl (superficielle)**		

***** EMBL:Embolie ** PHLB : Phlébite ***Situation médicale aigue

2.3 Exploration des données

Durant la période d'étude, on observe une fréquence de 54% de femmes d'âge moyen estimé à 49,7 ± 16,5 ans et 46% d'hommes dont l'âge moyen est de 51,96 ± 15,8 ans. Une TVP Proximale (Prxl) est survenue chez la moitié des patients, l'autre moitié se répartie entre la TV Superficielle (Sprl) et la TVP distale (Dstl) avec respectivement 31% et 16% . 80% des patients ont subi une TVP pour la première fois, 19% après un premier épisode. Parmi ces 19% de malades, 17% sont atteints par une phlébite et 2% par une embolie et seulement 1% ont des antécédents familiaux de phlébite. Une embolie pulmonaire (EP) est constatée chez 30 patients (10%). Par ailleurs, 7% des femmes ont été touchées par une TVP au cours de leur grossesse et 18% en post-partum.

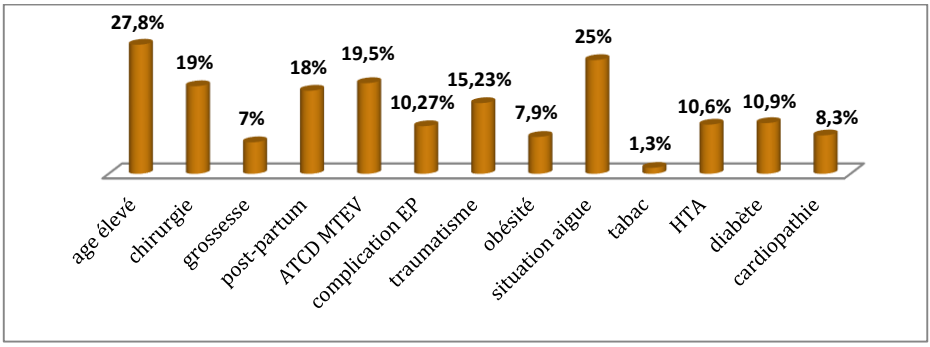
La variable obésité n'est pas renseignée pour beaucoup de patient et concerne seulement 8%. Le diabète et la HTA avec une proportion de 11% ne sont pas associés au risque de TVP dans notre échantillon, alors que dans les études mondiales se sont des facteurs de risque les plus fréquents.

2.3.1 Prévalence des facteurs de risque de la TVP

D'après la figure ci-dessous, on remarque que l'âge élevé est le facteur de risque le plus courant avec un effectif de 81 personnes (28%), suivi de la situation médicale aigue (25%). Ces deux facteurs sont corrélés entre eux car la majorité des patients âgés ont une situation médicale aigue. En 3ème place l'antécédent MTEV et la chirurgie, avec environ 19% (57 cas), puis le traumatisme (15%). En outre, une TVP est survenue chez 29 femmes en post-partum alors que pendant la grossesse on ne compte que 11femmes. Enfin, en dernière position, le diabète (11%), l'HTA (10,5%), la cardiopathie (8%) et l'obésité (8%). La plupart des facteurs sont fortement liées à la TVP proximale.

Figure1 : Prévalence des facteurs de risque de la TVP.

[§] MTEV : maladie thromboembolique veineuse



2.3.2 Répartition de la variable TOPOGRAPHIE selon l'âge et sexe

Le Tableau 2 montre que la Prxl est la pathologie la plus fréquente avec 161cas (> 50%). A noter que, la Prxl est le type le plus fréquent et à risque le plus grave d'EP[**]. En seconde position, on retrouve la Sprl qui survient chez 93 personnes et enfin la Dstl présente chez 48 cas. La comparaison de la topographie par classe d'âge fait apparaitre un accroissement de la Dstl chez femmes plutôt jeunes et les hommes âgés, alors que la Prxl concerne essentiellement les femmes âgées de [35; 55[et les hommes d'âge < 75ans et enfin la Sprl est prédominante chez les femmes avec 41%.

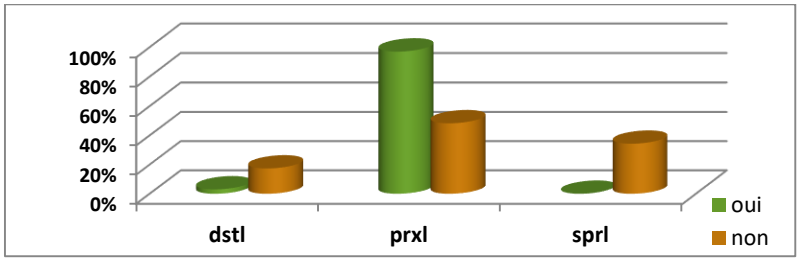
Tableau 2 Répartition des cas selon l'âge, le sexe et la topographie

AGE ➔	**[15;35[**	**[35;55[**	**[55;75[**	**≥75**	**Σ**	**%**		**[15;35[**	**[35;55[**	**[55;75[**	**≥75**	**Σ**	**%**
Femme	**44**	**60**	**33**	**25**	**162**	**53,6**	**Homme**	**34**	**40**	**48**	**18**	**140**	**46,4**
Dstl	11	11	4	2	28	17,3	Dstl	5	6	7	2	20	14,3
Prxl	20	**30**	19	19	**88**	**54,3**	Prxl	20	20	23	10	73	**52,1**
Sprl	13	**19**	10	4	46	28,4	Sprl	9	14	18	6	47	33,6

On constate 85% des patients ayant eu une complication d'EP sont atteints d'une Prxl. Les patients souffrant d'une Dstl sont beaucoup plus faible, environ 17%. Enfin, la fréquence de l'EP avec la Sprl est pratiquement nulle. D'après les études réalisées, on sait que 70 à 90 % des EP sont dues à une TVP des membres inférieurs. Ce qui est en harmonie avec les résultats obtenus à l'échelle internationale.

Figure2 : Répartition de la variable topographie selon la complication EP.

[**] Environ 50 % des patients ayant une TVP proximale ont aussi une EP sur l'angioscanner pulmonaire mais cliniquement asymptomatique ; https://sfcardio.fr/sites /Ref_Cardiologie/ch21_maladie_veineuse_te.pdf



3. BREVE DESCRIPTION THEORIQUE DES METHODES DE CLASSIFICATION UTILISEES

3.1 Les arbres de décision

Les arbres de décision font partie des méthodes supervisées en data mining. Le modèle établit sa décision en suivant une succession de tests basés sur les variables explicatives et construit sous la forme d'un arbre. Il a pour but de répartir un ensemble d'instance en classes homogènes relativement à une variable cible. Chaque nœud de l'arbre représente un test sur une variable explicatives. Les branches sortant des nœuds correspondent aux différentes modalités de la variable. À l'issue de ces tests, les nœuds de décision appelés feuilles sont étiquetées par la classe. A toute description complète est associée une et une seule feuille de l'arbre de décision.[††] Les deux algorithmes les plus utilisés sont CART (L. Breiman & al 1984) et C4.5 (J.R. Quinlan 1993). Pour sélectionner parmi les variables celle qui sépare le mieux les individus, des critères de qualités sont utilisés. CART adopte l'indice de Gini et C4.5 la notion d'entropie de Shannon. L'algorithme d'arbre de décision que nous avons choisi d'appliquer est C4.5 (J48 dans WEKA[‡‡]). Cet algorithme est basé sur le test de Gain Ratio comme mesure d'évaluation de partitionnement. Le gain est une mesure de segmentation, généré par la différence entre la répartition des classes dans le jeu d'apprentissage et la répartition des valeurs des attributs par rapport aux classes[§§].

Soient k classes disjointes de proportions respectives p_1, p_2, …, p_k, l'entropie I est définie par : $I = -\sum_{j=1}^{m} p_i \log_2 (p_i)$

La mesure du gain d'information que l'on obtient après partitionnement selon une variable est donnée par la formule : $Gain(S,T) = I(S) - \sum_{i=1}^{m} p(s_i) * I(S_i)$

[††] http://pageperso.lif.univ-mrs.fr/~francois.denis/IAAM1/chap2.pdf, consulté le 30/03/2018 à 20 :14.

[‡‡] http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html

[§§] ALAIN GIRARD, Extrapolation d'un algorithme génétique et d'un arbre de décision à des fins de catégorisation, thèse de doctorat publié, Université du Québec à Montréal, Canada, 2007, P.25.

- $T = \{S_1,\ S_2, \dots, S_m\}$: sous ensemble de partitionnement selon les classes $C_1,\ C_2, \dots, C_m$ *et* $p(s_i) = \frac{|S_i|}{|S|}$
- $I(S)$ *et* $I(S_i)$: correspondent aux entropies de *S* et de S_i.

Le gain est biaisé[***] en présence d'attributs ayant un grand nombre de valeurs, pour éviter cette contrainte C4 .5 appose le ratio de gain : $Gain_{Ratio}(S,T) = Gain(S,T)/\ I(S)$

3.2 La régression logistique multinomiale (RLM)

La RLM est une extension de la régression logistique binaire pour un problème à m classes (J. M. Hilbe 2009). Ce qui revient à effectuer (m-1) régressions logistiques binomiales(LOGIT) correspondantes aux combinaisons d'une modalité de référence (par exemple la dernière *Ym*) avec les m-1 autres modalités.
Ainsi, pour chacune de ces paires de classes (Y_j, Y_m), il existe une fonction décrite sous forme matricielle par l'équation suivante : $\log \frac{P(Yj/X=x)}{P(Ym/X=x)} = \alpha_j + \sum_k \beta_{jk} x \qquad j = 1, \dots \dots m\text{-}1$
Où X est le vecteur des observations, α_j et β_{jk} sont les paramètres du modèle pour la classe k.

L'équation traduit la probabilité d'appartenance d'un individu à une classe (Sheskin, 2007). L'objectif est d'estimer les coefficients inconnus β_m^T à partir des données. On utilise pour cela la méthode du maximum de vraisemblance (on préfère le log-vraisemblance). Le modèle produit la constante (α_m), les coefficients de régression, des statistiques Wald et un coefficient de détermination.

- L'évaluation globale du modèle de RLM se fait au moyen du Pseudo-R2 de Mac Fadden, du critère d'Akaike (AIC) et de Schwartz (noté BIC), de la matrice de confusion et du Test du rapport de vraisemblance.
- La statistique Wald (ratio entre β et le terme d'erreur) permet de tester la significativité des coefficients et la contribution de X = x sur Y, on retient Wald> 3.84.

4. EXTRACTION DES CONNAISSANCES

Nous avons commencé par appliquer le test du Khi2 afin de vérifier si les variables explicatives sont significatives et contribuent à expliquer réellement la variable à prédire *topographie*. Il en ressort que les variables *Complication EP*, *ATCD MTEV*, et *CHIRURGIE* sont corrélées avec la variable cible. C'est ce que nous allons tenter de vérifier en expérimentant les méthodes de classification.

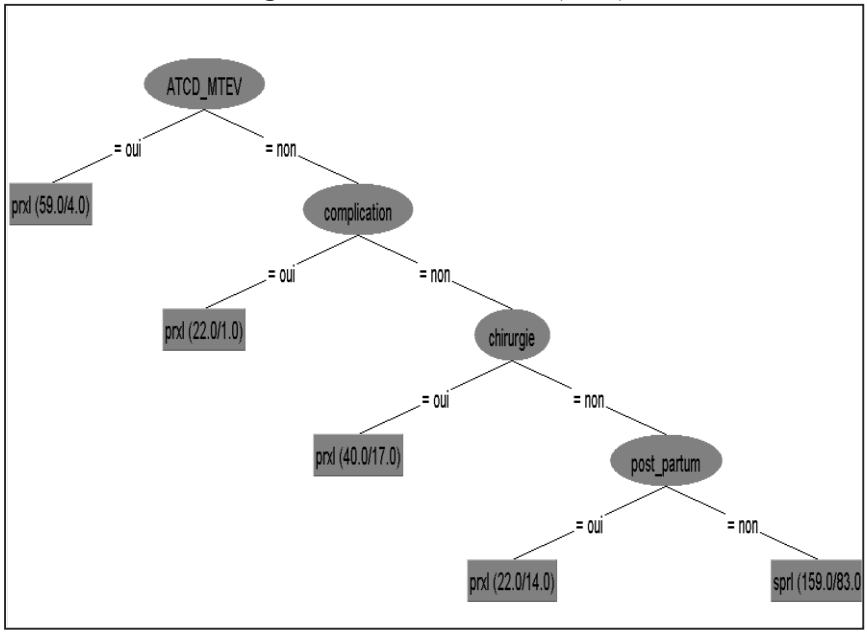
4.1 Les arbres de décision

[***] Le gain ration sert à pondérer le gain qui favorise les attributs ayant beaucoup de valeurs.

Pour l'élaboration de l'arbre de décision on a opté pour l'algorithme J48 de Weka. Cet algorithme est basé sur le test de Gain Ratio comme mesure d'évaluation de partitionnement.

- **Construction de l'arbre de décision :** Les arbres de décision sont des modèles non paramétriques, qui ne postulent d'aucune hypothèse à priori. De ce fait, nous n'excluons aucune des variables lors de l'analyse. Nous avons opéré par validation croisée[†††] pour sélectionner le meilleur modèle construit en phase apprentissage, ce qui a produit l'arbre suivant :

Figure 3 : Arbre de décision (weka)



On remarque que seules 4 variables ont été retenues par le classifieur. La variable de segmentation qui sépare au mieux les individus est *l'ATCD MTEV*, suivie de *complication EP*, *Chirurgie* et *post-partum* avec respectivement comme gains ratios : **0.192, 0.174, 0.028 et 0.019.**

L'arbre de décision donne lieu à **5 règles de décision** :
1) Si ATCD_MTEV = Oui **Alors** ***TVP = proximale (59 / 4)***
2) Si ATCD_MTEV = Non & EP = Oui **Alors** ***TVP = proximale (22 /1)***

[†††] Le recourt à la validation croisée a pour but d'estimer la vraie valeur de l'erreur

3) Si ATCD_MTEV = Non & EP = Non & chirurgie= Oui A**lors** ***TVP = proximale (40 /17)***
4) Si ATCD_MTEV= Non & EP=Non & chirurgi= Non & post-partum= Oui A**lors** ***TVP= proximal (22 /14)***
5) Si ATCD_MTEV=Non & EP=Non & chirurgi=Non & post-partum=Non A**lors** ***TVP=superficiel(159 /83)***

- **Évaluation du modèle :** Afin d'évaluer le modèle, nous calculons le taux de mauvais classement à partir de la matrice de confusion construite à partir de l'échantillon test

Tableau 4: Matrice de confusion.

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 99  0 63 |  a = prxl
 11  0 36 |  b = dstl
 11  0 82 |  c = sprl
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         181               59.9338 %
Incorrectly Classified Instances       121               40.0662 %
```

Le taux d'erreur est estimé à prés de 40%. Notre échantillon est relativement faible (302 instances), c'est ce qui a conduit à ce taux d'erreur. D'après ces résultats, nous déduisons que les patients avec des antécédents de *MTEV* ou une *complication EP*, ainsi que le facteur *chirurgie ou post-partum* ont un fort risque de développer une ***TVP proximale***.

4.2 Régression logistique multinomiale

Des modalités :Prxl, Dstl et Sprl de la variable dépendante *Topographie*, Tanagra[‡‡‡] choisit comme modalité de référence la dernière. Nous expérimentons le modèle d'abord sur l'ensemble des variables, puis nous réitérons l'analyse avec les variables pertinentes produites par les arbres de décision.

‡‡‡ http://eric.univ-lyon2.fr/_ricco/tanagra/en/tanagra/

- **Evaluation et test de significativité du modèle**

L'objectif des tests de significativité est d'éprouver le rôle d'une, de plusieurs ou de l'ensemble des variables explicatives. On commence par évaluer la qualité du modèle, puis on s'intéressera à l'extraction des variables dépendantes les plus significatives fournies par la RLM. Pour évaluer la qualité du modèle, on se base sur la matrice de confusion, qui confronte la classe réelle à la classe prédite par du modèle.

Tableau 5 Matrice de confusion

Cas1 : L'ensemble des variables **Cas2 :** Variables (arbres de décision)

Classifier performances

Error rate			0,4172				
Values prediction			Confusion matrix				
Value	Recall	1-Precision		prxl	dstl	sprl	Sum
prxl	0,8509	0,3911	prxl	137	2	22	161
dstl	0,1042	0,4444	dstl	31	5	12	48
sprl	0,3656	0,5000	sprl	57	2	34	93
			Sum	225	9	68	302

Classifier performances

Error rate			0,4669				
Values prediction			Confusion matrix				
Value	Recall	1-Precision		prxl	dstl	sprl	Sum
prxl	1,0000	0,4669	prxl	161	0	0	161
dstl	0,0000	1,0000	dstl	48	0	0	48
sprl	0,0000	1,0000	sprl	93	0	0	93
			Sum	302	0	0	302

CAS1 : La probabilité de mauvais classement du modèle est estimée à 41,7. Elle est moyennement élevée. 85,09% des patients atteints de « prxl », 36,56% de « sprl » et 10,42% de « dstl » ont été bien identifiés par le modèle. Par ailleurs, parmi les vraies positives 60% sont de modalité « sprl » et 55,5% de modalité « dstl ».

CAS2 : Le taux d'erreur est plus élevé, 53% des patients sont bien classés par le modèle. Il est impossible de se prononcer qu'on à la qualité du modèle, le taux tourne autour de la

moyenne. La capacité du modèle à retrouver les positifs pour la modalité« prxl » est de 100%, dont 53,3% le sont réellement.

- **Évaluation du modèle**

Pour établir la qualité d'ajustement, nous confrontant notre modèle avec le modèle trivial. On s'intéresse au rapport de vraisemblance.

L'interprétation des résultats (pseudo-R^2, du test du rapport de vraisemblance (CHI-2 test) et des critères AIC et SC (ou BIC)), figurant dans les trois parties du tableau ci-dessous mettent en balance la qualité de l'ajustement (-2LL) et la complexité du modèle.

Tableau III.10 : Evaluation du modèle.

Cas1 : L'ensemble des variables **Cas 2 :** Variables (arbres de décision)

Adjustement quality

Predicted attribute	TOPOGRAPHIE TVP	
Ref. value	sprl	
Number of examples	302	
Model Fit Statistics		
Criterion	Intercept	Model
AIC	602,187	593,069
SC	609,608	674,698
-2LL	598,187	549,069
Model Chi² test (LR)		
Chi-2		49,1180
d.f.		20
P(>Chi-2)		0,0003
R²-like		
McFadden's R²		0,0821
Cox and Snell's R²		0,1501
Nagelkerke's R²		0,1741

Adjustement quality

Predicted attribute	TOPOGRAPHIE TVP	
Ref. value	sprl	
Number of examples	302	
Model Fit Statistics		
Criterion	Intercept	Model
AIC	602,187	583,824
SC	609,608	620,928
-2LL	598,187	563,824
Model Chi² test (LR)		
Chi-2		34,3631
d.f.		8
P(>Chi-2)		0,0000
R²-like		
McFadden's R²		0,0574
Cox and Snell's R²		0,1076
Nagelkerke's R²		0,1248

Cas 1 :

- **Les Critères AIC et SC (BIC)** : comparent le modèle trivial «Intercept » avec le modèle testé[§§§]. L'AIC du *modèle testé* (*593.09)* est ***plus faible*** que celui du *modèle trivial (602,18)*, alors que le *SC (BIC)* du *modèle testé (674,698)* est ***plus élevé*** *(609,608)*. De ce fait, nous

[§§§] Le meilleur étant celui possédant AIC et /ou SC les plus bas

nous basons sur l'AIC pour évaluer le modèle[****] testé qui est considéré meilleur que l'Intercept. Par conséquent, les variables explicatives contribuent à l'explication de la Topographie TVP.

- Test LR : Le test du rapport des vraisemblances confronte la déviance du modèle étudiée à celle du modèle trivial. Sa statistique est basée sur le χ^2 [††††]: $\chi^2 = 598{,}187 - 549{,}069 = 48{,}118$**.** Avec un $ddl = 20$, la ρ-value = 0 ,0003 est nettement inferieure au seuil $\alpha = 5\%$, et $\chi^2_{tabulé} = 31{,}41 < \chi^2_{calculé} = 48{,}118$. *Le modèle est globalement très significatif.*

- R^2 **like :** $R^2 = 0.0821$ soit 8.21% des variables explicatives contribuent à l'explication de la variable cible. Ce qui est très faible. Le R^2 ne sera pas pris en compte comme critère d'évaluation.

Cas 2 :

- L'AIC du modèle testé est de 583,824 est inférieur à celui du modèle intercept (601,187). De ce fait, les variables explicatives contribuent à l'explication de la variable cible.

- Le Test LR donne comme valeur du χ^2 = 34,363. $Avec\ un\ ddl\ de$ 8, la $\rho - value = 0{,}0000$ et $\chi^2_{tabulée}(8) = 15{,}507 < \chi^2_{calculée}(8) = 34{,}3631$. Le modèle est globalement significatif.

- $\mathbf{R^2}$ (0,0574) a diminué un peu mais reste toujours faible.

- **Significativité des paramètres**

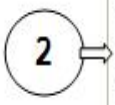
Cette partie, basée sur *le test de Wald* permet de contrôler la significativité des paramètres estimés pour chaque variable explicative.

[****] Le critère BIC est plus parcimonieux que l'AIC puisqu'il pénalise plus le nombre de variables présentent dans le modèle.

[††††] Chi2=-2LL [Intercept] – (-2LL[Model]

Attributes in the equation — Toutes les variables

Class.Value	prxl				dstl			
Pred.Att.	Coef.	Std.Err	Wald	p-value	Coef.	Std.Err	Wald	p-value
constant	-0,439190	-	-	-	-0,156346	-	-	-
age	0,007662	0,007309	1,099	0,2945	-0,016427	0,01023	2,578	0,1084
sexe	0,039305	0,3001	0,01716	0,8958	0,014659	0,4113	0,00127	0,9716
chirurgie	0,335110	0,3756	0,7958	0,3723	0,541751	0,4837	1,255	0,2627
grossesse	1,363655	1,136	1,442	0,2298	2,299139	1,141	4,057	0,0440
post-partum	0,432288	0,5328	0,6583	0,4172	0,653133	0,6362	1,054	0,3046
ATCD MTEV	0,579382	0,1989	8,482	0,0036	-0,115163	0,3226	0,1275	0,7211
complications EP	1,831747	0,6358	8,301	0,0040	-0,525598	1,178	0,1991	0,6554
traumatisme	0,226706	0,2007	1,277	0,2585	0,049722	0,2742	0,03288	0,8561
HTA	0,152627	0,4617	0,1093	0,7410	0,414175	0,6123	0,4575	0,4988
tabac	0,375353	0,7693	0,2381	0,6256	-0,298856	1,198	0,06227	0,8029



Variables détectées par les arbres de décision

Class.Value	prxl				dstl			
Pred.Att.	Coef.	Std.Err	Wald	p-value	Coef.	Std.Err	Wald	p-value
constant	0,118291	-	-	-	-0,801001	-	-	-
chirurgie	0,365094	0,3641	1,006	0,3159	0,499620	0,4653	1,153	0,2830
post-partum	0,312430	0,4967	0,3957	0,5293	0,827588	0,5746	2,074	0,1498
ATCD MTEV	0,543665	0,1945	7,81	0,0052	-0,139319	0,3136	0,1973	0,6569
complications EP	1,781920	0,6277	8,059	0,0045	-0,473845	1,171	0,1638	0,6857

La première partie du tableau, atteste que les variables les plus significatives (Wald calculée $>\chi^2_{tabulée}(1) = 3.84$) sont : *Grossesse, ATCD MTEV, et complication EP.* Les estimations par l'analyse du maximum de vraisemblance confirme que les variables explicatives *ATCD MTEV* ($\rho = 0{,}0036$) et *Complication EP* ($\rho = 0{,}004$) apporte une information significative au modèle par rapport à la modalité « prxl ».

Alors que ″grossesse ″ est significative pour la modalité « dstl ». En outre, les variables dont la ρ-value > 0,05 n'ont aucun effet sur la variable cible. **La partie 2** concernée par les variables détectées par les arbres de décision atteste que les variables les plus significatives sont : *ATCD MTEV, et complication EP* pour la modalité « prxl».

Les signes des *coefficients (modalité : prxl)* sont positifs. Ce qui signifie que l'augmentation de ces facteurs de risque augmente le risque d'avoir une TVP proximale. Après confrontation des résultats les trois facteurs de risque Age, ATCD MTEV et grossesse engendrent un grand risque d'avoir une TVP proximale qu'une TVP distale et que la Complication EP peut être observée plus fréquemment dans le cas d'une TVP proximale. Il convient donc d'adopter les équations de la RLM suivante :

Prxl = 0,579 * **ATCD MTEV** + 1,832 * **Complication EP**
Dstl= 2,299 * **Grossesse**

La première équation montre que pour l'ATCD MTEV, les patients ont 0,579 fois plus de chance de développer une TVP Proximale et 1,832 plus de chances de développer une complication EP. Quant à la grossesse, la chance de contracter une TVP Distale est de 2,299 fois par rapport à la référence.

5. DISCUSSION

Dans notre cohorte, la répartition de la TVP selon la typologie, est de 53% pour la TVP proximale, suivie de la TVP superficielle avec 31% et enfin la TVP distale (16%). On sait que le pronostic de la TVP proximale est sombre, notamment lorsqu'elle est associée à une embolie pulmonaire ou à un cancer.

La présente étude, suite aux différents tests réalisés, nous a révélé des résultats qui s'adaptent avec la réalité du terrain. Des quinze variables étudiées, *post-partum, chirurgie, antécédent de MTEV et complication d'embolie pulmonaire*, sont les variables qui contribuent de manière forte à expliquer la variable cible TOPOGRAPHIE TVP. Elles sont apparues dans chaque test et selon chaque méthode.

On peut dire donc que c'est des facteurs de risque permanant dans une TVP. L'effet de l'âge et la grossesse présente un risque moindre d'une TVP. Mais restent plus importants que les autres facteurs. De plus, nous avons noté une prédominance chez les femmes avec un ratio H/F = 0,86. cette prédominance pourrait s'expliquer par la présence de facteurs spécifiques aux femmes telles que les grossesses, le post-partum, etc .

La tranche d'âge des [35 – 55[est la plus représentée chez les femmes alors que celle des homme elle se situe dans l'intervalle [55, 75[. Selon la littérature la TVP est prédominante chez les femmes et les sujets âgés en général. Quant au facteur favorisant le plus souvent rencontré est sans conteste la chirurgie.

6. CONCLUSION

L'objectif de ce travail était d'étudier les facteurs de risque d'une TVP selon la Topologie de la TVP. La population est composée de 302 sujets hospitalisés, atteints de TVP colligés sur 9ans. Avant de démarrer l'analyse statistique, nous avons précédé au prétraitement des données afin de traiter les valeurs manquantes et supprimer le bruit pour obtenir des résultats précis. Des méthodes statistiques sont appliquées pour l'analyse des données hospitalières, elles permettent pour la plupart de découvrir les facteurs de risque, mais en aucun cas les relations entre les attributs (D. Tomar & S. Agarwal, 2013). La régression logistique multinomiale combinée aux arbres de décision semble être des méthodes appréciables pour dévoiler la relation entre les attributs. De plus, cette approche que nous proposons peut être appliquée à des échantillons de grande taille. En revanche, si on souhaite mener à bien un projet d'exploration de données dans le modèle médical, la

disponibilité de données de qualité est un préalable. Dans ce secteur, les données sont souvent incomplètes (nous avons été confronté à ce problème) soit par négligence ou pour des problèmes de confidentialité.et il faut dans ce cas chercher un moyen de traiter ces insuffisances.

BIBLIOGRAPHIE

1. SS. Anand, PS. Wells & al." Does this patient have deep vein thrombosis?" JAMA98 ; 279 .
2. CI. Lagerstedt CI, & al." Need for long-term anticoagulant treatment in symptomatic calf-vein thrombosis" Lancet 1985 ; 2 : 515-8.
3. V. Kakkar & al. "Natural history of postoperative deep-vein thrombosis", Lancet 1969 ; 2 : 230-2.
4. R. Moreno-Cabral R, & al. " Importance of calf vein thrombophlebitis" Surgery 1976 ; 80 : 735-42.
5. JM. Lohr & al. " Karmody Award. Calf vein thrombi are not a benign finding"Am J Surg 1995; 170 : 86-90.
    A. Cogo & al. "Compression ultrasonography for diagnostic management of patients with clinically suspected deep vein thrombosis: prospective cohort study". BMJ 1998 ; 316 : 17-20.
6. RA. Kraaijenhagen & al. ″Simplification of the diagnostic management of suspected deep vein thrombosis ″. Arch Intern Med 2002 ; 162 : 907-11.
7. M. Méan & D. Aujesky ″Maladie thromboembolique veineuse chez la personne âgée ″ ; Service de médecine interne CHUV, 1011 Lausanne ; Rev Med Suisse 2009; 5: 2142-6
8. S. Tuffery. ″ Data mining et statistique décisionnelle ″ ; 4ième ed, Broché – 21 août 2012.
9. D. Pyle. ″ Data preparation for data mining″ Morgan Kaufmann Publishers, Inc. San Francisco, USA. 1999
10. L. Breiman & al ″ Classification and Regression Tree″, California:Wadsworth International, 1984
11. J.R. Quinlan, ″ C4.5: Programfor machine learning ″, Morgan Kaufmann publisher 1993

12. J. M. Hilbe. ″ Logistic regression models ″. Chapman & Hall /CRC Texts in Statistical Science Series. CRC Press, Boca Raton, FL , (2009). ISBN978-1-4200-7575-5.
13. D. Tomar & S. Agarwal, ″A survey on Data Mining approaches for Healthcare″, International Journal of Bio-Science and Bio-Technology, Vol.5, No.5, (2013), pp. 241-266.