Ingeniería Investigación y Tecnología
volumen XVIII (número 4), octubre-diciembre 2017 457- 464
ISSN 1405-7743 FI-UNAM artículo arbitrado
Información del artículo: recibido: enero de 2016, aceptado: junio de 2017

Clasificación binaria para la predicción de trombosis

Binary classification to predict familiar thrombosis

Pérez-Salvador Blanca Rosa

Universidad Autónoma Metropolitana. Unidad Iztapalapa
Matematicas
Correo: psbr@xanum.uam.mx

Sánchez-Lordméndez Carlos Gabriel

Universidad Nacional Autónoma de México
Colegio de Ciencias y Humanidades
Correo: sanlordcag@gmail.com

Baptista-González Héctor Alfredo

Instituto Nacional de Perinatologia
Hematologia perinatal
Correo: hbaptista@medicasur.org.mx

Resumen

La trombosis, como otros problemas de salud, puede ser de difícil diagnóstico. El diagnostico corresponde a un problema de clasificación. En este trabajo se resuelve un problema de clasificación usando *software* libre sobre una base de datos de pacientes del Instituto Nacional de Perinatología Isidro Espinoza Reyes con *Síndrome de Anticuerpos Antifosfolípido* (SAAF) y pacientes con trombosis sin SAAF. El trabajo analizó el riesgo de que una persona padezca la enfermedad mencionada. El estudio comprende rutinas de selección de variables, de imputación y dos de clasificación, una basada en regresión logística y la otra en el cociente de verosimilitud. Los resultados muestran que la clasificación basada en el cociente de verosimilitud fue mejor que la clasificación hecha con la regresión logística.

Descriptores: clasificación binaria, regresión logística, lema de Neyman-Pearson, devianza.

Abstract

Thrombosis and other health problems can be difficult to diagnose. The diagnosis is a classification problem. In this work a classification problem is solved using free software on a database of patients from the National Institute of Perinatology Isidro Reyes Espinoza (Instituto Nacional de Perinatología Isidro Espinoza Reyes) with thrombosis problems. In this work the risk that a person will develop the disease mentioned was analyzed. The study includes, variables selection, imputation and analysis of two classification methods: one based on logistic regression and the other in the likelihood ratio. The results indicate that the likelihood ratio classification was better than logistic regression classification.

***Keywords*:** *binary classification, logistic regression, Neyman-Pearson lemma, deviance.*

Introducción

La trombosis es un problema mundial de salud pública. Tan sólo en México hay entre 400,000 y 500,000 casos de trombosis por año. La *Enfermedad Tromboembólica Venosa* (ETEV) constituye una de las mayores causas de morbilidad-mortalidad en el país (Secretaría de Salud, 2010). Debido a la importancia que reviste la trombosis, se han realizado diversos estudios respecto al riesgo de su padecimiento en general (Simioni *et al*., 2002), (Laporte *et al*., 2008), (Severinsen *et al*., 2009), (De Haan *et al*., 2012), su padecimiento durante el embarazo (Liu *et al*., 2009) y postparto (Heit *et al*., 2005), así como su reincidencia (Eichinger *et al*., 2010), (Vázquez *et al*., 2013), por mencionar algunos estudios. Dentro de las metodologías más empleadas para diagnosticar esta enfermedad se encuentra el análisis de supervivencia con estimación de Kaplan-Meier y regresión de Cox (Prandoni *et al*., 2007), (Eichinger *et al*., 2010), (Simioni *et al*., 2002), (Laporte *et al*., 2008), (Severinsen *et al*., 2009), (Heit *et al*., 2005), el modelo de regresión logística (Liu *et al*., 2009), (De Haan *et al*., 2012), (Laporte *et al*., 2008) y los modelos lineales generalizados (Heit *et al*., 2005).

El diagnóstico de la trombosis, como el de otras enfermedades, se puede tratar como un problema de clasificación binaria y, en este contexto, existen diferentes métodos estadísticos para resolver el problema, como son el método de discriminantes de Fisher (1936), el método del k-ésimo vecino más cercano de Fix *et al*. (1951), el método de la máquina de vector soporte propuesto por Cortés *et al*. (1995) y por Cristianini *et al.* (2000) y la regresión logística, Hosmer *et al*. (2000) y Escabias *et al*. (2007), así como el cociente de verosimilitud para el caso normal Zadora (2008).

Uno de los principales problemas en el diagnóstico clínico es clasificar correctamente a los pacientes susceptibles de padecer una enfermedad. Los profesionales de la salud requieren herramientas de diagnóstico confiable, herramientas que reduzcan el riesgo de tener falsos negativos o falsos positivos. En este sentido, el presente trabajo presenta un análisis comparativo de dos métodos de clasificación para diagnosticar a pacientes en riesgo de sufrir una trombosis. Las funciones clasificadoras se estimaron utilizando una base de datos de pacientes del Instituto Nacional de Perinatología Isidro Espinoza Reyes con *Síndrome de Anticuerpos Antifosfolípido* (SAAF) y pacientes con trombosis sin SAAF, característica que identifica a las personas que han sufrido trombosis.

La base de datos con la que se trabajó contiene 25 variables explicativas, que tenía datos faltantes, por lo que fue necesario realizar rutinas de imputación y selección de variables. Se hizo la comparación de dos métodos de clasificación, una basada en regresión logística y la otra en el cociente de verosimilitud para diagnosticar a los pacientes en riesgo de trombosis. Los resultados muestran que la clasificación basada en el cociente de verosimilitud clasificó mejor a los pacientes en la muestra que el método de regresión logística.

Desarrollo y métodos

La trombosis es una enfermedad de difícil diagnóstico. En virtud de ello, nos hemos dado a la tarea de desarrollar un sistema de clasificación que coadyuve a los diagnósticos realizados por los médicos. Para ello, supóngase que la población objetivo está particionada en dos conjuntos no vacíos y complementarios E y $\overline{E}$; E es el conjunto de individuos que tienen alto riesgo de padecer trombosis y $\overline{E}$ los individuos que tienen bajo riesgo de padecerla. Se puede pensar que una serie de variables susceptibles a ser medidas en los individuos son: el nivel de colesterol, presión sanguínea, etcétera, nos pueden dar información de su estatus de riesgo. En este sentido, es razonable pensar que existe una función de las variables observadas (vector x) que nos indica la probabilidad de que el individuo pertenezca al conjunto de alto riesgo E o al de bajo riesgo $\overline{E}$; a esta función se le denominará función clasificadora y nuestro interés es tener una estimación de ella. Con esta función estimada se obtienen dos conjuntos A_E y $A_{\overline{E}} = \overline{A_E}$. Si el vector **x** está en A_E al paciente se le clasifica en E, esto es, en alto riesgo de sufrir la trombosis, y si se encuentra en $A_{\overline{E}}$, al paciente se le clasifica en bajo riesgo de sufrir trombosis.

Al realizar la clasificación se observan cuatro posibles escenarios:

1) Que a un sujeto se le clasifique como elemento de E y realmente resida en E, lo cual es correcto. Esto ocurre con probabilidad $P\left(A_E \mid E\right)$.
2) Que a un sujeto se le clasifique como elemento de $\overline{E}$ y realmente se ubique en $\overline{E}$, lo cual también es correcto. Esto ocurre con probabilidad $P\left(A_{\overline{E}} \mid \overline{E}\right)$.
3) Que a un sujeto se le clasifique como elemento de E y realmente esté en $\overline{E}$, lo cual no es correcto; esto lo escribiremos como $P\left(A_{\overline{E}} \mid E\right)$ y lo llamaremos error tipo I.
4) Que a un sujeto se le clasifique como elemento de $\overline{E}$ y realmente se encuentre en E, lo cual tampoco es correcto. A este error lo llamaremos error tipo II y lo escribiremos como $P\left(A_{\overline{E}} \mid E\right)$.

La efectividad de un método se mide por la frecuencia de las buenas decisiones.

Clasificación mediante regresión logística

Sea Y_i la variable aleatoria definida como

$$Y_i = \begin{Bmatrix} 0 \text{ si el individuo } i \text{ pertenecen a } \overline{E} \\ 1 \text{ si el individuo } i \text{ pertenecen a } E \end{Bmatrix}$$

Entonces Y_i es una variable aleatoria binomial, por lo que $0 \le E(Y_i) = \pi_i \le 1$. El modelo que satisface esta restricción es el logístico, que tiene como función liga la función logit (Myers, Montgomery, & Vining, 2010), dada por

$$g(\mu_j) = ln\frac{\pi_j}{1-\pi_j}$$

De esta manera, la probabilidad de que el individuo *i* esté en riesgo de padecer una trombosis se estima con la ecuación $\hat{p}(x_i) = \frac{1}{1+e^{-\mathbf{x}_i^T\hat{\beta}}}$, donde x_i es el vector de variables explicativas del individuo *i*. En este trabajo para realizar la estimación de parámetros se empleó el *software* libre R.

Para clasificar a los pacientes se define el conjunto $A_E = \{x \mid \hat{p}(x) > \lambda\}$, con el criterio de clasificación siguiente: si los datos del paciente se encuentran en A_E al individuo se le diagnostica en riesgo, y si los datos del individuo se encuentran en $\overline{A_E} = A_{\overline{E}}$ al individuo se le clasifica en no riesgo.

Para un valor específico de λ se tiene que los datos de la base satisfacen las relaciones $k_1 = \#A_{\overline{E}} \cap$ y E y $k_0 = \#A_{\overline{E}} \cap \overline{E}$.

El valor de λ puede moverse; cuando λ disminuye, se reduce el valor de k_1 y aumenta el valor de $n_0 - k_0$, si λ aumenta ocurre lo contrario. Las probabilidades de cometer una mala clasificación se estiman como

$$\hat{P}(A_{\overline{E}} \mid E) = \frac{k_1}{n_1}$$

$$\hat{P}(A_E \mid \overline{E}) = \frac{n_0 - k_0}{n_0}$$

Para los efectos de este trabajo, se fijó el valor de λ en el punto donde se minimizó la suma $\frac{n_0 - k_0}{n_0} + \frac{k_1}{n_1}$, que corresponde a la suma de las estimaciones de los dos errores posibles.

Clasificación mediante el cociente de verosimilitud

Podría considerarse más grave clasificar mal a una persona que está en riesgo de padecer la enfermedad, porque no se le daría tratamiento preventivo y entonces sería deseable tener menor probabilidad de cometer este error, por lo que sería conveniente controlar la probabilidad de cometerlo en un nivel que llamaremos α, es decir $P(A_{\overline{E}} \mid E) = \alpha$. Cumpliéndose esta restricción, se busca que la probabilidad de cometer el otro error, $P(A_E \mid \overline{E})$, sea lo más pequeña posible. El siguiente teorema muestra el conjunto clasificador con esta propiedad.

Teorema. El conjunto

$$A_E = \left\{ x \in \mathbf{X} \left| \frac{P(X = x \mid \overline{E})}{P(X = x \mid E)} \le \lambda \right. \right\}$$

tal que $P(A_{\overline{E}} \mid E) = \alpha$, satisface la relación $P(A_E \mid \overline{E}) < P(C \mid \overline{E})$ para todo *C* tal que $P(\overline{C} \mid E) = \alpha$.

Esto significa que si se elige como criterio de clasificación la relación $\frac{P(X = x \mid \overline{E})}{P(X = x \mid E)} \le \lambda$ con un nivel α de cometer el error tipo I, necesariamente la probabilidad de cometer el error tipo II es mínimo. Este clasificador tiene una relación cercana al lema de Neyman-Pearson y su demostración es análoga; una prueba de este teorema se puede encontrar en Pérez (2013).

Estimación del clasificador por cociente de verosimilitud usando una base de datos

Se parte de una base de datos donde las variables explicativas son categóricas y se identifican los individuos que han padecido la enfermedad, así como los que no la han padecido. Consideramos que si un individuo ya presentó la enfermedad es un individuo de alto riesgo. El conjunto *E* se forma por todos los individuos de la base, quienes ya padecieron la enfermedad mientras que el conjunto $\overline{E}$ se forma por los individuos de la base que no la han padecido. Es importante señalar que alguno de los individuos que pertenecen a $\overline{E}$ en realidad podrían estar en riesgo de sufrir trombosis, solo que no han manifestado la enfermedad y se desconoce quiénes son; mientras que con los individuos que pertenecen a *E* se tiene certeza de que son de alto riesgo.

En una base de datos se desconoce la función de distribución conjunta de los vectores $\boldsymbol{x}_i$, por lo que para obtener este clasificador se estiman las probabilidades de los vectores $\boldsymbol{x}_i$ con la frecuencia relativa usando los datos de la base y se obtienen los cocientes.

$$c_i = \frac{\hat{\mathrm{p}}(\mathbf{x}_i \mid Y=0)}{\hat{\mathrm{p}}(\mathbf{x}_i \mid Y=1)} = \frac{f_{\mathbf{x}_i,0}/n_0}{f_{\mathbf{x}_i,1}/n_1}$$

Estos cocientes se ordenan de menor a mayor y se escoge un número λ para establecer la región crítica fijando un porcentaje de malas clasificaciones. En la tabla 1, en la primera columna se encuentran los valores del cociente c_i en orden creciente, en la segunda columna se encuentra el vector x_i correspondiente al cociente, en la tercera fila se encuentra, $f_{\mathbf{x}_{c(i)}1}$, la frecuencia del vector x_i en la base de datos para los pacientes con $Y = 1$; y finalmente en la cuarta columna se encuentra, $f_{\mathbf{x}_{c(i)}0}$, la frecuencia del mismo vector, pero para los pacientes en la base de datos con $Y = 0$. La línea horizontal a la mitad de la tabla representa el valor de λ, los valores de c_i por arriba de esa línea satisfacen la relación $c_i < \lambda$ y todos los pacientes que tengan los correspondientes vectores x_i se clasificarán como de bajo riesgo.

De esta manera, la probabilidad estimada para clasificar mal a un individuo que se encuentra en riesgo de sufrir trombosis se estima como

$$\hat{\mathrm{p}}\left(A_{\bar{E}} \mid E\right) = \frac{w_1}{n_1}$$

donde $w_1 = \sum_{i=p+1}^{k} f_{\mathbf{x}_{c(i)}1}$. Mientras que la probabilidad de clasificar mal a un individuo de no riesgo, se estima como

$$\hat{\mathrm{p}}\left(A_E \mid \bar{E}\right) = \frac{w_0}{n_0}$$

donde $w_0 = \sum_{i=1}^{p} f_{\mathbf{x}_{c(i)}0}$.

En este trabajo se propone seleccionar el valor de λ, con el cual la suma de las probabilidades estimadas de cometer los dos errores $\frac{w_0}{n_0} + \frac{w_1}{n_1}$ sea mínima; con esto se logra minimizar los dos errores simultáneamente y se puede comparar con el clasificador de regresión logística.

Tabla 1. Los cocientes de probabilidades condicionales de los vectores **x** en orden creciente junto con las frecuencias correspondientes tanto para los individuos que han sufrido de trombosis, como para los que no lo han sufrido

Cociente	Vector Asociado	Frecuencia para Y=1	Frecuencia para Y=0
$c_{(1)}$	$\mathbf{x}_{c_{(1)}}$	$f_{\mathbf{x}_{c(1)}}$	$f_{\mathbf{x}_{c_{(1)}}}$
$c_{(2)}$	$\mathbf{x}_{c_{(2)}}$	$f_{\mathbf{x}_{c(2)}}$	$f_{\mathbf{x}_{c_{(2)}}}$
$c_{(3)}$	$\mathbf{x}_{c_{(3)}}$	$f_{\mathbf{x}_{c(3)}}$	$f_{\mathbf{x}_{c_{(3)}}}$ $\Big\} w_0 = \sum_{i=1}^{p} f_{\mathbf{x}_{c(i)}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$c_{(p)}$	$\mathbf{x}_{c_{(p)}}$	$f_{\mathbf{x}_{c(p)}}$	$f_{\mathbf{x}_{c_{(p)}}}$
$c_{(p+1)}$	$\mathbf{x}_{c_{(p+1)}}$	$f_{\mathbf{x}_{c_{(p+1)}}}$	$f_{\mathbf{x}_{c_{(p+1)}}}$
$c_{(p+2)}$	$\mathbf{x}_{c_{(p+2)}}$	$f_{\mathbf{x}_{c_{(p+2)}}}$ $\Big\} w_1 = \sum_{i=p+1}^{k} f_{\mathbf{x}_{(c_i)}}$	$f_{\mathbf{x}_{c_{(p+2)}}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$c_{(k)}$	$\mathbf{x}_{c_{(k)}}$	$f_{\mathbf{x}_{c_{(k)}}}$	$f_{\mathbf{x}_{c_{(k)}}}$
		n_1	n_0

Finalmente, el conjunto $\{\mathbf{x}_{c_{(i)}} \mid c_i < \lambda\}$ define el conjunto de pacientes que se diagnosticarán como de riesgo. El proceso de clasificación identifica un punto crítico λ, que determina un conjunto de vectores asociados a un alto riesgo de padecer la enfermedad, a este conjunto lo denotamos como *R*, y a su complemento denotado como *NR*.

$$R = \left\{ \mathbf{x} \in base \quad de \quad datos \left| \frac{P(\mathbf{x}|0)}{P(\mathbf{x}|1)} \leq \lambda \right. \right\}$$

$$NR = \left\{ \mathbf{x} \in base \quad de \quad datos \left| \frac{P(\mathbf{x}|0)}{P(\mathbf{x}|1)} > \lambda \right. \right\}$$

Dado que *R* y *NR* no necesariamente cubren todos los posibles vectores $\boldsymbol{x}$, puede ocurrir que se presente un paciente con un vector $\boldsymbol{x}$ asociado, el cual no esté ni en *R* ni en *NR*. En estos casos, la solución que se propone es eliminar la variable menos significativa del vector $\boldsymbol{x}$ respecto al criterio de selección a través de la devianza, concepto que se explicará a continuación, para posteriormente repetir el procedimiento de clasificación. Al quitar una variable se busca reducir el número de posibles combinaciones de las variables explicativas y con ello reducir los casos en los cuales el vector asociado a un individuo no aparezca en la base de datos. Sin embargo, si la base es suficientemente grande, se tendrán muchas posibilidades de que contenga a todos los vectores factibles de ocurrir.

Selección de variables para los métodos de clasificación

La base de datos con la que se trabajó tenía 25 variables explicativas y la primera tarea fue depurarla para que solo quedaran las variables significativas para la variable de interés o variable respuesta; además que con menos variables se logra una reducción en el trabajo computacional y, por ende, en el tiempo de ejecución, asimismo una mejor interpretación de los resultados.

Se siguieron dos procesos para excluir variables. El primero y más sencillo fue excluir las variables que tenían más de 50% de datos faltantes. El segundo proceso fue excluir las variables usando la devianza en los modelos de regresión logística anidados con una variable adicional (Dobson, 2002). En este caso, el planteamiento de la prueba fue

$$H_0 : \beta_{r+1} = 0 \quad \text{vs} \quad H_1 : \beta_{r+1} \neq 0$$

La hipótesis nula significa que la variable X_{r+1} no es significativa en el modelo y la hipótesis alternativa significa que sí lo es. La estadística de prueba es la diferencia de las devianzas $\Delta D = D_r - D_{r+1}$, que se espera sea pequeña cuando H_0 es cierta y se espera que sea grande si H_0 es falsa. Entonces la regla de decisión de esta prueba es:

- Si $\Delta D \geq \chi^2_{\alpha,1}$ se rechaza la hipótesis nula y se incluye la variable X_{r+1}
- Si $\Delta D < \chi^2_{\alpha,1}$ no se rechaza la hipótesis nula y se excluye la variable X_{r+1}

Proceso de imputación

Debido a que la base de datos con la que se trabajó tenía datos faltantes se procedió a realizar un proceso de imputación y por cuestiones de simplicidad se eligió imputación simple a través de regresión lineal.

Resultados

Para aplicar los métodos de clasificación, la base de datos se partió en dos conjuntos, uno con 90% de las observaciones, denominado conjunto de entrenamiento para estimar la función clasificadora y otro denominado conjunto de prueba formado por 10% restante para evaluar los métodos de clasificación.

Con el conjunto de entrenamiento se obtienen los dos clasificadores, el clasificador por regresión logística y el clasificador por cociente de verosimilitud. Luego, estos clasificadores se aplican a los elementos del conjunto de prueba y se obtiene el valor predicho Y_{predi}, así se calcula el porcentaje de los elementos mal clasificados usando la fórmula

$$E_{M_j} = \frac{\sum_{i=1}^{n} \left| Y_{pred_i} - Y_{real_i} \right|}{n} \times 100$$

donde Y_{real_i} es el estatus real del individuo *i* y Y_{predi} es el estatus asignado al mismo individuo por el clasificador.

Para estabilizar la estimación del porcentaje de error, se repite mil veces el proceso de clasificación y evaluación procurando que en todos los casos el porcentaje de pacientes en riesgo y en no riesgo en los conjuntos de entrenamiento y de prueba permanezcan de manera semejante a la base de datos. Los resultados obtenidos se resumen en la tabla 2, donde E_{Mi} denota el error del método de clasificación con el método *i*. Si *i* = 1 el error corresponde al clasificador por regresión logística; y si

$i = 2$ y el error corresponde al clasificador por el cociente de verosimilitud.

Tabla 2. Comparación del error cometido por ambos métodos en mil ocasiones

	$E_{M1} > E_{M2}$	$E_{M1} = E_{M2}$	$E_{M1} < E_{M2}$
Frecuencia	807	46	147

En la figura 1 se muestra la distribución empírica de los errores en las mil repeticiones para los dos métodos de clasificación en el conjunto de prueba. La frecuencia corresponde al número de veces en que de las mil corridas la proporción de errores estuvo dentro del correspondiente intervalo de clase. Esto es, la frecuencia en [0%,10%] indica la cantidad de ocasiones en que el número de errores estuvo entre 0 y 10%.

Conclusiones

Del análisis de los datos, se concluye que:

- El clasificador por cociente de verosimilitud dio mejores resultados. Este clasificador puede programarse fácilmente, sin embargo, requiere almacenar en memoria todos los vectores del conjunto *R*, ya que el clasificador no se obtiene por una fórmula. De esta manera, mientras más niveles tenga cada variable, el número de posibles combinaciones (vectores **x**) aumentará, y en consecuencia, el almacenaje requerido será mayor.
- Para la aplicación del clasificador de máxima verosimilitud a los valores de cada variable se les dividió en dos clases, el punto de corte de las dos categorías se proporcionó por el personal de la Coordinación de Hematología Perinatal del Instituto Nacional de Perinatología Isidro Espinoza Reyes; se emplearon solo dos categorías en cada variable para que el conjunto crítico ocupara menos espacio en memoria. Aun así, el clasificador por cociente de verosimilitud resultó, en promedio, ser mejor que el clasificador por regresión logística, lo que hace pensar que con una clasificación de los valores de las variables más fina, se obtendrían mejores resultados.

E_{M1}	Frecuencia	E_{M2}	Frecuencia
(0%,10%]	2	(0%,10%]	0
(10%,20%]	304	(10%,20%]	840
(20%,30%]	485	(20%,30%]	160
(30%,40%]	163	(30%,40%]	0
(40%,50%]	45	(40%,50%]	0
(50%,60%]	1	(50%,60%]	0
(60%,70%]	0	(60%,70%]	0
(70%,80%]	0	(70%,80%]	0
(80%,90%]	0	(80%,90%]	0
(90%,100%]	0	(90%,100%]	0

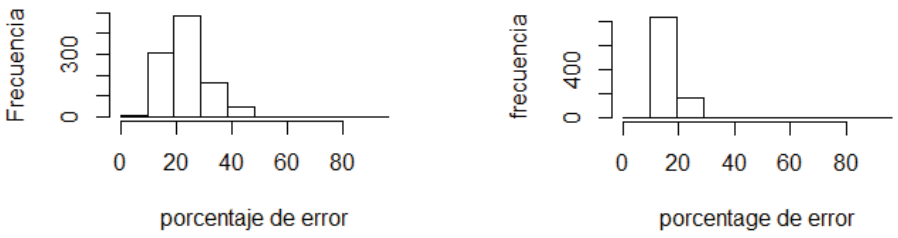

Figura 1. Distribución de los errores de cada método respecto al tamaño total del conjunto de prueba, a la izquierda los resultados para el clasificador mediante regresión logística y a la derecha los resultados para el clasificador del cociente de verosimilitud

Referencias

Cortes C. y Vapnik V. Support-vector networks. *Machine Leaming*, volumen 20, 1995: 273-297.

Cristianini N. y Shawe-Taylor J. Support vector machines and other kernel-based learning methods, Cambridge, Cambridge England, 2000.

De Haan-Hugoline G., Bezemer I.D., M.-Doggen C.J., Le-Cessie S., Reitsma-Pieter H., Arellano- Andre R., Tong C.H., Devlin J.J., Bare-Lance A., Rosendaal-Frits R., Vossen C.Y. Multiple SNP testing improves risk prediction of first venous thrombosis. *Blood*, volumen 120 (número 3), julio de 2012: 656-663 [en línea]. Disponible en: http://www.bloodjournal.org/content/120/3/656.long?sso-checked=true

Dobson A. *An introduction to generalized linear models,* 2a ed., Canada, CHAPMAN & HALL/CRC, 2002, pp. 82-87.

Eichinger S., Heinze G., M.-Jandeck L., Paul-A.K. Risk assessment of recurrence in patients with unprovoked deep vein thrombosis or pulmonary embolism. Circulation, volumen 121, 2010: 1630-1636 [en línea]. Disponible en: http://circ.ahajournals.org/content/121/14/1630.long

Escabias M., Aguilera A.M., Valderrama M.J., Functional PLS logit regression model. *Comput. Statist. Data Anal,* volumen 51, 2007: 4891-4902.

Fisher RA. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, volumen 7, 1936: 179-188.

Fix E. y Hodges J.L. An important contribution to nonparametric discriminant analysis and density estimation C. commentary on fix and hodges (1951). *International Statistical Review/Revue Internationale de Statistique*, volumen 57 (número 3), 1989: 233-238.

Heit J.A., Kobbervig C.E., James A.H., Petterson T.M., Bailey K.R., Melton L.J. Trends in the incidence of venous thromboembolism during pregnancy or postpartum: a 30-year population-based study. *Ann Intern Med*., volumen 143 (número 10), 15 de noviembre de 2005: 697-706 [en línea]. Disponible en: http://copacamu.org/IMG/pdf/Heit-ann_int_med.pdf

Hosmer D.W. y Stanley L. *Applied logistic regression*, 2a ed., New York, Chichester, Wiley, 2000, ISBN 0-471-35632-8.

Laporte S., Mismetti P., Décousus H., Uresandi F., Otero R., Lobo J.L., Monreal M., and the RIETE Investigators. Clinical predictors for fatal pulmonary embolism in 15 520 patients with venous thromboembolism. *Circulation,* volumen 117, 2008: 1711-1716 [en línea]. Disponible en: http://circ.ahajournals.org/content/117/13/1711.long

Liu S., Rouleau J., Joseph K.S., Sauve R., Liston R.M., Young D., Kramer M.S. Epidemiology of pregnancy-associated venous thromboembolism: a population-based study in Canada. *Journal of Obstetrics and Gynaecology Canada*, volumen 31 (número 7), 2009: 611-620 [en línea]. Disponible en: http://www.jogc.com/article/S1701-2163(16)34240-2/pdf

Myers R. et. al. *Generalized linear model with applications in engineering and the sciences,* 2a ed., Nueva Jersey, Wiley Interscience, 2010, pp. 120-125.

Pérez B.R. El riesgo crediticio. *Revista Contactos*, volumen 3 (número 90), octubre-diciembre, 2013: 23-30.

Prandoni P., Noventa F., Ghirarduzzi A., Pengo V., Bernardi E., Pesavento R., Iotti M., Tormene D., Simioni P., Pagnan A. The risk of recurrent venous thromboembolism after discontinuing anticoagulation in patients with acute proximal deep vein thrombosis or pulmonary embolism. A prospective cohort study in 1,626 patients. *Haematologica*, volumen 92 (número 2), febrero de 2007: 199-205 [en línea]. Disponible en: http://www.haematologica.org/content/92/2/199.full.pdf+html

Secretaría de Salud. *Guía de práctica clínica diagnóstico y tratamiento de la enfermedad tromboembólica venosa,* México, 2010 pp. 7.

Severinsen M.T., Kristensen-Søren R., Johnsen-Søren P., Dethlefsen C., Tjønneland A., Overvad K. Anthropometry, body fat, and venous thromboembolism. A danish follow-up study. *Circulation*, volumen 120, 2009: 1850-1857 [en línea]. Disponible en: http://circ.ahajournals.org/content/120/19/1850.long

Simioni P., Tormene D., Prandoni P., Zerbinati P., Gavasso S., Cefalo P., Girolami A. Incidence of venous thromboembolism in asymptomatic family members who are carriers of factor V Leiden: a prospective cohort study. *Blood Mar*, volumen 99 (número 6), 2002: 1938-1942[en línea]. Disponible en: http://www.bloodjournal.org/content/99/6/1938?sso-checked=true

Vázquez F.J., Posadas-Martínez M.L., Vicens J., González-Bernaldo de Quirós F., Giunta D.H. Incidence rate of symptomatic venous thromboembolic disease in patients from a medical care program in Buenos Aires, Argentina: a prospective cohort. *Thrombosis Journal*, 2013: 11-16 [en línea]. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3750410/pdf/1477-9560-11-16.pdf

Zadora G.L. y Neocleous T. Likelihood ratio model for classification of forensic evidence, Anal Chim Acta, 2009, pp. 642-266.

Citación sugerida:

Citación estilo Chicago

Pérez-Salvador, Blanca Rosa, Carlos Gabriel Sánchez-Lordméndez, Héctor Alfredo Baptista-González. Clasificación binaria para la predicción de trombosis. *Ingeniería Investigación y Tecnología*, XVIII, 04 (2017): 457-464.

Citación estilo ISO 690

Pérez-Salvador B.R., Sánchez-Lordméndez C.G., Baptista-González H.A. Clasificación binaria para la predicción de Trombosis. *Ingeniería Investigación y Tecnología*, volumen XVIII (número 4), octubre-diciembre 2017: 457-464.

Semblanzas de los autores

Blanca Rosa Pérez-Salvador. Es actuario egresada de la Facultad de Ciencias de la Universidad Nacional Autónoma de México (UNAM), maestra en ciencias matemáticas por la Universidad Autónoma Metropolitana (UAM), maestra en estadística e investigación de operaciones por el IIMAS, UNAM, doctorado en ciencias matemáticas por la Facultad de Ciencias, UNAM. Ha escrito libros y artículos en estadística y actualmente es profesor investigador en el Departamento de Matemáticas de la Universidad Autónoma Metropolitana Unidad Iztapalapa (UAM-I).

Carlos Gabriel Sánchez-Lordméndez. Es ingeniero eléctrico-electrónico egresado de la Facultad de Ingeniería de la Universidad Nacional Autónoma de México (UNAM). Realizó los estudios de maestría en matemáticas aplicadas e industriales en la unidad Iztapalapa de la Universidad Autónoma Metropolitana. Actualmente se desempeña como profesor en el plantel sur del Colegio de Ciencias y Humanidades y en la Facultad de Contaduría y Administración en la UNAM.

Héctor Alfredo Baptista-González. Es médico especialista en pedríatria médica con subespecialidad en hematología, cuenta con estudios de maestría en investigación clínica por la Universidad Autónoma del Estado de México (UAEM) y el doctorado en ciencias químicobiologicas por la Escuela Nacional de Ciencias Biologicas del INP. Miembro de la Academia Nacional de Medicina. Durante 30 años se ha dedicado al estudio de la trombosis y de las enfermedades hematológicas de la mujer embarazada, el feto y el recién nacido. Actualmente es director de investigación del Instituto Nacional de Perinatología.