VALORACIÓN DE RIESGO CARDIOVASCULAR MEDIANTE MODELOS DE CLASIFICACIÓN

Jorge Castellanos Vázquez*, Agustín Santiago Moreno[1], Carlos Bouza Herrera** y José Maclovio Sautto Vallejo*
*Universidad Autónoma de Guerrero
**Universidad de la Habana

ABSTRACT
According to the World Health Organization (WHO): every year, approximately 37 million people in the world suffer a cardiovascular event and approximately 46% of those people die from these causes. It is known that the occurrence of a cardiovascular event results from the interaction of different risk factors in which we can include hypertension, high levels of blood lipids, diabetes, obesity and smoking. The diagnosis of cardiovascular risk can be made from the presence of any of these risk factors; moreover, there are pocket calculators that allow estimating cardiovascular risk from a model proposed by Framingham, which is essentially a generalized linear model. In the present work, in addition to using the Framingham model, other classification models are used such as Decision Trees, logistic regression and Random Forest. The objective is to choose the best classification model based on goodness criteria such as the correct classifications rate, relative efficiency and deviance.

KEYWORDS: Classifiers, Statistical classification, Regression Tree, Logistic regression.

MSC: 62-07, 62H30

RESUMEN
Según la organización mundial de la salud (OMS): cada año, aproximadamente 37 millones de personas en el mundo sufren un evento cardiovascular y, aproximadamente 46% de esas personas mueren por dichas causas. Se sabe que la ocurrencia de un evento cardiovascular resulta de la interacción de diferentes factores de riesgo en los que podemos incluir la hipertensión, niveles altos de lípidos en la sangre, diabetes, obesidad y tabaquismo. El diagnóstico de riesgo cardiovascular puede hacerse a partir de la presencia de cualquiera de estos factores de riesgo, es más, existen calculadoras de bolsillo que permiten estimar el riesgo cardiovascular a partir de un modelo propuesto por Framingham, que en esencia es un modelo lineal generalizado. En el presente trabajo, además de utilizar el modelo de Framingham, se utilizan otros modelos de clasificación tales como Decision Trees, regresión logística y Random Forest. El objetivo consiste en elegir el mejor modelo de clasificación a partir de criterios de bondad tales como la tasa de clasificaciones correctas, la eficiencia relativa y la devianza.

PALABRAS CLAVE: Clasificadores, clasificación estadística, árbol de regresión, regresión logística.

1. INTRODUCCIÓN

1.1. El problema de Investigación

Las enfermedades cardiovasculares son la principal causa de muerte en adultos en el mundo (Molinero, 2003): (Ruiz, Segura, & Agustí, 2012): (OMS, 2015): El riesgo cardiovascular se define como la probabilidad de muerte cardiovascular de una persona en un periodo determinado (10 años) (Gobierno Federal, SALUD, SEDENA Y SEMAR): (Alcocer, Lozada, Fanghänel, Sánchez-Reyes & Campos-Franco, 2011): Existen diversos tipos de enfermedades cardiovasculares, entre ellas podemos mencionar la hipertensión arterial, enfermedad arterial coronaria, enfermedad valvular cardiaca, accidente cerebrovascular (trombosis o derrame cerebral) y fiebre reumática o enfermedad cardiaca reumática (DMEDICINA.com, Salud y bienestar, 2015): Cada año, aproximadamente 37 millones de personas en el mundo sufren un evento cardiovascular (ataque al corazón o una enfermedad vascular cerebral) (Alejandro, 2008): (López, 2013) y alrededor de 17.5 millones de personas mueren por dichas causas (OMS, 2015): Con frecuencia las enfermedades cardiovasculares

[1] asantiago@uagro.mx

resultan de la interacción de diferentes factores de Riesgo Cardiovascular, en donde se incluyen hipertensión, niveles elevados de lípidos en la sangre, diabetes, obesidad y tabaquismo (Ruiz, Segura, & Agustí, 2012): Por tanto, los principales factores de riesgo cardiovascular son: Presión arterial alta o hipertensión arterial (HTA): colesterol elevado o lipoproteína de baja densidad (LDL): las lipoproteínas de alta densidad (HDL, del inglés High density lipoprotein): diabetes, obesidad y sobrepeso, tabaquismo, inactividad física, género, herencia y la edad (Alcocer, Lozada, Fanghänel, Sánchez-Reyes, & Campos-Franco, 2011): Según afirmación de la OMS, la mayoría de las enfermedades cardiovasculares se pueden prevenir cuando el tratamiento se enfoca de manera global a todos los factores de riesgo que presenta una persona, por ejemplo, tratar la hipertensión arterial al mismo tiempo que la elevación del colesterol; el tener alta la presión arterial o el colesterol por separado representa un riesgo para la persona, pero si están juntos el riesgo se incrementa significativamente. Por lo que la Organización Mundial del Corazón (WHO, 2013): señala de acuerdo a los estudios realizados, que el tratamiento de los factores de riesgo se sigue enfocando en el manejo por separado de cada uno de ellos, en lugar de ser enfocado a la atención del riesgo global del individuo (WHO, 2003):
La literatura médica divide los factores de riesgo cardiovascular en dos categorías: principales y contribuyentes. Los principales factores de riesgo son aquellos cuyo efecto de aumentar el riesgo cardiovascular ha sido comprobado. Los factores contribuyentes son aquellos que los médicos piensan que pueden dar lugar a un mayor riesgo cardiovascular pero cuyo papel exacto no ha sido definido aún (Farmacias ABC, 2006): Los principales factores de riesgo cardiovascular son: Presión arterial alta o hipertensión arterial (T): colesterol elevado o lipoproteína de baja densidad (C): las lipoproteínas de alta densidad (H): diabetes (D): obesidad (OB) y sobrepeso (Sp): tabaquismo (Tab): inactividad física (IF): sexo (Hombre, Mujer): herencia (Her) y la edad (E) (Alcocer, Lozada, Fanghänel, Sánchez-Reyes & Campos-Franco, 2011): Por lo tanto, entre más factores de riesgo tenga una persona, mayores serán sus probabilidades de padecer una enfermedad del corazón. Algunos factores de riesgo pueden cambiarse, tratarse o modificarse y otros no. Pero el control del mayor número posible de factores de riesgo, mediante cambios en el estilo de vida y/o medicamentos, puede reducir el riesgo cardiovascular (Farmacias ABC, 2006):
Según los datos de la encuesta nacional de salud y nutrición del año 2012 (ENSANUT, 2012): estamos ante un verdadero problema de salud que día a día va creciendo, y que tanto las autoridades de salud como los usuarios de los servicios, debemos tomar conciencia de la realidad y hacer una verdadera medicina preventiva para enfrentar el problema (ENSANUT, 2012): Hasta el momento de la consulta, los resultados no están disponibles a nivel de entidad federativa o municipio, por lo que no tenemos un referente actual para los resultados obtenidos en el estudio realizado con pacientes del municipio de Zihuatanejo, Guerrero. En el presente trabajo se restringe a la descripción del segmento de la población que se encuentra en riesgo vascular, clasificado en esa situación por los factores de riesgo manifiestos en ellos, que ahora se sabe desempeñan un papel importante en las probabilidades de que una persona padezca de una enfermedad del corazón y desde la perspectiva médica, hacer las recomendaciones pertinentes para el control de la enfermedad y disminución del riesgo. La contribución del presente trabajo consiste en proponer modelos estadísticos que faciliten el diagnóstico de riesgo cardiovascular a partir de las probabilidades dadas por estos modelos.

1.2. Objetivos

El objetivo general consistió en seleccionar modelos estadísticos que faciliten el diagnóstico de riesgo cardiovascular a los médicos tratantes. Con el diagnostico precoz es posible prevenir un evento cardiovascular en aquellos pacientes con una probabilidad alta. Los objetivos específicos consistieron en

- Seleccionar los modelos que mejor se ajustan a los datos de RC.
- Identificar los mejores modelos a partir de la tasa de clasificaciones correctas y otros criterios de bondad.
- Hacer predicciones con el modelo seleccionado

2. METODOLOGÍA

2.1. Los datos.

Los datos utilizados en el presente trabajo se tomaron de un estudio realizado en la jurisdicción sanitaria 5 del Estado de Guerrero, dentro del programa Cruzada contra, sobrepeso, Obesidad y Diabetes y organizado por el

departamento de investigación, en coordinación con el departamento de enfermedades crónicas de los Servicios Estatales de Salud en Guerrero. Con estos datos se calcularon las puntuaciones del modelo de Framingham, que constituye un pilar básico, y en diferentes formas es ampliamente utilizado para la toma de decisiones terapéuticas en base a la estimación de riesgo proporcionada por el modelo al introducir las características de riesgo del paciente concreto (Rodríguez, 2014): El criterio clásico para la clasificación de RC es que el paciente tenga una puntuación de 0.2 o mayor, es decir, la probabilidad debe ser mayor o igual al 20%. De acuerdo con esto, los pacientes fueron clasificados en dos grupos excluyentes, los que están en RC y los que no están en RC, es decir, tenemos una variable dicotómica con dos opciones de respuesta, el primer grupo toma el valor 1, es decir, presencia de la condición y el segundo grupo toma el valor 0, o ausencia de la condición, dadas las características observadas en los pacientes, que en nuestro caso fueron las variables predictoras o covariables presión arterial alta o hipertensión arterial (T): colesterol elevado o lipoproteína de baja densidad (C): las lipoproteínas de alta densidad (H): diabetes (D): obesidad (OB) y sobrepeso (Sp): tabaquismo (Tab): inactividad física (IF): sexo (Hombre, Mujer): herencia (Her) y la edad (E): representadas genéricamente por $x_1, x_2, \dots, x_R$.

2.2. Modelos de clasificación y regresión.

2.2.1. Modelo de regresión logística.

El modelo de regresión logística, es quizá el más utilizado en la literatura médica para explicar el comportamiento de una variable dependiente dicotómica, dadas ciertas características, que generalmente son síntomas o factores de riesgo, Agresti (2007) & Hosmer y Lemeshow (2000): En este modelo las covariables o variables explicativas asociadas al riesgo cardiovascular consideradas fueron: hipertensión = T, colesterol elevado o lipoproteína de baja densidad = C, lipoproteína de alta densidad = H, diabetes mellitus tipo II = D, obesidad = OB, inactividad física = IF, sexo = (Hombre, Mujer) y edad = E. Por lo anterior, tenemos una variable respuesta dicotómica y variables explicativas también dicotómicas, lo que posibilita la formulación de un modelo estadístico para estimar una variable respuesta discreta (binaria) en función de una o varias variables explicativas que podrán ser cuantitativas o cualitativas. El modelo de regresión logística múltiple para la variable respuesta binaria Y en términos de valores observados $\mathrm{X_1 = x_1, \cdots, X_R = x_R}$ de las variables explicativas es de la forma $\mathrm{Y(x_1, x_2, \cdots, x_R) = p(x_1, x_2, \cdots, x_R) + e(x_1, x_2, \cdots, x_R)}$
donde $\mathrm{e(x_1, x_2, \cdots, x_R)}$ son errores aleatorios que se consideran centrados e independientes, de modo que$p(x_1, x_2, \cdots, x_R) = \frac{e^{\alpha+\Sigma_{r=1}^{R}\beta_r x_r}}{1+e^{\alpha+\Sigma_{r=1}^{R}\beta_r x_r}}$, (Seber, 1984)(Agresti, 2007):

Denotando a partir de ahora $\alpha = \beta_0$, $\mathrm{X = (X_0, X_1, \cdots, X_R)'}$ y $\mathrm{x = (x_0, x_1, \cdots, x_R)'}$ con $\mathrm{X_0 = 1}$, el modelo quedar resumido como sigue

$$\mathrm{p(x) = \frac{e^{\Sigma_{r=0}^{R}\beta_r x_r}}{1+e^{\Sigma_{r=0}^{R}\beta_r x_r}} = \frac{e^{\beta' x}}{1+e^{\beta' x}},} \quad (2.1)$$

(Hosmer, 2000): donde β es el vector columna de parámetros $(\beta_0, \beta_1, \dots, \beta_R)^T$.
Equivalentemente el modelo de regresión logística múltiple se puede ver como un modelo de regresión lineal múltiple para la transformación logit $\ln(p(x)/(1-p(x))) = \sum_{r=0}^{R}\beta_r x_r$.
Los parámetros del modelo (2.1) se estiman por máxima verosimilitud y se interpretan de la manera usual, ver Hosmer & Lemeshow (2000):

2.2.2. Modelo Naive Bayes

El modelo probabilístico puede ser formulado como $\mathrm{p(y|x_1, x_2, \dots x_R)}$ donde y una variable dependiente, con un pequeño número de resultados (o clases): en nuestro caso, dos. Esta variable está condicionada por varias variables independientes desde $\mathrm{x_1, \cdots, x_R}$. El problema es que si el número R de variables independientes es grande (o cuando ´estas pueden tomar muchos valores): entonces basar este modelo en tablas de probabilidad se vuelve imposible. Por lo tanto el modelo se reformula para hacerlo más manejable utilizando el teorema de Bayes (Rowe, 2003):

$$\mathrm{p(y|x_1, x_2, \dots x_R) = \frac{p(y)p(x_1, x_2 \dots x_R|y)}{p(x_1, x_2, \dots x_R)}} \quad (2.2)$$

Esto puede escribirse en lenguaje común como $Probabilidad\ posterior = \frac{Anterior * Probabilidad}{Evidencia}$

En la práctica sólo importa el numerador, ya que el denominador no depende de y y los valores de $x_1, \cdots, x_R$ son datos, por lo que el denominador es, en la práctica, constante. El numerador de la ecuación (2.2) es equivalente a la probabilidad compuesta $p(y, x_1, \cdots, x_R)$ que puede ser reescrita como sigue, aplicando repetidamente la definición de probabilidad condicional

$$\begin{aligned} p(y, x_1, \cdots, x_R) &= p(y)p(x_1, \cdots, x_R|y) = p(y)p(x_1|y)p(x_2, \cdots, x_R|y, x_1) \\ &= p(y)p(x_1|y)p(x_2|y, x_1)p(x_3, \cdots, x_R|y, x_1, x_2) \\ &= p(y)p(x_1|y)p(x_2|y, x_1)p(x_3|y, x_1, x_2)p(x_4, \cdots, x_R|y, x_1, x_2, x_3) \end{aligned}$$

y así sucesivamente. Ahora suponiendo independencia condicional entre las x_i y la x_j, $i \neq j$, entonces $p(x_i|y, x_j) = p(x_i|y)$ por lo que la probabilidad compuesta puede expresarse como

$$p(y, x_1, \cdots, x_R) = p(y)p(x_1|y)p(x_2|y) \cdots p(x_R|y) = p(y) \prod_{i=1}^{R} p\,(x_i|y)$$

Bajo estos supuestos, la distribución condicional sobre la variable clasificatoria y puede expresarse como:

$$p(y|x_1, \cdots, x_R) = \frac{1}{Z} p(y) \prod_{i=1}^{R} p\,(x_i|y)$$

Donde Z es una cantidad que depende solo de $x_1, \cdots, x_R$, es decir, es constante si los valores de x_i son conocidos.

Todos los parámetros del modelo, por ejemplo, clases prioris y características de las distribuciones de probabilidad, se puede aproximar con frecuencias relativas del conjunto de entrenamiento. Estas son las estimaciones de máxima verosimilitud de las probabilidades. Una clase priori se puede calcular asumiendo clases equiprobables, es decir, priori = 1/ (número de clases): o mediante el cálculo de una estimación de la probabilidad de clase del conjunto de entrenamiento, es decir, el priori de una clase dada = (número de muestras en la clase) / (número total de muestras): Para la estimación de los parámetros de la distribución de una característica, se debe asumir una distribución o generar modelos de estadística no paramétrica de las características del conjunto de entrenamiento. Para la estimación y clasificación se hace uso del software denominado Weka (José & Cesar, 2006).

Las hipótesis sobre las distribuciones de características son llamadas el modelo de eventos del Clasificador Naive Bayes. La distribución multinomial y la distribución de Bernoulli son populares para características discretas como las encontradas en la clasificación de enfermos a partir de los síntomas. Estas hipótesis conducen a dos modelos distintos, que a menudo se confunden.

2.2.3. Árboles de decisión.

Un árbol de decisión es una forma gráfica y analítica de representar todos los eventos (sucesos) que pueden surgir a partir de una decisión asumida en cierto momento. Nos ayudan a tomar la decisión más acertada, desde un punto de vista probabilístico, ante un abanico de posibles decisiones. Permite desplegar visualmente un problema y organizar el trabajo de cálculos que deben realizarse. Los modelos basados en ´arboles binarios por las facilidades que proporcionan en su implementación y explicación tienen aplicación en diversas ramas. Los modelos de árbol de Regresión y Clasificación (C&RT, Classification & Regression Tree): fueron introducidos en la Estadística por Breiman et al. (1984): La idea principal de dicho método consiste en particionar el espacio de las variables independientes, en forma tal que los valores de la variable de respuesta sean cada vez más homogéneos dentro de las clases de dicha partición. Los modelos C&RT son fáciles de aplicar e interpretar, por tal motivo se han hecho populares en diferentes áreas. Bajo el enfoque Bayesiano existen diversas propuestas. Chipman et al. (1998a) y Denison et al. (1998): proponen nuevas metodologías que utilizan los modelo C&RT en combinación con los métodos de Monte Carlo vía cadenas de Markov (MCMC, Markov Chain Monte Carlo): En estos trabajos utilizan los métodos MCMC para la exploración de la distribución a posteriori (Gamerman, 1997 y Guttorp, 1995):

Quinlan, J. R. (1993): desarrolla el algoritmo C4.5 que construye ´arboles de decisión desde un grupo de datos de entrenamiento, usando el concepto de entropía de información. Los datos de entrenamiento son un grupo $S = s_1, s_2, \cdots$ de ejemplos ya clasificados. Cada ejemplo $s_i = x_1, x_2, \cdots$ es un vector donde $x_1, x_2, \cdots$ representan los atributos o características del ejemplo. Los datos de entrenamiento son aumentados con un vector $C = c_1, c_2, \cdots$ donde $c_1, c_2, \cdots$ representan la clase a la que pertenece cada muestra

En cada nodo del árbol, C4.5 elige un atributo de los datos que más eficazmente dividen el conjunto de muestras en subconjuntos enriquecidos en una clase u otra. Su criterio es el normalizado para ganancia de información (diferencia de entropía) que resulta en la elección de un atributo para dividir los datos. El atributo

con la mayor ganancia de información normalizada se elige como parámetro de decisión. El algoritmo C4.5 divide recursivamente en sublistas más pequeñas. J48 es una implementación open source en lenguaje de programación Java del algoritmo C4.5 en la herramienta weka de minería de datos y es el que se utiliza en el presente trabajo para la clasificación.
Los ´arboles de decisión consisten en una estructura jerárquica, donde en cada nivel se aplica una prueba para uno o más valores de atributos que pueden tener uno o dos resultados. Clasificación de una instancia (una muestra) se inicia en la raíz del árbol. La instancia se evalúa en un nodo y toma la rama apropiada para su resultado. La clasificación está representada por las hojas. Las principales ventajas de los árboles de decisión son que no sólo pueden construir un modelo fácilmente interpretable, sino que también permiten la reducción automática de la selección de características y complejidad paso a paso. Ejemplos de métodos basados en ´arboles de decisión incluyen ID3, C4.5, J48, CART, mencionados antes.
Random Forest pertenece a la clase de métodos de conjunto. La idea básica detrás de la metodología es utilizar simultáneamente un conjunto de modelos de clasificadores individuales, combinando sus salidas para devolver una decisión. De esta manera, el conjunto puede superar a sus modelos constituyentes individuales (Bonissone, 2010; Breiman, 2001; Rokacth, 2010): Un conjunto es en sí mismo un algoritmo de aprendizaje supervisado, ya que puede ser entrenado y luego se usa para hacer predicciones. La introducción de Random Forest (RF) se hizo por primera vez en 2001, en un artículo de Breiman (Breiman, 2001): RF es un método de aprendizaje para la clasificación (y regresión) que operan mediante la construcción de una multitud de ´arboles de decisión sin podas individuales (modelos) en el momento de formación (un bosque de decisión) por muestreo bootstrap de los datos de entrenamiento y selección aleatoria de atributos. La salida de un RF es una predicción obtenida combinando apropiadamente las diferentes predicciones de las salidas de los árboles individuales. Por lo tanto, a través de la naturaleza de RF, la exactitud del modelo mejora significativamente sobre la de un solo árbol de clasificación

2.2.4. Medición de la calidad de clasificación y predicción.

A continuación, se describe como se realiza una tarea de clasificación y evaluado.

a) Se aplica el proceso de validación cruzada para evaluar la precisión de la predicción de los métodos de clasificación. El proceso es el siguiente. El conjunto determinado de datos E se divide en k conjuntos de ejemplos, $C_1, \cdots, C_k$. A continuación, se construye un conjunto de datos $D_i = E - C_i$, y la precisión de un modelo adquirido de D_i en los ejemplos de $C_i(D_i y\ C_i)$ son los conjuntos de entrenamiento y prueba, respectivamente) se pone a prueba. La precisión final del método se estima promediando la precisión en los k ensayos de validación cruzada.
b) Además, el conjunto de entrenamiento se puede aplicar para evaluar la precisión de la predicción de los métodos de clasificación. El proceso es el siguiente. La precisión de un modelo obtenido del conjunto de datos E en los ejemplos de E se prueban. En este caso, el modelo podrá estar sobreajustado.

Cuando se realiza una clasificación, se evalúan las siguientes medidas:

- El porcentaje de casos clasificados correctamente es a menudo llamado la precisión o exactitud de la muestra.
- Kappa es una medida de probabilidad corregida de acuerdo entre las clasificaciones y las clases verdaderas. Se calcula tomando el acuerdo esperado por azar fuera del acuerdo observado y dividiendo por el máximo acuerdo posible. Un valor mayor que 0 significa que el clasificador está haciendo la clasificación mejor que el azar (1 indica un acuerdo perfecto, 0 indica que no hay acuerdo de lo que cabría esperar por azar):
- TP Rate (tasa de verdaderos positivos): proporción de ejemplos que se clasifica como clase X, entre todos los ejemplos que realmente tienen clase X, es decir, cuanto fue capturado de la clase.
- FP Rate (tasa de falsos positivos): proporción de ejemplos que se clasifica como clase X, pero pertenecen a una clase diferente, entre todos los ejemplos que no son de la clase de X.
- TN Rate, FN Rate (tasas de verdaderos negativos y tasa de falsos negativos): las contrapartes de las definiciones anteriores.
- Precisión: proporción de ejemplos que son verdaderamente de una clase dividido por el total de los casos clasificados como esa clase. $TP/(TP + FP)$:

- Recall (sensibilidad): proporción de casos clasificados como una clase dada dividido por el total real de esa clase. $TP/(TP + FN)$:
- F-Measure:(2 × Precisión × sensibilidad)/(Precisión + Sensibilidad)
- Receiver Operating Characteristics (ROC) también se utilizan para evaluar el poder de un método de clasificación de los diferentes pesos asimétricos. Puesto que el área bajo la curva ROC (denotado por AUC) es una porción del área de la unidad cuadrada, su valor será siempre entre 0 y 1. Un clasificador realista no debe tener una AUC más baja que 0.5 (área bajo la línea diagonal entre (0,0) y (1,1)): El AUC tiene una importante propiedad estadística: el AUC de un clasificador es equivalente a la probabilidad de que el clasificador se ubicar un ejemplo positivo elegido al azar más alto que una instancia negativa elegida al azar.

3. RESULTADOS Y DISCUSIÓN

3.1. Selección de atributos y clasificación.

Siguiendo los métodos Genuer y Cadenas para la selección de atributos, en orden de mayor a menor capacidad de clasificación, los dos conjuntos de atributos se obtuvieron siguiendo el procedimiento de selección de atributos en WEKA, utilizando inicialmente un conjunto de entrenamiento y posteriormente el método de validación cruzada, con la finalidad de identificar los atributos estadísticamente significativos para explicar el riesgo cardiovascular con cada uno de los algoritmos considerados.
El conjunto de entrenamiento considera 172 instancias y las variables Edad (E): presión sistólica (PS): presión diastólica (PD): hipertensión arterial (HTA): lipoproteína de baja densidad (LDL): lipoproteína de alta densidad (HDL): prediabetes, Diabetes (D): obesidad (OB1): Ejer-treinta-minutos (IF): Sexo (Mujer, Hombre): edad mayor a 55 años, antecedentes familiares y la variable de clase RC (Sí, No): Para cada uno de los algoritmos considerados, el conjunto de entrenamiento incluyó las 11 variables y la variable de clase. Una vez ejecutado el experimento, se encontró que para los modelos considerados las variables seleccionadas por el procedimiento de búsqueda y que mejor explican el RC, son las que se indican en la tabla (1):

Tabla 1: Variables incluidas en los modelos.

Modelo	Variables
Reg. Logística	Edad, HTA, Diabetes, RCFraming
NaiveBayes	Edad, PD, HTA, LDL, Diabetes, RCFraming
Tree-J48	Edad, PS, LDL, Diabetes, OB1, Sexo, RCFraming
R-Tree	Edad, PS, HTA, Diabetes, Sexo, RCFraming
R-Forest	Edad, Diabetes, Sexo, RCFraming

Como se observa en la tabla (1): el primer modelo incluye tres atributos, el segundo incluye cinco, el tercero incluye seis, el cuarto cinco y solo tres el último modelo. Se observa que los atributos Edad y Diabetes han resultados seleccionados en todos los modelos.

3.1.1. Eficiencia en la clasificación.

En este procedimiento de clasificación, el conjunto de entrenamiento permitió identificar a las variables Edad, hipertensión arterial y diabetes como estadísticamente significativas, para el modelo logístico; para NaiveBayes son estadísticamente significativas Edad, PD, HTA, LDL y Diabetes; para Tree-J48 se tienen Edad, PS, LDL, Diabetes, OB1 y Sexo; para Random Tree, cuyas variables estadísticamente significativas son Edad, PS, HTA, Diabetes y Sexo; Random Forest solo considera los atributos Edad, Diabetes y Sexo. Si consideramos el número de atributos necesarios para explicar el riesgo cardiovascular, los mejores serán el modelo de regresión logística y Random Forest; si consideramos la precisión y el índice de acuerdo los mejores modelos son Random Fores y Random Tree, aunque indudablemente el mejor es Random Forest. La validación del modelo se realizó aplicando en cada caso un 10-fold cross-validation y el correspondiente índice Kappa de acuerdo en la clasificación. En las siguientes tablas se muestra la precisión media para cada algoritmo.

Tabla 2: Métodos de clasificación aplicados al riesgo cardiovascular.

	Reg. Logística	NaiveBayes	Tree-J48	R-Tree	R-Forest
10-fold cross-validation	94.186	94.186	94.186	96.512	97.043
Kappa statistic	0.6737	0.66	0.69	0.8042	0.841

Observe que en la tabla (2) la tasa de clasificaciones correctas para el modelo de regresión logística fue de 0.942 y sensibilidad (Recall) de 0.974, los modelos Naive Bayes, Tree.J48, R.Tree y R.Forest, la tasa de

clasificaciones correctas no sufrió cambio en la sensibilidad, igual que el modelo Random Forest, aunque la tasa de clasificaciones correctas inicial fue de 97.1% y se mantuvo en la resustitución en ese valor. Se observa que el estadístico de consenso en la clasificación para el modelo logístico es de 0.6737, mientras que para Random Forest es de 0.84, lo que indica que el modelo es mejor. Ambos modelos explican el riesgo cardiovascular a partir de tres variables, aunque coinciden en dos de ellas, Edad y Diabetes. El modelo logístico incluye la hipertensión arterial y Random Forest incluye Sexo.

Tabla 3: Precisión detallada por clase para los modelos.

Modelo	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
logistic	0.974	0.333	0.962	0.974	0.968	0.675	0.952	0.994	No
	0.667	0.026	0.750	0.667	0.706	0.675	0.952	0.629	Si
Weighted	0.942	0.333	0.962	0.974	0.968	0.675	0.952	0.994	
Naive Bayes	0.981	0.389	0.956	0.981	0.968	0.662	0.966	0.996	No
	0.611	0.019	0.786	0.611	0.688	0.662	0.966	0.763	Si
Weighted	0.942	0.350	0.938	0.942	0.939	0.662	0.966	0.972	
Trees.J48	0.968	0.278	0.968	0.968	0.968	0.690	0.902	0.978	No
	0.722	0.032	0.722	0.722	0.722	0.690	0.902	0.674	Si
Weighted	0.942	0.252	0.942	0.942	0.942	0.690	0.902	0.947	
Random Tree	0.987	0.222	0.974	0.987	0.981	0.806	0.882	0.973	No
	0.778	0.013	0.875	0.778	0.824	0.806	0.882	0.704	Si
Weighted	0.965	0.200	0.964	0.965	0.964	0.806	0.882	0.945	
Random	0.987	0.167	0.981	0.987	0.984	0.841	0.986	0.998	No
	0.833	0.013	0.882	0.833	0.857	0.841	0.986	0.905	Si
Weighted	0.971	0.151	0.970	0.971	0.971	0.841	0.986	0.989	

Tabla 4: Matriz de confusión para cada uno de los conjuntos.

RC-logistic		
No	Si	
150	4	No
6	12	Si

RC-Naive Bayes		
No	Si	
151	3	No
7	11	Si

RC-Trees.j48		
No	Si	
149	5	No
5	13	Si

RC-Random Tree		
No	Si	
152	2	No
4	14	Si

RC-Random Forest		
No	Si	
152	2	No
3	15	Si

4. CONCLUSIONES.

Los métodos de clasificación de riesgo cardiovascular considerados, tienen tasa de clasificaciones correctas relativamente altas, superiores al 94%, y la tasa de clasificaciones correctas mayor corresponde al modelo Random Forest (97.1%): que sin duda, debido al índice Kappa más alto (0.84) es el mejor. Este modelo, según lo establecido en la literatura, tiende a sobreajustar ciertos grupos de datos, además, a diferencia de los árboles de clasificación, es difícil de interpretar por el hombre. Por lo anterior, aunque la clasificación usando Random Forest es mejor, para fines prácticos de interpretación recomendamos el uso del modelo random tree o el modelo logístico, toda vez que con estos modelos es posible hacer predicciones de Riesgo cardiovascular en términos probabilísticos. Un valor de probabilidad cercano a 1 indicar alto riesgo cardiovascular, mientras que un valor cercano a 0 indicar bajo riesgo cardiovascular.

Por ejemplo, para el caso logístico, con una calculadora de bolsillo o una hoja de cálculo es posible, para un sujeto dado, hallar $\mathrm{p}[\mathrm{y}=1|\vec{\mathrm{x}}]=\frac{e^{\eta}}{1+e^{\eta}}$, donde $\eta = 9.2114 - 0.1083\mathrm{E} - 1.2586\mathrm{HTA} - 2.9441\mathrm{D}$ y e es la base de los logaritmos neperianos.

Acknowledgements. The authors thank the two anonymous referees for their helpful comments on the early version of these paper. One of the authors thanks the support of VLIR-Project JOINT2018SEL002

RECEIVED: JUNE, 2018
REVISED: JULY, 2018

REFERENCIAS.

[1] AGRESTI, A. (2007): **An introduction to Categorical Data Analysis**. Wiley, New York.
[2] ALCOCER, L., A., LOZADA, O., FANGHNEL, G., S´ANCHEZ-REYES, L.,y CAMPOS-FRANCO, E. (2011): Estratificación del riesgo cardiovascular global. Comparación de los métodos Framingham y SCORE en población Mexicana del estudio PRIT. **Cirugía y Cirujanos**, 79, 168- 174.
[3] ALEJANDRO, S. J., (2008): Chocolate bueno para el corazón, Recuperado el 2 de julio de 2016, Desde:http**://asisucede.com.mx/chocolate-bueno-para-el-corazn/.**
[4] BONISSONE, P., (2010): Soft Computing: A Continuously Evilving Concept, **Computational Intelligence System**, to appear, 2010.
[5] BREIMAN, L. (2001): Random Forests, Statistics Department University of California Berkeley.
[6] BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A.; and STONE, C.I. (1984): **Classification and regression trees**. Wadsworth , Belmont, Calif.
[7] CHIPMAN, H. A.; GEORGE, E. F.; and MCCULLOCH, R. E. (1998a): Bayesian CART model search (with discussion): **Journal of the American Statistical Association**, 93, 935-948.
[8] DENISON, D. G. T., MALLICK, B. K. and SMIT, A. F. M. (1998): Journal of the Royal Statistical Society. Series B (Statistical Methodology): Vol. 60, No. 2. (1998): pp. 333-350.
[9] DMEDICINA (2015): Factores de riesgo cardiovascular. Recuperado el 2 de Julio de 2016, de **http://www.dmedicina.com/enfermedades/enfermedades- vasculares-y-del-corazón/factores-de-riesgo-cardiovascular.html.**
[10] ENSANUT. (2012): **Módulo del Programa de Oportunidades en la Encuesta Nacional de Salud y Nutrición.** México: Instituto Nacional de salud Pública.
[11] FARMACIAS ABC. (2006): Riesgo cardiovascular, la importancia de la prevención. Recuperado el 06 de Julio de 2016, de **http://www.farmacias-abc.com.ar/contenido/Index624.asp.**
[12] GAMERMAN, D., (1997): **Markov Chain Monte Carlo, Stochastic Simulation for Bayesian Inference**, Chapman and Hall. New York.
[13] GOBIERNO FEDERAL, SALUD, SEDENA Y SEMAR. (2016): Detección y Estratificación de Factores de Riesgo Cardiovascular. **Recuperado el 02 de Julio de 2016, de internet**
[14] GUTTORP, P., (1995): **Stochastic modeling of scientific data**. Chapman and Hall, London;.
[15] HOSMER, D. W. and S. LEMESHOW (2000): **Applied Logistic Regression**, 2nd edn. Wiley, New York:.
[16] LÓPEZ, J. (2013): Enfermedades cardiovasculares. Recuperado el 2 de Julio de 2016, de Prezi **https://prezi.com/y8f6izd9jnks/enfermedades-cardiovasculares/.**
[17] MOLINERO, L. M. (2003): **Modelos de riesgo cardiovascular. Estudio de Framingham. Proyecto SCORE.** Asociación de la Sociedad Española de Hipertensión. Liga Española para la Lucha contra la Hipertensión Arterial.
[18] OMS. (2015): Enfermedades cardiovasculares. Recuperado el 2 de Julio de 2016, de Organización Mundial de la Salud: **http : //www.who.int/mediacentre/f actsheets/f s317/es/**
[19] QUINLAN, J. R. C4.5 (1993): **Programs for Machine Learning**. Morgan Kaufmann Publishers, London
[20] QUO. (2012): El treinta por ciento de los mexicanos son hipertensos. Recuperado el 2 de Julio de 2016, de **http : //quo.mx/noticias/2012/05/02/el − 30 − de − mexicanos − son − hipertensos**.
[21] ROKACH, L. (2010): Ensemble-based classifiers, Artif Intell Rev. 33: 1 − 39.
[22] ROWE, DANIEL, B., Multivariate Bayesian Statistics, Chapman and Hall/CRC, New York, Washington, D.C.
[23] RUIZ, E., SEGURA, L., and AGUSTÍ, R. (2012): Uso del Score de Framingham como Indicador de los Factores de Riesgo de las Enfermedades Cardiovasculares en la población Peruana. **Revista Peruana de Cardiología.** XXXVII, 3.
[24] SEBER, G.A. (1984): **Multivariate Observations**.: Wiley, New York.
[25] WHO. (2003): The World Health Report 2003. Recuperado el 2 de Julio de 2016, de World Healt Organization: **http : //www.who.int/whr/2003/en/whr03en.pdf ?ua = 1.**