UNIVERSIDAD AUTÓNOMA METROPOLITANA - IZTAPALAPA
DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERÍA

MODELO DE ANÁLISIS PARA LA PREDICCIÓN DE TROMBOSIS FAMILIAR

Tesis que presenta

Carlos Gabriel Sánchez Lordméndez

Para obtener el grado de

Maestro en Ciencias Matemáticas Aplicadas e Industriales

Asesores: Dra. Blanca Rosa Pérez Salvador

Dr. Héctor Alfredo Baptista González

Jurado Calificador:

Presidente:	Dr. Carlos Díaz Avalos
Secretario:	Dr. Alberto Castillo Morales
Vocal:	Dr. Héctor Alfredo Baptista González
Vocal:	Dra. Blanca Rosa Pérez Salvador

México D.F. Enero 2015

Dedicado a mis padres

Agradecimientos:

A Dios porque me ha hecho encontrar luz, aún en los momentos de más oscuridad.

A la MCMAI por creer en mí y darme la oportunidad de prepararme más y ser mejor persona.

Al CONACYT por financiar los estudios, esperando que la culminación de éste trabajo brinde la misma oportunidad a futuras generaciones, en particular de la MCMAI.

A la Dra. Blanca Rosa Pérez Salvador por la gran paciencia que tuvo para guiarme como maestra de la MCMAI, como tutora y asesora de mi tesis, por todas sus atenciones, consejos y enseñanzas a lo largo de todo el proceso.

Al doctor Baptista por todo su apoyo, la orientación brindada y la oportunidad de asistir al hospital de Perinatología. Agradezco también al equipo de trabajo del doctor Baptista por hacer muy gratas mis estancias en el hospital, especialmente a Fany Rosenfeld y Rocío Trueba.

Al ingeniero Jorge Solano por el apoyo en el manejo de RScript y a la licenciada Cinthia Rodríguez Maya por su orientación para subir al servidor los programas relacionados con la interfaz de usuario.

A mis papás, hermana y familiares por su apoyo desde siempre.

A Laura por su amor y compañía.

A mis alumnos cuya energía me motiva a superarme.

A todas aquellas personas que no he mencionado pero que con algún gesto gentil hicieron que las metas del día a día se hicieran más fáciles.

Índice general

Presentación.

El presente trabajo ha sido el resultado conjunto con el doctor Héctor Baptista y el personal que labora en el laboratorio de hematología perinatal del Instituto Nacional de Perinatología.

Los médicos de dicha institución estudian entre otras enfermedades la trombosis y su ocurrencia en personas de la misma familia. Dentro de estos estudios se presenta el diagnóstico o la clasificación de los individuos como propensos, o no, a padecer trombosis, situación que en ocasiones reviste la problemática de clasificar erróneamente a un sujeto; en particular, ocurre la situación en la cual la persona considerada como menos propensa a tener trombosis con los criterios usados, finalmente sí la padece. Con la finalidad de disminuir la probabilidad de que esto ocurra se decidió construir un par de clasificadores que coadyuven a los médicos en sus diagnósticos. Las metas que se persiguen al implementar dichos clasificadores son las siguientes:

- Usar software libre.
- Crear un programa que permita, dentro de ciertos límites, realizar la clasificación para distintas bases de datos y distintos padecimientos haciendo sólo pequeñas modificaciones.
- Hacer que el programa sea capaz de excluir variables de poca relevancia.
- Crear una herramienta que pueda usar datos del contexto mexicano.

Adicionalmente, dado que se trabajó con R, un software que carece de un constructor de

interfaces de usuario, un objetivo adicional es la construcción de una interfaz para que los médicos puedan hacer uso del programa con facilidad y tengan una herramienta de clasificación semejante a la ya existente llamada Vienna Prediction Model for Recurrent VTE [1].

En el primer capítulo se establece la teoría relacionada con los métodos de clasificación empleados, el capítulo 2 se dan detalles de la puesta en práctica de cada uno de estos métodos y finalmente en el tercer capítulo se presentan resultados obtenidos con la base de datos que nos proporcionó el Instituto Nacional de Perinatología Isidro Espinosa de los Reyes y se dan conclusiones.

[1] Este programa está disponible en internet en la página:
http://www.meduniwien.ac.at/user/georg.heinze/zipfile/ViennaPredictionModel.html.

Capítulo 1

Métodos de clasificación

En este capítulo se describirán brevemente las ideas relacionadas a los métodos de clasificación y se darán detalles de dos de ellos: la regresión logística y una propuesta a la cual se le ha llamado mejor clasificador de nivel α. También se describirán los conceptos de devianza en el caso particular de la regresión logística y la prueba de independencia ji-cuadrada, pues estos fueron los métodos que se usaron para determinar las variables de la base que serían incluidas en la estimación de los dos modelos estudiados.

1.1. Conceptos generales.

Supóngase que la población objetivo está particionada en dos conjuntos no vacíos y complementarios E y $\overline{E}$. Por ejemplo, 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 de medirse a los individuos como: su nivel de colesterol, presión sanguínea, etc., nos pueden dar información de su estatus de riesgo. En este sentido, es razonable pensar que se puede definir una función de las variables observadas que nos dé información sobre 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 con ella se podrá decidir si un sujeto en particular queda clasificado como

perteneciente a E o a $\overline{E}$.

Esto es, sea $x_i = (x_{11}, \ldots, x_{ik})$ el vector de variables explicativas asociadas al sujeto a_i en la población y X el conjunto de todos los posibles valores del vector x_i. Suponga que X se descompone en dos conjuntos no vacíos y complementarios A_E y $\overline{A_E} = A_{\overline{E}}$ tales que si un individuo a_0 tiene un vector específico x_0 que pertenece a A_E, $(x_0 \in A_E)$, al individuo a_0 se le clasifica dentro del conjunto de alto riesgo E, mientras que si su vector x_0 se asigna al conjunto $A_{\overline{E}}$, $(x_0 \in \overline{A_E})$, entonces se le reconocerá como perteneciente al conjunto de bajo riesgo $\overline{E}$.

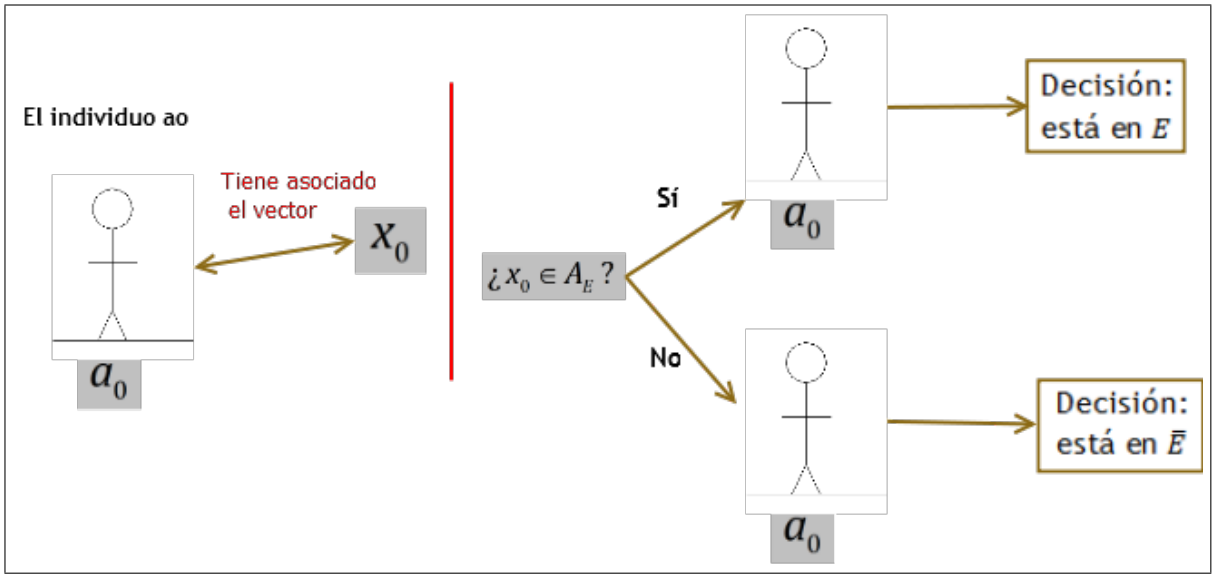


Figura 1.1: Un individuo a_0 tiene asociado un vector x_0, la clasificación de x_0 recae en la decisión sobre a_0.

Es importante considerar que ningún método de clasificación es perfecto, por lo que tendremos cuatro posibles escenarios:

1. Que a un sujeto se le clasifique como elemento de E y realmente esté en E, lo cual es correcto. Esto ocurre con probabilidad $P(A_E|E)$.

2. Que a un sujeto se le clasifique como elemento de $\overline{E}$ y realmente esté en $\overline{E}$, lo cual también es correcto. Esto ocurre con probabilidad $P\left(A_{\overline{E}}|\overline{E}\right)$.

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 $P\left(A_{\overline{E}}|E\right)$ y lo llamaremos error tipo I.

4. Que a un sujeto se le clasifique como elemento de $\overline{E}$ y realmente esté en E, lo cual tampoco es correcto. A este error lo llamaremos error tipo II y lo escribiremos como $P\left(A_E|\overline{E}\right)$.

Entonces, la calidad del método de clasificación se medirá en función de la probabilidad de cometer estos dos errores.

Existen varias técnicas para obtener la función clasificadora y en este trabajo se revisaran dos de ellas: la regresión logística y el mejor clasificador de nivel α.

1.2. Clasificación mediante regresión logística.

En esta sección se presenta el modelo de regresión logística como una técnica de clasificación. La regresión logística es el modelo lineal generalizado cuando la función de distribución de probabilidad asociada es una binomial. Comenzamos dando la definición de modelo lineal generalizado.

DEFINICIÓN **1.** *Sean $Y_1, \ldots, Y_n$ variables aleatorias independientes con medias $\mu_1, \ldots, \mu_n$ respectivamente y sea $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ un vector de variables explicativas para $Y_i, i = 1, 2, \ldots, n$. Se dice que Y_i y x_i se relacionan mediante un modelo lineal generalizado si*

$$g(\mu_i) = x_i^T \beta,$$

donde g es una función monótona diferenciable, β un vector de parámetros y la distribución de probabilidad de Y_i no es necesariamente normal pero sí un miembro de la familia exponencial. A la función g se le llama función vínculo.

Cuando la variable de respuesta Y es binaria es común que a uno de los valores se le codifique como 1 y al otro como 0; de tal forma que se tiene lo siguiente:

Y_i	$P(Y_i = y_i)$
0	$1 - \pi_i$
1	π_i

Y por consiguiente $0 \leq E(Y_i) = \pi \leq 1$ (Myers et al., 2002).

Podemos ver que estimar $E(Y_i)$ mediante una regresión lineal de la forma $E(Y_i) = \beta_0 + \sum \beta_i x_i$ presenta dificultades porque podría ser que el valor pronosticado por la regresión lineal fuera mayor a uno en algunos casos y menor que cero en otros, lo cual sería contrario con la restricción $0 \leq E(Y_i) = \pi_i \leq 1$ descrita arriba.

En estos casos, se debe recurrir a un modelo lineal generalizado para tener una transformación $g(\mu_i)$ que garantice que $0 \leq E(Y_i) = \pi_i \leq 1$

Dentro de las posibilidades comunes que satisfacen estas restricciones se pueden mencionar la función probit, la función complementaria log-log, y la función logit que es la que se verá con detalle en esta sección y que es la función vínculo asociada a la regresión logística, la cual presenta la siguiente estructura (Myers et al., 2002):

$$\pi_j = \frac{\exp(\beta_0 + \sum \beta_i x_{ij})}{1 + \exp(\beta_0 + \sum \beta_i x_{ij})} = \frac{1}{1 + \exp[-(\beta_0 + \sum \beta_i x_{ij})]}$$

Obsérvese que si se despeja el término $\beta_0 + \sum \beta_i x_{ij}$ se encuentra la función vínculo:

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

Por lo tanto

$$ln \frac{\pi_j}{1 - \pi_j} = \beta_0 + \sum \beta_i x_{ij}$$

En la siguiente sección se muestra que en el proceso de estimación de los parámetros del modelo de regresión logística se llega a un sistema de ecuaciones no lineal y que por lo tanto su solución se tendrá que obtener por métodos numéricos, con la ayuda de una computadora.

1.2.1. Estimación de parámetros en el modelo de regresión logística.

La estimación de los parámetros β_i en la regresión logística se hace con el método de máxima verosimilitud. Dado que cada observación tiene una distribución tipo $Bernoulli\,(\pi_i)$, la función de probabilidad de Y_i es:

$$f\,(y_i) = {\pi_i}^{y_i}(1-\pi_i)^{1-y_i}$$

Debido al supuesto que cada observación es independiente, la función de verosimilitud es (Myers et al., 2002):

$$L\,(\beta; y_1, y_2, ..., y_n) = \prod_{i=1}^{n} {\pi_i}^{y_i}(1-\pi_i)^{1-y_i}$$

Aplicando logaritmo natural y considerando que $1-\pi_i = \left[1+e^{x_i^T\beta}\right]^{-1}$ y que $\eta_i = \ln\frac{\pi_i}{1-\pi_i} = x_i^T\beta$, entonces se puede escribir:

$$\begin{aligned}
\ln L\,(\beta; y_1, y_2, ..., y_n) &= \ln\left[\prod_{i=1}^{n} {\pi_i}^{y_i}(1-\pi_i)^{1-y_i}\right] \\
&= \sum_{i=1}^{n} \ln\left[{\pi_i}^{y_i}(1-\pi_i)^{1-y_i}\right] \\
&= \sum_{i=1}^{n} y_i \ln\frac{\pi_i}{1-\pi_i} + \sum_{i=1}^{n} \ln\,(1-\pi_i) \\
&= \sum_{i=1}^{n} y_i x_i^T\beta - \sum_{i=1}^{n} \ln\left[1+e^{x_i^T\beta}\right] \\
&= \beta^T X^T y - \sum_{i=1}^{n} \ln\left[1+e^{x_i^T\beta}\right]
\end{aligned}$$

Al derivar respecto a los parámetros:

$$\begin{aligned}
\frac{\partial \ln L\,(\beta; y_1, y_2, ..., y_n)}{\partial\beta} &= X^T y - \sum_{i=1}^{n} \frac{e^{x_i^T\beta}x_i}{1+e^{x_i^T\beta}} \\
&= X^T y - \sum_{i=1}^{n} \pi_i x_i \\
&= X^T y - X^T\pi
\end{aligned}$$

Dado que $\mu = \pi$, al igualar la derivada anterior a cero, se puede escribir:

$$X^T (y - \mu) = 0$$

Que es un sistema de ecuaciones no lineal en el vector de parámetros β porque:

$$\mu_i = \frac{n_i}{1 + e^{x_i^T \beta}}$$

1.2.2. Regresión logística a través de R.

Como se vio en la sección anterior, tratar de encontrar los parámetros del modelo de regresión logística conduce a un sistema de ecuaciones no lineales cuya solución sin la ayuda de una computadora es inviable. En este trabajo, para encontrar dichos parámetros se decidió emplear el lenguaje de programación R, debido principalmente a que es un software libre especializado en análisis estadístico.

En R, el comando para encontrar los parámetros de la regresión logística es *glm(formula, binomial, ...)*. Por ejemplo, supóngase que se tienen los datos de la tabla que se muestra en la figura 1.2 y que se quiere ajustar un modelo de la forma $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Entonces para obtener el modelo en R escribiríamos *glm(Y ~ X1+X2,binomial)*. La respuesta proporcionada se muestra en la figura 1.2.

Y	X1	X2
1	2	3
1	3	3
0	2	0
0	5	-1
1	-4	2

```
Call:  glm(formula = Y ~ X1 + X2, family = binomial)

Coefficients:
(Intercept)           X1           X2
    -18.578       -2.355       16.543

Degrees of Freedom: 4 Total (i.e. Null);  2 Residual
Null Deviance:     6.73
Residual Deviance: 3.189e-10  AIC: 6
```

Figura 1.2: Tabla con valores de ejemplo y una respuesta típica de *glm(formula,binomial)*.

1.2.3. Clasificación por regresión logística

La probabilidad de que un individuo i padezca la enfermedad se estima con la ecuación $\widehat{p}(x_i) = \dfrac{1}{1+e^{-\mathbf{x}_i^T\widehat{\beta}}}$, donde x_i es el vector de variables explicativas del individuo i. Si $\widehat{p}(x_i)$ es "pequeña", se clasifica al individuo i como de bajo riesgo de sufrir la enfermedad, en cambio si $\widehat{p}(x_i)$ es "grande"se clasificará al individuo como de alto riesgo a padecer la enfermedad. La decisión de si $\widehat{p}(x_i)$ es "grande" o "pequeña"se determina a partir de si ésta es mayor o menor a un número predeterminado λ_α, al que se llamará valor crítico, el cual se va a encontrar siguiendo un criterio que se especificará más adelante en esta sección, y que se relaciona con que α sea la probabilidad de clasificar mal a un individuo de alto riesgo, esto es $P\left(A_{\overline{E}}|E\right) = \alpha$, es decir se considera que la probabilidad de cometer el error tipo I sea igual a α.

Para encontrar el valor crítico λ_α se parte la base de datos en dos conjuntos, uno con los datos de los individuos que ya padecieron la enfermedad (conjunto E) y otro con los individuos que no la han padecido (conjunto $\overline{E}$) y se estima la probabilidad condicional de riesgo en cada caso con la fórmula $\widehat{p}(x_i) = \dfrac{1}{1+e^{-\mathbf{x}_i^T\widehat{\beta}}}$. Luego se ordenan estas probabilidades de menor a mayor en los dos conjuntos.

Para ejemplificar este proceso, considérese una base de datos con n observaciones, n_1 de ellas correspondientes a personas que han padecido la enfermedad y n_0 de ellas correspondientes a personas que no han padecido la enfermedad. A este conjunto de observaciones se le aplica el modelo y se le ordena de acuerdo a su resultado y de acuerdo a su variable respuesta, como se muestra en la figura 1.3, en donde los paréntesis en los subíndices indican las estadísticas de orden; esto es para el conjunto de no enfermos para cualquier $r = 1, 2, \ldots, n_0 - 1$ se tiene que $\widehat{\mathrm{p}}\left(x_r\right) = \widehat{\mathrm{p}}_{(0,r)} \leq \widehat{\mathrm{p}}_{(0,r+1)} = \widehat{\mathrm{p}}\left(x_{r+1}\right)$ y para el grupo de los que enfermaron para toda $j = 1, 2, \ldots, n_1 - 1$ se satisface que $\widehat{\mathrm{p}}\left(x_j\right) = \widehat{\mathrm{p}}_{(0,j)} \leq \widehat{\mathrm{p}}_{(0,j+1)} = \widehat{\mathrm{p}}\left(x_{j+1}\right)$.

Ya ordenadas las probabilidades estimadas en cada uno de los conjuntos E y $\overline{E}$ se debe encontrar un valor de λ tal que si $\widehat{p}(x_i) > \lambda$, el individuo asociado se clasifica como

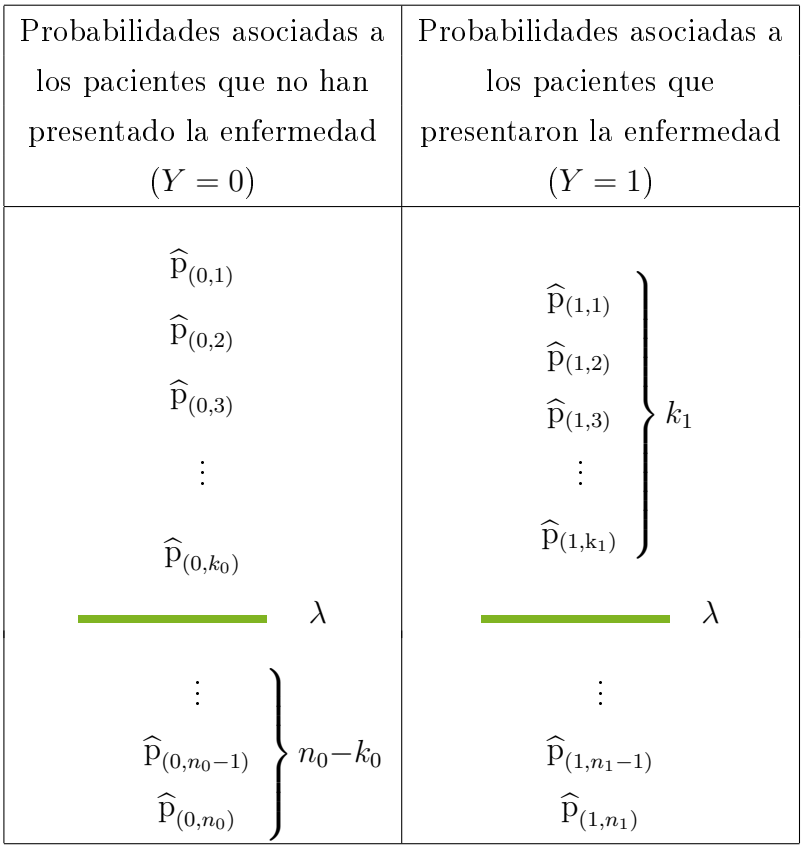


Figura 1.3: Probabilidades estimadas con el modelo de regresión logística de que los distintos individuos padezcan la enfermedad ya ordenadas.k_0 es el número de casos con $Y = 0$ y menores a λ y k_1 es el número de casos con $Y = 1$ y menores a λ.

de alto riesgo y si $\widehat{p}(x_i) \leq \lambda$ entonces el individuo se clasifica como de bajo riesgo. En la tabla de la figura 1.3 la raya horizontal representa el valor de λ.

Con estos datos se estima la probabilidad de cometer el error tipo I y la probabilidad de cometer el error tipo II como

- $\widehat{P}\left(A_{\overline{E}}|E\right) = \dfrac{k_1}{n_1}$
- $\widehat{P}\left(A_E|\overline{E}\right) = \dfrac{n_0 - k_0}{n_0}$

respectivamente, en donde k_0 es el número de casos con $Y = 0$ y menores a λ y k_1 es el número de casos con $Y = 1$ y menores a λ.

El valor de λ se puede cambiar; si se disminuye, disminuirá el valor de k_1 y aumentará el valor de $n_0 - k_0$, si lo aumentamos ocurrirá lo contrario. En estas condiciones se fija el valor de λ en el punto donde se minimice la suma $\dfrac{n_0 - k_0}{n_0} + \dfrac{k_1}{n_1}$.

Ahora se han mencionado algunos de los conceptos sobresalientes en el uso del modelo de regresión logística como clasificador. En la siguiente sección se detallan los conceptos relacionados al denominado mejor clasificador de nivel alfa.

1.3. Clasificación mediante el mejor clasificador de nivel alfa.

Para definir este clasificador se considera que uno de los dos errores es más grave. Si a una persona de alto riesgo de sufrir la enfermedad se le clasifica mal, no se le dará tratamiento preventivo poniendo en riesgo incluso su vida. Si a una persona de bajo riesgo se le clasifica mal entonces tendrá atención no necesaria pero que no le afectará en su salud, posiblemente se canalice la atención que podría brindársele a otros pacientes que lo necesitan más. En este sentido se podría considerar más grave clasificar mal a una persona de alto riesgo y entonces es conveniente poder controlar el nivel de la probabilidad de cometer ese error en un nivel que llamaremos α, es decir $P\left(A_{\overline{E}}|E\right) = \alpha$. Cumpliéndose

esta restricción se busca que la probabilidad del otro error, $P\left(A_E|\overline{E}\right)$, sea la más pequeña posible. El siguiente teorema establece la forma de la solución.

Teorema 1. *Dado el conjunto:*

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

Tal que $P\left(A_{\overline{E}}|E\right) = \alpha$*, entonces* $P\left(A_E|\overline{E}\right) < P\left(C|\overline{E}\right)$ *para todo* C *tal que* $P\left(\overline{C}|E\right) = \alpha$*.*

Esto significa que escogeremos como criterio de clasificación la ecuación

$$\frac{P\left(X = x|\overline{E}\right)}{P\left(X = x|E\right)} \leq \lambda$$

y que la probabilidad de cometer el error tipo II se minimiza cuando la probabilidad de cometer el error tipo I se fija a un valor α.

Demostración:

Sea C un conjunto definido por cualquier otro criterio clasificador cuya probabilidad de error tipo I sea también alfa, es decir $P\left(\overline{C}|E\right) = \alpha$.

Utilizando el hecho que la intersección es distributiva respecto a la unión, es decir que se satisface la relación $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ y dado que $A_E \cup A_{\overline{E}} = \mathbf{X} = C \cup \overline{C}$, podremos escribir al conjunto $A_{\overline{E}}$ como sigue:

$$A_{\overline{E}} = A_{\overline{E}} \cap X = A_{\overline{E}} \cap \left(C \cup \overline{C}\right) = \left(A_{\overline{E}} \cap C\right) \cup \left(A_{\overline{E}} \cap \overline{C}\right)$$

De forma similar también podremos escribir al conjunto $\overline{C}$ de la siguiente manera:

$$\overline{C} = \overline{C} \cap X = \overline{C} \cap \left(A_E \cup A_{\overline{E}}\right) = \left(\overline{C} \cap A_E\right) \cup \left(\overline{C} \cap A_{\overline{E}}\right)$$

Por otra parte observemos que:

$$\begin{aligned} P\left(A_{\overline{E}}|E\right) &= \frac{P\left(A_{\overline{E}} \cap E\right)}{P\left(E\right)} \\ &= \frac{P\left(\left[\left(A_{\overline{E}} \cap C\right) \cup \left(A_{\overline{E}} \cap \overline{C}\right)\right] \cap E\right)}{P\left(E\right)} \\ &= \frac{P\left[\left(A_{\overline{E}} \cap C \cap E\right) \cup \left(A_{\overline{E}} \cap \overline{C} \cap E\right)\right]}{P\left(E\right)} \end{aligned}$$

Y dado que $C \cap \overline{C} = \emptyset$, entonces $(A_{\overline{E}} \cap C \cap E) \cap \left(A_{\overline{E}} \cap \overline{C} \cap E\right) = \emptyset$ por lo que:

$$\begin{aligned} P\left(A_{\overline{E}}|E\right) &= \frac{P\left(A_{\overline{E}} \cap C \cap E\right) + P\left(A_{\overline{E}} \cap \overline{C} \cap E\right)}{P\left(E\right)} \\ &= \frac{P\left(A_{\overline{E}} \cap C \cap E\right)}{P\left(E\right)} + \frac{P\left(A_{\overline{E}} \cap \overline{C} \cap E\right)}{P\left(E\right)} \end{aligned}$$

En consecuencia:

$$P\left(A_{\overline{E}}|E\right) = P\left(A_{\overline{E}} \cap C|E\right) + P\left(A_{\overline{E}} \cap \overline{C}|E\right) \ldots (1)$$

De manera similar, se puede escribir:

$$P\left(\overline{C}|E\right) = P\left(\overline{C} \cap A_E|E\right) + P\left(\overline{C} \cap A_{\overline{E}}|E\right) \ldots (2)$$

Se reescriben las dos últimas ecuaciones nuevamente para tener un panorama más claro.

$$\begin{aligned} P\left(A_{\overline{E}}|E\right) &= P\left(A_{\overline{E}} \cap C|E\right) + P\left(A_{\overline{E}} \cap \overline{C}|E\right) \\ P\left(\overline{C}|E\right) &= P\left(\overline{C} \cap A_E|E\right) + P\left(\overline{C} \cap A_{\overline{E}}|E\right) \end{aligned}$$

Como $P\left(A_{\overline{E}}|E\right) = P\left(\overline{C}|E\right) = \alpha$ y evidentemente $P\left(A_{\overline{E}} \cap \overline{C}|E\right) = P\left(\overline{C} \cap A_{\overline{E}}|E\right)$ entonces de las dos ecuaciones anteriores se sigue que

$$P\left(A_{\overline{E}} \cap C|E\right) = P\left(\overline{C} \cap A_E|E\right) \ldots (3)$$

Si $x \in A_E \cap \overline{C}$, entonces $x \in A_E$ y por la definición de A_E se cumple

$$P\left(A_E \cap \overline{C}|\overline{E}\right) \leq \lambda P\left(A_E \cap \overline{C}|E\right) \ldots (4)$$

.

Por otra parte, cuando $x \in A_{\overline{E}} \cap C$ entonces $x \in A_{\overline{E}}$ y por lo tanto se cumple

$$P\left(A_{\overline{E}} \cap C|\overline{E}\right) > \lambda P\left(A_{\overline{E}} \cap C|E\right)$$

Considerando (1), (2), (3) y (4), se logra escribir:

$$P\left(A_E \cap \overline{C}|\overline{E}\right) \leq \lambda P\left(A_E \cap \overline{C}|E\right) = \lambda P\left(A_{\overline{E}} \cap C|E\right) < P\left(A_{\overline{E}} \cap C|\overline{E}\right)$$

Y por transitividad se tiene:

$$P\left(A_E \cap \overline{C}|\overline{E}\right) < P\left(A_{\overline{E}} \cap C|\overline{E}\right)$$

Sumando en ambos lados de la última desigualdad el término $P\left(A_E \cap C|\overline{E}\right)$ obtenemos

$$P\left(A_E \cap \overline{C}|\overline{E}\right) + P\left(A_E \cap C|\overline{E}\right) < P\left(A_{\overline{E}} \cap C|\overline{E}\right) + P\left(A_E \cap C|\overline{E}\right)$$

$$P\left(A_E|\overline{E}\right) < P\left(C|\overline{E}\right)$$

■

Esto es, la probabilidad de cometer el error tipo II es menor cuando elegimos como criterio de clasificación al conjunto A_E.

Ahora se explicará cómo utilizar este clasificador. Debe de tomarse en cuenta que desconocemos las probabilidades reales involucradas en la definición de A_E, por lo que se estimará este conjunto usando los estimadores de $P(X = x|E)$ y $P(X = x|\overline{E})$, donde $\widehat{P}(X = x|E)$ y $\widehat{P}(X = x|\overline{E})$ son las frecuencias relativas de cada vector de variables explicativas.

1.3.1. Cómo estimar el mejor clasificador con una base de datos.

Se parte de una base de datos donde las variables explicativas son categóricas y están identificados los individuos que no han padecido la enfermedad y los que sí la han padecido. Consideramos que si un individuo ya presentó la enfermedad es un individuo de alto riesgo. El conjunto E está formado por todos los individuos que ya padecieron la enfermedad mientras que el conjunto $\overline{E}$ está formado por los individuos de la base de datos que no han tenido la enfermedad. Los datos de la base se dividen en los conjuntos E y $\overline{E}$, como se muestra en la figura 1.4. En este punto cabe señalar que algunos de los sujetos que conforman a $\overline{E}$ en realidad podrían estar en riesgo sólo que no han manifestado la enfermedad y desconocemos quiénes son.

Después de hacer la partición de la base de datos, en cada conjunto se obtiene la frecuencia asociada a cada combinación de variables explicativas (para cada vector $\mathbf{x_i}$) y

Vectores asociados a los individuos que no han pedecido la enfermedad ($Y = 0$)	Vectores asociados a los individuos que sí padecieron la enfermedad ($Y = 1$)
$\mathbf{X}_{1,0}$	$\mathbf{X}_{1,1}$
$\mathbf{X}_{2,0}$	$\mathbf{X}_{2,1}$
$\mathbf{X}_{3,0}$	$\mathbf{X}_{3,1}$
$\vdots$	$\vdots$
$\mathbf{X}_{m,0}$	$\mathbf{X}_{n,1}$
n_0	n_1

Figura 1.4: Se separa la base de datos en dos conjuntos de acuerdo al valor de la variable respuesta. $\mathbf{X}_{i,j}$ es el vector de variables explicativas del sujeto i. El segundo subíndice j hace referencia al valor de la variable explicativa. n_0 es el total de individuos que no han padecido la enfermedad y n_1 es el total de individuos que presentaron la enfermedad.

Vector de variables	$Y=0$		$Y=1$		Cociente de probabilidades
	Frecuencia	$\widehat{\mathrm{p}}\,(\mathbf{x}_i \vert Y=0)$	Frecuencia	$\widehat{\mathrm{p}}\,(\mathbf{x}_i \vert Y=1)$	
$\mathbf{x}_1$	$f_{\mathbf{x}_{1,0}}$	$\dfrac{f_{\mathbf{x}_{1,0}}}{n_0}$	$f_{\mathbf{x}_{1,1}}$	$\dfrac{f_{\mathbf{x}_{1,1}}}{n_1}$	$c_1 = \dfrac{f_{\mathbf{x}_{1,0}}}{n_0} \div \dfrac{f_{\mathbf{x}_{1,1}}}{n_1}$
$\mathbf{x}_2$	$f_{\mathbf{x}_{2,0}}$	$\dfrac{f_{\mathbf{x}_{2,0}}}{n_0}$	$f_{\mathbf{x}_{2,1}}$	$\dfrac{f_{\mathbf{x}_{2,1}}}{n_1}$	$c_2 = \dfrac{f_{\mathbf{x}_{2,0}}}{n_0} \div \dfrac{f_{\mathbf{x}_{2,1}}}{n_1}$
$\mathbf{x}_3$	$f_{\mathbf{x}_{3,0}}$	$\dfrac{f_{\mathbf{x}_{3,0}}}{n_0}$	$f_{\mathbf{x}_{3,1}}$	$\dfrac{f_{\mathbf{x}_{3,1}}}{n_1}$	$c_3 = \dfrac{f_{\mathbf{x}_{3,0}}}{n_0} \div \dfrac{f_{\mathbf{x}_{3,1}}}{n_1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_k$	$f_{\mathbf{x}_{k,0}}$	$\dfrac{f_{\mathbf{x}_{k,0}}}{n_0}$	$f_{\mathbf{x}_{k,1}}$	$\dfrac{f_{\mathbf{x}_{k,1}}}{n_1}$	$c_k = \dfrac{f_{\mathbf{x}_{k,0}}}{n_0} \div \dfrac{f_{\mathbf{x}_{k,1}}}{n_1}$

Figura 1.5: Se realiza un conteo de la ocurrencia de cada vector explicativo y a partir de él se estima la probabilidad condicional de cada vector de variables explicativas.

a partir de éste se estima la probabilidad condicional de cada uno de estos vectores. Este proceso se ilustra en la figura 1.5 en donde se sigue la notación de que $\mathbf{x}_i$ representa una realización del vector de variables explicativas.

Una vez hecho este proceso se procede a obtener el cociente de probabilidades

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

y ordenarlo en forma ascendente sin perder la relación que cada cociente de probabilidades tiene con su vector de variables explicativas. Este proceso se ilustra en la figura 1.6.

La probabilidad del error tipo I se estima como

$$\widehat{\mathrm{p}}\,(A_{\overline{E}}|E) = \frac{w_1}{n_1}$$

donde $w_1 = \sum_{i=p+1}^{k} f_{\mathbf{x}_{c(i)}1}$ y $f_{\mathbf{x}_{c_{(i)}}1}$ representa la frecuencia de una combinación de variables $\mathbf{x}$ para un cociente $c(i)$ cuando $Y=1$.

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

Figura 1.6: Los cocientes de probabilidades condicionales de los vectores de variables explicativas se ordenan y a partir de ellos se estiman las probabilidades del error tipo I y tipo II. $f_{\mathbf{x}_{c_{(i)}}}$ representa la frecuencia relativa de una combinación de variables $\mathbf{x}$ para un cociente $c_{(i)}$.

Mientras que la probabilidad del error tipo II se estima como

$$\widehat{\mathrm{p}}\left(A_E|\overline{E}\right) = \frac{w_0}{n_0}$$

donde $w_0 = \sum_{i=1}^{p} f_{\mathbf{x}_{c_{(i)}}0}$ y $f_{\mathbf{x}_{c_{(i)}}0}$ representa la frecuencia de una combinación de variables $\mathbf{x}$ para un cociente $c(i)$ cuando $Y = 0$.

Entonces el valor crítico λ_α se determina como $\lambda_\alpha \geq c_{(p)}$. En este trabajo se ha propuesto que el valor crítico λ_α sea aquel en el cual la suma de los dos errores $\frac{w_0}{n_0} + \frac{w_1}{n_1}$ sea mínima. La propuesta surge luego de observar que con el método se hace mínima la probabilidad del valor del error tipo I pero la probabilidad del error tipo II puede permanecer elevada, entonces al considerar la mínima suma de ambos errores la idea es lograr tener un clasificador que minimice lo más posible ambos errores de manera simultánea.

Finalmente, el conjunto de vectores $\mathbf{x}_{c_{(i)}}$ tales que $c_{(i)} \leq c_{(p)}$ definen el conjunto de riesgo mientras que el conjunto de no riesgo lo forman los vectores $\mathbf{x}_{c_{(k)}}$ tales que $c_{(k)} > c_{(p)}$.

Para concluir el capítulo, a continuación se describe brevemente la teoría relacionada a los métodos de exclusión usados en este trabajo.

1.4. Selección de variables para los métodos de clasificación.

Siempre que se tenga una o más de una variables explicativas, no se podrá asegurar que las variables influyan en la variable de respuesta, por lo que se deben hacer pruebas para determinar cuáles variables se deben incluir y cuáles se deben excluir del modelo. En este trabajo se han seguido dos técnicas para seleccionar y excluir variables, la primera es una prueba de independencia y la segunda es mediante la devianza en el modelo de

		Variable X_i				
		Categoría 1 C_1	Categoría 2 C_2	...	Categoría m C_m	Total
Variable Respuesta: Y	Nivel 1: F_1	n_{11}	n_{12}	...	n_{1m}	$n_{1\bullet}$
	Nivel 2: F_2	n_{21}	n_{22}	...	n_{2m}	$n_{2\bullet}$
	Total	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet m}$	$n_{\bullet\bullet}$

Figura 1.7: Tabla de contingencia formada entre la variable respuesta y alguna variable explicativa.

regresión logística.

1.4.1. Selección de variables mediante una prueba de independencia.

Esta manera de seleccionar variables surge de la observación de que si la variable respuesta es independiente de alguna variable explicativa, entonces esta última no influye en la respuesta y por lo tanto puede ser excluida.

La prueba consiste en construir tablas de contingencia en donde los renglones o filas corresponden a la respuesta y las columnas a una variable explicativa, o viceversa, como se muestra en la figura 1.7.

Bajo la hipótesis de independencia se debe cumplir que $P(C_i \cap F_j) = P(C_i)\,P(F_j)$; en consecuencia las estimaciones de las dos partes de esta ecuación deben ser "cercanas" cuando la hipótesis de independencia se verifique. La manera en que se estiman cada una de las probabilidades involucradas en esta ecuación es:

$$\widehat{P}(C_i \cap F_j) = \frac{n_{ij}}{n_{\bullet\bullet}}$$
$$\widehat{P}(C_i) = \frac{n_{\bullet i}}{n_{\bullet\bullet}}$$
$$\widehat{P}(F_j) = \frac{n_{j\bullet}}{n_{\bullet\bullet}}$$

Al valor n_{ij} se le llama valor observado, O_{ij}, mientras que bajo la hipótesis de independencia se tiene otra estimación del mismo valor como

$$T\left[\widehat{P}\left(C_i\right)\widehat{P}\left(F_j\right)\right] = \frac{n_{\bullet i}n_{j\bullet}}{n_{\bullet\bullet}}$$

al cual se le suele llamarsele valor esperado, E_{ij}. En este caso, el estadístico $\chi_c^2 = \sum_{i,j} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$ tiene aproximadamente una distribución ji-cuadrada con $(2-1)\times(m-1)$ grados de libertad cuando la hipótesis de independencia es razonable, con lo cual se puede establecer la siguiente prueba de hipótesis:

H_0: X_i y Y son independientes.
H_1: X_i y Y no son independientes.

Y el criterio de decisión sobre qué hipótesis es cierta es:

- Si $\chi_c^2 > \chi_\alpha^2$ se rechaza la hipótesis de independencia y en consecuencia la variable explicativa se incluye en el modelo.
- Si $\chi_c^2 \leq \chi_\alpha^2$ no se rechaza la hipótesis de independencia y en consecuentemente la variable X_i se excluye del modelo.

1.4.2. Exclusión mediante la devianza en el modelo de regresión logística.

En esta sección se describirá brevemente el concepto de devianza para el caso particular de la regresión logística y posteriormente se comentará su uso para realizar pruebas de hipótesis sobre qué variables deben incluirse en el modelo logit; es decir, la devianza es un término que permite distinguir si una o más variables explicativas son influyentes en la variable de interés.

Suponga que se tienen $y_1, y_2, \ldots, y_N$ observaciones de la variable de interés y que para cada una de ellas existe un vector de variables explicativas $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ir})$. Las variables y_i y $\mathbf{x}_i$ están relacionadas mediante el modelo de regresión logística dado por $\mu_i = E(y_i) = \dfrac{1}{1 + \exp(-\mathbf{x}_i{}^T \beta)}$.

Se mencionó en la sección 1.2.1 que el logaritmo natural de la función de verosimilitud para el caso de un modelo Bernulli es $l(\pi, y) = \sum_{i=1}^{N} [y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)]$. Esta función tiene su máximo cuando $\pi_i = y_i$, es decir cuando

$$l(\pi, y) = \sum_{i=1}^{N} [y_i \log y_i + (1 - y_i) \log(1 - y_i)]$$

Mientras que si se aplica la función de verosimilitud a los datos estimados con el modelo $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_r x_r$, la función de verosimilitud tiene su máximo en $\widehat{\mu} = \dfrac{1}{1 + \exp\left(-\mathbf{x}_i{}^T \widehat{\beta}\right)}$ lo que implica que:

$$l(\beta_r, y) = \sum_{i=1}^{N} [y_i \log \widehat{\mu}_i + (1 - y_i) \log(1 - \widehat{\mu}_i)]$$

Entonces, la devianza del modelo $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_r x_r$ se define como:

$$\begin{aligned} D &= 2[l(\pi, y) - l(\beta_r, y)] \\ &= 2 \sum_{i=1}^{N} [y_i \log y_i + (1 - y_i) \log(1 - y_i) - y_i \log \widehat{\mu}_i - (1 - y_i) \log(1 - \widehat{\mu}_i)] \\ &= 2 \sum_{i=1}^{N} [y_i (\log y_i - \log \widehat{\mu}_i) + (1 - y_i) \{\log(1 - y_i) - \log(1 - \widehat{\mu}_i)\}] \\ &= 2 \sum_{i=1}^{N} \left[y_i \log \left(\frac{y_i}{\widehat{\mu}_i} \right) + (1 - y_i) \log \left(\frac{1 - y_i}{1 - \widehat{\mu}_i} \right) \right] \end{aligned}$$

Obsérvese que la devianza satisface que mientras mejor se ajuste el modelo, su función de verosimilitud $l(\beta_r, y)$ será mayor y en consecuencia su devianza será menor (Myers et al., 2002).

Además, según (Dobson, 2002) la devianza tiene la importante propiedad de que aproximadamente tiene una distribución ji cuadrada no centralizada con $N - r$ grados de libertad. Este hecho permite ejecutar la prueba de hipótesis sobre la influencia de subconjuntos de parámetros en un modelo de regresión logística.

Prueba de hipótesis para la inclusión de una variable usando la devianza.

Considérese un modelo lineal generalizado que se ha ajustado con r variables explicativas y el mismo modelo lineal generalizado con una variable más. La devianza del primer modelo sería $D_r = 2\left[l\left(\pi, y\right) - l\left(\beta_r, y\right)\right]$, mientras que la devianza del segundo modelo sería $D_{r+1} = 2\left[l\left(\pi, y\right) - l\left(\beta_{r+1}, y\right)\right]$. Dado que el modelo con una variable más tiene un parámetro adicional, su función de verosimilitud ajustará mejor y será mayor a la función de verosimilitud del modelo con una variable menos y como consecuencia la devianza del modelo con más variables será menor a la devianza del modelo con menos variables.

La diferencia de devianzas entre estos dos modelos es:

$$\begin{aligned} \Delta D &= D_r - D_{r+1} \\ &= 2\left[l\left(\pi, y\right) - l\left(\beta_r, y\right)\right] - 2\left[l\left(\pi, y\right) - l\left(\beta_{r+1}, y\right)\right] \\ &= 2\left[l\left(\beta_{r+1}, y\right) - l\left(\beta_r, y\right)\right] \end{aligned}$$

Y tiene la importante propiedad de tener aproximadamente una distribución ji cuadrada con un grado de libertad. Observe que en el planteamiento de la prueba

$$H_0 : \beta_{r+1} = 0 \qquad vs. \qquad H_1 : \beta_{r+1} \neq 0$$

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 la devianza $\Delta D = D_r - D_{r+1}$, que se espera sea pequeña si 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 $\Delta D \geq \chi^2_{\alpha,1}$ entonces se rechaza la hipótesis nula y se incluye la variable X_{r+1}.
- Si $\Delta D < \chi^2_{\alpha,1}$ entonces se acepta la hipótesis nula y se excluye la variable X_{r+1}.

Capítulo 2

Aplicación al problema de trombósis

En este capítulo se dará una pequeña descripción de la enfermedad de la trombosis y se describirá la manera en que se implementaron los métodos de clasificación en R. Se comenzará con la descripción de la enfermedad de la trombosis.

2.1. Conceptos básicos relacionados a la enfermedad de trombosis.

La sangre es un líquido que circula a través del cuerpo humano a través de venas y arterias cuya apariencia es viscosa y de color rojo; es un vehículo que transporta gran cantidad de sustancias entre órganos y tejidos (Higashida, 2008).

La sangre está formada por un líquido llamado plasma y por células que se pueden agrupar en tres conjuntos principales a saber: los eritrocitos o glóbulos rojos, los leucocitos o glóbulos blancos y los trombocitos o plaquetas. De manera muy general, las funciones de estos tres conjuntos de células sanguíneas son las siguientes: los glóbulos rojos sirven para transportar el oxígeno por medio de la hemoglobina, los glóbulos blancos sirven como defensa para combatir las infecciones y las plaquetas ayudan a la formación del coagulo cuando se rompen o lesionan vasos sanguíneos (Higashida, 2008).

Al conjunto de mecanismos para detener los procesos hemorrágicos se le llama hemostasia e involucra los procesos de coagulación sanguínea y la contracción de los vasos sanguíneos dañados (Martin, 2002).

La coagulación es el proceso por el cual la sangre pasa de estado líquido a sólido. Este proceso puede ser iniciado por el contacto de la sangre con una superficie extraña o por un tejido dañado. Una vez iniciado el proceso de coagulación se da la interacción de una serie de sustancias que dan lugar a una enzima llamada trombina, la cual convierte a la proteína fibrinógeno, que es soluble en la sangre, en la proteína fibrina, que no es soluble en la sangre (Martin, 2002).

Una manifestación anormal de la hemostasia es la trombosis que es la principal fuente de morbilidad y mortalidad en pacientes hospitalizados (Higashida, 2008). La trombosis es el proceso patológico por el cual las plaquetas y la fibrina interactúan con la pared vascular para formar un tapón que causa obstrucción vascular (Mehta and Hoffbrand, 2013).

La trombosis se produce cuando el sistema hemostásico se activa de manera inadecuada al grado que los procesos naturales anticoagulantes se superan y se permite la formación de un coagulo dentro de un vaso sanguíneo. Al crecer la masa o trombo, éste puede ocluir el vaso sanguíneo y producir la muerte de tejidos abastecidos normalmente por ese vaso. Además puede romperse una porción de un trombo, llamada émbolo y desplazarse a ramas vasculares más pequeñas obstruyéndolas causando subsecuentemente la destrucción del tejido. A esta obstrucción se le llama embolia o tromboembolia (Shirlyn and Mckenzie, 2000).

Se puede mencionar dos tipos de trombos principalmente, los trombos arteriales y los trombos venosos. Enseguida se da una explicación breve de cada uno de ellos.

Los trombos arteriales se producen en vasos sanguíneos en los cuales el flujo es rápido; generalmente ocurren en las arterias aunque pueden suceder en las venas. Estos trombos son de color blanco debido a su constitución principalmente de plaquetas y fibrina, entre las cuales quedan atrapados unos cuantos leucocitos y eritrocitos.

Dentro de los factores que aumentan la probabilidad de producción de trombos arteriales se encuentran dietas ricas en colesterol, el tabaquismo, el uso de anticonceptivos orales y la presencia de varias enfermedades entre las que se encuentran ateroesclerosis, hipertensión, diabetes y el síndrome nefrótico (Shirlyn and Mckenzie, 2000).

Por otra parte, la trombosis venosa se produce en vasos sanguíneos en los cuales el flujo de sangre es lento, lo cual ocurre por lo general en las venas. A diferencia del trombo arterial, el trombo venoso es de color rojo debido a que es más fácil que los eritrocitos queden atrapados entre la red de plaquetas y fibrina cuando el flujo sanguíneo es lento (Shirlyn and Mckenzie, 2000).

La trombosis venosa sintomática clínicamente significativa se denomina trombosis venosa profunda y por lo común se encuentra en las venas proximales de los miembros inferiores. Se aumenta el riesgo de padecer este tipo de trombosis en el caso de conductas, enfermedades y estados que causan estancamiento como por ejemplo, estar sentados por periodos largos de tiempo o estar inmovilizado por convalecer de una cirugía; de ahí se puede explicar que, como se mencionó con anterioridad, la trombosis sea la principal fuente de morbilidad y mortandad en pacientes hospitalizados. Otros factores que se han podido relacionar a la trombosis venosa profunda son la cirugía de cadera y del área pélvica, así como la insuficiencia cardiaca congestiva y el cáncer. También la obesidad y las deficiencias de inhibidores naturales de la coagulación también se acompañan con un aumento de la trombosis venosa profunda (Shirlyn and Mckenzie, 2000).

Dentro las complicaciones más graves de las trombosis venosas profundas se encuentran las embolias pulmonares, que se producen cuando los émbolos procedentes de las venas profundas en los miembros inferiores se desplazan a través de las venas más grandes hasta el lado derecho del corazón y finalmente alcanzan la circulación pulmonar y ocluyen los vasos pulmonares. Estas embolias pueden ser asintomáticas o mortales; entre 0.1 a 0.8 % de los pacientes sometidos a cirugía general y hasta 5 % de aquellos con cirugía de cadera mueren de embolia pulmonar cada año y aproximadamente 50 % de los pacientes con trombosis venosa profunda documentada presenta émbolos (Shirlyn and Mckenzie, 2000).

Finalmente, además de los factores que ya se han mencionado y que predisponen a la

trombosis, existen otro tipo de factores que son hereditarios y que también predisponen a este padecimiento. De entre ellos se pueden mencionar los siguientes: deficiencias de antitrombina III, de proteína C, de proteína S, en plasminógeno, así como disfibrinogenemia, homocistinuria y hemangioma cavernoso gigante (Shirlyn and Mckenzie, 2000).

2.2. Descripción de los datos empleados y de su tratamiento.

Los datos con los cuales se ha trabajado fueron proporcionados por el laboratorio de hematología perinatal del Instituto Nacional de Perinatología Isidro Espinosa de los Reyes.

La base de datos está escrita en Excel y cuenta con las observaciones de 960 pacientes con 25 variables explicativas. En ella se observa la presencia de datos faltantes en las cantidades mostradas en la figura 2.1, donde también se muestra el nombre de cada una de las variables.

En la figura 2.2 se muestra un esquema del manejo que se hizo sobre la información de la base de datos para finalmente obtener la clasificación de un paciente con alguno de los dos métodos.

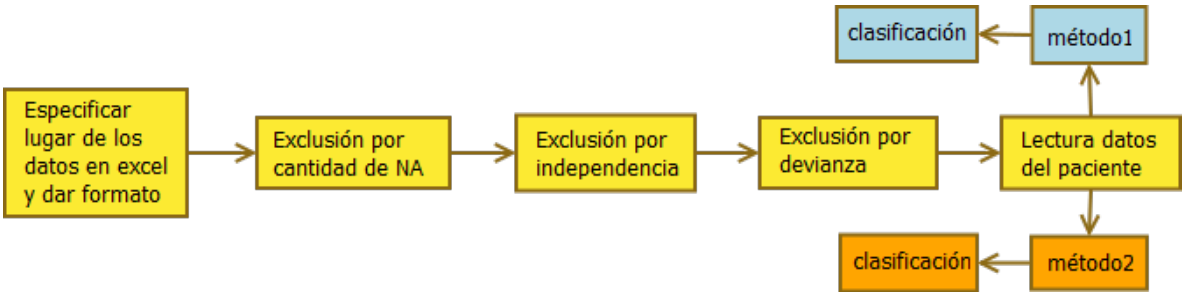


Figura 2.2: Diagrama de flujo del tratamiento que se lleva a cabo para implementar los métodos de clasificación.

A continuación se procederá a dar algunos detalles y pormenores de la implementación

Variable	PGR: *Pérdida gestacional recurrente*	PLAQ: *Plaquetas*	FVW: *Factor de Von Willebrand*	DIMEROS D	FS: *Ferretina sérica*
Número de datos faltantes	0	178	248	112	8
Variable	ACAIGM: *Anticuerpos anticadiolipina IgM*	ACAIGG: *Anticuerpos anti DNA de cadena sencilla*	ssDNA: *Anticuerpos anti SSA*	dsDNA: *Anticuerpos anti DNA de cadena doble*	RnP: *Anticuerpos anti Ribonucleoproteina*
Número de datos faltantes	0	1	2	2	4
Variable	Sm: *Anticuerpos anti Smith*	SSA: *Anticuerpos anti SSA*	SSB: *Anticuerpos anti SSB*	HIST: *Anticuerpos anti histonas*	scl70: *Anticuerpos anti scl70*
Número de datos faltantes	3	3	2	3	6
Variable	B2GP1IGM: *Anticuerpos anti beta 2*	B2GP1IGG: *Anticuerpos anti beta 2*	RPCAR 1:5: *Resistencia a la proteína C*	TTPA: *Tiempo de Tromboplastina*	TTPAMEZCLA
Número de datos faltantes	7	11	149	96	826
Variable	LAR: *Razón de anticoagulante lúpico*	FVL: G1691A: *Factor V Leiden*	PT G20210A: *Mutación de protrombina*	MTHFR C677T: *Mutación metilen tetrahidrofolato reductasa 677*	MTHFR A1298C: *Mutación metilen tetrahidrofolato reductasa 1298*
Número de datos faltantes	97	720	720	717	721

Figura 2.1: Cantidad de datos faltantes por cada variable.

de cada proceso y método de clasificación.

2.3. Implementación de los métodos de clasificación.

Para aplicar los métodos de clasificación a través de regresión logística y por el método de probabilidades, se ha empleado el software R versión 3.0.1. A continuación se detallan las ideas implicadas en cada una de ellas.

2.3.1. Preparación de la base de datos.

Debido a que la base de datos disponible se encuentra en formato .xlsx, ha sido necesario trasladarla al ambiente de R; para ello se ha cambiado al formato *.csv*. Sin embargo, al emplearlo, la base de datos resultante presenta dos problemas principales: 1) incluye muchas columnas sin nombre o vacías y 2) el formato de los datos se entiende de manera general como carácter, es decir, si se tiene el número 5.32, este se entiende como la palabra "5.32".

El primer problema se resuelve fácilmente quitando las columnas vacías; además se pide que la base de datos contenga sólo la información relevante, es decir que no contenga los nombres de los pacientes y los datos que se toman de manera rutinaria como números telefónicos o correos electrónicos.

Para solucionar la segunda problemática se han distinguido dos casos a saber: el primero es que la variable sea una variable numérica y su formato al aplicar el cambio de formato queda como numérica, en este caso no debe realizarse ninguna acción; el segundo caso ocurre cuando la variable es numérica pero queda como una palabra, en este caso hay que transformarla a su valor correspondiente.

El resultado final es una base de datos con las mismas características que la orginal completamente numérica. El procedimiento se resume en la figura 2.3.

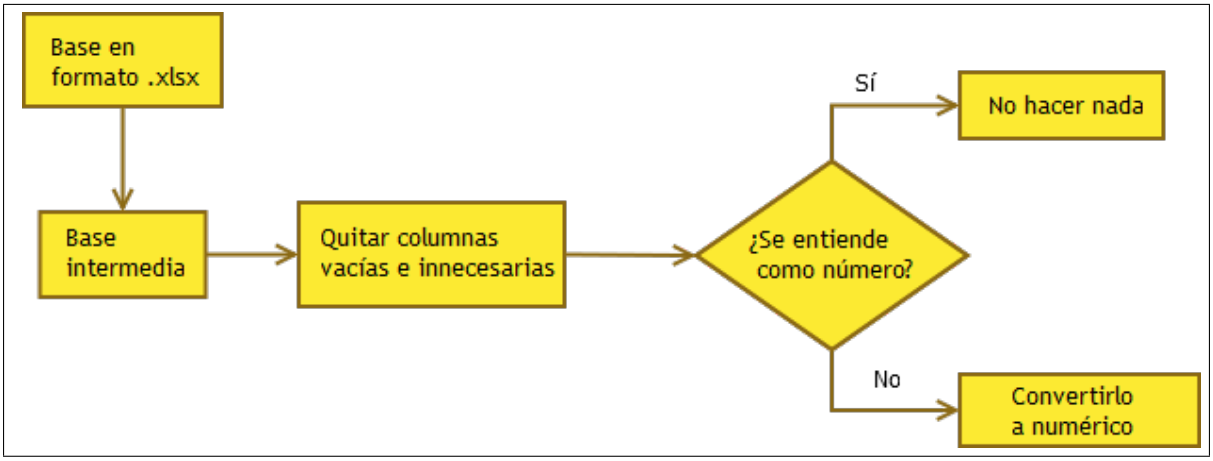


Figura 2.3: Procedimiento llevado a cabo para preparar la base de datos.

2.3.2. Exclusión de variables por cantidad de datos faltantes y a través de una prueba de independencia.

La base de datos con la que se cuenta tiene 25 variables explicativas y la primera tarea es depurarla para que sólo queden las variables que son significativas respecto a 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 y una mejor interpretación de los resultados.

Dentro de las opciones desarrolladas para excluir variables se siguieron dos. La primera y más sencilla fue excluir las variables que tienen más del 50 % de datos faltantes.

La segunda opción de exclusión de variables fue mediante una prueba de independencia entre cada variable explicativa con respecto a la variable respuesta basada en la distribución ji cuadrada. Si la hipótesis de independencia entre la variable respuesta y la variable explicativa en cuestión se sostiene, entonces dicha variable se descarta bajo el supuesto de que la variable respuesta tiene un comportamiento independiente a esta variable; en caso de que la hipótesis de independencia no se sostenga, entonces la variable

no se descarta bajo la suposición de que la variable respuesta sí depende de dicha variable.

Para implementar la prueba de independencia primero se categorizaron los datos de la variable en cuestión con base en su mediana y después se formaron dos conjuntos con base en la variable respuesta para finalmente realizar una tabla de contingencia como la mostrada en la figura 2.4.

Variable X_i	$Y = 0$	$Y = 1$
$x_i \leq Me\,(X_i)$	$f_{x_i \leq Me(X_i),0}$	$f_{x_i \leq Me(X_i),1}$
$x_i > Me\,(X_i)$	$f_{x_i > Me(X_i),0}$	$f_{x_i > Me(X_i),1}$

Figura 2.4: Tabla de contingencia formada para cada variable explicativa X_i. La mediana de la variable X_i se denota como $Me(X_i)$.

Finalmente, a la tabla de contingencia se le aplicó el comando *chisq.test* de R y el resultado de la prueba se encontraba a través de la comparación del valor p de esta prueba con respecto a un nivel de significancia preestablecido.

Para determinar el valor de significancia adecuado se hizo una simulación en la que se corrió mil veces sobre una base de datos con quince variables generadas aleatoriamente en el intervalo [1,5] el proceso siguiente:

1. Se eligieron aleatoriamente 5 variables explicativas distintas y se generó con ella la variable respuesta con la fórmula:

$$Y^* = \frac{1}{1 + \exp\left[-\left(c_1 X_a + c_2 X_b + c_3 X_c + c_4 X_d + c_5 X_e\right)\right]}, Y = \begin{cases} 0 & \text{si} \quad \text{Y}^* < 0.5 \\ 1 & \text{si} \quad \text{Y}^* \geq 0.5 \end{cases}$$

   Donde las c_i se eligieron aleatoriamente en el conjunto $c_i = \{c_i \in \mathbb{Z} | -9 \leq c_i \leq 9\}$. Además, los datos seleccionados fueron tales que se forzó a que el número de unos obtenidos estuviera entre el 20 % y 80 %.

2. Se corrió el proceso de exclusión de variables y se contabilizó el número de veces que las variables que generaron el modelo fueron seleccionadas. Los resultados obtenidos se muestran en la figura 2.5.

Número de variables correctas detectadas	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$
Cinco	223	247	211	195	170
Cuatro	227	240	250	223	240
Tres	271	268	259	284	276
Dos	196	181	198	207	208
Una	72	56	75	81	86
Cero	11	8	7	10	20

Figura 2.5: En la tabla se observa que las 5 variables fueron incluidas en más ocasiones cuando el nivel de significancia fue de 0.4.

Dado los resultados obtenidos con la simulación, se consideró para las pruebas de independencia una significancia de 0.4.

Cabe mencionar que como alternativa no paramétrica a la eliminación de variables a través de una prueba de independencia se consideró una prueba de rangos de Mann-Whitney, sin embargo los resultados obtenidos de ella fueron significativamente inferiores. Los resultados obtenidos con esta prueba se muestran en la figura 2.6.

Número de variables correctas detectadas	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$
Cinco	3	1	2	0	2
Cuatro	24	37	43	17	26
Tres	136	149	129	110	128
Dos	265	268	265	276	283
Una	282	287	248	303	277
Cero	290	258	313	294	284

Figura 2.6: Resultados de la simulación para excluir variables con la prueba de Mann-Whitney.

Es importante señalar que la significancia seleccionada con este método es muy grande;

esto significa que es probable que algunas variables que son independientes se incluyan en el modelo, sin embargo se utilizará la devianza con el modelo de regresión logística para refinar la selección de variables. En esta primera prueba la meta es quitar variables aun con una alta probabilidad de que queden las variables que influencian en el modelo.

Cabe mencionar que se hizo una prueba utilizando únicamente la devianza para la selección de variables, los resultados son semejantes a los que se obtuvieron usando la prueba de independencia pero más tardados computacionalmente.

2.3.3. Selección de variables a través de la devianza del modelo de regresión logística.

Luego de excluir las variables que se manifiestan como independientes se procede a excluir a otras más usando la devianza de modelos de regresión logística anidados. Para ello se hace una introducción progresiva de variables explicativas en la regresión logística, comenzando con una sola variable explicativa y terminando cuando la devianza no excede al nivel de significancia establecido o bien cuando todas las variables son incluidas. Este proceso se ilustra en la figura 2.7 y se puede resumir en los siguientes pasos:

1. Se realizan todas las regresiones logísticas de Y con cada una de las variables explicativas; se calculan las devianzas en cada caso y se selecciona la variable explicativa que proporciona la menor devianza. Sea W_1 la variable seleccionada; se hace $n = 2$.

2. Se realizan todas las regresiones logísticas de Y con las variables seleccionadas W_i, $i = 1, 2, \ldots, n-1$, y una variable explicativa más de las que no han sido seleccionadas. Se calcula la devianza en cada caso y se selecciona la variable adicional del modelo con menor devianza, esta variable se denota como W_n.

3. Se realiza la diferencia entre la devianza del modelo con W_n y W_{n-1}; si esta diferencia no es mayor al nivel de significancia, entonces el proceso se detiene; en caso contrario se hace $n = n + 1$ y se vuelve al paso dos.

Luego de completar este procedimiento se procede a recuperar los índices j de las variables

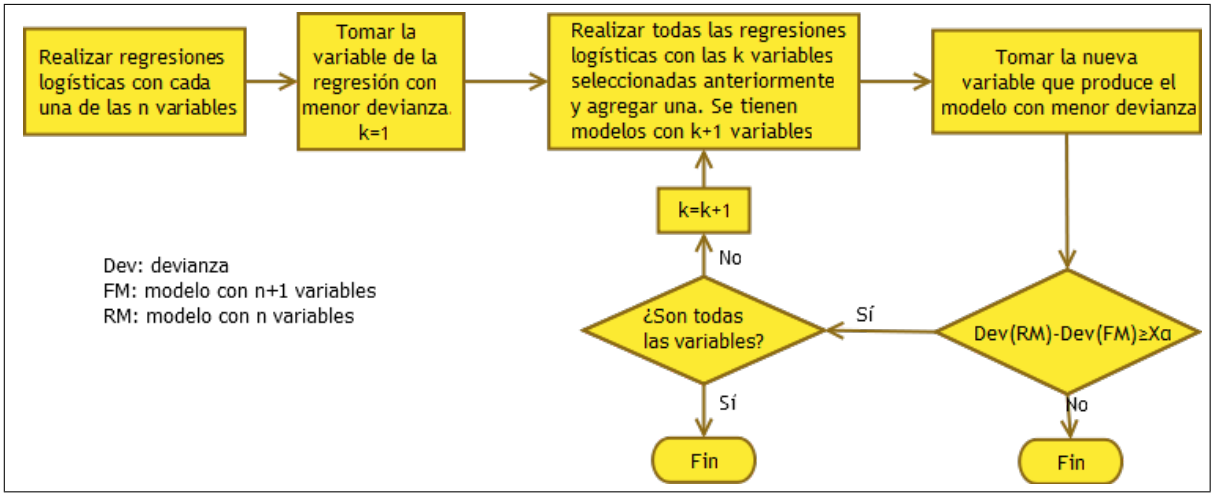


Figura 2.7: Procedimiento para implementar la clasificación a través de la regresión logística.

$W_i = X_j$ para toda $i = 1, \ldots, n$ a considerar y sus coeficientes para efectos de uso en las subrutinas. Una respuesta típica de este proceso se muestra en la figura 2.8.

2.3.4. Proceso de imputación.

Debido a que la base de datos con la que se ha trabajado presenta datos faltantes se ha procedido a realizar un proceso de imputación y por cuestiones de simplicidad se ha elegido imputación simple a través de regresión lineal.

	indices2	coeficientes
Intercept	0	-2.3940932368
ACAIGM	6	-0.0077306028
dsDNA	9	-0.0005541177
SSB	13	0.0038090761
ssDNA	8	0.0012850828
FS	5	-0.0004646587
B2GP1IGM	16	0.0043323640
Sm	11	-0.0124843292
HIST	14	-0.0042686934
SSA	12	-0.0053441052

Figura 2.8: Respuesta típica producida por el buscador a través de la devianza por regresión logística.

Se analizaron dos métodos de imputación, en ambos previamente se ordenaron los datos con base en el número de datos faltantes, en cada caso se imputa primero la variable con menos datos con lo cual se logra una nueva variable "completamente observada". Luego el proceso se realiza sucesivamente hasta "rellenar" la base de datos.

Para "rellenar" los datos faltantes usando el primer método de imputación se utiliza como variable dependiente la variable que se va a imputar y como variables explicativas las variables que están completamente llenas, se estima la variable dependiente con un modelo de regresión lineal y se imputan los valores faltantes con el modelo estimado.

Para "rellenar" los datos faltantes usando el segundo método de imputación se utiliza también como variable dependiente la variable a imputar y como variables independientes todas las demás considerando sólo los datos de los individuos que tienen todos los datos llenos. Se estima la variable dependiente con un modelo de regresión lineal y se imputan los valores faltantes con el modelo estimado.

Para decidir sobre qué método de imputación elegir, se hizo nuevamente un estudio de simulación en el cual se repitió cada proceso de imputación en mil ocasiones tomando como referencia la base de datos *mtcars* que el paquete R incluye y considerando como variable respuesta a la variable vs. Un fragmento de esta base se muestra en la figura 2.9.

	vs	mpg	cyl	disp	hp	drat	wt	qsec	am	gear	carb
Mazda RX4	0	21.0	6	160	110	3.90	2.620	16.46	1	4	4
Mazda RX4 Wag	0	21.0	6	160	110	3.90	2.875	17.02	1	4	4
Datsun 710	1	22.8	4	108	93	3.85	2.320	18.61	1	4	1
Hornet 4 Drive	1	21.4	6	258	110	3.08	3.215	19.44	0	3	1

Figura 2.9: Fragmento de la base de datos empleada para decidir sobre el método de imputación.

En cada iteración, a la base de datos se le insertó aleatoriamente datos faltantes. Un ejemplo de una de las tablas creadas en alguna iteración se muestra en la figura 2.10.

También, en cada iteración se tomó la diferencia en valor absoluto entre cada dato imputado y su correspondiente valor real, almacenándose la diferencia máxima para cada

	vs	mpg	cyl	disp	hp	drat	wt	qsec	am	gear	carb
Mazda RX4	0	21.0	6	NA	110	NA	2.620	16.46	1	NA	4
Mazda RX4 Wag	0	21.0	6	160	110	3.90	2.875	17.02	1	4	4
Datsun 710	1	22.8	4	108	93	3.85	2.320	18.61	1	4	1
Hornet 4 Drive	1	21.4	6	258	110	3.08	NA	19.44	NA	3	1

Figura 2.10: Fragmento de la base de datos empleada para decidir sobre el método de imputación con datos faltantes insertados aleatoriamente.

variable. Luego se contó el número de variables para las cuales el método 1 estuvo más cercano al valor real. Los resultados obtenidos se resumen en la figura 2.11.

Número de veces en que la diferencia máxima del método 1 estuvo más cercana al valor real	0	1	2	3	4	5	6	7	8	9	10
Frecuencia	0	0	0	0	3	9	34	76	202	358	318

Figura 2.11: Resultado obtenido luego de hacer cada opción de imputación en mil ocasiones y comparar los resultados.

Con base en los resultados obtenidos, se eligió la opción de imputación 1.

2.3.5. Implementación del método de clasificación mediante regresión logística.

Una vez que se han seleccionado las variables explicativas significativas y que se han imputado los valores faltantes a las mismas, se procede a estimar la función clasificadora para poder realizar la clasificación de un paciente luego de que sus datos se han leído.

Para construir el clasificador mediante la regresión logística se utilizaron los datos como estaban en la base. La estimación de la función clasificadora se realiza con el procedimiento descrito en la sección 1.2.3.

2.3.6. Implementación del mejor clasificador de nivel alfa.

Para estimar el mejor clasificador de nivel alfa, una vez que la base está lista, es necesario categorizar las variables explicativas, preferentemente en pocas clases para que el número de posibles combinaciones sea manejable; por simplicidad en este trabajo se codificaron todas las variables en dos categorías. Para conseguir esto se construyó una rutina con la cual al proporcionarle los puntos de corte para cada una de las variables explicativas transforma la variable en categorías que van de 0 a PC_i, donde PC_i es el número de costes que tiene la variable i.

Para ejemplificar este proceso, supónganse tres variables explicativas X_1, X_2 y X_3, la primera con un punto de corte denotado por C_{1X_1} (este punto de corte origina dos categorías, la 0 y la 1); la segunda con dos puntos de corte denotados por C_{1X_2} y C_{2X_2} (que generan tres categorías, la 0, la 1 y la 2), y la tercera con tres puntos de corte denotados por C_{1X_3}, C_{2X_3} y C_{3X_3} (los cuales generan cuatro categorías, la 0, la 1, la 2 y la 3). De esta manera un vector con valores observados originales $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$ tales que $C_{1X_1} < x_{i1}$, $C_{1X_2} < x_{i2} \leq C_{2X_2}$ y $x_{i3} \leq C_{1X_3}$, se convierte en un vector $\mathbf{x}^*$ con valores categóricos de la forma $\mathbf{x}^*{}_i = (1, 1, 0)$; ver figura 2.12.

Una vez que se han categorizado todos los elementos de la base de datos el resultado es una matriz de enteros $\mathbf{X}^*$ cuyas filas $\mathbf{x}_i^*$ representan la categorización de cada observación. Si bien, a partir de los puntos categorizados $\mathbf{x_i^*}$ ya se puede aplicar el método de la mejor clasificación, se pueden simplificar los cálculos computacionales y su almacenamiento con una asignación inyectiva de los vectores $\mathbf{x_i^*}$ a los números naturales de la siguiente manera:

1. Se define el vector $\mathbf{PC} = (PC_1, PC_2, \ldots, PC_n)$ donde PC_i indica el número de puntos de corte que tiene la variable X_i, $i = 1, \ldots, n$.

2. A partir del vector $\mathbf{PC}$ se construye el vector

$$\mathbf{K} = \left(\prod_{i=2}^{n} [PC_i + 1], \prod_{i=3}^{n} [PC_i + 1], \ldots, \prod_{i=n}^{n} [PC_i + 1], 1 \right)$$

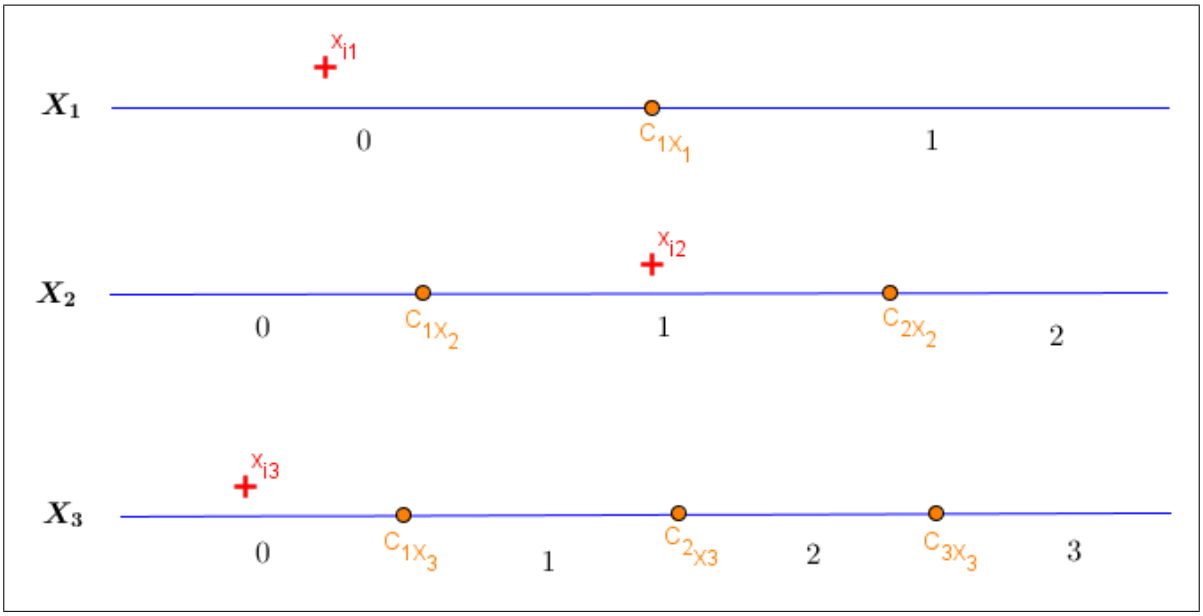


Figura 2.12: En esta figura se muestra el proceso de categorización para un vector $\mathbf{x_i}$. Sus coordenadas se muestran con una cruz y se comparan con los puntos de corte mostrados con un punto; dependiendo de la localización de las coordenadas de $\mathbf{x_i}$ se hará la categorización.

3. Finalmente la codificación se obtiene mediante la operación $\langle \mathbf{K}, \mathbf{x}_i^* \rangle$, donde $\mathbf{x}_i^*$ es una fila de la matriz $\mathbf{X}^*$ y $\langle , \rangle$ representa el producto punto ordinario.

Como ejemplo, considérese nuevamente al elemento categorizado $\mathbf{x}^*{}_i = (1, 1, 0)$, el correspondiente vector $\mathbf{PC} = (1, 2, 3)$, con el que se obtiene

$$\mathbf{K} = ([2+1]\,[3+1]\,, 3+1, 1) = (3 \times 4, 4, 1) = (12, 4, 1)$$

De esta manera se tiene que la codificación del elemento sería $(12, 4, 1) \cdot (1, 1, 0) = 16$. Se puede observar el proceso de la transformación

$$x_i = (x_{i1}, x_{i2}, x_{i3}) \to x_i^* = (x_{i1}*, x_{i2}*, x_{i3}*) \to \langle K, x_i^* \rangle$$

Para ejemplificar que la función $x_i^* \to k_i$ es inyectiva se escriben todos los puntos imagen para el ejemplo donde $\mathbf{K} = (12, 4, 1)$ en la figura 2.13. En ella debe observarse que las posiciones de interés son aquellas en las cuales la norma del vector vale 1, es decir cuando todos los elementos son cero excepto uno que vale 1. También puede observarse que precisamente para el último dígito esta combinación siempre ocurre en la posición 1, mientras que para el penúltimo dígito esta combinación se presenta en la posición PC_n+1; de manera análoga puede apreciarse que la combinación de ceros con un uno en la primera posición se presenta justamente en la posición $(PC_n + 1)\,(PC_{n-1} + 1)$.

Clasificación de un elemento con el mejor clasificador de nivel alfa.

Una vez que se han categorizado y codificado las variables, se procede a aplicar el mejor clasificador de nivel alfa como se indica en la sección 1.3.1. Debemos notar que la cardinalidad del conjunto de posibles vectores codificados es igual a $\prod_{i=1}^{n} [PC_i + 1]$, pero en la base de datos sólo aparecen por lo general unas cuantas combinaciones, por ejemplo, si se tienen 5 variables, la primera con 4 categorías, la segunda con 5 categorías, la tercera con 5, la cuarta con 6 y la quinta con 3 categorías, el total de combinaciones posibles es $4 \times 5 \times 5 \times 6 \times 3 = 1800$; suponga que en la base están los datos de 200 individuos, entonces el número de combinaciones $\mathbf{x}^*$ en la base es menor o igual a 200 y sólo de estas

Categorización $\mathbf{X}_i^*$	$\langle \mathbf{K}, \mathbf{X^*}_i \rangle$	Categorización $\mathbf{X}_i^*$	$\langle \mathbf{K}, \mathbf{X^*}_i \rangle$
000	$12\,(0) + 4\,(0) + 1\,(0) = 0$	100	$12\,(1) + 4\,(0) + 1\,(0) = 12$
001	$12\,(0) + 4\,(0) + 1\,(1) = 1$	101	$12\,(1) + 4\,(0) + 1\,(1) = 13$
002	$12\,(0) + 4\,(0) + 1\,(2) = 2$	102	$12\,(1) + 4\,(0) + 1\,(2) = 14$
003	$12\,(0) + 4\,(0) + 1\,(3) = 3$	103	$12\,(1) + 4\,(0) + 1\,(3) = 15$
010	$12\,(0) + 4\,(1) + 1\,(0) = 4$	110	$12\,(1) + 4\,(1) + 1\,(0) = 16$
011	$12\,(0) + 4\,(1) + 1\,(1) = 5$	111	$12\,(1) + 4\,(1) + 1\,(1) = 17$
012	$12\,(0) + 4\,(1) + 1\,(2) = 6$	112	$12\,(1) + 4\,(1) + 1\,(2) = 18$
013	$12\,(0) + 4\,(1) + 1\,(3) = 7$	113	$12\,(1) + 4\,(1) + 1\,(3) = 19$
020	$12\,(0) + 4\,(2) + 1\,(0) = 8$	120	$12\,(1) + 4\,(2) + 1\,(0) = 20$
021	$12\,(0) + 4\,(2) + 1\,(1) = 9$	121	$12\,(1) + 4\,(2) + 1\,(1) = 21$
022	$12\,(0) + 4\,(2) + 1\,(2) = 10$	122	$12\,(1) + 4\,(2) + 1\,(2) = 22$
023	$12\,(0) + 4\,(2) + 1\,(3) = 11$	120	$12\,(1) + 4\,(2) + 1\,(3) = 23$

Figura 2.13: Al aplicar el vector $\mathbf{K}$ se genera un sistema de numeración en donde cada dígito tiene distinta base.

posibles combinaciones tenemos información, por lo tanto nuestro clasificador sólo podrá considerar estos vectores. Cuando la base es grande, la ausencia de algunos vectores $\mathbf{x}^*$ puede ser debido a que la probabilidad de tener individuos asociados a esos vectores es cero o casi cero, por lo que no afecta al método de clasificación. Así que sólo se consideran los vectores codificados que aparecen en la base de datos, y la probabilidad estimada se calcula con la frecuencia relativa de los pacientes en la misma base de datos. El proceso define un punto crítico $c_{(p)}$ y un conjunto de vectores o combinaciones de variables que se clasifican en un estado de riesgo a padecer la enfermedad, al cual denotaremos como R, así como un conjunto de vectores que se clasifican como un estado de bajo riesgo a padecer la enfermedad al cual denotaremos por NR.

$$R = \left\{ x^* \in \text{base de datos} \left| \frac{P\left(x* | 0\right)}{P\left(x* | 1\right)} \leq c_{(p)} \right. \right\}$$

$$NR = \left\{ x^* \in \text{base de datos} \left| \frac{P\left(x* | 0\right)}{P\left(x* | 1\right)} > c_{(p)} \right. \right\}$$

Para clasificar a un sujeto se leen sus datos, se categorizan y codifican y finalmente se observa si su vector de datos pertenece a R o a NR.

R y NR no cubren todos los posibles vectores de x^*, entonces puede ocurrir que se presente un paciente con un vector x^* asociado el cual no esté ni en R ni en NR. En estos casos la solución que se propone seguir es eliminar la última variable incluida por el proceso de selección a través de la devianza y repetir el procedimiento de clasificación. Al quitar una variable lo que se busca es 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.

Para ejemplificar este proceso considérese el conjunto de diez observaciones sobre cinco variables explicativas binarias mostradas en la figura 2.14 y a un individuo cuyo vector de observaciones y su respectiva codificación son $\mathbf{x}^* = (1, 0, 1, 1, 1)$ y $\mathbf{x}^{**} = 23$.

Como puede observarse, no hay información relacionada a los datos de la persona y entonces se procede a eliminar la variable X_5 con lo cual el vector con los datos de la persona en cuestión pasa a ser $\mathbf{x}^* = (1, 0, 1, 1)$ y su codificación $\mathbf{x}^{**} = 11$. El nuevo

Y	X_1	X_2	X_3	X_4	X_5	Codificación
$*$	1	0	1	0	1	21
$*$	1	0	0	1	1	19
$*$	1	1	1	1	1	31
$*$	1	1	0	1	0	26
$*$	0	1	1	1	0	14
$*$	1	1	1	0	1	29
$*$	0	1	1	0	1	13
$*$	1	1	1	0	0	28
$*$	1	0	0	1	1	19
$*$	1	0	1	1	0	22

Figura 2.14: Un conjunto de diez observaciones en el caso de cinco variables explicativas binarias. El asterisco denota cualquiera de los valores entre 1 ó 0.

Y	X_1	X_2	X_3	X_4	Codificación
$*$	1	0	1	0	20
$*$	1	0	0	1	5
$*$	1	1	1	1	27
$*$	1	1	0	1	22
$*$	0	1	1	1	7
$*$	1	1	1	0	28
$*$	0	1	1	0	24
$*$	1	1	1	0	19
$*$	1	0	0	1	20
$*$	1	0	1	1	11

Figura 2.15: conjunto resultante luego de excluir a la variable X_5. El asterisco denota cualquiera de los valores entre 1 ó 0.

conjunto de observaciones se resume en la figura 2.15. Como podrá notarse, en este caso ya se posee información relacionada al paciente en cuestión y por lo tanto se podrá proceder a catalogarlo dentro del conjunto de bajo o alto riesgo a padecer trombosis.

2.4. Desarrollo de una interfaz de usuario.

Una de las principales dificultades para utilizar R es que se requiere conocer su sintaxis y que esta es desconocida por la mayoría de las personas para quienes se ha pensado la utilización de estos métodos de clasificación. Adicionalmente, para muchos usuarios actuales es poco atractivo ejecutar un programa escribiendo directamente comandos. Por estas razones se pensó en construir una interfaz de usuario que facilitara el uso del programa en el sector médico.

Dentro de las alternativas que se exploraron para generar la interfaz de usuario estuvieron JAVA, Visual Basic y PHP, siendo elegida la última por parecer la más fácil de usar, pues de manera general requiere solamente de conocimientos básicos de PHP, el cual es muy similar a C, y HTML. Además, la utilización de PHP permite que el usuario no tenga que instalar programas adicionales y tenga acceso a los métodos de clasificación en cualquier momento a través de diversos dispositivos y sistemas operativos.

En términos más específicos, los requerimientos para construir la interfaz de usuario usando PHP son:

- Instalar el compilador de R.
- Instalar Apache y ejecutarlo durante el uso del programa.
- Hacer RScript una variable de entorno para que sea siempre visible a la computadora.

Ahora, de manera muy breve, se mencionarán los detalles fundamentales del funcionamiento de la interfaz de usuario, mismos que se ilustran en la figura 2.16.

- El aspecto gráfico se construyó usando formularios HTML e insertando el código

PHP en él.

- Para vincular la interfaz con el programa escrito en R se usó la instrucción "exec" de PHP para invocar RScript y así poder ejecutar los programas de R en segundo plano.
- Para vincular los resultados producidos mediante el código escrito en R con la interfaz, la información resultante de los scripts de R se guardó en archivos con extensión *.txt*, los cuales posteriormente se leyeron y desplegaron con PHP.

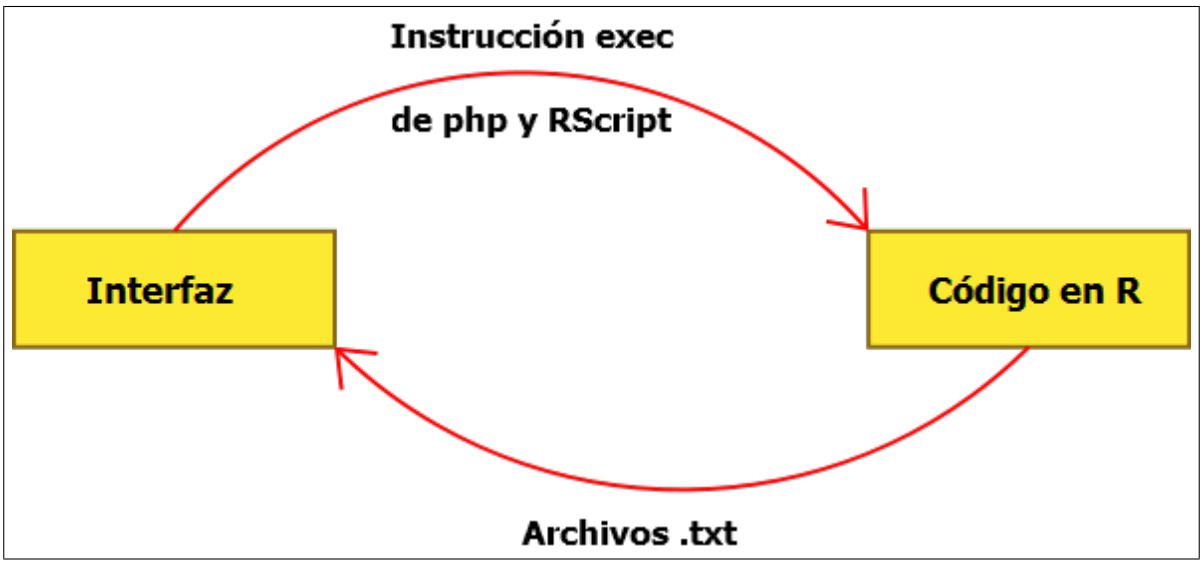


Figura 2.16: Esquema de funcionamiento de la interfaz de usuario.

La interfaz de usuario se encuentra alojada en los servidores del posgrado en matemáticas de la UAMI en la siguiente dirección

http://posmat.izt.uam.mx/csanchez/

Capítulo 3

Evaluación de los resultados.

Para evaluar los resultados obtenidos con los métodos de clasificación se ha hecho una división aleatoria de la base de datos en dos conjuntos, uno de ellos conformado por el 90 % de las observaciones, denominado conjunto de entrenamiento, y otro denominado conjunto de prueba formado por el 10 % restante.

Con el conjunto de entrenamiento se ha corrido el programa y se ha obtenido tanto el clasificador mediante regresión logística como el mejor clasificador de nivel alfa. Estos clasificadores se han aplicado a las observaciones del conjunto de prueba para obtener un vector de predicciones Y_{pred}. Este vector de predicciones asigna 0 a una persona catalogada en bajo riesgo y 1 a una persona catalogada en alto riesgo.

Con el vector de predicciones de cada clasificador se ha procedido a obtener el porcentaje de error de cada clasificador mediante la diferencia en valor absoluto entre el valor predicho y el real a través de

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

en donde Y_{real_i} es la componente i del vector con las observaciones reales de cada paciente en el conjunto de prueba. Por ejemplo, supóngase que en el conjunto de prueba se tuvieran las siguientes observaciones de la variable respuesta $Y_{real} = (1, 0, 1, 1, 0, 1, 0)$, mientras

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

Figura 3.1: Comparación del error cometido por ambos métodos en mil ocasiones. E_{Mi} representa el error del método i.

que los valores predichos por alguno de los clasificadores fueran $Y_{pred} = (0, 1, 1, 1, 0, 1, 0)$; entonces el porcentaje de error sería

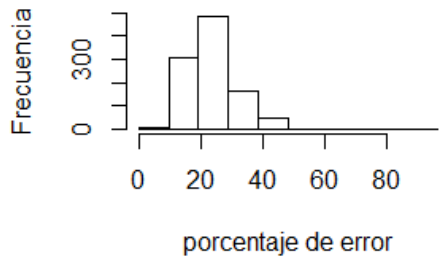
$$\frac{|0-1| + |1-0| + |1-1| + |1-1| + |0-0| + |1-1| + |0-0|}{7} \times 100 = \frac{2}{7} \times 100 \approx 28.5\,\%$$

El proceso de obtención del error de cada clasificador se ha repetido en mil ocasiones procurando que el conjunto de prueba y de entrenamiento tengan la misma proporción de pacientes sanos y enfermos que hay en la base de datos. Los resultados se resumen en la figura 3.1 en donde E_{Mi} denota el error del método de clasificación mediante regresión logística si $i = 1$ y el error del método del mejor clasificador de nivel alfa si $i = 2$. En dicha figura se puede observar que de manera global el mejor clasificador de nivel alfa resulta ser mejor pues el número conjunto de sus equivocaciones es menor a las equivocaciones del método de clasificación por regresión logística.

En la figura 3.2 se muestra con más detalle la distribución de los errores de cada método con respecto al tamaño total del conjunto de prueba. Por ejemplo, [0 %,10 %] indica la cantidad de ocasiones en que el número de errores estuvo entre el 0 y el 10 % con respecto al tamaño total del conjunto de prueba, es decir, [0 %,10 %] indicaría cuantas veces el número de errores estuvo entre 0 y 9, pues el conjunto de prueba contiene 96 elementos.

En la misma figura 3.2 se aprecia que el error para el mejor clasificador de nivel alfa tiende a estar mayormente por debajo del 20 % mientras que los errores cometidos por el clasificador mediante regresión logística están predominantemente por debajo del 30 %. Es importante señalar que para obtener el clasificador de regresión logística se utilizaron valores de las variables explicativas sin categorizar mientras que para obtener el mejor clasificador de nivel alfa los valores de las variables explicativas se categorizaron en dos

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



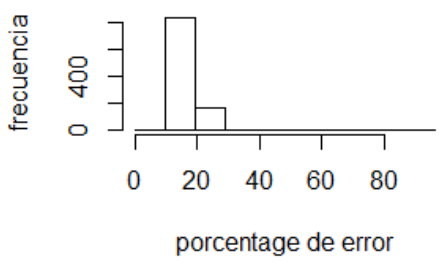


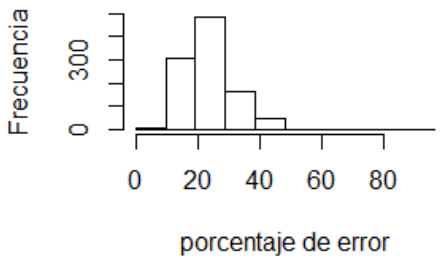
Figura 3.2: Distribución de los errores de cada método con respecto al tamaño total del conjunto de prueba; a la izquierda los resultados para el clasificador mediante regresión logística y a la derecha los resultados para el mejor clasificador de nivel alfa.

niveles, lo que conlleva a una pérdida de información y en consecuencia a una desventaja del mejor clasificador de nivel alfa.

Para sondear el efecto de la categorización, se utilizó la misma que se empleo en el mejor clasificador de nivel alfa para obtener el clasificador por regresión logística; los resultados se muestran en la figura 3.3 donde se aprecia que el método de regresión logística empeora.

Se han graficado las curvas de Lorenz de la función de distribución empírica de los sujetos clasificados en riesgo y bajo riesgo para el clasificador por regresión logística y el mejor clasificador de nivel alfa. En el caso del clasificador por regresión logística esto se ha hecho ordenando los resultados con respecto a las estimaciones del modelo $\widehat{p}(x_i) = \dfrac{1}{1+e^{-\mathbf{X}_i^T\widehat{\beta}}}$ y después realizando un conteo cuando $Y=0$ y cuando $Y=1$ para finalmente construir la función de distribución. Este mismo proceso se ha repetido para el mejor clasificador de nivel alfa pero usando el cociente de probabilidades $c_i = \dfrac{\widehat{\mathrm{p}}\left(\mathbf{x}_i|Y=0\right)}{\widehat{\mathrm{p}}\left(\mathbf{x}_i|Y=1\right)} = \dfrac{f_{\mathbf{x}_i,0}/n_0}{f_{\mathbf{x}_i,1}/n_1}$. Los resultados se muestran en la figura 3.4.

E_{M1}	Frecuencia	E_{M1}	Frecuencia
$[0\,\%, 10\,\%]$	2	$(0\,\%, 10\,\%]$	0
$(10\,\%, 20\,\%]$	304	$(10\,\%, 20\,\%]$	126
$(20\,\%, 30\,\%]$	485	$(20\,\%, 30\,\%]$	468
$(30\,\%, 40\,\%]$	163	$(30\,\%, 40\,\%]$	231
$(40\,\%, 50\,\%]$	45	$(40\,\%, 50\,\%]$	161
$(50\,\%, 60\,\%]$	1	$(50\,\%, 60\,\%]$	14
$(60\,\%, 70\,\%]$	0	$(60\,\%, 70\,\%]$	0
$(70\,\%, 80\,\%]$	0	$(70\,\%, 80\,\%]$	0
$(80\,\%, 90\,\%]$	0	$(80\,\%, 90\,\%]$	0
$(90\,\%, 100\,\%]$	0	$(90\,\%, 100\,\%]$	0



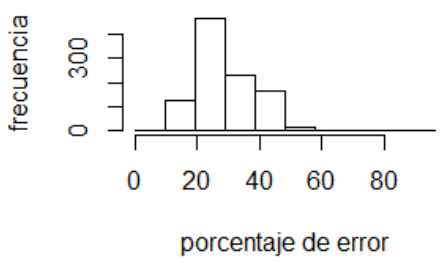


Figura 3.3: Comparación de la distribución de los errores del método de clasificación mediante regresión logística. A la izquierda sin categorizar las variables y a la derecha categorizando las variables.

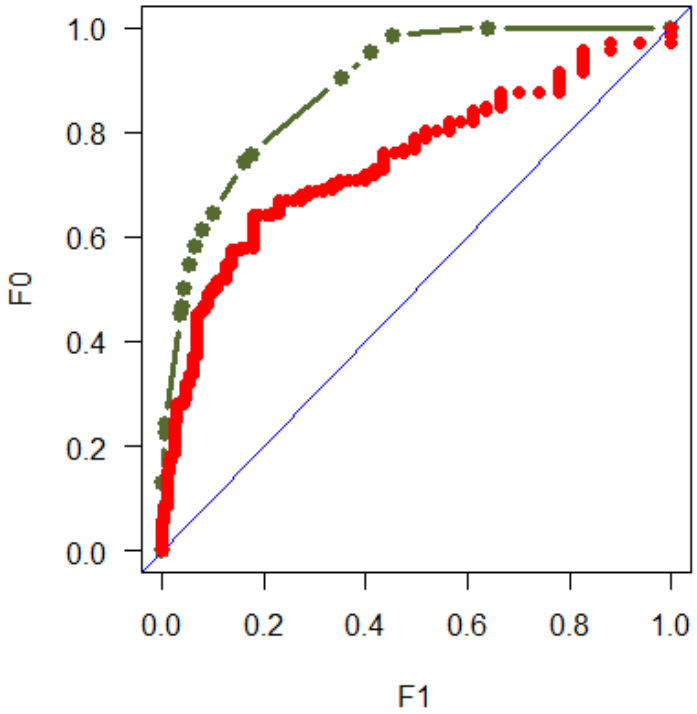


Figura 3.4: Gráfica de la Lorenz. En verde se muestra la gráfica correspondiente al mejor clasificador de nivel alfa y en rojo la correspondiente al clasificador por regresión logística.

Por otra parte, en la figura 3.5 se muestra la matriz de correlación de las variables seleccionadas para realizar la clasificación antes de realizar la imputación, mientras que en la figura 3.6 se muestra la matriz de correlación de las mismas variables luego de realizar la imputación. Finalmente, en la figura 3.7 se muestra la diferencia en valor absoluto de las matrices de las figuras 3.5 y 3.6; la mayor diferencia entre las dos matrices es de 0.1837. Lo cual nos hace concluir que la imputación no modifica de manera sustantiva los valores de las variables.

	ETE	FVW..VIDAS.	RPCAR.1.5	LAR	DIMEROS.D	RnP	PGR	SSA	SSB
ETE	1.0000	0.1130	-0.1007	0.1034	0.1463	-0.0443	-0.1264	-0.0477	-0.0508
FVW..VIDAS.	0.1130	1.0000	0.0049	0.1372	0.4119	0.1362	-0.0987	0.1689	0.0569
RPCAR.1.5	-0.1007	0.0049	1.0000	-0.1642	-0.0634	0.0062	0.0296	0.0046	-0.0790
LAR	0.1034	0.1372	-0.1642	1.0000	0.0340	0.1223	-0.0372	0.1189	0.0660
DIMEROS.D	0.1463	0.4119	-0.0634	0.0340	1.0000	0.2845	-0.1736	0.0206	-0.0525
RnP	-0.0443	0.1362	0.0062	0.1223	0.2845	1.0000	0.0051	0.1154	0.0178
PGR	-0.1264	-0.0987	0.0296	-0.0372	-0.1736	0.0051	1.0000	-0.0051	-0.0068
SSA	-0.0477	0.1689	0.0046	0.1189	0.0206	0.1154	-0.0051	1.0000	0.2285
SSB	-0.0508	0.0569	-0.0790	0.0660	-0.0525	0.0178	-0.0068	0.2285	1.0000

Figura 3.5: Matriz de correlación de las variables seleccionadas sin realizar imputación.

	ETE	FVW..VIDAS.	RPCAR.1.5	LAR	DIMEROS.D	RnP	PGR	SSA	SSB
ETE	1.0000	0.1622	-0.1353	0.1168	0.1966	-0.0483	-0.1468	-0.0454	0.0129
FVW..VIDAS.	0.1622	1.0000	-0.0268	0.1664	0.4323	0.1165	-0.1332	0.1795	0.1096
RPCAR.1.5	-0.1353	-0.0268	1.0000	-0.0931	-0.1011	0.0154	0.0875	-0.0851	-0.0656
LAR	0.1168	0.1664	-0.0931	1.0000	0.0149	0.0585	-0.0246	0.0904	0.0711
DIMEROS.D	0.1966	0.4323	-0.1011	0.0149	1.0000	0.2048	-0.1715	0.0124	-0.0477
RnP	-0.0483	0.1165	0.0154	0.0585	0.2048	1.0000	-0.0057	0.1243	0.2015
PGR	-0.1468	-0.1332	0.0875	-0.0246	-0.1715	-0.0057	1.0000	-0.0132	-0.0205
SSA	-0.0454	0.1795	-0.0851	0.0904	0.0124	0.1243	-0.0132	1.0000	0.2582
SSB	0.0129	0.1096	-0.0656	0.0711	-0.0477	0.2015	-0.0205	0.2582	1.0000

Figura 3.6: Matriz de correlación de las variables seleccionadas después de realizar la imputación.

	ETE	FVW..VIDAS.	RPCAR.1.5	LAR	DIMEROS.D	RnP	PGR	SSA	SSB
ETE	0.0000	0.0492	0.0346	0.0134	0.0503	0.0040	0.0204	0.0023	0.0637
FVW..VIDAS.	0.0492	0.0000	0.0317	0.0292	0.0204	0.0197	0.0345	0.0106	0.0527
RPCAR.1.5	0.0346	0.0317	0.0000	0.0711	0.0377	0.0092	0.0579	0.0897	0.0134
LAR	0.0134	0.0292	0.0711	0.0000	0.0191	0.0638	0.0126	0.0285	0.0051
DIMEROS.D	0.0503	0.0204	0.0377	0.0191	0.0000	0.0797	0.0021	0.0082	0.0048
RnP	0.0040	0.0197	0.0092	0.0638	0.0797	0.0000	0.0108	0.0089	0.1837
PGR	0.0204	0.0345	0.0579	0.0126	0.0021	0.0108	0.0000	0.0081	0.0137
SSA	0.0023	0.0106	0.0897	0.0285	0.0082	0.0089	0.0081	0.0000	0.0297
SSB	0.0637	0.0527	0.0134	0.0051	0.0048	0.1837	0.0137	0.0297	0.0000

Figura 3.7: Diferencia en valor absoluto entre las matrices de correlación con y sin imputación.

Capítulo 4

Conclusiones.

Con base en la información mostrada anteriormente y en el proceso de programación de los métodos descritos en este trabajo, llegamos a las siguientes conclusiones y observaciones:

- Se ha probado que el mejor clasificador de nivel alfa es factible a ser programado y generalizado, incluso para usar más de un punto de corte lo cual conllevaría a hacer una modificación en la interfaz, la cual trabaja sólo con un punto de corte, lo cual sería posible, pero se deja para estudios posteriores.

- El mejor clasificador de nivel alfa resulta ser mejor opción respecto al clasificador mediante regresión logística pese a tener la desventaja haber usado una categorización binaria por lo que inferimos que si se tuviese una categorización más fina se lograría una clasificación más exacta.

- Una posible debilidad del mejor clasificador de nivel alfa es su alta sensibilidad a los puntos de corte. Un caso extremo se tiene cuando se da un punto de corte para una variable en donde todos los valores quedan en una categoría; en este caso el punto de corte no será congruente con la información de la base de datos, lo cual llevará a malos resultados. Por esta razón, se debe hacer un análisis meticuloso y asesorarse con expertos en el tema para tener una categorización adecuada.

- Un ligero inconveniente que presenta el mejor clasificador de nivel alfa es que las combinaciones de variables observadas se deben almacenar en memoria y que estas podrían ser demasiadas, especialmente para grandes bases de datos; sin embargo, esta última "inconveniencia" será seguramente superada dado el ritmo vertiginoso del desarrollo computacional.

- Dado que el mejor clasificador de nivel alfa se construye con base en la información contenida en la base de datos puede llegar un individuo cuyas características no estén dentro del conjunto de resultados almacenado. La probabilidad de que esto ocurra disminuye con una base de datos grande pero se incrementa cuando hay muchas variables y muchas categorías por variable; sin embargo en el cuerpo del trabajo se dio una posible solución que fue quitar una variable y repetir la construcción del clasificador hasta lograr clasificar al individuo; también se pueden utilizar técnicas de selección de variables como se hizo en este trabajo.

- Con la construcción de la interfaz de usuario se ha visto la posibilidad de emplear R desde un entorno más gráfico y amigable, lo cual abre muchísimas posibilidades para el manejo de este programa.

Apéndice A

Descripción de la interfaz de usuario.

En esta sección se detallará brevemente la experiencia del usuario al utilizar la interfaz; también se encuentra una guía de su uso en la liga

http://posmat.izt.uam.mx/csanchez/

En la primera ventana se solicita adjuntar un archivo, este debe estar en formato *.csv*. Si no se adjunta un archivo con la extensión correcta, entonces se reconviene al usuario de hacerlo y se regresa a la página inicial (ver figuras A.1 y A.2).

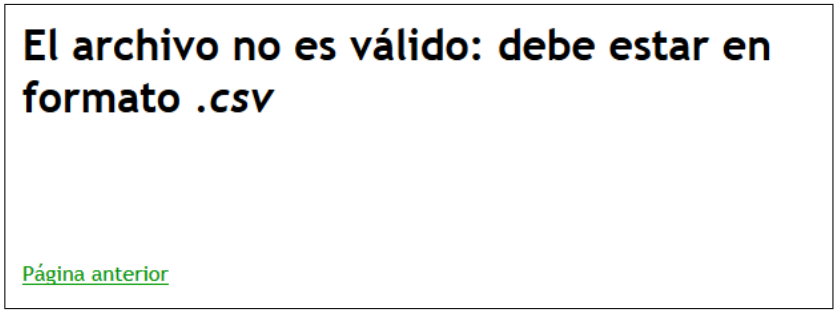


Figura A.2: Mensaje que aparece en caso de que el usuario no adjunte un archivo .csv. Esta página devuelve automáticamente a la página de inicio.

Figura A.1: Página inicial de la interfaz de usuario donde se solicita adjuntar un archivo.

Luego de que se ha adjuntado un archivo con la extensión apropiada se verá un formulario en donde se pedirá al usuario rellenar todos los campos (figura A.3). Luego de completar esta acción se mostrará una página con los resultados de la clasificación a través de regresión logística en donde aparecerán las opciones de volver al inicio o utilizar el mejor clasificador de nivel alfa, esto se muestra en la figura A.4.

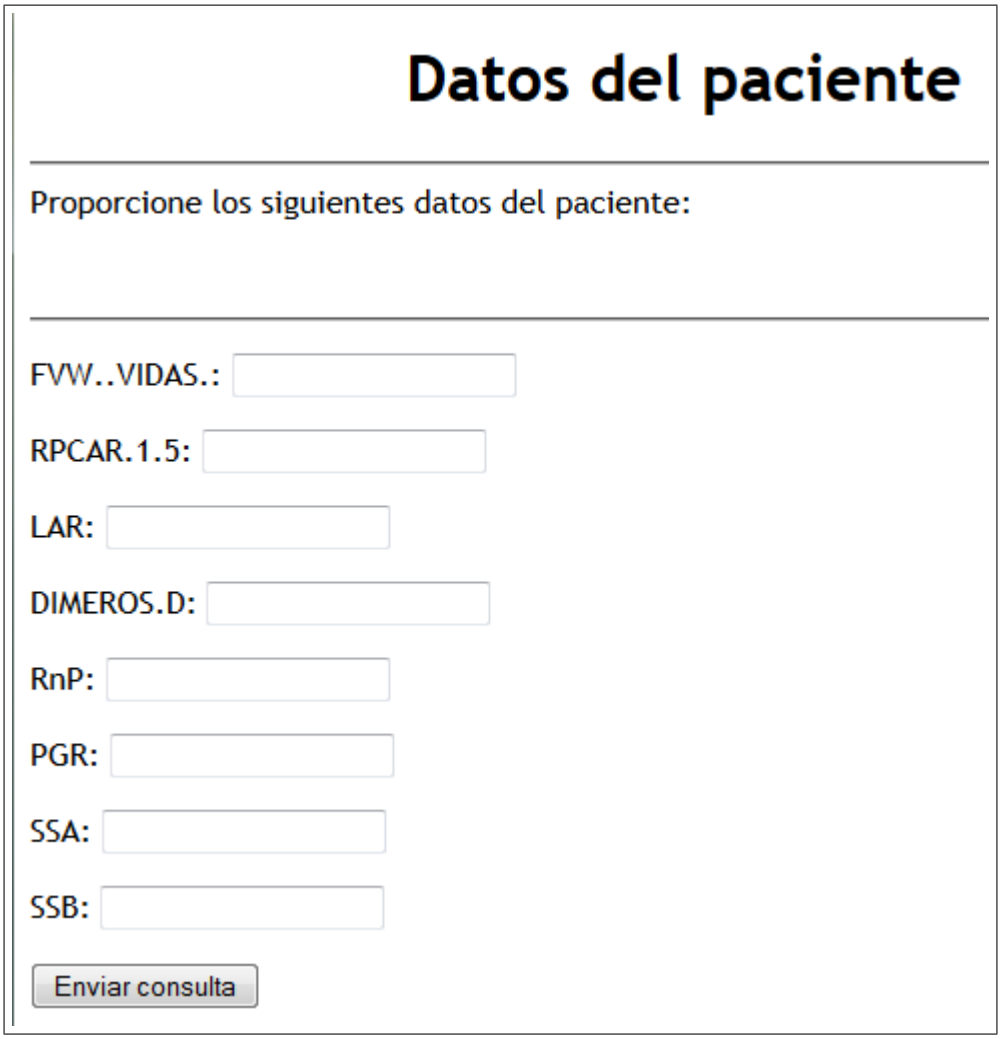


Figura A.3: Mensaje que aparece en caso de que el usuario no adjunte un archivo .csv. Esta página devuelve automáticamente a la página de inicio.

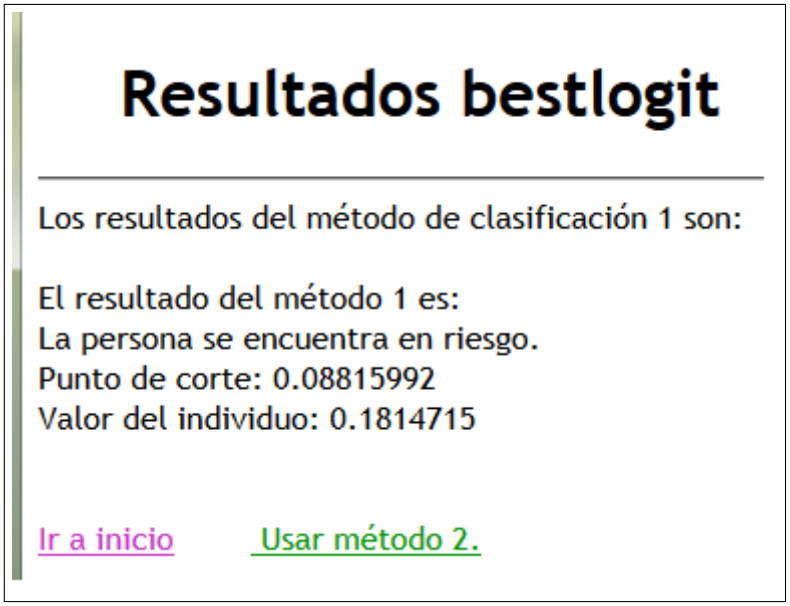


Figura A.4: Ventana con los resultados del método de clasificación a través de regresión logística.

Si en la ventana con los resultados de la clasificación mediante regresión logística se selecciona usar el método 2 entonces aparecerá otro formulario similar al anterior pero en él no se solicitarán los datos del paciente sino los puntos de corte a emplear por cada variable; nuevamente todos los campos tienen que ser llenados para poder proseguir (figura A.5). Finalmente al oprimir el botón de enviar consulta se mandará a la página con los resultados de la clasificación con el mejor clasificador de nivel alfa y con la opción de volver al inicio; esto se muestra en la figura A.6.

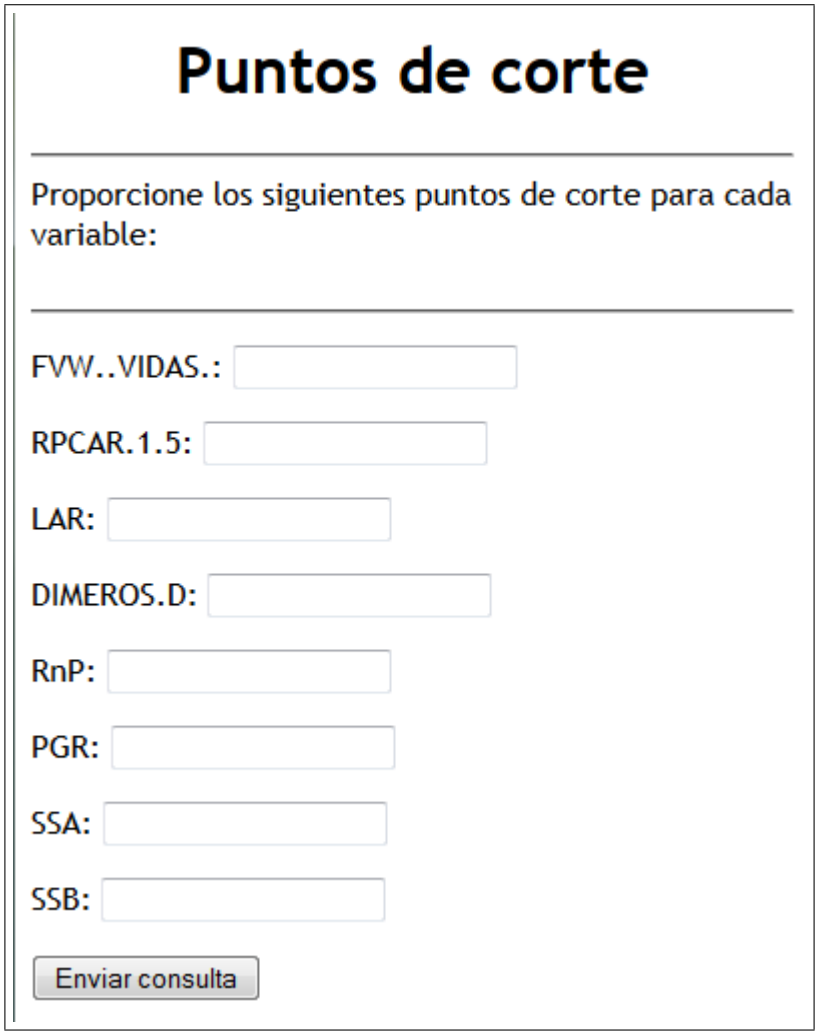


Figura A.5: Formulario en donde se requieren los puntos de corte a usar por cada variable.

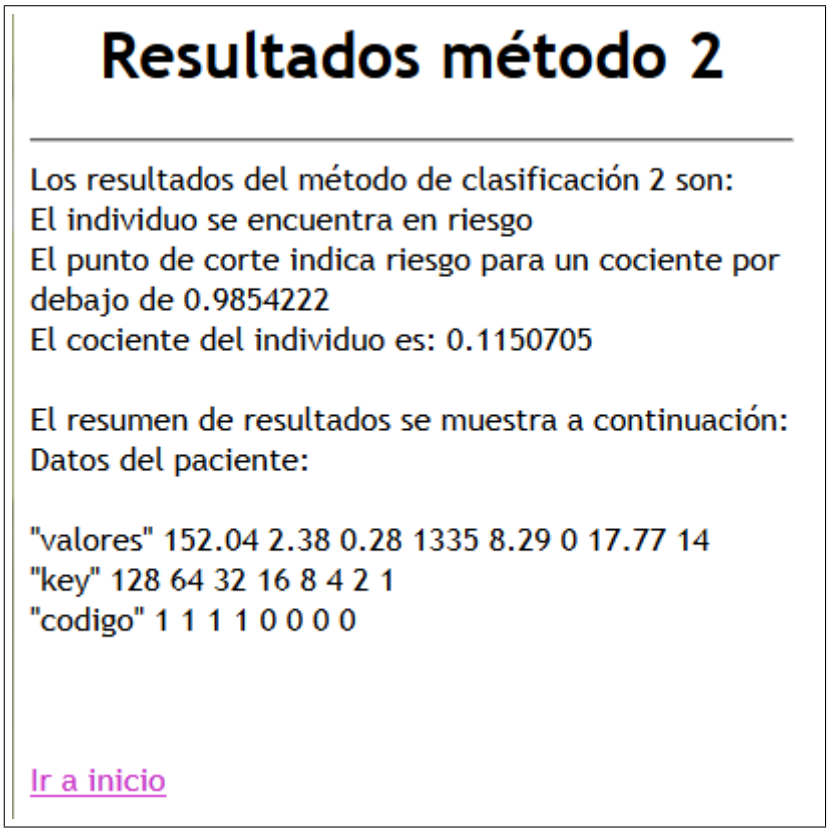


Figura A.6: Página con los resultados del mejor clasificador de nivel alfa.

Bibliografía

Annette J Dobson. *An introduction to generalized linear models*. CRC press, second edition, 2002.

Bertha Higashida. Isnb 9789701063811 ciencias de la salud. ed, 2008.

Elizabeth A Martin. *Oxford concise colour medical dictionary*. Oxford University Press, 2002.

Atul Mehta and Victor Hoffbrand. *Haematology at a Glance*. John Wiley & Sons, 2013.

Raymond H Myers, Douglas C Montgomery, G Geoffrey Vining, and Timothy J Robinson. *Generalized linear models: with applications in engineering and the sciences*. John Wiley & Sons, second edition, 2002.

B Shirlyn and M Mckenzie. *Hematología Clínica*. Manual Moderno, 2000.

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE EXAMEN DE GRADO

No. 00118
Matrícula: 2123802784

MODELO DE ANALISIS PARA LA PREDICCION DE TROMBOSIS FAMILIAR.

En México, D.F., se presentaron a las 11:00 horas del día 7 del mes de enero del año 2015 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DR. CARLOS DIAZ AVALOS
DRA. BLANCA ROSA PEREZ SALVADOR
DR. HECTOR ALFREDO BAPTISTA GONZALEZ
DR. ALBERTO CASTILLO MORALES

Bajo la Presidencia del primero y con carácter de Secretario el último, se reunieron para proceder al Examen de Grado cuya denominación aparece al margen, para la obtención del grado de:

MAESTRO EN CIENCIAS (MATEMÁTICAS APLICADAS E INDUSTRIALES)

DE: CARLOS GABRIEL SANCHEZ LORDMENDEZ

y de acuerdo con el artículo 78 fracción III del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

APROBAR

Acto continuo, el presidente del jurado comunicó al interesado el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.

CARLOS GABRIEL SANCHEZ LORDMENDEZ
ALUMNO

REVISÓ

LIC. JULIO CESAR DE LARA ISASSI
DIRECTOR DE SISTEMAS ESCOLARES

DIRECTOR DE LA DIVISIÓN DE CBI

DR. JOSE GILBERTO CORDOBA HERRERA

PRESIDENTE

DR. CARLOS DIAZ AVALOS

VOCAL

DRA. BLANCA ROSA PEREZ SALVADOR

VOCAL

CANCELADO
DR. HECTOR ALFREDO BAPTISTA GONZALEZ

SECRETARIO

DR. ALBERTO CASTILLO MORALES