Modelos de clasificación binaria bayesianos aplicados a la detección de enfermedades cardíacas.

Yael Rene Santiago Cruz

11 de junio de 2022

Índice general

Capítulo 1

Introducción

Las enfermedades cardiovasculares (ECV) son la principal causa de muerte en todo el mundo y cada año mueren más personas por ECV que por cualquier otra causa. Se calcula que en 2015 murieron por esta causa 17,7 millones de personas, lo cual representa un 31 % de todas las muertes registradas en el mundo y más de tres cuartas partes de estas muertes se producen en países de bajos ingresos.

La mayoría de las ECV pueden prevenirse actuando sobre factores de riesgo comportamentales, como el consumo de tabaco, las dietas malsanas y la obesidad, la inactividad física o el consumo nocivo de alcohol, utilizando estrategias que abarquen a toda la población.

El aprendizaje automático o machine learning es la ciencia de programar computadoras que sean capaces de aprender por si solas utilizando datos. Dentro del aprendizaje automático se encuentra una rama llamada aprendizaje supervisado, esta permite aprender el comportamiento de un conjunto de datos que contienen "características" y sus respectivas "etiquetas" para tratar de predecir en un futuro dichas etiquetas.

En la actualidad, es más frecuente utilizar algoritmos de aprendizaje supervisado como pruebas para la detección de enfermedades, dentro de este tipo de algoritmos se encuentran los basados en la teoría de la estadística bayesiana.

En este proyecto se pretende elegir un modelo de clasificación binaria bayesiano para estimar la probabilidad de una posible enfermedad cardíaca dado un conjunto de variables explicativas. Los modelos que se pretende ajustar al conjunto de datos son los modelos lineales generalizados bayesianos en donde la variable respuesta es binaria. Además, me apoyaré de las técnicas del aprendizaje supervisado para la elección del mejor modelo.

Capítulo 2

Modelos de clasificación binaria bayesianos

Un problema de aprendizaje supervisado consiste en tratar de predecir el valor de una variable, llamada "dependiente" o "respuesta", a partir de otras variables, llamadas "independientes" o "explicativas", cuando la variable respuesta puede tomar únicamente dos valores, el problema se denomina como de clasificación binaria. El modelo de regresión lineal no es el más adecuado para dar solución a este tipo de problemas, debido a que asume normalidad de la variable respuesta, y por lo tanto se debe recurrir a otro tipo de técnicas, dentro de las cuales se encuentran los modelos lineales generalizados.

2.1. Modelos lineales generalizados

2.1.1. La familia exponencial

Considere una variable aleatoria X que pertenece a la familia exponencial, entonces su función de densidad de probabilidad tiene la forma

$$f_X(x;\phi,\theta) = c(x,\phi)\exp\left(\frac{x\theta - a(\theta)}{\phi}\right),$$

en donde θ es el parámetro canónico y ϕ es el parámetro de dispersión. Las funciones $a(\theta)$ y $c(x,\phi)$ determinan las funciones de densidad y distribución de X.

2.1.2. Definición de los modelos lineales generalizados

Un modelo lineal generalizado se caracteriza por los siguientes componentes:

- *Componente aleatorio*: Un vector de observaciones $y = (y_1, y_2, ..y_n)$, que son una realización del vector $Y = (Y_1, Y_2, .., Y_n)$. Los componentes de Y son independientes con distribución perteneciente a la familia exponencial, de tal manera que $E[Y] = \mu = (\mu_1, \mu_2, .., \mu_n)$.

- *Componente sistemático:* Un conjunto de variables $X_1, X_2, ..X_p$, parámetros $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_p)$ y una matriz

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{np} \end{pmatrix},$$

  tal que x_{ij} representa la i-ésima observación de la variable X_{j-1} para toda $i = 1, 2, .., n$ y $j = 1, .., p$.

- *Función liga o enlace*: Una función g que relaciona los componentes aleatorio y sistemático de la siguiente manera:

$$g(\mu) = X\beta^T.$$

Para el enfoque bayesiano, además supondremos que β es un vector aleatorio con función de densidad conjunta $p(\beta)$, llamada densidad "prior" o "inicial". Para la densidad inicial se puede usar una no informativa, es decir,

2.1.3. Funciones enlace para modelos de variable respuesta binaria

Si la variable respuesta Y es binaria, entonces $E(Y) = P(Y = 1)$ y por tanto la función enlace del modelo debe ser la inversa de una función de distribución de probabilidad. Dentro del conjunto de las funciones de enlace más utilizadas para los modelos en donde la variable respuesta es binaria están:

- *Logit o función logística*: $g(p) = \log\left(\frac{p}{1-p}\right)$.
- *Probit o función normal inversa*: $g(p) = \Phi^{-1}(p)$, en donde Φ es la función de distribución de probabilidad asociada a una variable aleatoria normal estándar.
- *Cloglog o función loglog complementario*: $g(p) = \log[-\log(1-p)]$.
- *Cauchit o función Cauchy-Lorentz:*: $g(p) = \tan[\pi(p - 0.5)]$.

2.1.4. Estimación de los parámetros

La estimación de los parámetros del modelo consiste en resolver el problema de optimización

$$\max_{\hat{\beta} \in \Theta} E_\beta[U(\hat{\beta}, \beta)|X, Y],$$

en donde Θ es el espacio parametral y $U(\hat{\beta}, \beta)$ es una función de utilidad, si dicha función es la pérdida cuadrática, $U(\hat{\beta}, \beta) = -(\hat{\beta} - \beta)^2$, entonces $\hat{\beta} = E(\beta|X, Y)$.

No siempre es fácil calcular la $E(\beta|X, Y)$, y por tanto se debe recurrir a simulaciones Monte Carlo vía Cadenas de Markov para estimar la distribución posterior $p(\beta|X, Y)$.

2.2. Elección de modelos

La mayoría de los modelos de clasificación binaria, como los modelos lineales generalizados, estiman $P(Y = 1|X)$ y por tanto necesitamos definir una regla de decisión para clasificar un elemento de la muestra dado un conjunto de características X. Usualmente se suele elegir un umbral $\alpha \in (0, 1)$ y utilizar la regla de clasificación

$$\hat{Y_i} = \begin{cases} 1 & \text{si } P(Y_i = 1|X_i) > \alpha, \\ 0 & \text{en otro caso.} \end{cases}$$

Cuando se ajusta un modelo lineal generalizado la regla anterior se reduce a

$$\hat{Y_i} = \begin{cases} 1 & \text{si } X_i\beta^T > g(\alpha), \\ 0 & \text{en otro caso.} \end{cases}$$

2.2.1. Matriz de confusión

Un muy buena forma de evaluar el desempeño de un modelo es construir la matriz de confusión. La idea general es contar el número de veces que la instancia de la clase A es clasificada como de la clase B, en donde $A, B \in \{0, 1\}$. La matriz de confusión tiene dimensión 2×2 y se define como

$$MC = \begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix},$$

donde

- TN es el número de verdaderos negativos, es decir, los valores de la clase 0 que el modelo asignó a esa misma clase,
- FP es el número de falsos positivos, es decir, los valores de la clase 0 que el modelo asignó a la clase 1,
- FN es el número de falsos negativos, es decir, los valores de la clase 1 que el modelo asignó a la clase 0,
- TP es el número de verdaderos positivos, es decir, los valores de la clase 1 que el modelo asignó a esa misma clase.

2.2.2. Métricas derivadas de la matriz de confusión

Las métricas que se derivan de la matriz de confusión más usadas son:

- *Precisión* $= \dfrac{TP + TN}{TN + FP + FN + TP}$. Explica el porcentaje de datos correctamente clasificados.

- $Sensibilidad = \dfrac{TP}{FN + TP}$. Frecuencia relativa de predecir un evento como positivo cuando el evento observado es positivo. Es la fracción de verdaderos positivos.
- $Especificidad = \dfrac{TN}{TN + FP}$. Frecuencia relativa de predecir un evento como negativo cuando el evento observado es negativo. Es la fracción de verdaderos negativos.
- *AUC Score*. Calcula el área bajo la curva ROC *(Receiver Operating Characteristic)*, esta curva grafica $1 - Especificidad$ (eje horiontal) vs $Sensibilidad$ (eje vertical). Con esta orientación de los ejes,un valor del eje x cercano a cero (alta especificidad) generalmente implica un valor del eje y bajo (baja sensibilidad), y viceversa. Un modelo que predice adecuadamente resulta en una curva ROC que crece rápidamente a 1: cuanto mas cercana esté la curva a la parte superior izquierda, mejor serán sus predicciones, o equivalentemente, el área bajo la curva se aproxima a 1.

2.2.3. Train and test split y cross-validation

El principal objetivo al utilizar un modelo de clasificación binaria es realizar buenas predicciones para datos que no se encontraban en la muestra al momento del ajuste. Para verificar que lo antes mencionado ocurra, existen varias técnicas que podemos usar, entre ellas se encuentran: *train and test split* y *cross validation*.

Train and test split

Supongamos que para ajustar un modelo se tiene un conjunto de datos $D = \{(X_1, y_1), .., (X_n, y_n)\}$, en donde (X_i, y_i) representa las variables explicativas y respuesta correspondientes a la i-ésima observación.

El método de *train and test split* consiste en obtener a partir del conjunto D, una muestra aleatoria sin remplazo de tamaño m. Los elementos que cayeron dentro de la muestra conforman el conjunto de entrenamiento, mientras que los elementos que quedaron fuera de la muestra conforman al conjunto de prueba. Una vez realizada la muestra, se ajusta el modelo al conjunto de entrenamiento y se evalúa el desempeño del mismo, con alguna métrica, en el conjunto de prueba. Se recomienda utilizar entre el 70 % y 85 % de los datos para el conjunto de entrenamiento.

Cross validation

Para el método de *cross validation* se necesita dividir el conjunto de datos en K bloques, después se debe utilizar el i-ésimo bloque como conjunto de prueba y el resto como conjunto de entrenamiento, esto se debe realizar para toda $i = 1, 2, ..., K$. Para evaluar que tan bueno es el modelo utilizaremos la métrica

$$M = \frac{1}{K} \sum_{i=1}^{K} m_i,$$

en donde m_i es el valor de la métrica, como por ejemplo: *Accuracy* o *AUC Score*, obtenida para el i-ésimo conjunto de prueba.

Sobreajuste y subajuste de un modelo

Supongamos que tenemos un modelo de clasificación binaria de tal manera que se ajusta muy bien a los datos de entrenamiento, y por tanto clasifica de manera correcta al 100 % de estos, pero en cambio, al momento de predecir el valor de los datos prueba, solamente el 5 % de ellos son clarificados correctamente, a este efecto se le denomina sobreajuste de un modelo. Por otro lado, el subajuste ocurre cuando el modelo clasifica de manera correcta una cantidad muy pequeña de los datos tanto de entrenamiento como de prueba.

2.3. Prepocesamiento de datos para la mejora de los modelos

2.3.1. Escalamiento de variables numéricas

El escalamiento de variables numéricas consiste en aplicarles una transformación, de tal manera que las variables transformadas tomen valores dentro de un mismo rango de longitud pequeña. Los dos métodos más comunes para escalar las variables numéricas son: *Escalador mín-míx* y *Escalador estándar*.

Escalador mín-míx

El *Escalador mín-míx* transforma cada variable numérica individualmente, de tal manera que las nuevas variables estén dentro del intervalo $[0, 1]$. Suponga que se tiene una muestra $x_1, x_2, .., x_n$ de una variable numérica, entonces la transformación esta dada por:

$$z_i = \frac{x_i - x_{\text{mín}}}{x_{\text{máx}} - x_{\text{mín}}},$$

donde $x_{\text{mín}} = \text{mín}\{x_1, x_2, ..x_n\}$ y $x_{\text{máx}} = \text{máx}\{x_1, x_2, ..x_n\}$.

Escalador estándar

El *Escalador estándar* transforma cada variable numérica individualmente, de tal manera que las nuevas variables tengan media cero y varianza uno. Suponga que se tiene una muestra $x_1, x_2, .., x_n$ de una variable numérica, entonces la transformación esta dada por:

$$z_i = \frac{x_i - \bar{X}}{S_x},$$

donde $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ y $S_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$.

2.3.2. Detección de datos atípicos

A veces, un conjunto de datos puede contener valores extremos que están fuera del rango esperado y diferentes a los demás datos. Estos se denominan valores atípicos y, a menudo, un modelo puede mejorar al comprender e incluso eliminar estos valores atípicos, ya que estos se pueden deber a errores de medición, pero esto también puede provocar perdida de información, así que se debe ser muy cuidadoso al momento de trabajar con datos atípicos.

Una forma de detectar valores atípicos es mediante el rango intercuartílico. Sea $x_1, x_2, .., x_n$ una muestra de la variable X, un dato x se considera atípico si

$$x < q_{25} - IQR * 1.5 \text{ o } x > q_{75} + IQR * 1.5,$$

en donde q_{25} y q_{75} son los percentiles 25 y 75 de la muestra respectivamente, e $IQR = q_{75} - q_{25}$.

Capítulo 3

Análisis descriptivo de los datos

El conjunto de datos a trabajar se puede obtener en la siguiente liga: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction. Este conjunto de datos tiene un total de 918 observaciones recolectadas de distintas regiones: Cleveland, Hungarian, Switzerland, Long Beach VA y Stalog.

Variable	**Tipo**	**Descripción**
Age	Numérica	Edad del paciente en año.
Sex	Categórica	Sexo del paciente(M: Hombre, H: Mujer).
ChestPainType	Categórica	Tipo de dolor de pecho(TA: Angina típica, ATA: Angina atípica, NAP: Dolor no anginoso, ASY: Asintomático).
RestingBP	Numérica	Presión arterial en reposo(mmHg).
Cholesterol	Numérica	Colesterol sérico(mm/dl).
FastingBS	Categórica	Azúcar en la sangre durante el ayuno(1: si FastingBS > 120 mg/dl, 0: en caso contrario).
RestingECG	Categórica	Resultados del electrocardiograma en reposo(Normal: Normal, ST: Con anomalías en la onda ST-T, HVI: Hipertrofia ventricular izquierda probable o definitiva según el criterio de Estes).
MaxHR	Numérica	Frecuencia cardíaca máxima alcanzada(valores entre 60 y 202) .
ExerciseAngina	Categórica	Angina inducida por el ejercicio (Y: Sí, N: No).
Oldpeak	Numérica	oldpeak = ST(Valor numérico medido en depresión)
ST_Slope	Categórica	La pendiente del segmento ST del ejercicio máximo (Up: Pendiente ascendente, Plana: Flat, Down: Pendiente descendente)
HeartDisease	Categórica	Variable respuesta (1:Enfermedad cardíaca, 0: Normal).

Tabla 3.1: Variables que conforman al conjunto de datos.

El conjunto está conformado por 12 variables, once explicativas y una respuesta, en la Tabla 3.1 se muestra una descripción de cada una de ellas.

Algo a destacar es que las variables numéricas no están correlacionadas, con excepción de las variables MaxHR y Age(ver Figura 3.1), esto ocasionaría un problema al momento de ajustar el modelo, por tanto es conveniente eliminar alguna o realizar un análisis de componentes principales y utilizar el primer componente para sustituir a ambas variables.

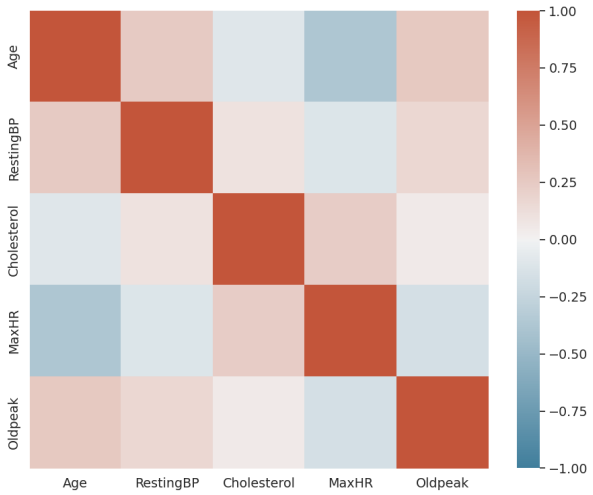

Figura 3.1: Correlograma de las variables numéricas.

Al momento de agrupar los datos dependiendo si son normales o tienen alguna enfermedad cardíaca y obtuvieron los datos de la Tabla 3.2 y de la Tabla 3.3, además también se pueden visualizar de manera gráfica en la Figura 3.2:

	Age	**RestingBP**	**Cholesterol**	**MaxHR**	**Oldpeak**
mean	50.55	130.18	227.12	148.15	0.40
std	9.44	16.5	74.63	23.28	0.7
min	28	80	0	69	-1.1
25 %	43	120	197.25	134	0
50 %	51	130	227	150	0
75 %	57	140	266.75	165	0.6
max	76	190	564	202	4.2

Tabla 3.2: Resumen de la distribucion de las variables numéricas para el grupo de personas Normales.

	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
mean	55.9	134.18	175.94	127.65	1.27
std	8.72	19.82	126.39	23.38	1.15
min	31	0	0	60	-2.6
25 %	51	120	0	112	0
50 %	57	132	217	126	1.2
75 %	62	145	267	144.25	2
max	77	200	603	195	6.2

Tabla 3.3: Resumen de la distribución de las variables numéricas para el grupo de personas con enfermedad del corazón.

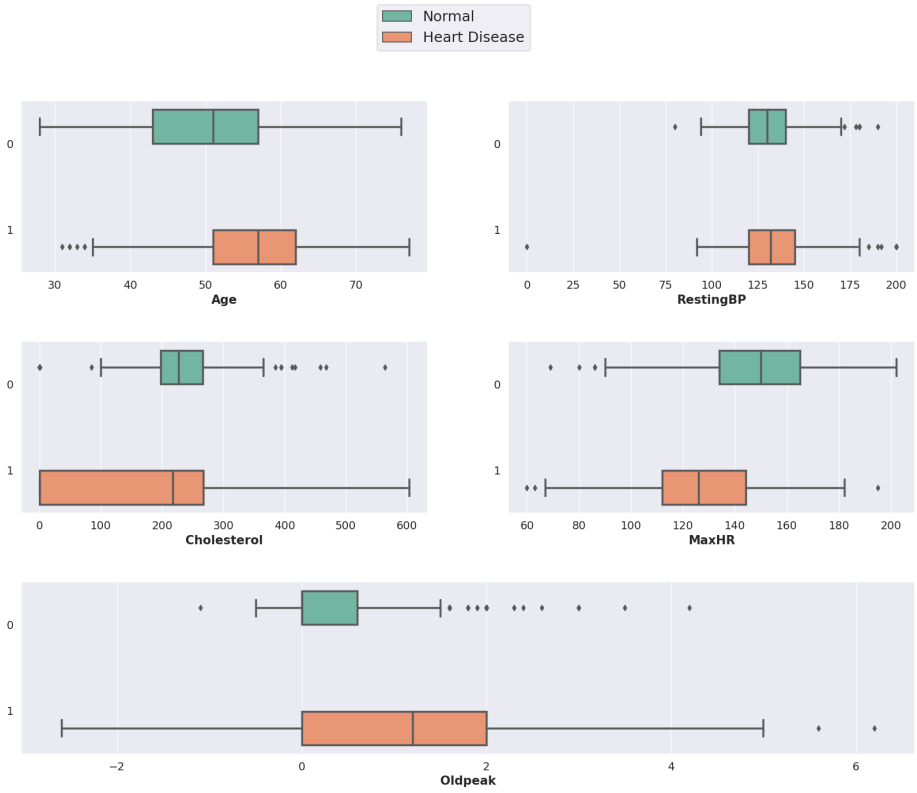

Figura 3.2: Gráfico de cajas de las variables numéricas agrupadas por HeartDisease.

De las variables categóricas se deduce lo siguiente(ver Figura 3.3):

- El conjunto de datos tiene un buen balanceo, ya que aunque la cantidad de personas normales es menor al número de personas enfermas, estas son muy cercanas.
- Existe un sesgo hacia el valor 0 en la variable FastingBS, lo mismo sucede con las mujeres en la variable Sex, los asintomáticos en la variable ChestPainType, los normales en la variable RestingECG, los negativos en la variable ExerciseAngina y la pendiente plana en la variable ST_Slope.

- Dado que una persona tiene una cantidad de azúcar en la sangre durante el ayuno mayor a 120 mg/dl, es más probable que este enferma a que no lo esté, en cambio, si esta cantidad es menor a 120 mg/dl, la probabilidad anterior es aproximada al 50 %.
- El número de hombres enfermos es mayor a las que no lo están, para las mujeres sucede el caso contrario.
- Mayormente las personas enfermas tienden a no presentar síntomas de dolor en el pecho.
- Las personas que están enfermas presentan en mayor cantidad una angina inducida por el ejercicio, mientras que las no están enfermas presentan en una minoría esta angina.
- Dado que la pendiente del segmento ST es ascendente, es más probable que no tenga una enfermedad del corazón a que la tenga, el caso contrario sucede si la pendiente es plana o descendiente.

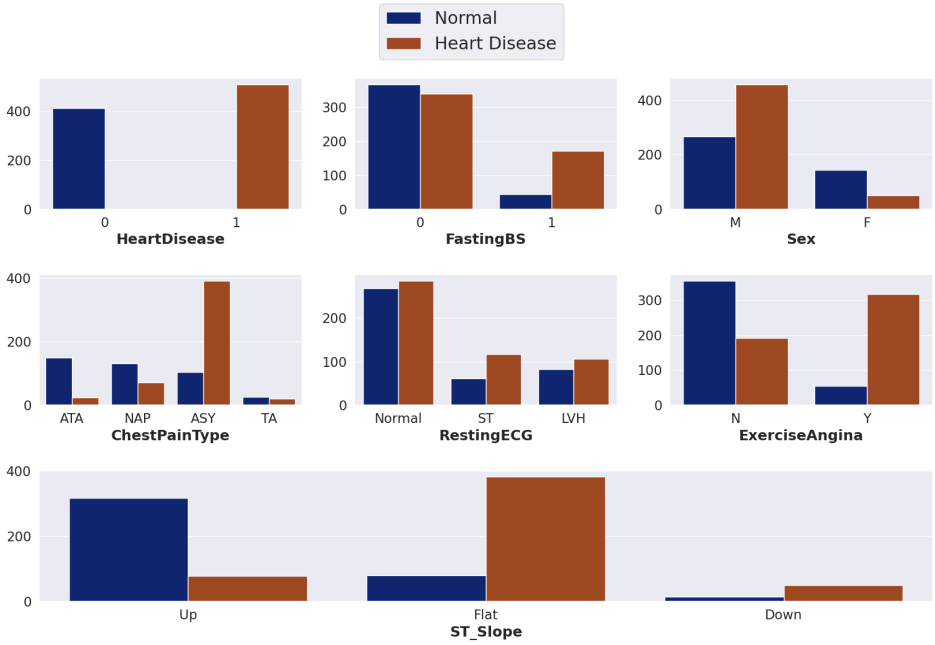

Figura 3.3

Capítulo 4

Ajuste de los modelos y resultados

Al no tener información sobre los datos atípicos, es decir, no sabemos si pueden provenir de errores al momento de la recolección, se decidió eliminarlos, posteriormente se construyeron variables indicadoras para trabajar con las variables categóricas y se aplico el escalador estándar a las variables numéricas, al final se obtuvieron 15 variables explicativas .

Una vez terminado el preprocesamiento de los datos, se ajustó un modelo por cada una de las funciones enlace mencionadas en la sección 2.1.3, además se utilizaron distribuciones no informativas normales para cada uno de los parámetros, es decir, $\beta_0, \beta_1, .., \beta_{15}$ son variables aleatorias i.i.d $N(0, 1000)$. La estimación de las densidades posteriores se obtuvieron en jags, simulando 3 cadenas, con un total de 5000 iteraciones, un periodo de quemado de 2000 y tomando datos para la muestra de 5 en 5 (thin=5).

Para el conjunto de entrenamiento se usaron el 85 % de los datos y el resto se dejaron para la prueba. Se dividió el conjunto de prueba en 10 bloques y se aplicó *cross validation*, para medir el desempeño de los modelos, se calculó el *AUC Score*, la Tabla 4.1 muestra los resultados obtenidos.

	Entrenamiento		**Prueba**	
Modelo	Media	Desviación estándar	Media	Desviación estándar
logit	0.9351	0.0053	0.9157	0.0634
cloglog	0.9336	0.0048	0.9179	0.0585
probit	0.9353	0.0054	0.9157	0.0630
cauchit	0.9322	0.0058	0.9121	0.0639

Tabla 4.1: Media y desviación estándar del *AUC Score*

En general, la media y desviación estándar del *AUC Score*, para ambos conjuntos, son muy parecidas, solamente difiriendo por pocos decimales. También hay que señala que la media del *AUCScore* en todos los modelos es muy buena, ya que es cercana a 1. Por la simplicidad para la interpretación de los parámetros se escogió el modelo *logit*. Las trazas del muestreo y las densidades posteriores se pueden ver en las Figuras 4.1 y 4.2 .

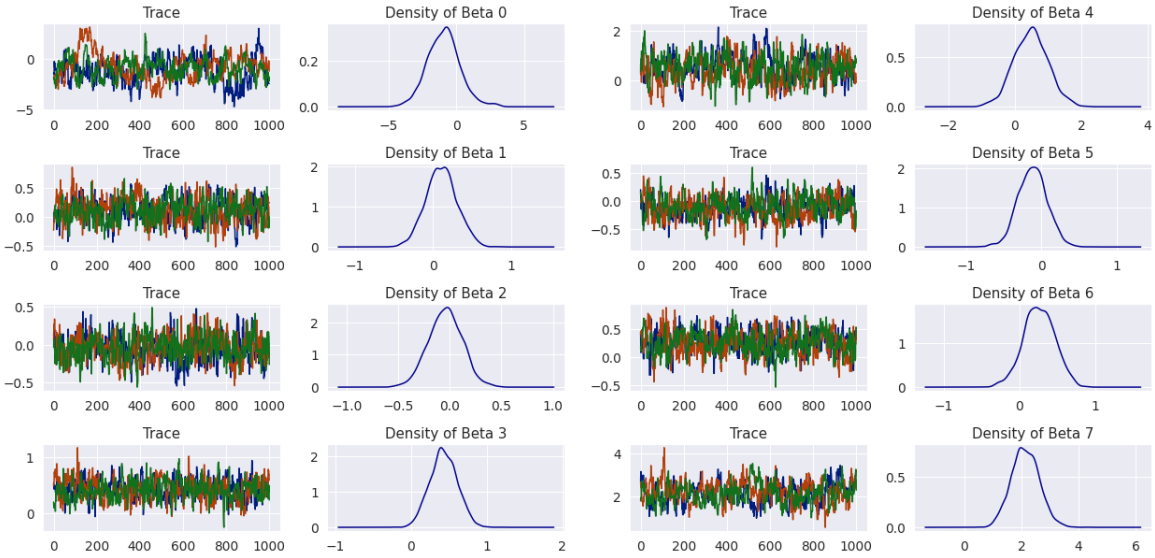

Figura 4.1: Trazas y densidades para los parámetros del modelo.

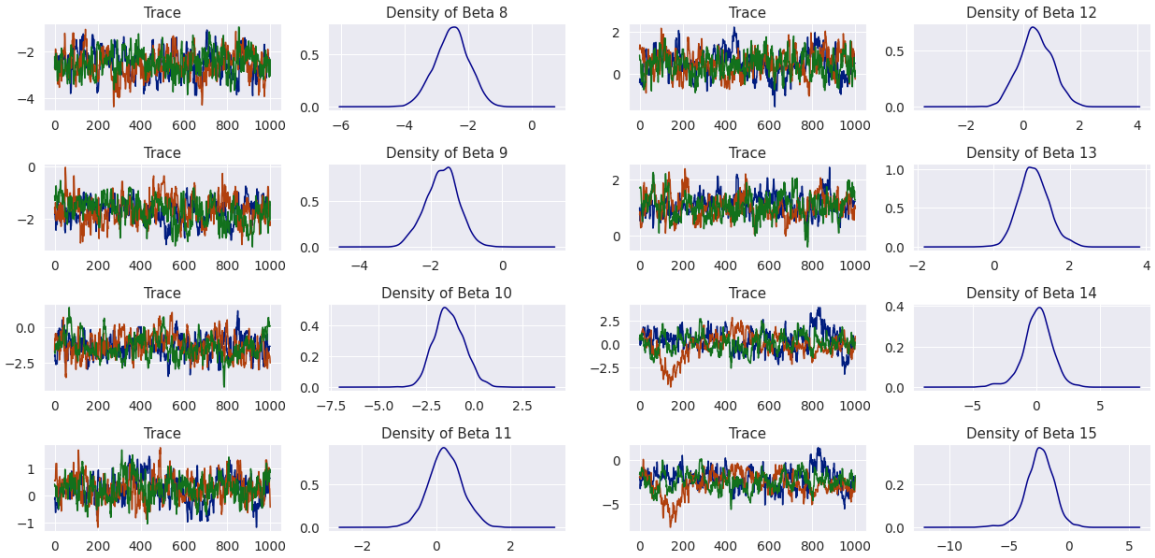

Figura 4.2: Trazas y densidades para los parámetros del modelo.

Al ver la forma de las trazas y las densidades nos damos cuenta que el algoritmo converge, al menos para la mayoría de los parámetros. Otra cosa importante son los estimadores, que se obtuvieron a partir de la media, y los intervalos de credibilidad (ver Tabla 4.2).

Parámetro	**Estimador**	**[0.025**	**0.975]**
Beta 0	-0.94	-3.223	1.615
Beta 1	0.106	-0.273	0.500
Beta 2	-0.042	-0.355	0.272
Beta 3	0.433	0.096	0.779
Beta 4	0.481	-0.546	1.523
Beta 5	-0.114	-0.490	0.251
Beta 6	0.262	-0.154	0.658
Beta 7	2.141	1.213	3.104
Beta 8	-2.505	-3.569	-1.503
Beta 9	-1.683	-2.597	-0.827
Beta 10	-1.317	-2.724	0.219
Beta 11	0.253	-0.596	1.145
Beta 12	0.445	-0.642	1.541
Beta 13	1.056	0.320	1.890
Beta 14	0.116	-2.441	2.224
Beta 15	-2.418	-4.914	-0.264

Tabla 4.2: Estimadores e intervalos del 95 % de credibilidad de los parámetros.

Los intervalos de credibilidad muestran que algunos parámetros, como β_0 o β_4 no son significativos, debido a que sus intervalos de credibilidad contienen al cero. Por último , se calcularon algunas métricas para una secuencia de umbrales(ver Tabla 4.3) y se escogió el que maximizara tanto la especificidad como la sensibilidad para el conjunto de prueba.

Umbral	*Precisión*	*Sensibilidad*	*Especificidad*
0.1	0.862	0.780	0.955
0.2	0.851	0.780	0.932
0.3	**0.851**	**0.840**	**0.864**
0.4	0.851	0.880	0.818
0.5	0.862	0.920	0.795
0.6	0.851	0.920	0.773

Tabla 4.3: *Precisión*, *Sensibilidad*, *Especificidad* calculada para el conjunto de prueba utilizando distintos umbrales.

Capítulo 5

Conclusiones

Después de ajustar distintos modelos, se puede concluir que en general todos son muy buenos, ya que sus valores para la media del *AUC Score* se aproximan al 1. Por la simplicidad que otorga el modelo *logit* para interpretar el valor de los parámetros, se escogió a este como el mejor. Al momento del ajuste del modelo elegido, se concluyó que la mayoría de las trazas convergía, aunque algunos parámetros no eran significativos utilizando un intervalo del 95 % de credibilidad. Además se dedujo que el umbral adecuado para hacer las predicciones es de 0.30, el cual maximizaba la especificidad y la sensibilidad.

También hay que señalar el hecho de que al momento de eliminar los datos atípicos podríamos haber descartando información importante, pero al no saber la procedencia de estos se decidió descartarlos.

Bibliografía

[1] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Second edition. O'Reilly Media, Inc.

[2] Dobson, J.A. & Barnett, A.G. (2008) *An Introduction to Generalized Linear Models*. Third Edition. CHAPMAN & HALL/CRC.

[3] James, G., Witten, D., Hastie, T. & Tibshirani, R. (2017). *An Introduction to Statistical Learning*. Seventh Printing . Springer.

[4] Kuschke, J.K. (2015). *Doing Bayesian Data Analysis: A tutorial with R, JAGS, and Stan*. Second Edition. Elsevier.

[5] Robert, C.P. & Casella, G. (2010) *Introducing Monte Carlo Methods with R*. Springer.

[6] Albarrán Naranjo, L. (2022). *Métodos de Simulación Estocástica*. Universidad Nacional Autónoma de México

[7] Albarrán Naranjo, L. (2022). *Inferencia Bayesiana*. Universidad Nacional Autónoma de México.

[8] Inglis, A., Ahmed, A., Wundervald, B. & Prado, E. (2018) *JAGS: Just Another Gibbs Sampler*.

[9] Plummer, M. (2017). *JAGS Version 4.3.0 user manual*.

[10] Ravelo, M.C. & Balaguer, E.P. (2019). *Funciones de enlace alternativas en modelos de respuesta binomial*.

[11] Gelman, A., Jakulin, A., Pittau, M.G., & Su, Y.S. (2008).*A WEAKLY INFORMATIVE DEFAULT PRIOR DISTRIBUTION FOR LOGISTIC AND OTHER REGRESSION MODELS*. Columbia University, Columbia University, University of Rome, and City University of New York.

[12] (2017). *Enfermedades cardiovasculares*. Organización Mundial de la Salud. https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular\protect\discretionary{\char\hyphenchar\font}{}{}diseases\protect\discretionary{\char\hyphenchar\font}{}{}cvds)#:~:text=Las%20ECV%20son%20la%20principal,muertes%20registradas%20en%20el%20mundo.

[13] SckitLearn. https://scikit\protect\discretionary{\char\hyphenchar\font}{}{}learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html#sklearn.preprocessing.StandardScaler.