LA REGRESÓN LOGÍSTICA PARA LA ESTIMACIÓN DE LA PROBABILIDAD DE EGRESO DE LOS ESTUDIANTES DE INGENIERÍA DE LA UNVES

Prof. MSc. Mario Damián Vázquez

.

mario.vazquez@unves.edu.py

Universidad Nacional de Villarrica del Espíritu Santo – Paraguay
Dirección General de Investigación

Junio – 2019

Índice

Índice

Planteo del Problema

En el Paraguay, hay una enorme diferencia entre la matrícula y el egreso, pero ninguna institución se encarga de analizar los motivos de esta gran diferencia, salvo realizar estadística descriptiva.Con esta trabajo de investigación lo que se realiza es estimar la probabilidad que un estudiante termine su carrera universitaria teniendo en cuenta variables académicas y demográficas que se encuentran en las fichas académicas del estudiante.

Índice

Marco Teórico

Para la modelización de las variables se utiliza la Regresión Logística.

La Regresión Logística

Una transformación de p_i, que es fundamental en la **Regresión Logística**, es la transformación **Logit**. Esta transformación se define en término de p_i, como:

$$g_i = log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k \, x_{ki} \tag{1}$$

donde:

$$1 - p_i = 1 - \frac{e^{\left(\beta_0 + \sum_{k=1}^{p} \beta_k \, x_{ki}\right)}}{1 + e^{\left(\beta_0 + \sum_{k=1}^{p} \beta_k \, x_{ki}\right)}} = \frac{1}{1 + e^{\left(\beta_0 + \sum_{k=1}^{p} \beta_k \, x_{ki}\right)}} \tag{2}$$

De modo que, al hacer la transformación, se tiene un modelo lineal que se denomina **Logit**. Así, el modelo de **Regresión Logística** también se llama **Modelo Logit**.

La Regresión Logística

Una transformación de p_i, que es fundamental en la **Regresión Logística**, es la transformación **Logit**. Esta transformación se define en término de p_i, como:

$$g_i = log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k \, x_{ki} \tag{1}$$

donde:

$$1 - p_i = 1 - \frac{e^{\left(\beta_0 + \sum_{k=1}^{p} \beta_k \, x_{ki}\right)}}{1 + e^{\left(\beta_0 + \sum_{k=1}^{p} \beta_k \, x_{ki}\right)}} = \frac{1}{1 + e^{\left(\beta_0 + \sum_{k=1}^{p} \beta_k \, x_{ki}\right)}} \tag{2}$$

De modo que, al hacer la transformación, se tiene un modelo lineal que se denomina **Logit**. Así, el modelo de **Regresión Logística** también se llama **Modelo Logit**.

La Regresión Logística

Una transformación de p_i, que es fundamental en la **Regresión Logística**, es la transformación **Logit**. Esta transformación se define en término de p_i, como:

$$g_i = log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{k=1}^{p} \beta_k \, x_{ki} \tag{1}$$

donde:

$$1 - p_i = 1 - \frac{e^{\left(\beta_0 + \sum_{k=1}^{p} \beta_k \, x_{ki}\right)}}{1 + e^{\left(\beta_0 + \sum_{k=1}^{p} \beta_k \, x_{ki}\right)}} = \frac{1}{1 + e^{\left(\beta_0 + \sum_{k=1}^{p} \beta_k \, x_{ki}\right)}} \tag{2}$$

De modo que, al hacer la transformación, se tiene un modelo lineal que se denomina **Logit**. Así, el modelo de **Regresión Logística** también se llama **Modelo Logit**.

Estimación por Máxima Verosimilitud

La estimación de los parámetros $\beta_0, \beta_1, \ldots, \beta_p$ de un Modelo de Regresión Logística se realiza por medio del método de estimación de Máxima Verosimilitud (MV), que consiste en obtener los estimadores máximo verosímiles que hacen que se maximice la función de verosimilitud asociada a la muestra.

Prueba de Wald para la significación de un parámetro del modelo

Bajo régimen asintótico, se puede usar la prueba de Wald, basada en la distribución normal, para decidir sobre la significación de la asociación entre la covariable X_k y la variable respuesta Y.
Esta decisión se basa en la prueba de hipótesis

$$H_0: \quad \beta_k = 0$$

$$para\ k = 0, 1, \ , \ldots \ , p$$

$$H_1: \quad \beta_k \neq 0$$

Prueba de Wald para la significación de un parámetro del modelo

Bajo régimen asintótico, se puede usar la prueba de Wald, basada en la distribución normal, para decidir sobre la significación de la asociación entre la covariable X_k y la variable respuesta Y.
Esta decisión se basa en la prueba de hipótesis

$$H_0: \quad \beta_k = 0$$
$$H_1: \quad \beta_k \neq 0$$
$$para\ k = 0, 1, \ , \ldots \ , p$$

Medidas de asociación entre variables categóricas

Considérese, primeramente el siguiente cuadro.

Cuadro : Tabla de Contingencia con probabilidades conjuntas π_{ij} y frecuencias observadas n_{ij} con $i = 1, 2, \ldots, r$; $j = 1, 2, \ldots, c$

X \ Y	$Columna\ 1$	$Columna\ 2$	$\cdots$	$Columna\ c$	$Total$
$Fila\ 1$	$\pi_{11}\,(n_{11})$	$\pi_{12}\,(n_{12})$	$\cdots$	$\pi_{1c}\,(n_{1c})$	$\pi_{1\bullet}\,(n_{1\bullet})$
$Fila\ 2$	$\pi_{21}\,(n_{21})$	$\pi_{22}\,(n_{22})$	$\cdots$	$\pi_{2c}\,(n_{2c})$	$\pi_{2\bullet}\,(n_{2\bullet})$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
$Fila\ r$	$\pi_{r1}\,(n_{r1})$	$\pi_{r2}\,(n_{r2})$	$\cdots$	$\pi_{rc}\,(n_{rc})$	$\pi_{r\bullet}\,(n_{r\bullet})$
$Total$	$\pi_{\bullet 1}\,(n_{\bullet 1})$	$\pi_{\bullet 2}\,(n_{\bullet 2})$	$\cdots$	$\pi_{\bullet c}\,(n_{\bullet c})$	$1\,(n)$

Tau de Goodman y Kruskal

Los autores, sugieren una medida de asociación para variables categóricas nominales, definida por:

$$\tau = \frac{\sum_i \sum_j \frac{\pi_{ij}^2}{\pi_{i\bullet}} - \sum_j \pi_{\bullet j}^2}{1 - \sum_j \pi_{\bullet j}^2} \tag{3}$$

siendo su estimador muestral:

$$\widehat{\tau} = \frac{\sum_i \sum_j \frac{n_{ij}^2}{n_{i\bullet}} - \sum_j n_{\bullet j}^2}{1 - \sum_j n_{\bullet j}^2} \tag{4}$$

$\tau \in [0, 1]$ y su interpretación es similar al coeficiente de determinación.

Tau de Goodman y Kruskal

Los autores, sugieren una medida de asociación para variables categóricas nominales, definida por:

$$\tau = \frac{\sum_i \sum_j \frac{\pi_{ij}^2}{\pi_{i\bullet}} - \sum_j \pi_{\bullet j}^2}{1 - \sum_j \pi_{\bullet j}^2} \tag{3}$$

siendo su estimador muestral:

$$\widehat{\tau} = \frac{\sum_i \sum_j \frac{n_{ij}^2}{n_{i\bullet}} - \sum_j n_{\bullet j}^2}{1 - \sum_j n_{\bullet j}^2} \tag{4}$$

$\tau \in [0, 1]$ y su interpretación es similar al coeficiente de determinación.

Tau de Goodman y Kruskal

Los autores, sugieren una medida de asociación para variables categóricas nominales, definida por:

$$\tau = \frac{\sum_i \sum_j \frac{\pi_{ij}^2}{\pi_{i\bullet}} - \sum_j \pi_{\bullet j}^2}{1 - \sum_j \pi_{\bullet j}^2} \tag{3}$$

siendo su estimador muestral:

$$\widehat{\tau} = \frac{\sum_i \sum_j \frac{n_{ij}^2}{n_{i\bullet}} - \sum_j n_{\bullet j}^2}{1 - \sum_j n_{\bullet j}^2} \tag{4}$$

$\tau \in [0, 1]$ y su interpretación es similar al coeficiente de determinación.

Selección paso a paso (*Stepwise*)

- Según Hosmer, D. y Lemeshow, S. (2000), los criterios para la inclusión de una variable en un modelo pueden variar de un problema a otro y de una disciplina científica a otra. Sin embargo, todos los criterios comparten el mismo enfoque para la construcción de modelos estadísticos, el cual implica buscar el modelo con parsimonia, consistente en un modelo que ajuste bien a los datos con el menor número de variables posibles, logrando de esta manera un equilibrio entre complejidad y precisión.
- Entre los principales procedimientos de selección de variables se tiene el método de selección paso a paso (*Stepwise*), el cual engloba una serie de procedimientos de selección automática de variables significativas, ya sea para la inclusión o exclusión de las mismas en el modelo de forma secuencial, basado únicamente en criterios estadísticos. Este método combina la selección hacia adelante (*Forward*) para incluir una nueva variable y la selección hacia atrás (*Backward*) para la eliminación de una variable.

Selección paso a paso (*Stepwise*)

- Según Hosmer, D. y Lemeshow, S. (2000), los criterios para la inclusión de una variable en un modelo pueden variar de un problema a otro y de una disciplina científica a otra. Sin embargo, todos los criterios comparten el mismo enfoque para la construcción de modelos estadísticos, el cual implica buscar el modelo con parsimonia, consistente en un modelo que ajuste bien a los datos con el menor número de variables posibles, logrando de esta manera un equilibrio entre complejidad y precisión.
- Entre los principales procedimientos de selección de variables se tiene el método de selección paso a paso (*Stepwise*), el cual engloba una serie de procedimientos de selección automática de variables significativas, ya sea para la inclusión o exclusión de las mismas en el modelo de forma secuencial, basado únicamente en criterios estadísticos. Este método combina la selección hacia adelante (*Forward*) para incluir una nueva variable y la selección hacia atrás (*Backward*) para la eliminación de una variable.

Selección paso a paso (*Stepwise*)

- Según Hosmer, D. y Lemeshow, S. (2000), los criterios para la inclusión de una variable en un modelo pueden variar de un problema a otro y de una disciplina científica a otra. Sin embargo, todos los criterios comparten el mismo enfoque para la construcción de modelos estadísticos, el cual implica buscar el modelo con parsimonia, consistente en un modelo que ajuste bien a los datos con el menor número de variables posibles, logrando de esta manera un equilibrio entre complejidad y precisión.
- Entre los principales procedimientos de selección de variables se tiene el método de selección paso a paso (*Stepwise*), el cual engloba una serie de procedimientos de selección automática de variables significativas, ya sea para la inclusión o exclusión de las mismas en el modelo de forma secuencial, basado únicamente en criterios estadísticos. Este método combina la selección hacia adelante (*Forward*) para incluir una nueva variable y la selección hacia atrás (*Backward*) para la eliminación de una variable.

Interpretación de los parámetros en términos de OR

En la formulación del modelo se tiene una serie de coeficientes que son los parámetros, a saber:

- β_0, la ordenada en el origen, y
- $(\beta_1, \ \ldots \ , \beta_p)$, las pendientes, donde p es el número de variables explicativas.

A partir de estos parámetros pueden calcularse los denominados cocientes de ventaja (OR), que serán de mucha utilidad a la hora de interpretar el modelo, definida por:

$$OR = e^{\beta_k} \tag{5}$$

Interpretación de los parámetros en términos de OR

En la formulación del modelo se tiene una serie de coeficientes que son los parámetros, a saber:

- β_0, la ordenada en el origen, y
- $(\beta_1, \ \ldots \ , \beta_p)$, las pendientes, donde p es el número de variables explicativas.

A partir de estos parámetros pueden calcularse los denominados cocientes de ventaja (OR), que serán de mucha utilidad a la hora de interpretar el modelo, definida por:

$$OR = e^{\beta_k} \tag{5}$$

Test de Hosmer–Lemeshow

Hosmer, D. y Lemeshow, S. (2000) proponen hacer uso de un estadístico que lleva sus nombres. Para la construcción de este estadístico, se agrupan las variables explicativas en g grupos o clases (los autores recomiendan 10 grupos basados en los deciles de las probabilidades estimadas $\widehat{p}_i$). Sean n_j el número total de observaciones en el g–ésimo grupo, O_j el número de respuestas $Y = 1$ para el j–ésimo grupo. El estadístico está dado por:

$$\chi^2_{HL} = \sum_{j=1}^{g} \frac{(O_j - E_j)^2}{v_j} \overset{H_0}{\sim} \chi^2_{g-2} \tag{6}$$

donde:

- $E_j = n_j \widehat{\pi}_j$,
- $v_j = n_j \widehat{\pi}_j (1 - \widehat{\pi}_j)$,
- $\widehat{\pi}_j$: es el promedio de las probabilidades estimadas en el j–ésimo grupo, es decir la frecuencia esperada.

La hipótesis nula en este test es que el modelo propuesto ajusta al conjunto de datos observados, por lo tanto, cuanto mayor sea el valor de χ^2_{HL}, peor será el ajuste del modelo.

Test de Hosmer–Lemeshow

Hosmer, D. y Lemeshow, S. (2000) proponen hacer uso de un estadístico que lleva sus nombres. Para la construcción de este estadístico, se agrupan las variables explicativas en g grupos o clases (los autores recomiendan 10 grupos basados en los deciles de las probabilidades estimadas $\widehat{p}_i$). Sean n_j el número total de observaciones en el g–ésimo grupo, O_j el número de respuestas $Y = 1$ para el j–ésimo grupo. El estadístico está dado por:

$$\chi^2_{HL} = \sum_{j=1}^{g} \frac{(O_j - E_j)^2}{v_j} \overset{H_0}{\sim} \chi^2_{g-2} \tag{6}$$

donde:

- $E_j = n_j \widehat{\pi}_j$,
- $v_j = n_j \widehat{\pi}_j (1 - \widehat{\pi}_j)$,
- $\widehat{\pi}_j$: es el promedio de las probabilidades estimadas en el j–ésimo grupo, es decir la frecuencia esperada.

La hipótesis nula en este test es que el modelo propuesto ajusta al conjunto de datos observados, por lo tanto, cuanto mayor sea el valor de χ^2_{HL}, peor será el ajuste del modelo.

Matriz de Confusión

La Matriz de Confusión, también conocida como **Tabla de Clasificación**, se utiliza para evaluar la capacidad de discriminación del modelo ajustado como un indicador de bondad de ajuste. Esta Matriz resulta de la clasificación cruzada de la variable respuesta, Y_i, con una variable dicotómica cuyos valores se derivan de las probabilidades estimadas.
Considérese el Cuadro. En él, se tiene:

Cuadro : Matriz de Confusión

O_i \ E_i	$Y = 1$	$Y = 0$	$Total$
$Y = 1$	n_{11}	n_{12}	$n_{1\bullet}$
$Y = 0$	n_{21}	n_{22}	$n_{2\bullet}$
$Total$	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Índices para medir la bondad de ajuste

- **Tasa de aciertos o precisión**: es el cociente entre las predicciones correctas y el total de predicciones:

$$a = \frac{n_{11} + n_{22}}{n}. \tag{7}$$

- **Sensibilidad**: es la razón entre los valores 1 correctos y el total de valores 1 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 1)$:

$$s = \frac{n_{11}}{n_{11} + n_{12}}. \tag{8}$$

- **Especificidad**: es la razón entre la frecuencia de valores 0 correctos y el total de valores 0 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 0)$:

$$e = \frac{n_{22}}{n_{21} + n_{22}}. \tag{9}$$

Índices para medir la bondad de ajuste

- **Tasa de aciertos o precisión**: es el cociente entre las predicciones correctas y el total de predicciones:

$$a = \frac{n_{11} + n_{22}}{n}. \tag{7}$$

- **Sensibilidad**: es la razón entre los valores 1 correctos y el total de valores 1 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 1)$:

$$s = \frac{n_{11}}{n_{11} + n_{12}}. \tag{8}$$

- **Especificidad**: es la razón entre la frecuencia de valores 0 correctos y el total de valores 0 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 0)$:

$$e = \frac{n_{22}}{n_{21} + n_{22}}. \tag{9}$$

Índices para medir la bondad de ajuste

- **Tasa de aciertos o precisión**: es el cociente entre las predicciones correctas y el total de predicciones:

$$a = \frac{n_{11} + n_{22}}{n}. \tag{7}$$

- **Sensibilidad**: es la razón entre los valores 1 correctos y el total de valores 1 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 1)$:

$$s = \frac{n_{11}}{n_{11} + n_{12}}. \tag{8}$$

- **Especificidad**: es la razón entre la frecuencia de valores 0 correctos y el total de valores 0 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 0)$:

$$e = \frac{n_{22}}{n_{21} + n_{22}}. \tag{9}$$

Índices para medir la bondad de ajuste

- **Tasa de aciertos o precisión**: es el cociente entre las predicciones correctas y el total de predicciones:

$$a = \frac{n_{11} + n_{22}}{n}. \tag{7}$$

- **Sensibilidad**: es la razón entre los valores 1 correctos y el total de valores 1 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 1)$:

$$s = \frac{n_{11}}{n_{11} + n_{12}}. \tag{8}$$

- **Especificidad**: es la razón entre la frecuencia de valores 0 correctos y el total de valores 0 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 0)$:

$$e = \frac{n_{22}}{n_{21} + n_{22}}. \tag{9}$$

Índices para medir la bondad de ajuste

- **Tasa de aciertos o precisión**: es el cociente entre las predicciones correctas y el total de predicciones:

$$a = \frac{n_{11} + n_{22}}{n}. \tag{7}$$

- **Sensibilidad**: es la razón entre los valores 1 correctos y el total de valores 1 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 1)$:

$$s = \frac{n_{11}}{n_{11} + n_{12}}. \tag{8}$$

- **Especificidad**: es la razón entre la frecuencia de valores 0 correctos y el total de valores 0 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 0)$:

$$e = \frac{n_{22}}{n_{21} + n_{22}}. \tag{9}$$

Índices para medir la bondad de ajuste

- **Tasa de aciertos o precisión**: es el cociente entre las predicciones correctas y el total de predicciones:

$$a = \frac{n_{11} + n_{22}}{n}. \tag{7}$$

- **Sensibilidad**: es la razón entre los valores 1 correctos y el total de valores 1 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 1)$:

$$s = \frac{n_{11}}{n_{11} + n_{12}}. \tag{8}$$

- **Especificidad**: es la razón entre la frecuencia de valores 0 correctos y el total de valores 0 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 0)$:

Índices para medir la bondad de ajuste

- **Tasa de aciertos o precisión**: es el cociente entre las predicciones correctas y el total de predicciones:

$$a = \frac{n_{11} + n_{22}}{n}. \tag{7}$$

- **Sensibilidad**: es la razón entre los valores 1 correctos y el total de valores 1 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 1)$:

$$s = \frac{n_{11}}{n_{11} + n_{12}}. \tag{8}$$

- **Especificidad**: es la razón entre la frecuencia de valores 0 correctos y el total de valores 0 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 0)$:

$$e = \frac{n_{22}}{n_{21} + n_{22}}. \tag{9}$$

Índices para medir la bondad de ajuste

- **Tasa de aciertos o precisión**: es el cociente entre las predicciones correctas y el total de predicciones:

$$a = \frac{n_{11} + n_{22}}{n}. \tag{7}$$

- **Sensibilidad**: es la razón entre los valores 1 correctos y el total de valores 1 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 1)$:

$$s = \frac{n_{11}}{n_{11} + n_{12}}. \tag{8}$$

- **Especificidad**: es la razón entre la frecuencia de valores 0 correctos y el total de valores 0 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 0)$:

$$e = \frac{n_{22}}{n_{21} + n_{22}}. \tag{9}$$

Índices para medir la bondad de ajuste

- **Tasa de aciertos o precisión**: es el cociente entre las predicciones correctas y el total de predicciones:

$$a = \frac{n_{11} + n_{22}}{n}. \tag{7}$$

- **Sensibilidad**: es la razón entre los valores 1 correctos y el total de valores 1 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 1)$:

$$s = \frac{n_{11}}{n_{11} + n_{12}}. \tag{8}$$

- **Especificidad**: es la razón entre la frecuencia de valores 0 correctos y el total de valores 0 observados. En otras palabras, es la probabilidad de clasificación correcta $(Y = 0)$:

$$e = \frac{n_{22}}{n_{21} + n_{22}}. \tag{9}$$

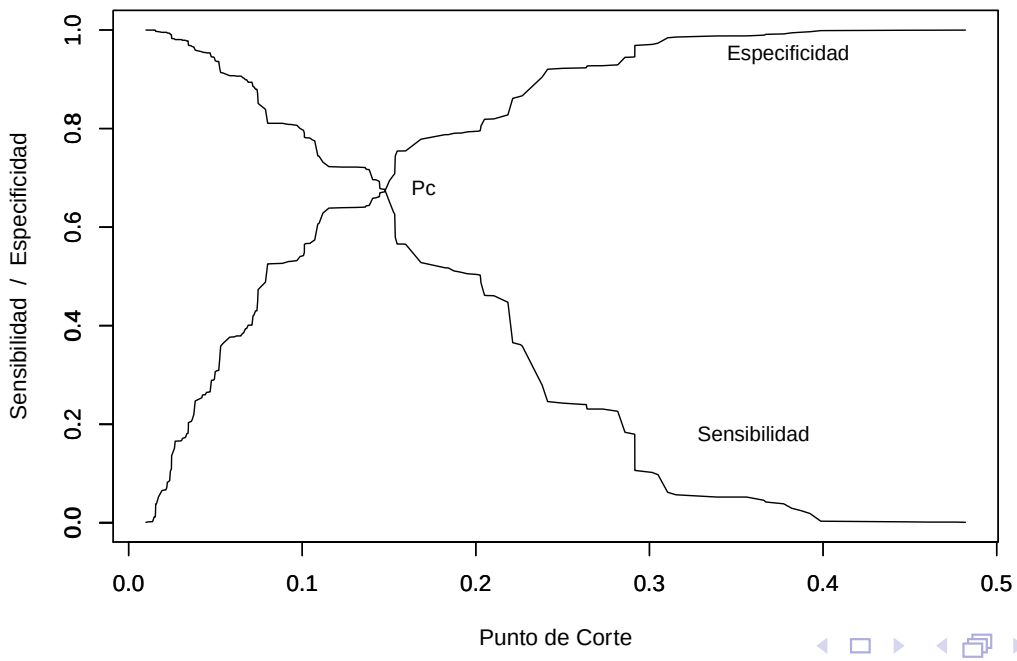
Punto de Corte

Para clasificar a los individuos se fija un punto de corte (p_c) tal que si la probabilidad estimada por el modelo para un individuo es mayor, se clasifica como $Y = 1$, en caso contrario se clasifica como $Y = 0$. Aunque muchas veces se toma $0{,}5$ como el punto de corte, Hosmer, D. y Lemeshow, S. (2000) sugieren que si el objetivo es elegir un punto de corte óptimo para los fines de clasificación, se puede seleccionar un punto de corte que maximiza tanto la sensibilidad como la especificidad.

Punto de Corte

Esta elección se facilita a través de un gráfico como el que se muestra en la siguiente Figura donde se observa la opción óptima para un punto de corte donde aproximadamente la sensibilidad y especificidad se intersecan.

Figura : Ilustración de Sensibilidad y Especificidad versus Punto de Corte de Hosmer–Lemeshow



Análisis de los residuos

Si algunos de los residuos dados por la ecuación

$$\widehat{e}_i = \frac{y_i - \widehat{p}_i}{\sqrt{\widehat{p}_i(1-\widehat{p}_i)}} \tag{10}$$

resultan ser significativos, esto es, residuos superiores a ± 2, debe estudiarse su influencia sobre el ajuste del modelo. Una medida para estudiar la influencia de los residuos significativos es conocida como *Distancia de Cook*.

Análisis de los residuos

Si algunos de los residuos dados por la ecuación

$$\widehat{e}_i = \frac{y_i - \widehat{p}_i}{\sqrt{\widehat{p}_i(1-\widehat{p}_i)}} \tag{10}$$

resultan ser significativos, esto es, residuos superiores a ± 2, debe estudiarse su influencia sobre el ajuste del modelo. Una medida para estudiar la influencia de los residuos significativos es conocida como *Distancia de Cook*.

Distancia de Cook

Cook, R. (1977) introduce una estadística para indicar la influencia de una observación con respecto a un modelo particular. Para una única observación, esta estadística proporciona también información sobre si dicha observación es un outlier, queda definida por:

$$D_i = \frac{(\widehat{\beta} - \widehat{\beta}_i)' \, X'X \, (\widehat{\beta} - \widehat{\beta}_i)}{pS^2} \tag{11}$$

Distancia de Cook

Cook, R. (1977) introduce una estadística para indicar la influencia de una observación con respecto a un modelo particular. Para una única observación, esta estadística proporciona también información sobre si dicha observación es un outlier, queda definida por:

$$D_i = \frac{(\widehat{\beta} - \widehat{\beta}_i)' \, X'X \, (\widehat{\beta} - \widehat{\beta}_i)}{pS^2} \tag{11}$$

Índice

Para el presente trabajo se cuenta con la base de datos de la Dirección General Académica de la UNVES.
En esta base,

- La Muestra está conformada por 1532 estudiantes en el período enero 2008 – diciembre 2018.
- Como la muestra está constituida por personas, no se incluyeron en la base de datos las siguientes variables (nombre, apellido, número de documento de identidad, etc.), con el fin de cuidar el aspecto ético de la investigación.

Para el presente trabajo se cuenta con la base de datos de la Dirección General Académica de la UNVES.
En esta base,

- La Muestra está conformada por 1532 estudiantes en el período enero 2008 – diciembre 2018.
- Como la muestra está constituida por personas, no se incluyeron en la base de datos las siguientes variables (nombre, apellido, número de documento de identidad, etc.), con el fin de cuidar el aspecto ético de la investigación.

Para el presente trabajo se cuenta con la base de datos de la Dirección General Académica de la UNVES.
En esta base,

- La Muestra está conformada por 1532 estudiantes en el período enero 2008 – diciembre 2018.
- Como la muestra está constituida por personas, no se incluyeron en la base de datos las siguientes variables (nombre, apellido, número de documento de identidad, etc.), con el fin de cuidar el aspecto ético de la investigación.

Variable Dependiente

El Modelo de Regresión Logística de respuesta binaria se aplica de modo a clasificar en dos poblaciones (los egresados y los que no son egresados).
Así, la variable dependiente se define por:

$$Y_i = \begin{cases} 1 & \text{si el individuo ``}i\text{'' egresado} \\ 0 & \text{si el individuo ``}i\text{'' no es egresado} \end{cases} \quad \text{con } i = 1, 2, 3, \ \ldots \ , 1532$$

Por tanto, Y_i sigue una distribución Bernoulli con parámetro p_i, con $p_i = P(Y_i = 1)$.

Variables Independientes

- **Sexo**
  - Masculino
  - Femenino

Variables Independientes

- **Estado Civil**
  - Soltero
  - Casado
  - Separado

Variables Independientes

- **Cantidad de materias aprobadas en el primer curso**: Se considera la cantidad de la cantidad de materias aprobadas por el estudiante en el primer año la carrera. Para los fines del estudio, esta variable se ha categorizado en los siguientes niveles:
  - Sin materias aprobadas.
  - 1 o 2 materias aprobadas.
  - 3 o 4 materias aprobadas.
  - 5 o más materias aprobadas.

Variables Independientes

- **Edad del estudiante**: Es el tiempo transcurrido, medido en años, desde la fecha que nació el socio hasta la fecha que fue su 'ultima matriculación en la carrera. Para los fines del estudio, esta variable que originalmente es continua, se ha categorizado en los siguientes niveles:
  - Edad entre 18 y 20 años.
  - Edad entre 21 y 25 años.
  - Edad entre 26 y 30 años.
  - Edad $\geq$ 31 años.

Variables Independientes

- **Tipo de Ingeniería**: Es el tipo de ingeniería en el cual el estudiante se matricula:
  - Eléctrica.
  - Informática.
  - Ambiental.
  - Agronindustria.
  - Zootecnia.

El software R y la Regresión Logística

El análisis de los resultados se realiza a través del software libre `R` en su versión 3.5.3, con los siguientes paquetes:
`MASS`,
`xtable`,
`ROCR` y
`ResourceSelection`([a]).

[a] Para más información ver la página oficial https://www.r-project.org

Índice

Medidas de asociación entre las variables

Cuadro : Medidas de asociación entre las variables explicativas categóricas.

	Materias Ap.	Sexo	Edad	Estado Civil	Ingeniería
Materias Ap.	$\widehat{\gamma} = 1$	$\widehat{\gamma} = 0{,}05$	$\widehat{\gamma} = 0{,}08$	$\widehat{\gamma} = 0{,}1$	$\widehat{\gamma} = 0{,}11$
Sexo		$\widehat{\gamma} = 1$	$\widehat{\gamma} = 0{,}1$	$\widehat{\gamma} = 0{,}03$	$\widehat{\gamma} = 0{,}001$
Edad			$\widehat{\gamma} = 1$	$\widehat{\gamma} = 0{,}12$	$\widehat{\gamma} = 0{,}02$
Estado Civil				$\widehat{\gamma} = 1$	$\widehat{\gamma} = 0{,}03$
Ingeniería					$\widehat{\gamma} = 1$

Selección *Stepwise* de las variables en el Modelo

Para seleccionar las variables que definen el mejor modelo se utiliza el procedimiento conocido como *Stepwise*, siguiendo las recomendaciones de Hosmer, D. y Lemeshow, S. (2000). Para facilitar la escritura del código de los diferentes modelos en el paquete estadístico R, se realiza la siguiente reparametrización de variables.

Selección *Stepwise* de las variables en el Modelo

- $Y =$ Estudiante egresado o no egresado
- $X_1 =$ Número de Materias aprobadas en el primer curso
- $X_2 =$ Sexo
- $X_3 =$ Edad
- $X_4 =$ Estado Civil
- $X_5 =$ Tipo de Ingeniería

Selección *Stepwise* de las variables en el Modelo

- $Y =$ Estudiante egresado o no egresado
- $X_1 =$ Número de Materias aprobadas en el primer curso
- $X_2 =$ Sexo
- $X_3 =$ Edad
- $X_4 =$ Estado Civil
- $X_5 =$Tipo de Ingeniería

Selección *Stepwise* de las variables en el Modelo

- $Y =$ Estudiante egresado o no egresado
- $X_1 =$ Número de Materias aprobadas en el primer curso
- $X_2 =$ Sexo
- $X_3 =$ Edad
- $X_4 =$ Estado Civil
- $X_5 =$Tipo de Ingeniería

Selección *Stepwise* de las variables en el Modelo

- $Y =$ Estudiante egresado o no egresado
- $X_1 =$ Número de Materias aprobadas en el primer curso
- $X_2 =$ Sexo
- $X_3 =$ Edad
- $X_4 =$ Estado Civil
- $X_5 =$Tipo de Ingeniería

Selección *Stepwise* de las variables en el Modelo

- $Y =$ Estudiante egresado o no egresado
- $X_1 =$ Número de Materias aprobadas en el primer curso
- $X_2 =$ Sexo
- $X_3 =$ Edad
- $X_4 =$ Estado Civil
- $X_5 =$Tipo de Ingeniería

Selección *Stepwise* de las variables en el Modelo

- $Y =$ Estudiante egresado o no egresado
- $X_1 =$ Número de Materias aprobadas en el primer curso
- $X_2 =$ Sexo
- $X_3 =$ Edad
- $X_4 =$ Estado Civil
- $X_5 =$Tipo de Ingeniería

Selección *Stepwise* de las variables en el Modelo

- $Y =$ Estudiante egresado o no egresado
- $X_1 =$ Número de Materias aprobadas en el primer curso
- $X_2 =$ Sexo
- $X_3 =$ Edad
- $X_4 =$ Estado Civil
- $X_5 =$Tipo de Ingeniería

En resumen

La Regresión Logística incorpora cuatro (4) variables al considerar su relación con la variable dependiente. Que son la cantidad de materias aprobadas en el primer semestre, el sexo, el estado civil y el tipo de ingeniería.

Test de Hosmer–Lemeshow

La hipótesis nula de este test es que el modelo propuesto es el apropiado para explicar la probabilidad que un estudiante sea egresado, con lo cual lo conveniente es no rechazarla.

Siendo $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{4i} + \beta_4 X_{5i}$, el resumen del test se muestra en el siguiente cuadro:

Cuadro : Test de Hosmer–Lemeshow para la bondad de ajuste del modelo final propuesto

χ^2	gl	Pr(> Chi)
12,1862	8	0,1431

Test de Hosmer–Lemeshow

La hipótesis nula de este test es que el modelo propuesto es el apropiado para explicar la probabilidad que un estudiante sea egresado, con lo cual lo conveniente es no rechazarla.

Siendo $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{4i} + \beta_4 X_{5i}$, el resumen del test se muestra en el siguiente cuadro:

Cuadro : Test de Hosmer–Lemeshow para la bondad de ajuste del modelo final propuesto

χ^2	gl	Pr(> Chi)
12,1862	8	0,1431

Test de Hosmer–Lemeshow

La hipótesis nula de este test es que el modelo propuesto es el apropiado para explicar la probabilidad que un estudiante sea egresado, con lo cual lo conveniente es no rechazarla.

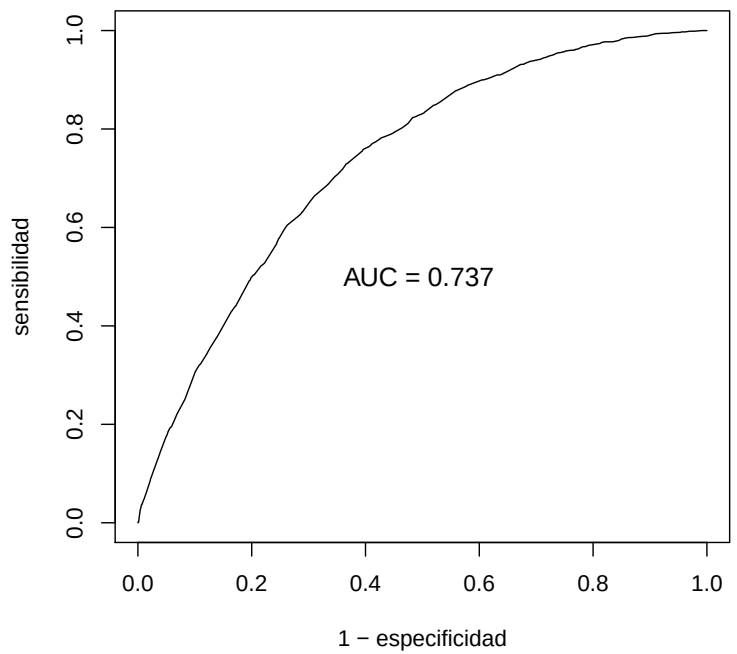
Siendo $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{4i} + \beta_4 X_{5i}$, el resumen del test se muestra en el siguiente cuadro:

Cuadro : Test de Hosmer–Lemeshow para la bondad de ajuste del modelo final propuesto

χ^2	gl	Pr(> Chi)
12,1862	8	0,1431

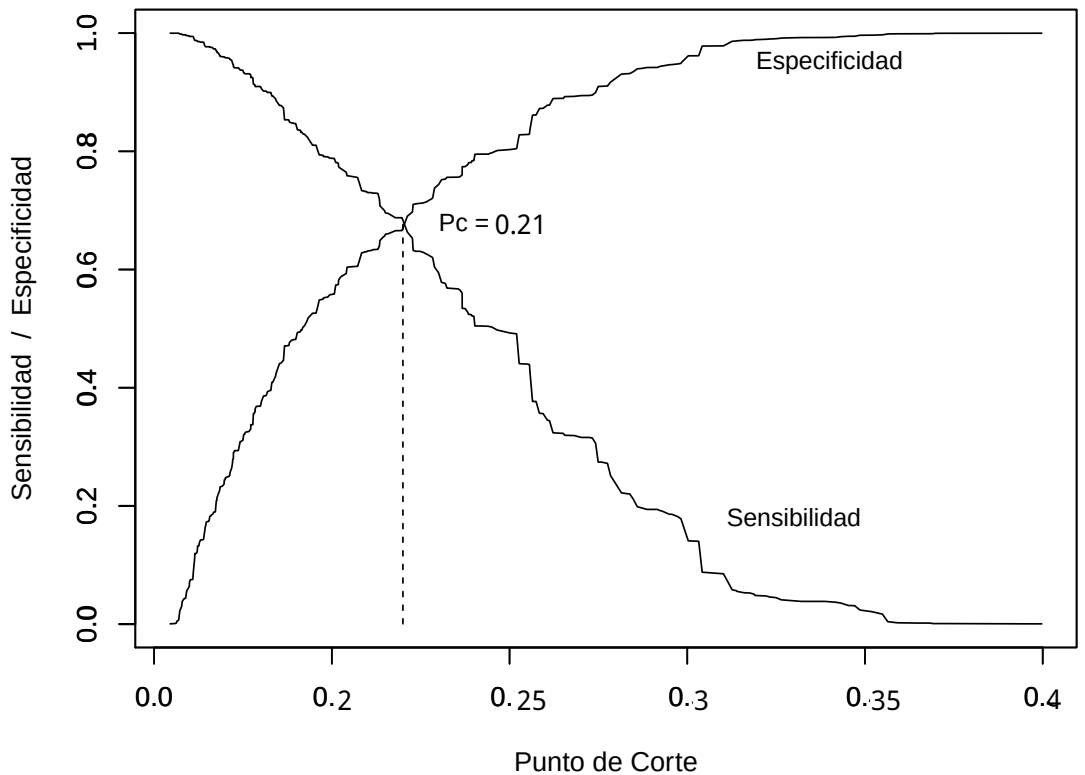
Área bajo la curva ROC

Figura : Curva ROC del modelo final ajustado por Regresión Logística.



Punto de corte

Figura : Sensibilidad y Especificidad versus Punto de Corte del modelo final ajustado por Regresión Logística.



Tasa de clasificaciones correctas

Cuadro : Índices para medir bondad de ajuste del modelo final ajustado por Regresión Logística.

Precisión	*Sensibilidad*	*Especificidad*
0,7254	0,7257	0,7252

Punto de corte = 0,21

Tasa de clasificaciones correctas

Cuadro : Índices para medir bondad de ajuste del modelo final ajustado por Regresión Logística.

Precisión	*Sensibilidad*	*Especificidad*
$0{,}7254$	$0{,}7257$	$0{,}7252$
Punto de corte $= 0{,}21$		

Análisis de los Residuos

Cuadro : Cuartiles de los residuos estimados del modelo final ajustado por Regresión Logística

Deviance Residuals				
Min	1Q	Median	3Q	Max
$-1{,}1772$	$-0{,}5736$	$-0{,}3768$	$-0{,}2375$	$2{,}9342$

Análisis de los Residuos

Cuadro : Cuartiles de los residuos estimados del modelo final ajustado por Regresión Logística

Deviance Residuals				
Min	1Q	Median	3Q	Max
$-1{,}1772$	$-0{,}5736$	$-0{,}3768$	$-0{,}2375$	$2{,}9342$

Análisis de los Residuos

Al observar el Cuadro, el valor máximo indica que existen residuos (en valor absoluto), que exceden el valor 2, éstos a su vez, podrían señalar que existen valores observados que afectan el ajuste global del modelo.
Más puntualmente, de 1532 residuos, 8 son mayores a 2 en valor absoluto, es decir, alrededor de un $0{,}52\,\%$.

Distancia de Cook

El valor máximo hallado es aproximadamente $0{,}0034$, menor al valor límite. Por tanto, ninguna observación es potencialmente influyente en el buen ajuste del modelo final estimado por Regresión Logística, y podemos dar por validado el modelo.

Análisis de los Residuos

Al observar el Cuadro, el valor máximo indica que existen residuos (en valor absoluto), que exceden el valor 2, éstos a su vez, podrían señalar que existen valores observados que afectan el ajuste global del modelo.
Más puntualmente, de 1532 residuos, 8 son mayores a 2 en valor absoluto, es decir, alrededor de un $0{,}52\,\%$.

Distancia de Cook

El valor máximo hallado es aproximadamente $0{,}0034$, menor al valor límite. Por tanto, ninguna observación es potencialmente influyente en el buen ajuste del modelo final estimado por Regresión Logística, y podemos dar por validado el modelo.

Análisis de los Residuos

Al observar el Cuadro, el valor máximo indica que existen residuos (en valor absoluto), que exceden el valor 2, éstos a su vez, podrían señalar que existen valores observados que afectan el ajuste global del modelo.
Más puntualmente, de 1532 residuos, 8 son mayores a 2 en valor absoluto, es decir, alrededor de un $0{,}52\,\%$.

Distancia de Cook

El valor máximo hallado es aproximadamente $0{,}0034$, menor al valor límite. Por tanto, ninguna observación es potencialmente influyente en el buen ajuste del modelo final estimado por Regresión Logística, y podemos dar por validado el modelo.

Índice

Conclusiones

Desde el punto de vista inferencial queda demostrado que las variables (*Número de materias aprobadas en el primer curso, el sexo, el estado civil y el tipo de ingeniería*) son las que contribuyen a la construcción de un modelo matemático (*Modelo de Regresión Logística*) que permite estimar la probabilidad de egreso de un estudiante de ingeniería de la UNVES.

Conclusiones

Desde el punto de vista inferencial queda demostrado que las variables (*Número de materias aprobadas en el primer curso, el sexo, el estado civil y el tipo de ingeniería*) son las que contribuyen a la construcción de un modelo matemático (*Modelo de Regresión Logística*) que permite estimar la probabilidad de egreso de un estudiante de ingeniería de la UNVES.

Conclusiones

- La probabilidad de ser egresado aumenta en todas las categorías de la variable *Número de materias aprobadas en el primer curso* lo que significa que los estudiantes con más materias aprobadas tienen mayor probabilidad de ser egresados que aquellos con poca materias aprobadas. Esto puede interpretarse de la siguiente manera, las personas que aprueban la mayoría o todas las materias, tienen mayor chance de ser egresados.
- Con respecto a la variable *Tipo de Ingeniería*, la probabilidad de ser egresado aumenta cuando pasa de la categoría de referencia "Eléctrica" a las otras 4 categorías "Informática", "Ambiental", "Agroindustria", "Zootecnia" , lo que significa que los estudiantes que no sean de Ingeniería Eléctrica tienen mayor chance de ser egresados.
- Con respecto a la variable *Estado Civil*, la probabilidad de ser egresado aumenta cuando pasa de la categoría de referencia "Casado" a las otras dos categorías "Soltero" y "Separado"; y en la variable *Sexo* cuando también cuando pasa de la categoría de referencia "Femenino" a "Masculino" lo que significa que los estudiantes masculinos solteros o separados tienen mayor chance de ser egresados que aquellas que son casadas.

Conclusiones

- La probabilidad de ser egresado aumenta en todas las categorías de la variable *Número de materias aprobadas en el primer curso* lo que significa que los estudiantes con más materias aprobadas tienen mayor probabilidad de ser egresados que aquellos con poca materias aprobadas. Esto puede interpretarse de la siguiente manera, las personas que aprueban la mayoría o todas las materias, tienen mayor chance de ser egresados.
- Con respecto a la variable *Tipo de Ingeniería*, la probabilidad de ser egresado aumenta cuando pasa de la categoría de referencia "Eléctrica" a las otras 4 categorías "Informática", "Ambiental", "Agroindustria", "Zootecnia", lo que significa que los estudiantes que no sean de Ingeniería Eléctrica tienen mayor chance de ser egresados.
- Con respecto a la variable *Estado Civil*, la probabilidad de ser egresado aumenta cuando pasa de la categoría de referencia "Casado" a las otras dos categorías "Soltero" y "Separado"; y en la variable *Sexo* cuando también cuando pasa de la categoría de referencia "Femenino" a "Masculino" lo que significa que los estudiantes masculinos solteros o separados tienen mayor chance de ser egresados que aquellas que son casadas.

Conclusiones

- La probabilidad de ser egresado aumenta en todas las categorías de la variable *Número de materias aprobadas en el primer curso* lo que significa que los estudiantes con más materias aprobadas tienen mayor probabilidad de ser egresados que aquellos con poca materias aprobadas. Esto puede interpretarse de la siguiente manera, las personas que aprueban la mayoría o todas las materias, tienen mayor chance de ser egresados.
- Con respecto a la variable *Tipo de Ingeniería*, la probabilidad de ser egresado aumenta cuando pasa de la categoría de referencia "Eléctrica" a las otras 4 categorías "Informática", "Ambiental", "Agroindustria", "Zootecnia" , lo que significa que los estudiantes que no sean de Ingeniería Eléctrica tienen mayor chance de ser egresados.
- Con respecto a la variable *Estado Civil*, la probabilidad de ser egresado aumenta cuando pasa de la categoría de referencia "Casado" a las otras dos categorías "Soltero" y "Separado"; y en la variable *Sexo* cuando también cuando pasa de la categoría de referencia "Femenino" a "Masculino" lo que significa que los estudiantes masculinos solteros o separados tienen mayor chance de ser egresados que aquellas que son casadas.

Conclusiones

- La probabilidad de ser egresado aumenta en todas las categorías de la variable *Número de materias aprobadas en el primer curso* lo que significa que los estudiantes con más materias aprobadas tienen mayor probabilidad de ser egresados que aquellos con poca materias aprobadas. Esto puede interpretarse de la siguiente manera, las personas que aprueban la mayoría o todas las materias, tienen mayor chance de ser egresados.
- Con respecto a la variable *Tipo de Ingeniería*, la probabilidad de ser egresado aumenta cuando pasa de la categoría de referencia "Eléctrica" a las otras 4 categorías "Informática", "Ambiental", "Agroindustria", "Zootecnia" , lo que significa que los estudiantes que no sean de Ingeniería Eléctrica tienen mayor chance de ser egresados.
- Con respecto a la variable *Estado Civil*, la probabilidad de ser egresado aumenta cuando pasa de la categoría de referencia "Casado" a las otras dos categorías "Soltero" y "Separado"; y en la variable *Sexo* cuando también cuando pasa de la categoría de referencia "Femenino" a "Masculino" lo que significa que los estudiantes masculinos solteros o separados tienen mayor chance de ser egresados que aquellas que son casadas.

Conclusiones

- Clasificando a los socios con probabilidad de ser morosos mayor a $0{,}21$ como socios morosos, se obtiene para el modelo los índices para medir la bondad de ajuste del modelo final estimado por Regresión Logística, siendo la Precisión igual a $67{,}64\,\%$, la Especificidad igual a $67{,}63\,\%$ y la Sensibilidad igual a $67{,}69\,\%$, esto es, el modelo identifica 2 de cada 3 morosos aproximadamente, por lo que la capacidad predictiva del modelo matemático propuesto es aceptable. Similarmente, el valor del área bajo la curva ROC, que es igual a $0{,}737$, supera el valor mínimo de $0{,}7$ referenciado, el que permite considerar que el modelo tiene capacidad de discriminación también aceptable.
- El modelo propuesto es validado mediante el análisis de los residuos. Sólo el $0{,}55\,\%$ de estos residuos ajustados están fuera del intervalo ± 2, pero ninguno de estos errores es potencialmente influyente en el modelo, ya que el cálculo de las distancias de $Cook$ arroja un valor máximo de $0{,}0034$ (menor al valor límite dado por 1).

Conclusiones

- Clasificando a los socios con probabilidad de ser morosos mayor a $0{,}21$ como socios morosos, se obtiene para el modelo los índices para medir la bondad de ajuste del modelo final estimado por Regresión Logística, siendo la Precisión igual a $67{,}64\,\%$, la Especificidad igual a $67{,}63\,\%$ y la Sensibilidad igual a $67{,}69\,\%$, esto es, el modelo identifica 2 de cada 3 morosos aproximadamente, por lo que la capacidad predictiva del modelo matemático propuesto es aceptable. Similarmente, el valor del área bajo la curva ROC, que es igual a $0{,}737$, supera el valor mínimo de $0{,}7$ referenciado, el que permite considerar que el modelo tiene capacidad de discriminación también aceptable.
- El modelo propuesto es validado mediante el análisis de los residuos. Sólo el $0{,}55\,\%$ de estos residuos ajustados están fuera del intervalo ± 2, pero ninguno de estos errores es potencialmente influyente en el modelo, ya que el cálculo de las distancias de $Cook$ arroja un valor máximo de $0{,}0034$ (menor al valor límite dado por 1).

Conclusiones

- Clasificando a los socios con probabilidad de ser morosos mayor a $0{,}21$ como socios morosos, se obtiene para el modelo los índices para medir la bondad de ajuste del modelo final estimado por Regresión Logística, siendo la Precisión igual a $67{,}64\,\%$, la Especificidad igual a $67{,}63\,\%$ y la Sensibilidad igual a $67{,}69\,\%$, esto es, el modelo identifica 2 de cada 3 morosos aproximadamente, por lo que la capacidad predictiva del modelo matemático propuesto es aceptable. Similarmente, el valor del área bajo la curva ROC, que es igual a $0{,}737$, supera el valor mínimo de $0{,}7$ referenciado, el que permite considerar que el modelo tiene capacidad de discriminación también aceptable.
- El modelo propuesto es validado mediante el análisis de los residuos. Sólo el $0{,}55\,\%$ de estos residuos ajustados están fuera del intervalo ± 2, pero ninguno de estos errores es potencialmente influyente en el modelo, ya que el cálculo de las distancias de $Cook$ arroja un valor máximo de $0{,}0034$ (menor al valor límite dado por 1).

Conclusiones

Finalmente, para dar cumplimiento a otro de los objetivos de este trabajo, el de plantear posibles acciones para evitar el bajo egreso de los estudiantes, es razonable realizar las siguientes sugerencias:

- Instaurar en la Universidad *Políticas de utilización de Modelos Matemáticos para la Predicción de la Egreso de los estudiantes.*
- Realizar, en años venideros *estudios similares a los efectos de verificar la permanencia (o no) de las variables que componen el modelo*. Si se instauran políticas de control, pueden modificarse y/o agregarse nuevas variables.

Conclusiones

Finalmente, para dar cumplimiento a otro de los objetivos de este trabajo, el de plantear posibles acciones para evitar el bajo egreso de los estudiantes, es razonable realizar las siguientes sugerencias:

- Instaurar en la Universidad *Políticas de utilización de Modelos Matemáticos para la Predicción de la Egreso de los estudiantes*.
- Realizar, en años venideros *estudios similares a los efectos de verificar la permanencia (o no) de las variables que componen el modelo*. Si se instauran políticas de control, pueden modificarse y/o agregarse nuevas variables.

Conclusiones

Finalmente, para dar cumplimiento a otro de los objetivos de este trabajo, el de plantear posibles acciones para evitar el bajo egreso de los estudiantes, es razonable realizar las siguientes sugerencias:

- Instaurar en la Universidad *Políticas de utilización de Modelos Matemáticos para la Predicción de la Egreso de los estudiantes*.
- Realizar, en años venideros *estudios similares a los efectos de verificar la permanencia (o no) de las variables que componen el modelo*. Si se instauran políticas de control, pueden modificarse y/o agregarse nuevas variables.

AGUYJE (MUCHAS GRACIAS(

AGUYJE (MUCHAS GRACIAS(

Índice

Referencias Bibliográficas

AGRESTI, A. (2002)
"Data Categorical Analysis"(Second Edition)
Addison-Wesley.

HOSMER, D. y LEMESHOW, S. (2000)
"Applied Logistic Regression"(Second Edition)
Addison-Wesley.

CONTACTO

Prof. MSc. Mario Vázquez
mario.vazquez@unves.edu.py
mscientiae@gmail.com
+595 972 284 338 (Whatsapp)