UNIVERSIDAD DE EL SALVADOR
FACULTAD MULDISCIPLINARIA DE OCCIDENTE

DEPARTAMENTO DE MATEMÁTICA

LICENCIATURA EN ESTADÍSTICA

APLICACIÓN DE REGRESIÓN LOGÍSTICA MULTINOMIAL

DOCENTE:

Licdo. Jaime Isacc Peña

CÁTEDRA: Proyecto de Estudios Estadísticos

Autor:

Oscar Mauricio Rodríguez Reyes

Santa Ana, 11 de Agosto de 2023

Indice de contenidos

Regresión Logística Multinomial

La regresión logística multinomial es un tipo de modelo de regresión utilizado para predecir una variable categórica con más de dos categorías. A diferencia de la regresión logística binomial, que se utiliza para predecir dos categorías, la regresión logística multinomial se utiliza cuando la variable dependiente tiene tres o más categorías discretas. Es una extensión de la regresión logística que permite modelar relaciones entre variables predictoras y una variable de respuesta categórica con múltiples niveles. Por ejemplo, puede usarse para predecir el tipo de persona de acuerdo a la preferenca alimenticia que han sido catalogadas como vegano, omnívoro y carnívoro.

Criterios a Considerar para Aplicar la Técnica

1. Es recomendable que la variable dependiente sea nominal.
2. Las variables independientes o predictoras que se vayan a incluir en el modelo deben ser numéricas, si se dispone de variables categóricas, deben ser transformadas en variables indicadoras.
3. No debe haber presencia de multicolinealidad.
4. Debe haber linealidad entre la variable dependiente con las variables independientes.

Definición de Conceptos Básicos

Prueba de hipótesis de coeficientes

En regresión logística, las hipótesis son de interés:

La hipótesis nula, que es cuando todos los coeficientes de la ecuación de regresión toman el valor cero, y

La hipótesis alternativa de que el modelo actualmente en consideración es preciso y difiere significativamente del nulo de cero, es decir, da significativamente mejor que el nivel de predicción aleatoria o aleatoria de la hipótesis nula.

Prueba de razón de verosimilitud

La razón de verosimilitud es una estadística utilizada en la regresión logística multinomial para comparar dos modelos diferentes: el modelo nulo (sin variables predictoras) y el modelo con las variables predictoras. Se utiliza para evaluar si el modelo con las variables predictoras es significativamente mejor que el modelo nulo.

La razón de verosimilitud se basa en la razón -2LL y ayuda a determinar si el conjunto de variables predictoras proporciona información útil para predecir la variable dependiente. Un valor alto de la razón de verosimilitud sugiere que el modelo con las variables predictoras es una mejora significativa sobre el modelo nulo.

Las hipótesis a considerar son las siguientes

- H_0 : No hay diferencia entre el modelo nulo y el modelo final.
- H_1 : Hay diferencia entre el modelo nulo y el modelo final.

Supuesto L: multicolinealidad

Simplemente ejecute una "regresión lineal" después de asumir la variable dependiente categórica como variable continua.

Si el VIF (factor de inflación de la varianza) más grande es mayor que 10, entonces hay un motivo de preocupación (Bowerman y O'Connell, 1990).

- La tolerancia por debajo de 0,1 indica un problema grave.
- La tolerancia por debajo de 0,2 indica un problema potencial (Menard, 1995).
- Si el índice de condición es mayor que 15, se asume la multicolinealidad.

Características de la regresión logística multinomial

Regresión logística multinomial para predecir la pertenencia a más de dos categorías. (Básicamente) funciona de la misma manera que la regresión logística binaria. El análisis divide la variable de resultado en una serie de comparaciones entre dos categorías.

Por ejemplo, si tiene tres categorías de resultados (A, B y C), el análisis consistirá en dos comparaciones que elija:

- Compare todo con su primera categoría (por ejemplo, A frente a B y A frente a C),
- O su última categoría (por ejemplo, A frente a C y B frente a C),
- O una categoría personalizada (por ejemplo, B frente a A y B frente a C). (Peña, 2021)

Odds Ratio

El Odds Ratio (OR) es una medida que compara las probabilidades de que un evento ocurra en dos grupos diferentes. En la regresión logística multinomial, se utilizan Odds Ratios para medir la relación entre las categorías de la variable dependiente y las variables predictoras. Ayuda a entender cómo la presencia o ausencia de una categoría en una variable dependiente está asociada con las variables predictoras. Un OR mayor que 1 indica que la probabilidad de una categoría es más alta en un grupo en comparación con otro, mientras que un OR menor que 1 indica lo contrario.

Aplicación de La Técnica

La base de datos que se trabajará corresponde a información relacionada al rendimiento académico que los estudiantes del Instituto Nacional Cornelio Azenón Sierra llevan en la asignatura de Matemática.

El objetivo es determinar si el tipo de ambiente familiar en que vive un alumno, la nota previa que este lleve en Matemática y su percepción de temas matemáticos explican el rendimiento que el alumno presenta en la asignatura de matemática; clasificado como rendimiento bajo, medio o alto.

Las variables que contiene la base se describen a continuación:

- **Rendimiento:** Corresponde a la clasificación en que se encuentra el estudiante respecto a la nota promedio de matemática (niveles: 1 - bajo, 2 - medio y 3 - alto).
- **I9:** Ambiente familiar que el estudiante vive en su hogar (1 - agradable, 2 - autoritario y 3 - violento)
- **I17:** Nota promedio en matemática que el estudiante obtuvo en su año de estudio anterior.
- **I21:** Contiene las respuestas a la pregunta ¿Cómo considera que es su percepción para entender los contenidos matemáticos? (1 - Excelente, 2 - Muy Buena, 3 - Buena, 4 - Mala)

Aplicación del Método Usando el Software Estadístico SPSS

Inspeción de los datos

La base de datos cuenta con 106 observaciones, tal como se muestra en la siguente figura, consta de 4 variables dónde solo una es numérica mientras que el resto son categóricas, además, no tiene presencia de datos clasificados como *missing.*

Selección del modelo

Para la aplicación de la técnica en este software se debe seguir la siguiente secuencia: *Analizar>Regresión>Logística Multinomial.* Luego se abre la ventana que se muestra en la Figura 2; en esta se debe agregar la variable dependiente junto a la categoría de referencia en este caso es Rendimiento Medio (2); en el área de factores se deben agregar las variables independientes categóricas nominales y en la parte de covariables se agregan las variables independientes numéricas.

	Rendimiento	I9	I17	I21
1	Medio	Agradable	7.00	Muy Buena
2	Medio	Autoritario	7.00	Buena
3	Medio	Agradable	7.00	Muy Buena
4	Bajo	Violento	6.00	Buena
5	Medio	Autoritario	7.00	Buena
6	Medio	Autoritario	7.50	Buena
7	Medio	Agradable	9.00	Muy Buena
8	Medio	Agradable	7.00	Muy Buena
9	Medio	Agradable	7.00	Buena
10	Medio	Autoritario	8.00	Buena
11	Bajo	Violento	6.50	Mala
12	Medio	Agradable	7.00	Buena
13	Medio	Agradable	7.00	Muy Buena
14	Medio	Autoritario	6.50	Buena
15	Alto	Agradable	10.00	Excelente
16	Bajo	Autoritario	7.00	Buena

Figura 1: Inspección de la Base de Datos

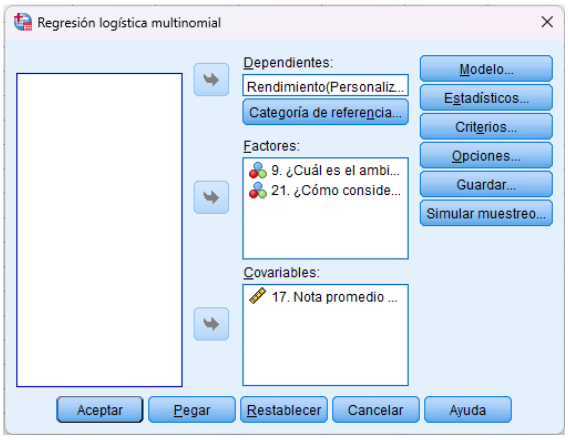

Figura 2: Selección de las variables en el modelo

Para el presente caso, se trabajará un modelo que considera solo los efectos principales de las variables. Ahora lo que sigue es la interpretación de los resultados.

Comprobacioón de Multicolinealidad

De acuerdo a la figura 3, se observa que el valor de VIF para cada variable predictora es menor que 10, por tanto se concluye que no hay multicolinealidad.

Ajuste del Modelo

En la figura 4 se muestra una tabla que contiene el ajuste del modelo seleccionado, con base al criterio AIC es evidente que el agregar variables se explica de mejor manera la variabilidad de los datos de la variable dependiente que el modelo que contiene solo la constante. Lo expuesto también se puede verificar con base al p valor que se muestra en la última columna, pues este es menor que 0.05 lo cual indica que el modelo final se ajusta significativamente mejor que el modelo sin predictores.

Modelo		Estadísticas de colinealidad	
		Tolerancia	VIF
1	9. ¿Cuál es el ambiente familiar que se vive dentro de su hogar?	.827	1.210
	17. Nota promedio que obtuvo en Matemática en su año de estudio anterior	.767	1.305
	21. ¿Cómo considera su grado percepción de los temas trabajados en Matemática?	.701	1.426

a. Variable dependiente: Rendimiento

Figura 3: Prueba de Multicolinealidad

	Criterios de ajuste de modelo			Pruebas de la razón de verosimilitud		
Modelo	AIC	normalizado	Logaritmo de la verosimilitud -2	Chi-cuadrado	gl	Sig.
Sólo intersección	165.598	170.925	161.598			
Final	56.497	93.785	28.497	133.102	12	.000

Figura 4: Información del Ajuste del Modelo

En la figura 5 se muestra una prueba de bondad de ajuste en el cual se busca que el p valor que se encuentra en la última columna sea mayor que 0.05. Con base a la desvianza, se obtiene un p valor prácticamente de 1, por tanto, esto implica que los valores predichos mediante el modelo no difieren significativamente de los valores observados. Por tanto, existe un buen ajuste del modelo.

	Chi-cuadrado	gl	Sig.
Pearson	27.254	68	1.000
Desvianza	19.858	68	1.000

Figura 5: Bondad de Ajuste

Pseudo R-Cuadrado

En la figura 6, con base al valor que se presenta en el índice de corrección de Nagelkerke, se concluye que existe una fuerte relación del 86.5 % entre los predictores y la predicción, es decir, el ambiente familiar, la nota promedio de Matemática en años previos y la percepción de los contenidos matemáticos están relacionados en un 86.5% con el rendimiento académico que un alumno presente en la asignatura de Matemática.

Razón de Verosimilitud por variable

Es necesario verificar si las variables incluidas al modelo, son significativas para este. Con base a las pruebas de razón de verosimilitud que se muestran en la figura 7, teniendo como referencia la última columna, se concluye que tanto la nota previa, el ambiente familiar y la percepción de temas matemáticos aportan información significativa al modelo dado que los p valores que se obtuvieron para estas es menor que 0.05.

Cox y Snell	.715
Nagelkerke	.865
McFadden	.717

Figura 6: Pseudo R-Cuadrado

Efecto	Criterios de ajuste de modelo: AIC de modelo reducido	BIC de modelo reducido	Logaritmo de la verosimilitud -2 de modelo reducido	Pruebas de la razón de verosimilitud: Chi-cuadrado	gl	Sig.
Intersección	56.497	93.785	28.497[a]	.000	0	.
17. Nota promedio que obtuvo en Matemática en su año de estudio anterior	64.798	96.759	40.798	12.301	2	.002
9. ¿Cuál es el ambiente familiar que se vive dentro de su hogar?	61.628	88.262	41.628	13.131	4	.011
21. ¿Cómo considera su grado percepción de los temas trabajados en Matemática?	122.039	143.347	106.039	77.543	6	.000

El estadístico de chi-cuadrado es la diferencia de la log-verosimilitud -2 entre el modelo final y el modelo reducido. El modelo reducido se forma omitiendo un efecto del modelo final. La hipótesis nula es que todos los parámetros de dicho efecto son 0.

a. Este modelo reducido es equivalente al modelo final porque omitir el efecto no aumenta los grados de libertad.

Figura 7: Razón de Verosimilitud para las variables

Parámetros del modelo

En la figura 8 y 9 se muestran los valores correspondientes a los β, Wald, significancia estadística y los Old ratio (Exp(B)) de cada variable. En este caso se busca que los valores de Wald no sean cero y que el p valor sea menor a 0.05 para que el predictor haga un aporte significativo a la variable dependiente que en este caso es el Rendimiento Académico en Matemática

En la relación de Rendimiento Bajo con Rendimiento medio, se ha obtenido que la nota previa es un predictor significativo. El valor de β que le corresponde es de -1.273, Wald de 5.186 y Old Ratio de 0.280 (Exp(B)) , dado que el valor de β es negativo, implica que a medida que el promedio en Matemática aumenta en una unidad, existe una razón de probabilidad de 0.28 de pertenecer a un grupo de rendimiento bajo. En otras palabras, entre mayor sea el promedio en matemática es menos probable que tengamos un rendimiento bajo.

Rendimiento[a]		B	Desv. Error	Wald	gl	Sig.	Exp(B)	95% de intervalo de confianza para Exp(B): Límite inferior	Límite superior
Bajo	Intersección	49.094	6230.740	.000	1	.994			
	17. Nota promedio que obtuvo en Matemática en su año de estudio anterior	-1.273	.559	5.186	1	.023	.280	.094	.837
	[9. ¿Cuál es el ambiente familiar que se vive dentro de su hogar?=1]	-20.305	6230.738	.000	1	.997	1.519E-9	.000	.[b]
	[9. ¿Cuál es el ambiente familiar que se vive dentro de su hogar?=2]	-21.197	6230.738	.000	1	.997	6.228E-10	.000	.[b]

Figura 8: Estimaciones de los Parámetros (Rendimiento Bajo con Rendimiento Medio)

En la relación de Rendimiento Medio con Rendimiento Alto (Figura 8), se ha obtenido que el ambiente

familiar agradable (1) es un predictor significativo. El valor de β que le corresponde es de -19.805, Wald de 97.563 y Old Ratio de $2.504E-9$ (Exp(B)) , dado que el valor de β es negativo, significa que si el ambiente familiar en que vive el alumno cambia a un ambiente que no sea agradable, existe una razón de probabilidad de $2.504E-9$ de pertenecer a un grupo de rendimiento Alto. En otras palabras, si el ambiente familiar en que vive el alumno es agradable, es bastante probable que este presente un rendimiento académico alto.

Rendimiento[a]		B	Desv. Error	Wald	gl	Sig.	Exp(B)	95% de intervalo de confianza para Exp(B)	
								Límite inferior	Límite superior
Alto	Intersección	2.815	14533.620	.000	1	1.000			
	17. Nota promedio que obtuvo en Matemática en su año de estudio anterior	1.793	1.041	2.968	1	.085	6.009	.781	46.220
	[9. ¿Cuál es el ambiente familiar que se vive dentro de su hogar?=1]	-19.805	2.005	97.563	1	.000	2.504E-9	4.919E-11	1.275E-7

Figura 9: Estimaciones de los Parámetros (Rendimiento Medio con Rendimiento Alto)

Precisión de aciertos del modelo

Finalmente, en la figura 10 se presenta una tabla resumen de la precisión del modelo. Se observa que se clasificó correctamente el 91.5% de los datos.

	Pronosticado			
Observado	Bajo	Medio	Alto	Porcentaje correcto
Bajo	9	8	0	52.9%
Medio	0	70	0	100.0%
Alto	0	1	18	94.7%
Porcentaje global	8.5%	74.5%	17.0%	91.5%

Figura 10: Resumen de Clasificación

En los siguientes 2 capítulos se aplicará siempre una Regresión Logística Multinomial a la misma base de datos con la diferencia que en este caso se usará el Software Estadístico R y el lenguaje de programación de Python.

Aplicación del Método Usando el Software Estadístico R

Librerías

Las librerías utilizadas para este caso son:

```
# Tratamiento de datos
library(foreign)
library(jmv)
library(summarytools)
library(gmodels)

# Modelado y Ajuste
library(nnet)
library(DescTools)
library(lmtest)

# Gráficos
library(ggplot2)
```

Datos

Se importa la base de datos que se encuentra en el directorio de trabajo y se guarda en un objeto que se llamará *RendimientoAcademico*, luego se crea una copia de estos datos.

```
RendimientoAcademico <- read.spss("RendAcad_Seminario.sav")
RendAcad <- as.data.frame(RendimientoAcademico)
```

Exploración de la base de datos

```
head(RendAcad)
```

```
  Rendimiento          I9 I17       I21
1       Medio   Agradable 7.0 Muy Buena
2       Medio Autoritario 7.0     Buena
3       Medio   Agradable 7.0 Muy Buena
4        Bajo    Violento 6.0     Buena
5       Medio Autoritario 7.0     Buena
6       Medio Autoritario 7.5     Buena
```

Se muestra un resumen con medidas estadísticas de las variables, recordando que solo la variable *I*17 es numérica, la cual corresponde a la nota previa en la asignatura de matemática. Se observa que se tiene una nota previa promedio de 7.79

```
summary(RendAcad)
```

```
 Rendimiento          I9          I17               I21
 Bajo :17    Agradable  :83   Min.   : 5.000   Excelente:16
 Medio:70    Autoritario:18   1st Qu.: 7.000   Muy Buena:33
 Alto :19    Violento   : 5   Median : 8.000   Buena    :51
                              Mean   : 7.794   Mala     : 6
                              3rd Qu.: 8.375
                              Max.   :10.000
```

Otra forma de tener un análisis descriptivo de las variables es de la siguiente manera

```
descriptives(RendAcad, freq = T)
```

```
DESCRIPTIVES

Descriptives

                        Rendimiento    I9     I17         I21

  N                             106    106         106    106
  Missing                         0      0           0      0
  Mean                                         7.794340
  Median                                       8.000000
  Standard deviation                           1.175450
  Minimum                                      5.000000
  Maximum                                      10.00000



FREQUENCIES

Frequencies of Rendimiento

  Levels    Counts    % of Total    Cumulative %

  Bajo          17      16.03774        16.03774
  Medio         70      66.03774        82.07547
  Alto          19      17.92453       100.00000



Frequencies of I9

  Levels         Counts    % of Total    Cumulative %

  Agradable          83      78.30189        78.30189
  Autoritario        18      16.98113        95.28302
  Violento            5       4.71698       100.00000



Frequencies of I21

  Levels       Counts    % of Total    Cumulative %

  Excelente        16      15.09434        15.09434
  Muy Buena        33      31.13208        46.22642
  Buena            51      48.11321        94.33962
  Mala              6       5.66038       100.00000
```

Tabla de contingencia para la variable dependiente (Rendimiento) y la variable independiente I9 (Ambiente Familiar)

```
CrossTable(RendAcad$I9, RendAcad$Rendimiento)
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  106


             | RendAcad$Rendimiento
 RendAcad$I9 |      Bajo |     Medio |      Alto | Row Total |
-------------|-----------|-----------|-----------|-----------|
   Agradable |         8 |        58 |        17 |        83 |
             |     2.119 |     0.186 |     0.303 |           |
             |     0.096 |     0.699 |     0.205 |     0.783 |
             |     0.471 |     0.829 |     0.895 |           |
             |     0.075 |     0.547 |     0.160 |           |
-------------|-----------|-----------|-----------|-----------|
 Autoritario |         4 |        12 |         2 |        18 |
             |     0.429 |     0.001 |     0.466 |           |
             |     0.222 |     0.667 |     0.111 |     0.170 |
             |     0.235 |     0.171 |     0.105 |           |
             |     0.038 |     0.113 |     0.019 |           |
-------------|-----------|-----------|-----------|-----------|
    Violento |         5 |         0 |         0 |         5 |
             |    21.978 |     3.302 |     0.896 |           |
             |     1.000 |     0.000 |     0.000 |     0.047 |
             |     0.294 |     0.000 |     0.000 |           |
             |     0.047 |     0.000 |     0.000 |           |
-------------|-----------|-----------|-----------|-----------|
Column Total |        17 |        70 |        19 |       106 |
             |     0.160 |     0.660 |     0.179 |           |
-------------|-----------|-----------|-----------|-----------|
```

Tabla de contingencia para la variable dependiente (Rendimiento) y la variable independiente I21 (Percepción de temas matemáticos)

```
CrossTable(RendAcad$I21, RendAcad$Rendimiento)
```

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
```

```
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  106


             | RendAcad$Rendimiento
RendAcad$I21 |      Bajo |     Medio |      Alto | Row Total |
-------------|-----------|-----------|-----------|-----------|
   Excelente |         0 |         0 |        16 |        16 |
             |     2.566 |    10.566 |    60.131 |           |
             |     0.000 |     0.000 |     1.000 |     0.151 |
             |     0.000 |     0.000 |     0.842 |           |
             |     0.000 |     0.000 |     0.151 |           |
-------------|-----------|-----------|-----------|-----------|
   Muy Buena |         2 |        28 |         3 |        33 |
             |     2.048 |     1.768 |     1.437 |           |
             |     0.061 |     0.848 |     0.091 |     0.311 |
             |     0.118 |     0.400 |     0.158 |           |
             |     0.019 |     0.264 |     0.028 |           |
-------------|-----------|-----------|-----------|-----------|
       Buena |         9 |        42 |         0 |        51 |
             |     0.082 |     2.056 |     9.142 |           |
             |     0.176 |     0.824 |     0.000 |     0.481 |
             |     0.529 |     0.600 |     0.000 |           |
             |     0.085 |     0.396 |     0.000 |           |
-------------|-----------|-----------|-----------|-----------|
        Mala |         6 |         0 |         0 |         6 |
             |    26.374 |     3.962 |     1.075 |           |
             |     1.000 |     0.000 |     0.000 |     0.057 |
             |     0.353 |     0.000 |     0.000 |           |
             |     0.057 |     0.000 |     0.000 |           |
-------------|-----------|-----------|-----------|-----------|
Column Total |        17 |        70 |        19 |       106 |
             |     0.160 |     0.660 |     0.179 |           |
-------------|-----------|-----------|-----------|-----------|
```

Gráfico de Cajas y Bigotes para el rendimiento y la nota previa en matemática (I17)

```
ggplot(RendAcad, aes(x = Rendimiento, y = I17, fill=Rendimiento)) +
  geom_boxplot() + labs(x='Rendimiento', y='Nota Previa en Matemática') +
  theme_bw() + theme(legend.position = "none")
```

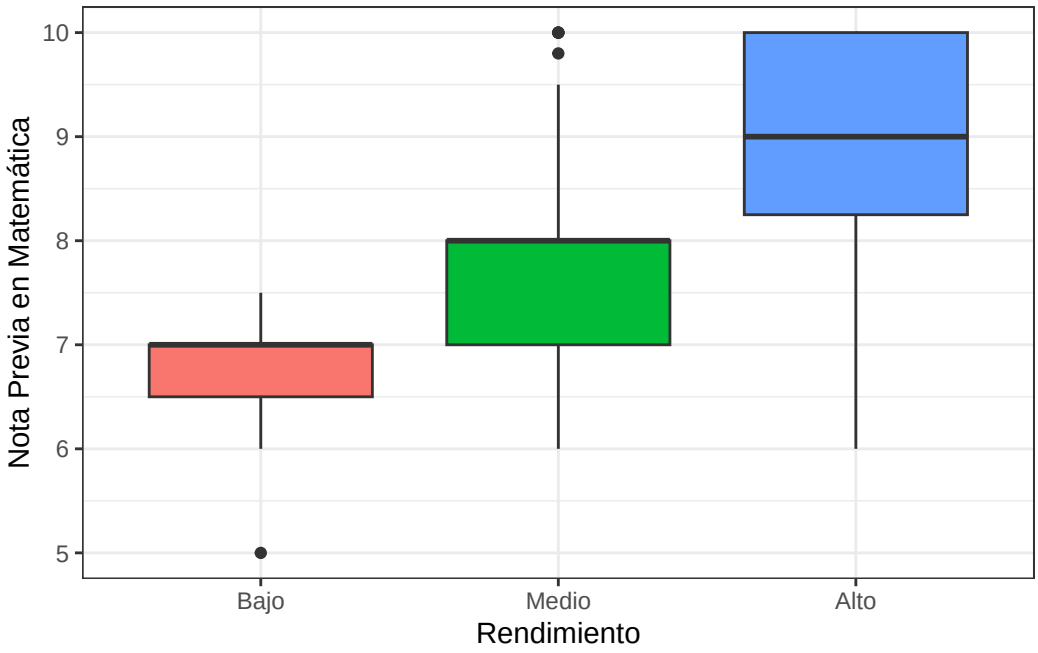

Figura 11: Gráfico de cajas y bigotes para el Rendimiento y Nota Previa

Gráfico de Cajas y Bigotes para la nota previa en matemática (I17)

```
ggplot(RendAcad, aes(y = I17)) +
  geom_boxplot(fill = 'orange') + labs(y='Nota Previa en Matemática') +
  theme_bw()
```

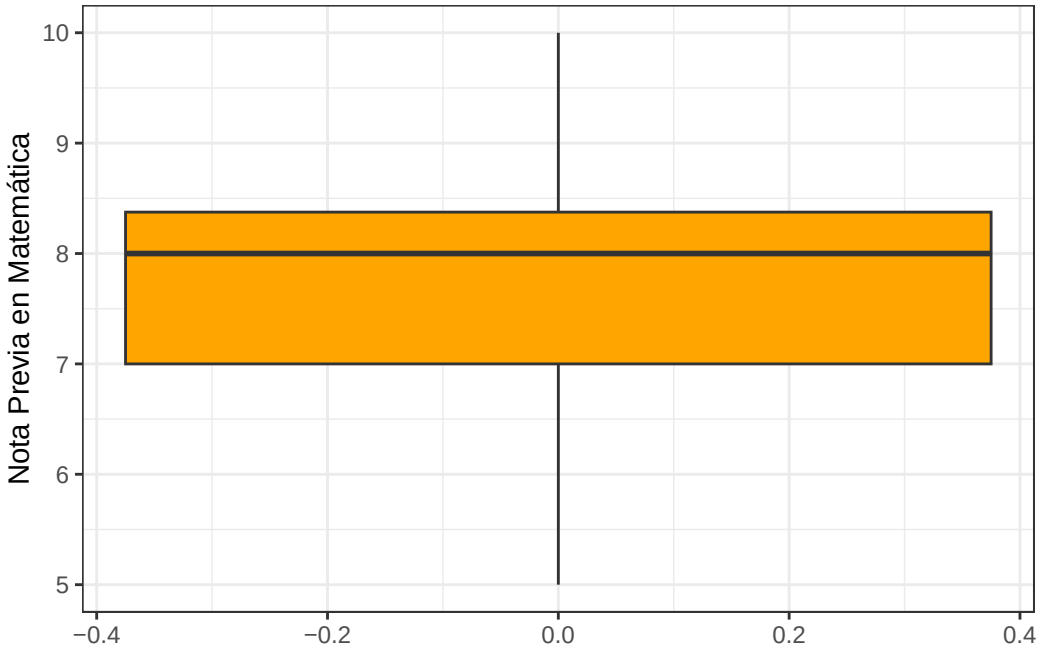

Figura 12: Gráfico de cajas y bigotes para la Nota Previa

Preparación de los datos

Dado que solo una variable es numérica, es necesario que las variables categóricas sean transformadas en nuevas variables indicadoras. Dado que se trabajará con el Rendimiento Medio como referencia, es necesario la variable dependiente sea recodificada.

```
# Recodificar (2 - Medio)
RendAcad$Rendimiento <- relevel(as.factor(RendAcad$Rendimiento), ref = 2)

# Creando variables indicadoras
RendAcad$I9 <-  relevel(as.factor(RendAcad$I9), ref = 2)
RendAcad$I21 <- relevel(as.factor(RendAcad$I21), ref = 3)

# Asignar etiquetas
levels(RendAcad$Rendimiento) <- c("Medio", "Bajo", "Alto")
levels(RendAcad$I9) <- c("Autoritario","Agradable", "Violento")
levels(RendAcad$I21) <- c("Buena", "Excelente", "Muy Buena", "Mala")
```

Ajuste del Modelo

Se crea un modelo sin predictores para posteriormente realizar comparaciones

```
ModIni <- multinom(Rendimiento ~ 1, data = RendAcad)
```

```
# weights:  6 (2 variable)
initial  value 116.452903
final  value 92.820910
converged
```

```
summary(ModIni)
```

```
Call:
multinom(formula = Rendimiento ~ 1, data = RendAcad)

Coefficients:
     (Intercept)
Bajo   -1.415133
Alto   -1.303932

Std. Errors:
     (Intercept)
Bajo   0.2703773
Alto   0.2586771

Residual Deviance: 185.6418
AIC: 189.6418
```

Se crea el modelo de interés con los predictores

```
ModFin <- multinom(Rendimiento ~ I9 + I17 + I21, data = RendAcad, model = T)
```

```
# weights:  24 (14 variable)
initial  value 116.452903
iter  10 value 27.292456
```

```
iter  20 value 26.318583
iter  30 value 26.272201
iter  40 value 26.272015
iter  50 value 26.271766
final  value 26.269992
converged
```

```
summary(ModFin)
```

```
Call:
multinom(formula = Rendimiento ~ I9 + I17 + I21, data = RendAcad,
    model = T)

Coefficients:
     (Intercept) I9Agradable I9Violento       I17 I21Excelente I21Muy Buena
Bajo    6.565651    0.888481  14.421190 -1.273215    -3.075926   -0.7379364
Alto  -26.966089   -3.158467   1.784618  1.794149    34.651071   11.0147561
       I21Mala
Bajo 14.515895
Alto  2.807975

Std. Errors:
     (Intercept) I9Agradable   I9Violento       I17 I21Excelente I21Muy Buena
Bajo    3.989851    1.178823 2.351728e+02 0.5589838    0.1243594    0.9095305
Alto  190.384750    2.006145 2.023721e-03 1.0414538    3.8039013  190.1904142
          I21Mala
Bajo 2.487367e+02
Alto 2.949532e-03

Residual Deviance: 52.53998
AIC: 80.53998
```

```
anova(ModIni, ModFin)
```

```
Likelihood ratio tests of Multinomial Models

Response: Rendimiento
           Model Resid. df Resid. Dev   Test    Df LR stat. Pr(Chi)
1              1       210  185.64182
2 I9 + I17 + I21       198   52.53998 1 vs 2    12 133.1018       0
```

En el modelo sin predictores se obtuvo un valor para AIC de 189.64 mientras que para el modelo con predictores un valor de 80.54 esto indica que es mejor el modelo con los predictores que sin ellos. Esto también se verifica en la prueba de razón de verosimilitud chi-cuadrado porque se obtuvo un valor de 133.10 con un p valor menor que 0.05.

Ahora bien, se sabe que resultó más significativo el modelo con los predictores. Lo que sigue es analizar este modelo.

Con la siguiente prueba de bondad de ajuste Chi Cuadrado, como el p_valor obtenido es menor que 0.05, existe evidencia estadística para concluir que los valores predichos no difieren significativamente de los reales.

```
# Prueba de bondad de ajuste
chisq.test(RendAcad$Rendimiento,predict(ModFin))
```

```
    Pearson's Chi-squared test

data:  RendAcad$Rendimiento and predict(ModFin)
X-squared = 149.58, df = 4, p-value < 2.2e-16
```

Pseudo R-Cuadrado

```
PseudoR2(ModFin, which = c ("CoxSnell", "Nagelkerke", "McFadden"))
```

```
  CoxSnell Nagelkerke   McFadden
 0.7151173  0.8652792  0.7169820
```

Con base al valor que se presenta en el índice de corrección de Nagelkerke, se concluye que existe una fuerte relación del 86.5 % entre los predictores y la predicción.

Pruebas de razón de verosimilitud por variable

```
lrtest(ModFin, "I9")
```

```
# weights:  18 (10 variable)
initial  value 116.452903
iter  10 value 34.266998
iter  20 value 32.853501
iter  30 value 32.845142
iter  40 value 32.840627
final  value 32.836532
converged
Likelihood ratio test

Model 1: Rendimiento ~ I9 + I17 + I21
Model 2: Rendimiento ~ I17 + I21
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  14 -26.270
2  10 -32.837 -4 13.133    0.01064 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En el nivel de significancia del 0.05, dado que se obtuvo un p valor de 0.01 se concluye que la variable $I9$ aporta información significativa para el modelo.

```
lrtest(ModFin, "I17")
```

```
# weights:  21 (12 variable)
initial  value 116.452903
iter  10 value 32.736624
iter  20 value 32.421311
final  value 32.420510
converged
Likelihood ratio test

Model 1: Rendimiento ~ I9 + I17 + I21
Model 2: Rendimiento ~ I9 + I21
  #Df  LogLik Df  Chisq Pr(>Chisq)
```

```
1  14 -26.270
2  12 -32.421 -2 12.301   0.002132 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En el nivel de significancia del 0.05, dado que se obtuvo un p valor de 0.002 se concluye que la variable $I17$ aporta información significativa para el modelo.

```
lrtest(ModFin, "I21")
```

```
# weights:  15 (8 variable)
initial  value 116.452903
iter  10 value 65.181896
iter  20 value 65.053657
iter  30 value 65.042897
final  value 65.041330
converged
Likelihood ratio test

Model 1: Rendimiento ~ I9 + I17 + I21
Model 2: Rendimiento ~ I9 + I17
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  14 -26.270
2   8 -65.041 -6 77.543  1.149e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En el nivel de significancia del 0.05, dado que se obtuvo un p valor de $1.15e-14$ se concluye que la variable $I21$ aporta información significativa para el modelo.

Estimación de Parámetros

```
summary(ModFin)
```

```
Call:
multinom(formula = Rendimiento ~ I9 + I17 + I21, data = RendAcad,
    model = T)

Coefficients:
     (Intercept) I9Agradable I9Violento       I17 I21Excelente I21Muy Buena
Bajo    6.565651    0.888481  14.421190 -1.273215    -3.075926   -0.7379364
Alto  -26.966089   -3.158467   1.784618  1.794149    34.651071   11.0147561
       I21Mala
Bajo 14.515895
Alto  2.807975

Std. Errors:
     (Intercept) I9Agradable   I9Violento       I17 I21Excelente I21Muy Buena
Bajo    3.989851    1.178823 2.351728e+02 0.5589838    0.1243594    0.9095305
Alto  190.384750    2.006145 2.023721e-03 1.0414538    3.8039013  190.1904142
         I21Mala
Bajo 2.487367e+02
Alto 2.949532e-03
```

```
Residual Deviance: 52.53998
AIC: 80.53998
```

```
# Valor Z asociado a Wald
z <- summary(ModFin)$coefficients/summary(ModFin)$standard.errors
z
```

```
     (Intercept) I9Agradable   I9Violento       I17 I21Excelente I21Muy Buena
Bajo    1.645588   0.7537021   0.06132168 -2.277731   -24.734166  -0.81133772
Alto   -0.141640  -1.5743958 881.84976764  1.722735     9.109351   0.05791436
          I21Mala
Bajo   0.05835848
Alto 952.00693205
```

```
# Prueba de 2 colas
p <- (1 - pnorm(abs(z), 0, 1)) * 2
p
```

```
     (Intercept) I9Agradable I9Violento        I17 I21Excelente I21Muy Buena
Bajo   0.0998486   0.4510281   0.951103 0.02274259            0    0.4171718
Alto   0.8873644   0.1153960   0.000000 0.08493642            0    0.9538168
       I21Mala
Bajo 0.9534631
Alto 0.0000000
```

Es de recordar que en este caso se trabajó con el Rendimiento Medio como categoría de referencia, por tal motivo los resultados devuelven 2 filas; en la primera relaciona Rendimiento Bajo con Rendimiento Medio y en la segunda Rendimiento Medio con Rendimiento Alto. Por otro lado, en la prueba de 2 colas (p), se observa que la variable I17 (*Nota Previa*) es significativa ($pvalor = 0.02$) para clasificar entre Rendimiento Bajo y Rendimiento Medio; de igual manera, el ambiente familiar violento ($I9Violento$) es significativa ($pvalor < 0.05$) para clasificar entre Rendimiento Medio y Rendimiento Alto. Con respecto a la variable I9Violento, esta corresponde a una variable indicadora dónde el valor de 1 significa que hay un ambiente familiar violento y el 0 en otro caso.

```
# Risk Ratio
round(exp(coef(ModFin)), 2)
```

```
     (Intercept) I9Agradable I9Violento  I17 I21Excelente I21Muy Buena
Bajo      710.27        2.43 1832496.64 0.28 5.000000e-02         0.48
Alto        0.00        0.04       5.96 6.01 1.118842e+15     60764.20
       I21Mala
Bajo 2014527.58
Alto      16.58
```

Resultados de Clasificación

Puede observarse que se obtuvieron los mismos resultados de clasificación que en SPSS

```
table(RendAcad$Rendimiento, predict(ModFin))
```

```
       Medio Bajo Alto
  Medio    70    0    0
  Bajo      8    9    0
  Alto      1    0   18
```

Predicción de casos

Por ejemplo, suponer que hay dos estudiantes y las respuestas que dieron para las preguntas correspondientes a las variables predictoras son las siguientes:

Estudiante 1. - Nota Previa en Matemática: 9 - Ambiente Familiar: Agradable - Percepción de Temas Matemáticos: Excelente

Estudiante 2. - Nota Previa en Matemática: 6 - Ambiente Familiar: Autoritario - Percepción de Temas Matemáticos: Muy Buena

```
caso = data.frame(I9 =  factor(c("Agradable", "Autoritario"),
                               levels = levels(RendAcad$I9)),
                  I17  = c(9, 6),
                  I21 = factor(c("Excelente", "Muy Buena"),
                               levels = levels(RendAcad$I21)))
prediccion <- data.frame(t(predict(ModFin, newdata = caso, "probs")))

round(prediccion, 2)
```

```
      X1   X2
Medio  0 0.86
Bajo   0 0.14
Alto   1 0.00
```

Puede apreciarse que, para el estudiante 1 se predice un rendimiento alto mientras que para el estudiante 2 se predice un rendimiento medio.

Aplicación del Método Usando el Lenguaje de Programación Python

Librerías

```
# Tratamiento de datos
import pandas as pd
import pyreadstat as pr
import numpy as np

# Procesado y modelado de datos
from sklearn.metrics import accuracy_score
import statsmodels.api as sm

# Gráficos
import matplotlib.pyplot as plt
import seaborn as sns
from plotnine import *

# Configuración warnings
import warnings
warnings.filterwarnings('ignore')
```

Datos

```
# Leer los datos
data = pd.read_spss('RendAcad_Seminario.sav')
```

Explorción de los datos

Inspección de los datos

```
data.head()
```

```
  Rendimiento           I9  I17        I21
0       Medio    Agradable  7.0  Muy Buena
1       Medio  Autoritario  7.0      Buena
2       Medio    Agradable  7.0  Muy Buena
3        Bajo     Violento  6.0      Buena
4       Medio  Autoritario  7.0      Buena
```

Tipo de variables

```
data.dtypes
```

```
Rendimiento    category
I9             category
I17             float64
I21            category
dtype: object
```

Frecuencia de los datos

```
# Rendimiento
print(data.Rendimiento.value_counts(), '\n')
```

```
Rendimiento
Medio    70
Alto     19
Bajo     17
Name: count, dtype: int64
```

```
# Ambiente Familiar
print(data.I9.value_counts(), '\n')
```

```
I9
Agradable      83
Autoritario    18
Violento        5
Name: count, dtype: int64
```

```
# Nota Previa
print('Descriptivos de la Nota Previa\n',data.I17.describe(), '\n')
```

```
Descriptivos de la Nota Previa
 count    106.00000
mean       7.79434
```

```
std        1.17545
min        5.00000
25%        7.00000
50%        8.00000
75%        8.37500
max       10.00000
Name: I17, dtype: float64
```

```
# Percepción de Temas Matemáticos
print(data.I21.value_counts(), '\n')
```

```
I21
Buena        51
Muy Buena    33
Excelente    16
Mala          6
Name: count, dtype: int64
```

```
# Tabla cruzada
data[['Rendimiento', 'I17']].groupby('Rendimiento').mean()
```

```
                I17
Rendimiento
Alto       9.010526
Bajo       6.705882
Medio      7.728571
```

Con la siguientes líneas de código se crea un gráfico similar al de la figura 11.

```
#plt.figure(figsize= (6,6))
#sns.boxplot(data= data, x= 'Rendimiento',y = 'I17')
#plt.ylabel('Rendimiento Previo en Matemática')
#plt.show()
```

Preparación de los datos

Seleccionar y recodificar la variable dependiente de acuerdo a la categoría de referencia (Rendimiento Medio)

```
y = data['Rendimiento'].astype('category').cat.reorder_categories(['Medio', 'Bajo',
'Alto'], ordered=True)
```

Seleccionar las variables predictoras categóricas y transformarlas en variables indicadoras, luego crear un solo conjunto de variables predictoras.

```
recod = pd.get_dummies(data.drop(['Rendimiento', 'I17'], axis = 1), drop_first = False)
recod = recod.drop(['I9_Violento', 'I21_Mala'], axis = 1)
recod = pd.DataFrame(np.where(recod ,1,0), columns = recod.columns)

x = pd.concat((data['I17'], recod), axis = 1)
```

Ajuste del Modelo

Se debe añadir un vector de constantes al conjunto de variables predictoras para el intercepto del modelo.

```
Modelo = sm.MNLogit(y, sm.add_constant(x))
Resultado = Modelo.fit()
```

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.247829
         Iterations: 35
```

Parámetros Estimados

Se muestran los resultados del modelo

```
print(Resultado.summary(), '\n')
```

```
                          MNLogit Regression Results
==============================================================================
Dep. Variable:            Rendimiento   No. Observations:                  106
Model:                        MNLogit   Df Residuals:                       92
Method:                           MLE   Df Model:                           12
Date:                lun., 25 sep. 2023   Pseudo R-squ.:                 0.7170
Time:                        10:13:24   Log-Likelihood:                -26.270
converged:                     False   LL-Null:                       -92.821
Covariance Type:            nonrobust   LLR p-value:                 1.470e-22
================================================================================================
Rendimiento=Bajo       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const               53.9192   3.98e+05      0.000      1.000    -7.8e+05     7.8e+05
I17                 -1.2728      0.559     -2.277      0.023      -2.368      -0.177
I9_Agradable       -17.1944   1467.753     -0.012      0.991   -2893.938    2859.549
I9_Autoritario     -18.0859   1467.754     -0.012      0.990   -2894.830    2858.658
I21_Buena          -29.2729   3.98e+05  -7.36e-05      1.000    -7.8e+05     7.8e+05
I21_Excelente      -23.2082    2.2e+07  -1.06e-06      1.000   -4.31e+07    4.31e+07
I21_Muy Buena      -30.0121   3.98e+05  -7.54e-05      1.000    -7.8e+05     7.8e+05
------------------------------------------------------------------------------------------------
Rendimiento=Alto       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const               10.1997   4.24e+05   2.41e-05      1.000    -8.3e+05     8.3e+05
I17                  1.7933      1.041      1.723      0.085      -0.247       3.833
I9_Agradable       -23.4123   1.86e+05     -0.000      1.000   -3.65e+05    3.65e+05
I9_Autoritario     -20.2558   1.86e+05     -0.000      1.000   -3.65e+05    3.65e+05
I21_Buena          -38.5156   2.81e+06  -1.37e-05      1.000    -5.5e+06     5.5e+06
I21_Excelente       36.3832    2.2e+07   1.66e-06      1.000    -4.3e+07     4.3e+07
I21_Muy Buena       -5.8886   4.14e+05  -1.42e-05      1.000   -8.12e+05    8.12e+05
================================================================================================
```

```
print(Resultado.summary2())
```

```
                                Results: MNLogit
================================================================================
Model:                    MNLogit             Method:                 MLE
Dependent Variable:       Rendimiento         Pseudo R-squared:       0.717
Date:                     2023-09-25 10:13    AIC:                    80.5399
No. Observations:         106                 BIC:                    117.8280
Df Model:                 12                  Log-Likelihood:         -26.270
```

```
Df Residuals:              92                   LL-Null:               -92.821
Converged:                 0.0000               LLR p-value:           1.4697e-22
No. Iterations:            35.0000              Scale:                 1.0000
------------------------------------------------------------------------------------
Rendimiento = 0  Coef.      Std.Err.      t    P>|t|      [0.025         0.975]
------------------------------------------------------------------------------------
          const  53.9192   397844.6733  0.0001 0.9999   -779707.3119   779815.1504
            I17  -1.2728        0.5589 -2.2773 0.0228        -2.3683       -0.1774
   I9_Agradable -17.1944     1467.7532 -0.0117 0.9907     -2893.9379     2859.5491
 I9_Autoritario -18.0859     1467.7535 -0.0123 0.9902     -2894.8299     2858.6582
      I21_Buena -29.2729   397841.9650 -0.0001 0.9999   -779785.1959   779726.6501
  I21_Excelente -23.2082 21989761.3713 -0.0000 1.0000 -43099163.5246 43099117.1081
  I21_Muy Buena -30.0121   397841.9650 -0.0001 0.9999   -779785.9351   779725.9109
------------------------------------------------------------------------------------
Rendimiento = 1  Coef.      Std.Err.      t    P>|t|      [0.025         0.975]
------------------------------------------------------------------------------------
          const  10.1997   423703.5272  0.0000 1.0000   -830433.4538   830453.8531
            I17   1.7933        1.0409  1.7229 0.0849        -0.2468        3.8334
   I9_Agradable -23.4123   186440.9785 -0.0001 0.9999   -365441.0154   365394.1908
 I9_Autoritario -20.2558   186440.9785 -0.0001 0.9999   -365437.8589   365397.3472
      I21_Buena -38.5156  2805880.7876 -0.0000 1.0000  -5499463.8043  5499386.7731
  I21_Excelente  36.3832 21955780.8410  0.0000 1.0000 -43032503.3175 43032576.0839
  I21_Muy Buena  -5.8886   414486.8728 -0.0000 1.0000   -812385.2313   812373.4540
====================================================================================
```

Risk Ratio

```
ratio = pd.DataFrame({"Bajo": np.exp(Resultado.params[0]),
                      "Alto": np.exp(Resultado.params[1])})
print(np.round(ratio, decimals = 2))
```

```
                      Bajo          Alto
const           2.611132e+23  2.689480e+04
I17             2.800000e-01  6.010000e+00
I9_Agradable    0.000000e+00  0.000000e+00
I9_Autoritario  0.000000e+00  0.000000e+00
I21_Buena       0.000000e+00  0.000000e+00
I21_Excelente   0.000000e+00  6.324437e+15
I21_Muy Buena   0.000000e+00  0.000000e+00
```

Resultado de Clasificación

```
Clasificacion = pd.DataFrame(Resultado.pred_table(), columns = ['Medio', 'Bajo', 'Alto'],
                index = ['Medio', 'Bajo', 'Alto'], dtype = np.int8)
print(Clasificacion)
```

```
       Medio  Bajo  Alto
Medio     70     0     0
Bajo       8     9     0
Alto       1     0    18
```

Se observa que la clasificación final de los casos, ha sido la misma en los 3 software.

Predicción de casos

Por ejemplo, suponer que hay dos estudiantes y las respuestas que dieron para las preguntas correspondientes a las variables predictoras son las siguientes:

Estudiante 1. - Nota Previa en Matemática: 9 - Ambiente Familiar: Agradable - Percepción de Temas Matemáticos: Mala

Estudiante 2. - Nota Previa en Matemática: 6 - Ambiente Familiar: Autoritario - Percepción de Temas Matemáticos: Muy Buena

```
caso = pd.DataFrame({'I17': [9, 6],
                     'I9_Agradable': [1, 0],
                     'I9_Autoritario': [0, 1],
                     'I21_Buena': [0, 0],
                     'I21_Excelente': [0, 0],
                     'I21_Muy Buena': [0, 1]})

prediccion = np.argmax(Resultado.predict(sm.add_constant(caso)), axis=1)

categoria = ['Bajo', 'Alto', 'Medio']
predic = [categoria[i] for i in prediccion]
predic
```

```
['Alto', 'Bajo']
```

Con base al resultado, se observa que, para el estudiante 1 se predice que tendrá un rendimiento alto, mientras que para el estudiante 2, será un rendimiento bajo.

Colnclusión

La replicación de los resultados en diferentes software aumenta la replicabilidad del estudio. El hecho de que SPSS, R y Python hayan producido resultados similares al aplicar la regresión logística multinomial sugiere que el modelo es robusto y confiable. La consistencia entre los resultados en diferentes software confirma que el modelo de regresión logística multinomial utilizado es apropiado para analizar el rendimiento académico y que las variables predictoras seleccionadas son relevantes para hacer predicciones precisas.