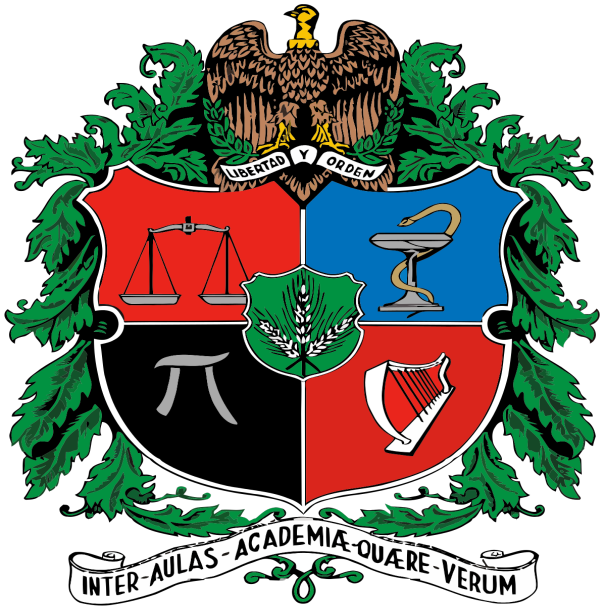
Parcial 2 I.A. datos categoricos

Oliver Orley Rodriguez Berrocal

Agosto 2021



Universidad nacional de Colombia sede Medellín

Facultad de ciencias

Estadística

1 Marco Teórico.

1.1 Tema: Modelos logísticos:

A menudo son encontrados problemas en el campo de aprendizaje estadístico que corresponde al método de clasificación. Existen muchos tipos de modelos clasificadores, como árboles de clasificación, análisis discriminante lineal etc. pero en este caso nos centraremos en el modelo logístico, este es uno de los modelos más populares y usados en el campo del aprendizaje estadístico y por tanto una excelente herramienta para modelado de datos categóricos. Este modelo puede ser usado para el caso donde su variable respuesta es dicótoma o donde son más de dos. También Este modelo proviene de la familia de modelos llamados **GLM** que traduce Generaliced Linear Model, que para este caso tendrá una componente binomial o multinomial. El modelo siempre dará como resultado una probabilidad de que la observación pertenezca a determinada clase, podemos denotar esta probabilidad y su función (llamada función de regresión logística) así:

$$\pi(X) = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}} \quad ó \quad Logit: \ log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \alpha + \beta X$$

Para el caso binario la curva de su gráfica de probabilidad será así:

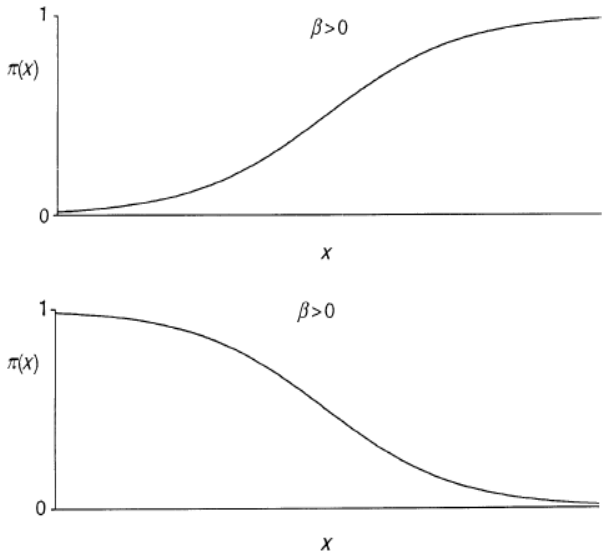


Figure 1: Curva de probabilidad de modelo logístico. tomado de [2]

Para el caso donde hay más de dos categorías(distribución multinomial sobre Y) tenemos que, para J clases nominales y tomando a J como el caso base tenemos que su logit es:

$$log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j x, \quad j = 1, ..., J-1$$

Es decir que cuando la respuesta está en alguna categoría j, este será el log odds que la respuesta sea j. De la ecuación se desprende el siguiente resultado :

$$\begin{aligned}\log\left(\frac{\pi_a}{\pi_b}\right) &= \log\left(\frac{\pi_a/\pi_J}{\pi_b/\pi_J}\right) = \log\left(\frac{\pi_a}{\pi_J}\right) - \log\left(\frac{\pi_b}{\pi_J}\right) \\ &= (\alpha_a + \beta_a x) - (\alpha_b + \beta_b x) \\ &= (\alpha_a - \alpha_b) + (\beta_a - \beta_b)x\end{aligned}$$

Figure 2: Ecuación logit(log odds), modelo logístico multinomial. Tomado de [2]

Pero logit multi-categórico tiene una expresión alternativa en términos de la probabilidad de respuesta, así :

$$\pi_j = \frac{e^{\alpha_j+\beta_j x}}{\sum_h e^{\alpha_h+\beta_h x}}, \quad j = 1, ..., J. \quad También: \quad \sum_j \pi_j = 1.$$

De esta manera podemos obtener sus gráficas de las estimaciones de respuestas de probabilidad para un ejemplo con j =3 clases:

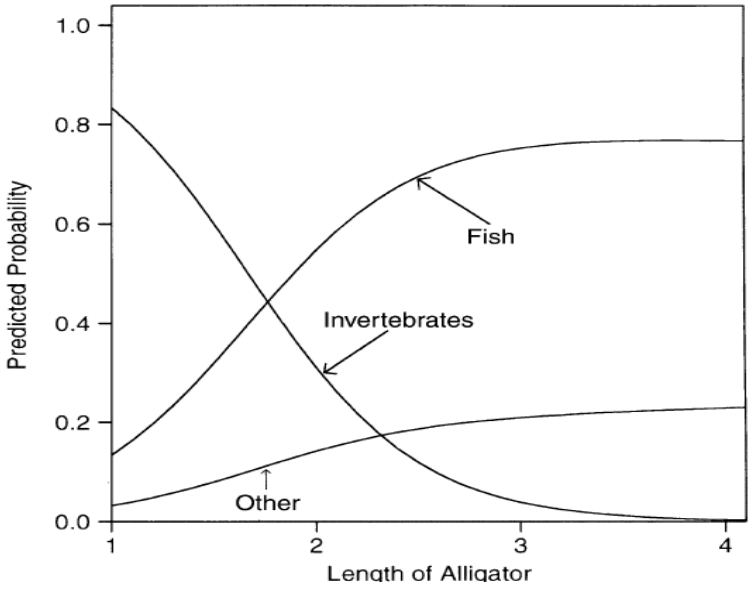


Figure 3: Curvas de probabilidad logístico multinomial. Tomado de [2]

1.2 Antecedentes:

Se tiene la base de datos proporcionada con el ministerio de educación de los estudiantes que presentaron la prueba ICFES del 2013-2, con algunos de sus datos económicos, demográficos y de puntajes del desempeño de cada alumno. Para esta base de datos se realizarán algunos modelos logísticos con distribución binomial para el caso binario o multinomial para más de dos categorías, sobre la variable respuesta.

1.3 Bases teóricas:

Para la siguiente aplicación de modelos clasificadores, se realizará primero la organización y limpieza de los datos y su correspondiente análisis para comprobar que no se encuentren resultados extraños o datos faltantes. Una vez limpia y funcional la base de datos se realizará un particionamiento de la base de datos con el fin de realizar la técnica de entrenamiento y validación para medir el desempeño de los modelos que se desarrollarán, con la métrica $Accurracy = \frac{TP+TN}{TP+TN+FP+FN}$. Se deberán generar modelos logísticos cuya variable respuesta será una categorización binaria de los puestos obtenidos en los resultados de los exámenes de estado ICFES. Para este caso se realizarán tres particiones de los puestos en sus percentiles 25, 50 y 75. Se realizará un análisis de los resultados obtenidos para la significancia de las variables explicativas en el modelo, y se plantean algunas hipótesis al respecto de ellas. También se realizarán modelos multinomiales dividiendo los puestos en más de dos categorías y serán analizarán sus resultados de manera similar a los modelos logísticos binomiales. Para realizar estos procedimientos se utilizará el paquete estadístico R. La intención final es proponer soluciones a través de generar modelos que nos ayuden predecir de la mejor manera la clase a la que pertenezca un determinado estudiante, con fines o propósitos académico-administrativos.

1.4 Conceptos claves:

Modelo logístico, distribución binomial y multinomial, probabilidad, precisión(Accuracy), logit.

2 Datos

Después de realizar todo el procedimiento de limpieza y organización de los datos, de imputación de datos faltantes, y de excluir algunas de las variables que no pueden ser significativas en la realización del modelo tales como la nacionalidad del estudiante (ESTU_PAIS_ RESID), ya que el examen se realiza para estudiantes colombianos, ESTU_CONSECUTIVO que identifica al estudiante en la base de datos, etc. se obtienen el conjunto de datos para la modelación, con 474996 observaciones, a continuación una breve vista del resultado:

	ind_colegio	ESTU_EDAD	ESTU_TIPODOCUMENTO	ESTU_GENERO	ESTU_RESIDE_DEPT
1	10	16	T	M	ANTIOQUIA
2	10	16	T	M	SANTANDER
3	10	17	T	M	ANTIOQUIA
4	10	16	T	M	SANTANDER
5	10	17	T	M	ANTIOQUIA
6	10	16	T	F	ANTIOQUIA

El conjunto de todas las variables será el siguiente, con su respectiva estructura o clase, donde format son variables categóricas, integer son númericas enteras, y numeric valores numéricos:

	Variables	tipo
1	ind_colegio	factor
2	ESTU_EDAD	factor
3	ESTU_TIPODOCUMENTO	factor
4	ESTU_GENERO	factor
5	ESTU_RESIDE_DEPT	factor
6	COLE_CALENDARIO_COLEGIO	factor
7	COLE_GENERO_POBLACION	factor
8	COLE_NATURALEZA	factor
9	COLE_ES_BILINGUE	factor
10	COLE_INST_JORNADA	factor
11	COLE_CARACTER_COLEGIO	factor
12	COLE_INST_VLR_PENSION	factor
13	FAMI_COD_EDUCA_PADRE	factor
14	FAMI_COD_EDUCA_MADRE	factor
15	FAMI_COD_OCUP_PADRE	factor
16	FAMI_COD_OCUP_MADRE	factor
17	ESTU_ESTRATO	factor
18	ESTU_TRABAJA	factor
19	ESTU_HORAS_TRABAJO	numeric
20	ESTU_PUESTO	integer
21	indice_discapacidad	factor

Se crearon dos índices, uno que se llama ind_colegio que es una propuesta de la riqueza de los colegios a lo que pertenece el estudiante, y índice_discapacidad que hace un resumen de todas las variables de discapacidad para simplificar al hecho de que un estudiante tenga o no alguna discapacidad.

A continuación la lista de las variables con datos faltantes y su respectiva imputación:

variable	Acción
ESTU_EDAD:	Elimino o pongo el promedio.
COLE_ES_BILINGUE:	Se pondrán la media 17 años
COLE_INST_VLR_PENSION:	Elimino.
FAMI_COD_EDUCA_PADRE:	Elimino.
FAMI_COD_EDUCA_MADRE:	Elimino.
FAMI_COD_OCUP_PADRE:	Elimino.
FAMI_COD_OCUP_MADRE:	Elimino.
ESTU_ESTRATO:	Elimino.
ESTU_TRABAJA:	Elimino.
ESTU_HORAS_TRABAJO:	NA´s = 0 i.e. no trabajan.

Pueden realizarse otro tipo de propuestas, pero para este caso se realizaron estas asignaciones.

Otro descubrimiento interesante fue que al ver la distribución de la edad de los estudiantes, se pudo notar que, existían estudiantes que tenían mucho más de 100 años o menos de 10 años.

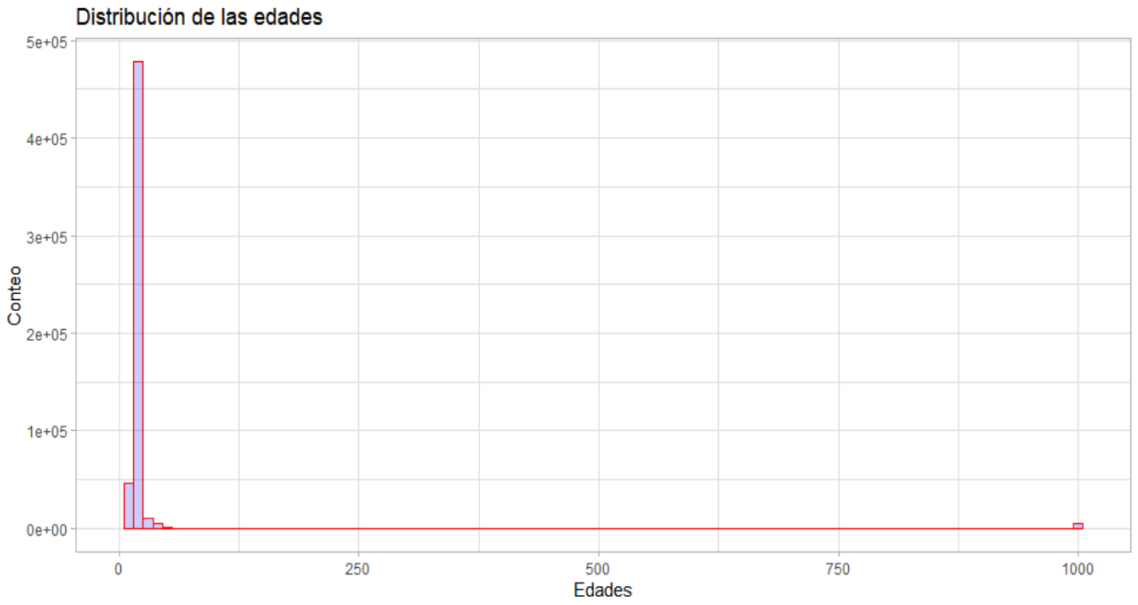


Figure 4: Distribución de la variable edad

La cantidad de estos errores es poca, menos aún entre estudiantes de menores a 15 años. Para el caso de los menores de 15 años se utiliza la mediana para modificarlos que en este caso es 17 años, asumiendo que fue un error haber puesto, por ejemplo 6 años. Para finalizar el estudio fue reducido a los estudiantes entre 15 y 18 años, ya que según [3], la edad más común según Mineducación es de 16 años al momento de graduarse, por lo que la distribución de las edades quedará así:.

Figure 5: Distribución de la variable edad organizada, idea complementada por [3]

Finalmente, se decide fijar la variable edad del estudiante como variable categórica nominal con 4 clases.

3 Modelos logísticos

Se realizaron 3 modelos logísticos, con el fin de observar cuál es su comportamiento en relación con su variable respuesta. La variable respuesta son los puestos categorizados, en el primer cuartil, la mediana y el percentil 75, por ejemplo, modelo uno(dividiendo la variable respuesta en el primer cuartil) tiene dos clases estudiantes que pertenecen a los puestos [0, 250] y (250,1000], y así sucesivamente. Se podrá observar la significancia de los parámetros de las variables explicativas, luego se muestra la matriz de confusión y su correspondiente métrica Accuracy, para medir la calidad de las predicciones usando la técnica de validación.

3.1 Modelo 1:

Para este modelo se realizó la división de la variable respuesta en la mediana es decir (0, 500] y (500, 1000]. A continuación se puede observar una parte de la tabla de valores estimados dado que son 164 variables resultantes por la gran cantidad de factores en el modelo, y sus correspondientes estadísticos:

	Estimate	Std..Error	z.value	Pr...z..
(Intercept)	38.32	60064.42	0.00	1.00
ind_colegio20	-0.21	0.01	-24.04	0.00
ind_colegio30	-0.30	0.01	-21.17	0.00
ind_colegio40	-0.46	0.03	-16.68	0.00
ind_colegio50	-0.27	0.04	-7.24	0.00
ind_colegio60	-0.33	0.08	-4.12	0.00
ind_colegio70	-0.52	0.06	-9.23	0.00
ind_colegio80	-0.20	0.12	-1.73	0.08
ind_colegio90	0.06	0.10	0.63	0.53
ind_colegio100	-1.14	0.08	-13.72	0.00
ESTU_EDAD16	0.17	0.01	13.58	0.00
ESTU_EDAD17	0.56	0.01	42.40	0.00
ESTU_EDAD18	0.81	0.02	45.20	0.00
ESTU_TIPODOCUMENTOE	-0.03	0.21	-0.12	0.90
ESTU_TIPODOCUMENTOP	-0.94	0.45	-2.10	0.04
ESTU_TIPODOCUMENTOQ	1.23	0.90	1.38	0.17
ESTU_TIPODOCUMENTOR	0.09	0.02	3.92	0.00
ESTU_TIPODOCUMENTOT	-0.03	0.02	-1.75	0.08
ESTU_TIPODOCUMENTOV	0.40	0.71	0.56	0.58
ESTU_GENEROM	-0.39	0.01	-54.72	0.00
ESTU_RESIDE_DEPTANTIOQUIA	-0.71	0.10	-6.86	0.00
ESTU_RESIDE_DEPTARAUCA	-1.14	0.11	-10.00	0.00
ESTU_RESIDE_DEPTATLÃNTICO	-0.33	0.10	-3.12	0.00
ESTU_RESIDE_DEPTBOGOTÃ	-1.10	0.10	-10.59	0.00
ESTU_RESIDE_DEPTBOLÃVAR	-0.22	0.10	-2.15	0.03

Como conclusión de este acercamiento, el modelo muestra que las variables que corresponden a valores de la pensión no

son significativas, extrañamente tampoco el modelo encuentra significativo la educación del padre y de la madre, tampoco es significativo si el estudiante trabaja, por lo mismo las horas que pueda trabajar el estudiante no son significativas.

Ahora para poder entender el comportamiento de estas variables no significativas en relación con la variable respuesta, se observarán las siguientes gráficas:

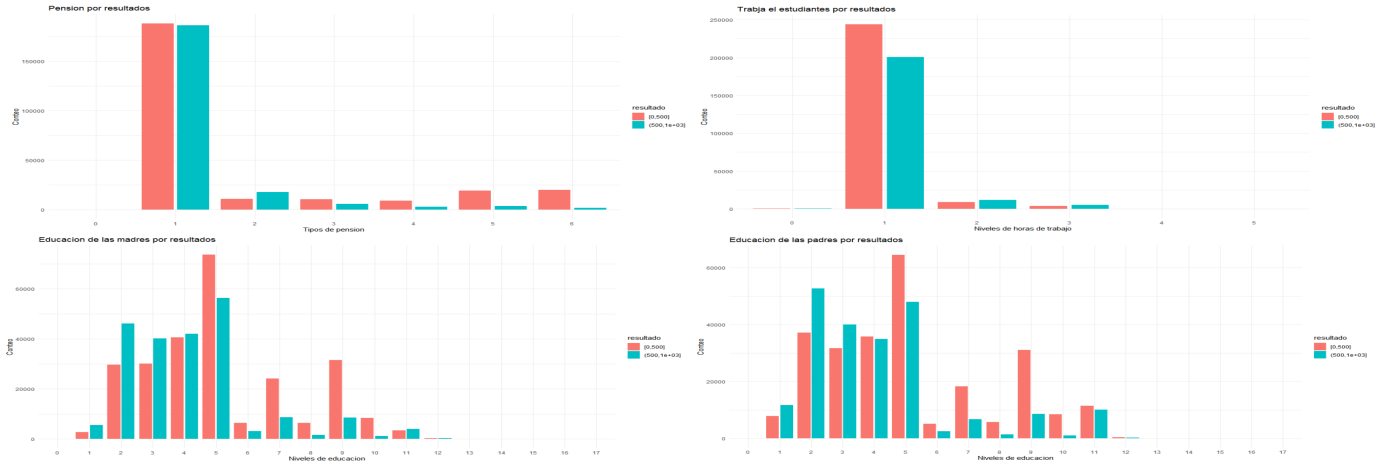


Figure 6: Superior izquierda: Pensión que pagan los estudiantes con categoría de puesto. Superior derecha:Cargarías de trabajo con categoría de puesto. Inferior Izquierda: Educación de madres con categoría de puesto. Inferior derecha: Educación de Padres con categoría de puesto.

En esta gráfica superior derecha podemos observar que la mayor parte de los estudiantes se concentran en la clase 1 de pensión con resultados similares para ambos niveles del resultado. Solo en la pensión 2 se observa que hay más en la clase (500, 1000] que en la otra. Quizá lo que percibió el modelo fue carencia de diferencias significativas para darle significativa a esta variable. Análisis similares pueden ser observados para las demás variables.

Ahora se observará la matriz de confusión:

	(0,500]	(500,1000]
(500,1000]	6900	13710
(0,500]	18769	8121

Con lo que podemos observar que el 68.37684% de los estudiantes estuvieron correctamente clasificados por el modelo. También se obtiene la tasa de clasificación incorrecta que es 31.62316%.

3.2 modelo 2:

Como se mencionó anteriormente, la variable respuesta consta de los puestos [0, 250] y (250,1000]. Una parte los resultados (dada la gran cantidad de variables involucradas) del modelo se pueden observar a continuación:

	Estimate	Std..Error	z.value	Pr...z..
(Intercept)	10601391272814306.00	107312481.15	98789918.56	0.00
ind_colegio20	-29696051705970.16	262879.87	-112964342.35	0.00
ind_colegio30	-33161693084461.24	411999.68	-80489608.38	0.00
ind_colegio40	-326229722860927.94	795832.92	-409922378.88	0.00
ind_colegio50	-302192505613549.81	1079061.38	-280051264.82	0.00
ind_colegio60	-179410888849082.69	2477613.58	-72412780.60	0.00
ind_colegio70	-159988375828908.28	1447785.14	-110505607.45	0.00
ind_colegio80	-505180596407358.00	2807661.29	-179929323.21	0.00
ind_colegio90	-632801761077549.88	3009990.97	-210233774.14	0.00
ind_colegio100	202540492848787.12	2342668.24	86457181.40	0.00
ESTU_EDAD16	161228666332901.78	373321.27	431876452.04	0.00
ESTU_EDAD17	122831164928793.41	393077.98	312485487.90	0.00
ESTU_EDAD18	-68183137273566.04	535727.13	-127272138.48	0.00
ESTU_TIPODOCUMENTOE	202412484833621.62	5773326.87	35059938.44	0.00
ESTU_TIPODOCUMENTOP	-235783498975399.31	9606163.08	-24545023.55	0.00
ESTU_TIPODOCUMENTOQ	-235314932772135.94	25375464.37	-9273325.18	0.00
ESTU_TIPODOCUMENTOR	18824856907617.41	680813.83	27650520.89	0.00
ESTU_TIPODOCUMENTOT	7607514395952.23	534149.26	14242300.85	0.00
ESTU_TIPODOCUMENTOV	375317880624201.44	20240453.93	18542957.68	0.00
ESTU_GENEROM	-130690642294469.94	214099.77	-610419338.97	0.00
ESTU_RESIDE_DEPTANTIOQUIA	439557185071325.00	3132580.45	140317923.91	0.00

Para este modelo, todas las variables muestran ser significativas para predecir la clase a la pertenecen los estudiantes.

Para tener una percepción del porqué las anteriores variables del modelo 1, ahora son significativas en el modelo 2 podemos observar las siguientes gráficas:

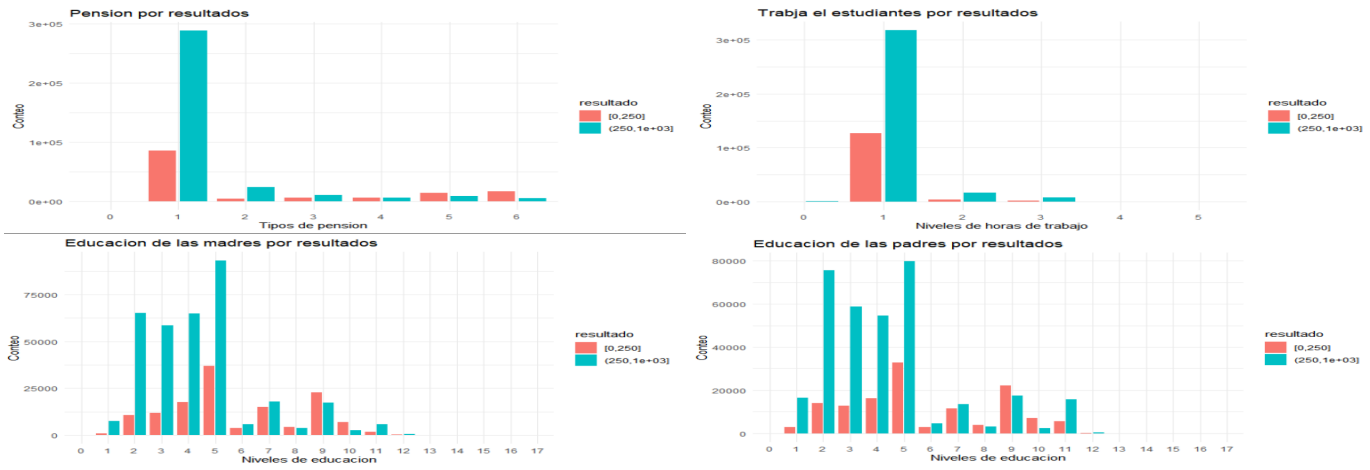


Figure 7: Superior izquierda: Pensión que pagan los estudiantes con categoría de puesto. Superior derecha: Categorías de trabajo con categoría de puesto. Inferior Izquierda: Educación de madres con categoría de puesto. Inferior derecha: Educación de Padres con categoría de puesto.

Podemos ver en cada gráfica una significativa diferencia en algunos niveles de la barra verde sobres la roja, sobre todo donde las barras tienen mayor altura(mayor cantidad de estudiantes pertenecientes a esta combinación de clases). Esto nos permite suponer que el modelo encuentra un patrón significativo de cada una de estas variables con la variable respuesta.

A continuación su matriz de confusión:

	(0,250]	(250,1000]
(250,1000]	10701	32908
(0,250]	2524	1367

De esta matriz obtenemos que la precisión es de aproximadamente 74.59% y de este se deriva la tasa de clasificación incorrecta del 25.4%.

3.3 Modelo 3:

Para el siguiente modelo se dividió la variable respuesta en el percentil 75 o tercer cuartil, es decir las clases son (0, 750] y (750, 1000]. A continuación una parte de la tabla de estimaciones y estadísticos:

	Estimate	Std..Error	z.value	Pr...z..
(Intercept)	15.21	254.96	0.06	0.95
ind_colegio20	-0.16	0.01	-15.29	0.00
ind_colegio30	-0.28	0.02	-15.74	0.00
ind_colegio40	-0.37	0.04	-10.15	0.00
ind_colegio50	-0.30	0.05	-6.07	0.00
ind_colegio60	-0.42	0.10	-4.01	0.00
ind_colegio70	-0.48	0.08	-5.92	0.00
ind_colegio80	-0.32	0.20	-1.63	0.10
ind_colegio90	-0.04	0.11	-0.40	0.69
ind_colegio100	-1.02	0.13	-8.03	0.00
ESTU_EDAD16	0.14	0.02	8.90	0.00
ESTU_EDAD17	0.51	0.02	30.89	0.00
ESTU_EDAD18	0.72	0.02	35.43	0.00
ESTU_TIPODOCUMENTOE	-0.17	0.23	-0.74	0.46
ESTU_TIPODOCUMENTOP	-0.50	0.56	-0.90	0.37
ESTU_TIPODOCUMENTOQ	0.56	1.17	0.48	0.63
ESTU_TIPODOCUMENTOR	0.11	0.02	4.83	0.00
ESTU_TIPODOCUMENTOT	-0.03	0.02	-1.65	0.10
ESTU_TIPODOCUMENTOV	0.52	0.63	0.82	0.41
ESTU_GENEROM	-0.27	0.01	-32.97	0.00

Para este modelo, se observa que no son significativos la variable de pensión del colegio, tampoco muestra significancia para la educación de ninguno de los padres, tampoco es útil para explicar la variable respuesta el hecho de que el estudiante trabaje y por ello las horas que de trabajo que realice. Este modelo es muy similar en estos resultados al modelo 1.

Miremos las correspondientes gráficas nuevamente:

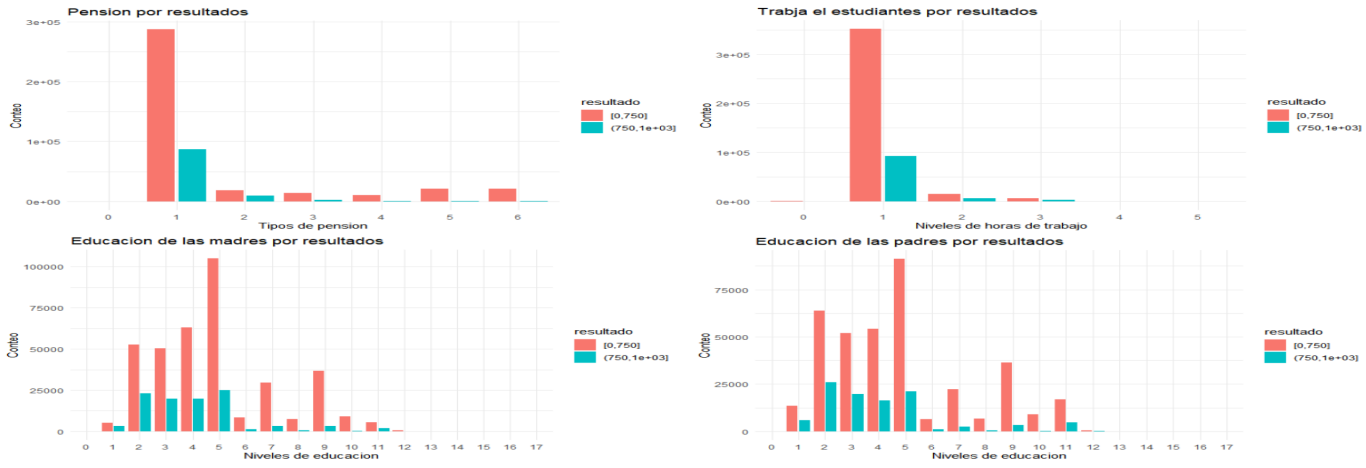


Figure 8: Superior izquierda: Pensión que pagan los estudiantes con categoría de puesto. Superior derecha:Categorías de trabajo con categoría de puesto. Infiero Izquierda: Educación de madres con categoría de puesto. Inferior derecha: Educación de Padres con categoría de puesto.

Al comparar estas gráficas con las gráficas del modelo dos, la primera impresión de notar que las barras rojas son significativamente mayores a las verdes, y de pensar que el modelo está rechazando las variables como variables significativas, uno podría pensar, que algo malo está sucediendo en el procedimiento de modulación o que el algoritmo glm se encuentra defectuoso de alguna manera. Pero, si se observa detenidamente, para las barras más grandes, las barras entre verdes y rojas de cada gráfica tienen una desproporción menor (en la mayoría de los casos) favoreciendo las rojas en este caso, y verdes el caso del modelo 2, en el modelo 2 estas desproporciones son mayores. Por esta razón podríamos considerar que el modelo 3 en este caso, define que dicha desproporción no es suficiente para clasificar las variables como significativas.

Ahora su matriz de confusión:

	(0,750]	(750,1000]
(750,1000]	1069	1362
(0,750]	36216	8853

Para este caso tenemos que la métrica Accurracy o precisión es de aproximadamente el 79.11% y su tasa de clasificación incorrecta es del 20.9% aproximadamente.

Conclusión: Podemos observar que con las mismas variables, y con el mismo con junto de entrenamiento, que simplemente cambiando los percentiles de la variable respuesta podemos llegar a modelos completamente diferentes, con precisiones de clasificación en su predicción diferentes. Por ello, al momento de realizar una propuesta para categorizar una variable respuesta, sería recomendable tener completamente claro el objetivo del estudio para elegir mejor su partición y poder probar con otras particiones cercanas al objetivo para finalmente obtener el mejor modelo posible.

4 Modelos logísticos multinomiales:

Para estos tipos de modelos se considerarán 2 modelos diferentes, uno cuya variable respuesta (la cual es resultado), se dividirá en 3 partes iguales y el otro modelo considerará 4 clases haciendo la división en 4 clases diferentes en sus correspondientes cuartiles. Se utilizará la técnica de validación y de prueba, como se observó anteriormente y con la matriz de confusión miraremos la calidad del modelo para predecir correctamente las clases con la métrica accuracy o precisión. Para el ajuste de estos modelos se utiliza la librería de R nnet que ajusta modelos logísticos multinomiales vía redes neuronales.

4.1 Modelo 1:

En este modelo se realiza una partición de los datos en tres clases, (0,333], (333,667] y (667,1000]. Se establece el caso base (0,333] y después se ajusta el modelo se muestra una pequeña parte de su resultado:

	(Intercept)	ind_colegio20	ind_colegio30	ind_colegio40	ind_colegio50	ind_colegio60	ind_colegio70
(333,667]	-0.7855115	-0.1991492	-0.2232351	-0.3773156	-0.2588000	-0.2698861	-0.5010101
(667,1e+03]	-0.6740005	-0.2973010	-0.4371333	-0.5903934	-0.4336856	-0.5545058	-0.6975340

Ahora obtenemos su matriz de confusión:

	(0,333]	(333,667]	(667,1000]
(0,333]	11402	5421	2495
(333,667]	3817	5095	3736
(667,1000]	2288	5518	7728

Ahora para este modelo obtenemos una precisión del 51%, lo cual quiere decir que es levemente mejor que hacer darle igual probabilidad a cada clase de ocurrir y realizar muestras aleatorias. Es decir el modelo no es bueno.

4.2 Modelo 2:

Para el modelo 2 dividiremos los puestos de los estudiantes es sus cuartiles, esto es (0,250], (250,500], (500,750], (750,1000], de esta manera obtendremos los siguientes parámetros estimados:

	(Intercept)	ind_colegio20	ind_colegio30	ind_colegio40	ind_colegio50	ind_colegio60	ind_colegio70
(250,500]	-1.728528	-0.1606452	-0.1820420	-0.4260102	-0.2214784	0.2838469	-0.5300050
(500,750]	-1.713541	-0.2657089	-0.3131852	-0.6565923	-0.2623722	0.1078883	-0.7611131
(750,1000]	-1.390967	-0.3032186	-0.4504886	-0.7650623	-0.4451874	-0.1930231	-0.9261322

Si se quieren observar estos parámetros dejando el modelo en términos de odds solo necesitaremos exponenciarlos.

Ahora obtendremos la matriz de confusión:

	(0,250]	(250,500]	(500,750]	(750,1000]
(0,250]	8516	4314	2450	1311
(250,500]	2785	3776	3195	2068
(500,750]	1165	2236	2472	2170
(750,1000]	759	2118	3499	4666

Al calcular su Accurracy tenemos que su precisión será 40.90526%, es decir que su desempeño es bastante malo para predecir la clase para nuevas observaciones.

Desafortunadamente, la librería nnet no cuenta con algunos procedimientos como el cálculo del valor p para la significancia de las variables y presenta problemas para presentar resúmenes con modelos de muchas variables y muchas observaciones como este caso.

Conclusión: Podemos ver que los modelos multinomiales ajustados no son buenos para predecir, este pude ser debido a la alta cantidad de variables productoras y sobre todo la gran cantidad de variables indicadoras. Es términos generales es necesario tener cuidado a la hora de realizar categorizaciones de variables numéricas, como es recomendado por expertos en el manejo de datos categóricos.

5 Recomendaciones y conclusiones generales:

Como recomendaciones para posteriores estudios, realizar nuevas categorizaciones de las variables, como los cambios realizados a la variable edad, podría tenerse un rango más amplio de edades. Proponer nuevos indicadores, que permitan simplificar la complejidad del modelo.

También, se propone realizar otros acercamientos tales como modelos VGAM, que son Vector Generalized Linear and Additive Models [5], que son otra propuesta para los modelos multinomiales, también realizar modelación vía GSK como una alternativa, esto a pesar de su complejidad podría ser interesante. Otra opción para los modelos multinomiales podría realizar aproximaciones de modelos glm vía log poisson lineal[6].

Como conclusión final, estos modelos para el manejar de datos categóricos son herramientas fundamentales a la hora de proponer modelos, pueden ser vistos como puntos de partida para avanzar a partir de estos a comparar con otras propuestas. También es necesario nuevamente, resaltar el cuidado que se debe tener a la hora de categorizar una variable numérica, ya que como se pudo observar, más claramente en el modelo logístico, que diferentes tipos de divisiones, llevan a modelo que pueden ser significativamente diferentes.

6 Código:

Para más detalle del código, podrá visitar el script llamado parcial1.Rmd en el siguiente repositorio de git aquí.

7 Referencias:

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2021) an Introduction to Statistical Learning second edition, Springer Texts in Statistics.

[2] Alan Agresti, (2007) An Introduction to Categorical Data Analysis. New Jersey Published by JohnWiley Sons, Inc., Hoboken.

[3] GONZÁLEZ, L. (diciembre 2008) El dilema de un bachiller. mineducacion. https://www.mineducacion.gov.co/1621/article-183908.html: :text=A

[4] Ripley B. , Venables W. (2021-05-03) Package 'nnet'. cran.r-project. https://cran.r-project.org/web/packages/nnet/nnet.pdf

[5] Yee T., Moler C. (2021-01-13) Package 'VGAM'. cran.r-project. https://cran.r-project.org/web/packages/VGAM/VGAM.pdf

[6] Raghvendra. (2012-03-15) Can I use glm algorithms to do a multinomial logistic regression?. stats.stackexchange. https://stats.stackexchange.com/q/24705

Nota: Si está observando en formato pdf, puede hacer clic sobre los enlaces web, y lo llevarán directamente a la página.