Modelo de regresión logístico para encuesta de percepción ciudadana

Uribe Giraldo Jimena, Puentes Rocha Daniel Felipe
jiuribeg@unal.edu.co, dapuentes@unal.edu.co

***Resumen*—En esta investigación se trazó como objetivo explorar la vulnerabilidad en la población de Medellín mayores a 18 años; para esto se obtuvo la base de datos de la Encuesta de Percepción Ciudadana. El estudio cuenta con limitaciones para inferir los resultados a toda la población de la ciudad de Medellín; sin embargo, aporta información útil para identificar las características de la comunidad vulnerables.**

***Palabras clave*—EPC, Medellín, Vulnerabilidad, Modelo logístico bayesiano**

Medellín, una ciudad ubicada en Colombia, se ha enfrentado a diversos desafíos urbanos en las últimas décadas. Aunque ha experimentado avances significativos en términos de desarrollo y transformación, también ha enfrentado problemáticas relacionadas con la seguridad, movilidad, infraestructura y calidad de vida de sus habitantes. Ante esta situación, la encuesta de percepción ciudadana se ha convertido en una herramienta fundamental para comprender la realidad de la ciudad y mejorar la toma de decisiones en la planificación urbana.

La encuesta de percepción ciudadana en Medellín tiene como objetivo principal recopilar datos sobre las opiniones, actitudes y percepciones de los residentes en relación con distintos aspectos de su entorno urbano. A través de una metodología de muestreo representativa, se abordan temas como seguridad ciudadana, calidad de los servicios públicos, movilidad, espacios verdes, cultura y recreación, entre otros. Los participantes responden a una serie de preguntas estructuradas que les permiten expresar su satisfacción, inquietudes, percepción de su propia vulnerabilidad y expectativas en relación con su vida en la ciudad.

Los resultados de la encuesta proporcionan información valiosa para comprender las necesidades y preferencias de la población de Medellín. Permiten identificar áreas problemáticas y evaluar la efectividad de las políticas públicas implementadas. Además, la encuesta promueve la participación ciudadana y fortalece el diálogo entre los habitantes de la ciudad.

En el presente estudio se llevó a cabo una investigación sobre la vulnerabilidad de las personas en función de su edad, estrato socioeconómico, comuna y sexo. Se utilizó un modelo de regresión logístico bayesiano para identificar los factores asociados a la vulnerabilidad y comprender los determinantes de este estado, con el objetivo de desarrollar estrategias de intervención efectivas.

El presente documento corresponde al articulo final del proyecto de investigación del programa de Estadística

Los resultados obtenidos del modelo pueden ser utilizados por los responsables de la toma de decisiones para diseñar políticas y programas específicos dirigidos a abordar las causas subyacentes de la vulnerabilidad. Por ejemplo, si se identifica que el género o el estrato socioeconómico son factores importantes, se pueden implementar acciones focalizadas para abordar las desigualdades y promover la equidad en la sociedad.

I. Introducción

La vulnerabilidad es un fenómeno de gran importancia social y económica, ya que afecta de manera desigual a diferentes segmentos de la población. Comprender los factores que contribuyen a la vulnerabilidad es fundamental para desarrollar intervenciones efectivas dirigidas a reducir las desigualdades.

Se han propuesto diversas soluciones para abordar la vulnerabilidad en diferentes contextos. Algunas investigaciones han explorado la relación entre la edad y la vulnerabilidad, demostrando que los grupos de mayor edad suelen enfrentar mayores riesgos y desafíos. Además, se han analizado los efectos del estrato socioeconómico, evidenciando que aquellos en situaciones de menor estatus económico tienen una mayor probabilidad de experimentar vulnerabilidad. Asimismo, se ha investigado la influencia de la comuna y el entorno local en la vulnerabilidad, destacando la importancia de las condiciones socioambientales. Por último, se han examinado las diferencias de género en la vulnerabilidad, revelando que las mujeres a menudo se encuentran en situaciones más precarias.

En este estudio, se propone utilizar un modelo de regresión logística bayesiana para analizar la relación entre las variables de interés (edad, estrato socioeconómico, comuna y sexo) y la vulnerabilidad. Esta metodología permitirá identificar los factores que están asociados con un mayor riesgo de encontrarse en un estado de vulnerabilidad. Al obtener resultados significativos, se podrá mejorar la comprensión de los determinantes de la vulnerabilidad y se podrán desarrollar estrategias de intervención más efectivas. [1] [2] [3]

La estructura del documento se organizará de la siguiente manera: en la siguiente sección se mostrará la descripción de los datos y metodología utilizada. A continuación, se presentarán y analizarán los resultados obtenidos del modelo de regresión logística bayesiana, destacando las variables que mostraron una influencia significativa en la vulnerabilidad.

Por último, se presentarán las conclusiones principales y se propondrán posibles direcciones para investigaciones futuras en el campo de la vulnerabilidad y sus determinantes.

II. Descripción de datos

Mediante una metodología de muestreo representativa, se recopilan datos sobre temas como seguridad, movilidad, infraestructura, servicios públicos, espacios verdes, cultura y recreación, entre otros. Los participantes de la encuesta responden a una serie de preguntas estructuradas, que les permiten expresar su nivel de satisfacción, preocupaciones y expectativas en relación con los diferentes aspectos de su vida en la ciudad.

Con la base de datos obtenida de dicho estudio, se implementó un análisis exploratorio, con el fin de encontrar posibles relaciones entre las variables, detectar valores atípicos y de esta forma conocer los datos a los que nos enfrentamos y depurar los datos para garantizar que los resultados obtenidos con el modelo serán válidos y aplicables. [4]

La base inicial contenía 558 variables, luego de filtrar la base se redujo a 5 variables, ya que 4 se consideraron significativas a la hora de explicar la vulnerabilidad. De modo que se procedió a elegir las variables más significativas que conformaran un modelo parsimonioso.

Las 5 variables se muestran a continuación:

II-A. Tabla de variables

Tabla I
Denominación de Variables, Tipos y Niveles

Variables	Tipos	Niveles
Comuna	Categórica	1: Popular 2: Santa Cruz 3: Manrique 4: Aranjuez 5: Castilla 6: Robledo 7: Villa Hermosa 8: Buenos Aires 9: La Candelaria 10: Laureles-Estadio 11: La América 12: San Javier 13: El Poblado 14: Guayabal 15: Belén 16: Altavista.
Sexo	Categórica	0: Femenino 1: Masculino
Edad	Numérica	
Estrato	Categórica	1: Estrato 1 2: Estrato 2 3: Estrato 3 4: Estrato 4 5: Estrato 5 6: Estrato 6
Vulnerabilidad	Categórica	0: No se considera vulnerable 1: Se considera vulnerable

A continuación unas gráficas descriptivas de algunas variables:

Figura 1.

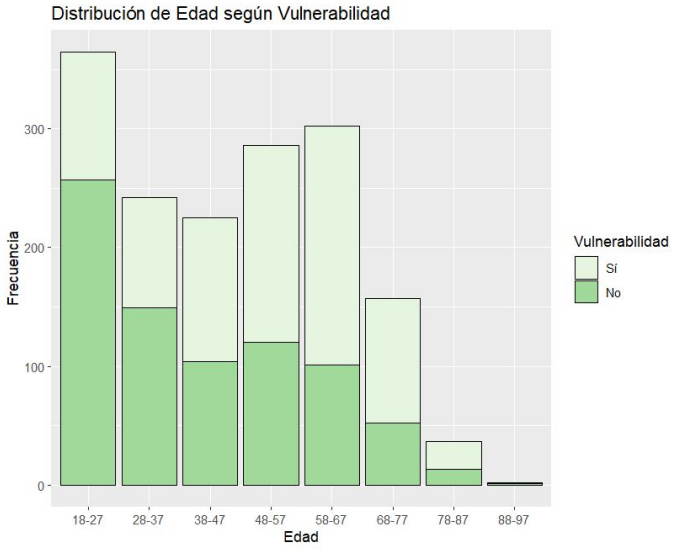
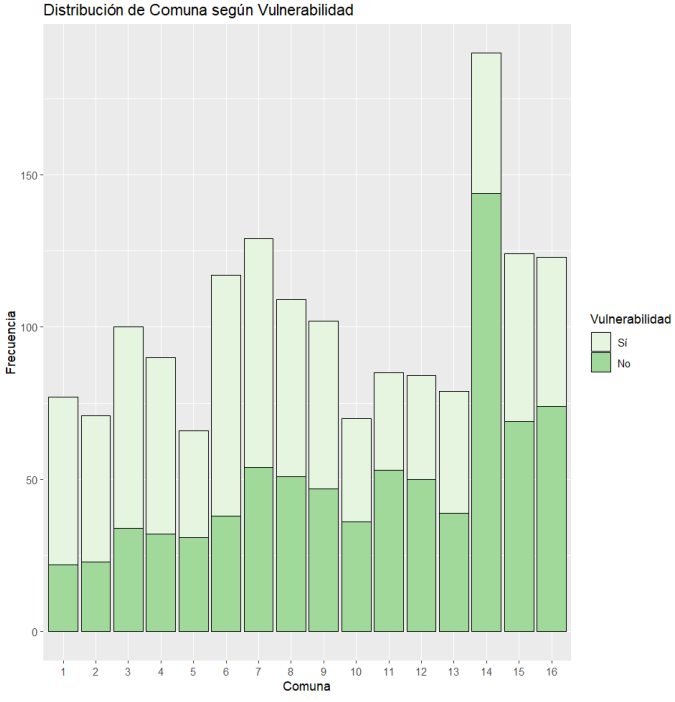


Figura 2.



III. Métodos

Todo se realizó por medio de RStudio con librerías como rstan para procesos como MCMC para ver la convergencia de cada parámetro. [5]

III-A. MCMC

Los métodos de Markov Chain Monte Carlo (MCMC) se basan en la creación de una adecuada cadena de Markov. Las cadenas de Markov, bajo ciertas condiciones, convergen

a una densidad estacionaria invariante en el tiempo. La idea fundamental de los métodos MCMC es la de diseñar una cadena de Markov cuya densidad sea estacionaria. [6]

III-B. Regresión Logística

La regresión logística es un instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica y un conjunto de variables predictoras o independientes, que pueden ser cuantitativas o categóricas. [7]

III-C. LOOCV

El método LOOCV, o Validación Cruzada dejando uno fuera (Leave-One-Out Cross-Validation, en inglés), es una técnica utilizada para estimar la precisión puntual de la predicción fuera de la muestra a partir de un modelo bayesiano ajustado utilizando la probabilidad logarítmica evaluada en las simulaciones posteriores de los valores de los parámetros.

El propósito principal del LOOCV es estimar qué tan bien un modelo entrenado se generalizará a nuevos datos no vistos. La idea central detrás de este método es simular una situación en la que tenemos un conjunto de datos limitado y queremos evaluar qué tan bien nuestro modelo se comportará cuando se le presente un nuevo punto de datos. [8]

III-D. Curva ROC

La curva ROC es una herramienta estadística que se utiliza para evaluar la capacidad discriminativa de una prueba diagnóstica dicotómica. Se trata de curvas en las que se presenta la sensibilidad en función de los falsos positivos para distintos puntos de corte. Son útiles para elegir el punto de corte más adecuado de una prueba, conocer el rendimiento global de esta y comparar la capacidad discriminativa de 2 o más pruebas diagnósticas. [9]

III-E. Factor de Bayes

El factor de Bayes es una medida de la evidencia a favor o en contra de una hipótesis, basada en los datos observados y el conocimiento previo. Se calcula como el cociente entre la probabilidad de los datos bajo la hipótesis alternativa y la probabilidad de los datos bajo la hipótesis nula. Un factor de Bayes mayor que uno indica que los datos son más compatibles con la hipótesis alternativa, mientras que un factor de Bayes menor que uno indica que los datos son más compatibles con la hipótesis nula. El factor de Bayes se puede interpretar como una actualización de las creencias sobre las hipótesis después de ver los datos. [10]

IV. Solución propuesta

Teniendo en cuenta las definiciones anteriores, nuestra variable respuesta Vulnerabilidad es dicotómica, y nuestras variables predictoras son Sexo, Edad, Comuna, Estrato y Comuna, realizamos un modelo de regresión logístico bayesiano.

Para este caso se tiene que:

$$Y_i = \begin{cases} 1, & \text{si el individuo i es vulnerable} \\ 0, & \text{si el individuo i no es vulnerable} \end{cases} \quad (1)$$

Donde Y_i sigue una distribución de probabilidad de Bernoulli

$$Y_i \sim \text{Bernoulli-logit}(\theta_i), \quad i = 1, 2, \ldots, 1616$$

$$P(Y_i = 1) = \theta_i, \quad P(Y_i = 0) = 1 - \theta_i.$$

El predictor lineal está dado por:

$$\theta_i = \frac{e^{\beta^T X_i}}{1 + e^{\beta^T X_i}}$$

donde $\beta^T X_i = \beta_0 + \sum_{j=1}^{4} \beta_j x_{ij}$.

En el contexto bayesiano, la inferencia está basada en la información proporcionada por la distribución a posteriori de los parámetros, la cual está dada por:

$$f(\beta|y, X) = \frac{1}{\prod_{i=1}^{1616} \left(1 + e^{\beta^T X_i}\right)} \times \left(\prod_{i=1}^{1616} e^{\beta^T X_i y_i} \right) \times f(\beta)$$

Con $f(\beta)$ la distribución apriori para el vector de parámetros β. Para este estudio se optó por utilizar una distribución apriori poco informativa para cada parámetro, dada por:

$$\beta_j \sim N(0, \sigma^2), \quad j = 0, 1, \ldots, 4$$

Bajo el supuesto de independencia de las componentes del vector β

V. Selección del modelo

En este estudio, se utilizaron dos modelos diferentes para analizar los datos. El primer modelo tuvo en cuenta las variables enumeradas en la Tabla 1. Por otro lado, el segundo modelo es reducido, no se consideró a la variable Comuna. De esta manera, se compararon los resultados obtenidos con ambos modelos para evaluar la influencia de las variables adicionales en las medidas de interés.

V-1. Modelo 1:

- Y: Vulnerabilidad
- x_1**:** Comuna
- x_2**:** Sexo
- x_3**:** Edad
- x_4**:** Estrato

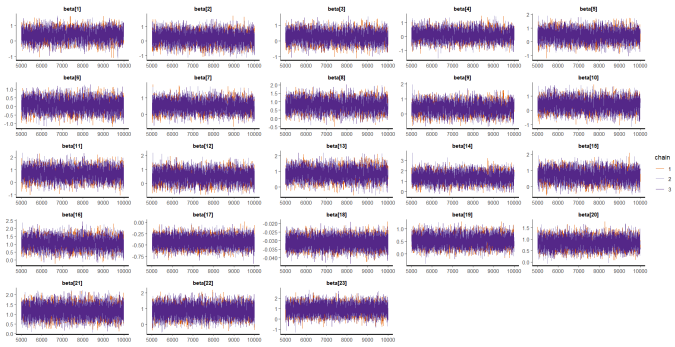
Con ayuda del software estadistico R [5], se obtuvo los siguientes resultados para verificar convergencia y estacionariedad:

Tabla II
RESUMEN DEL AJUSTE MODELO 1

β_i	Mean	se mean	sd	2.5 %	97.5 %	neff	$\hat{R}$
β_1	0.39	0.01	0.35	-0.29	1.06	3592	1
β_2	0.22	0.01	0.38	-0.52	0.97	4471	1
β_3	0.27	0.01	0.34	-0.40	0.95	3732	1
β_4	0.10	0.01	0.36	-0.61	0.82	3743	1
β_5	0.46	0.01	0.39	-0.31	1.23	4211	1
β_6	0.06	0.01	0.34	-0.60	0.74	3672	1
β_7	0.49	0.01	0.33	-0.15	1.13	3216	1
β_8	0.79	0.01	0.34	0.14	1.46	3462	1
β_9	0.38	0.01	0.35	-0.30	1.08	3542	1
β_{10}	0.38	0.01	0.39	-0.39	1.15	3581	1
β_{11}	0.68	0.01	0.44	-0.18	1.55	3557	1
β_{12}	0.50	0.01	0.40	-0.28	1.30	3188	1
β_{13}	0.84	0.01	0.36	0.13	1.55	4046	1
β_{14}	1.36	0.01	0.46	0.47	2.28	3743	1
β_{15}	0.68	0.01	0.35	0.01	1.38	3051	1
β_{16}	1.06	0.01	0.35	0.39	1.75	3083	1
β_{17}	-0.43	0.00	0.11	-0.64	-0.21	11495	1
β_{18}	-0.03	0.00	0.00	-0.04	-0.02	11691	1
β_{19}	0.52	0.00	0.21	0.11	0.94	6453	1
β_{20}	0.83	0.00	0.24	0.37	1.30	5046	1
β_{21}	1.17	0.00	0.30	0.58	1.78	5723	1
β_{22}	0.87	0.01	0.37	0.14	1.58	5279	1
β_{23}	0.94	0.01	0.48	-0.01	1.88	6716	1

En la tabla II, siendo β_1 el intercepto, β_2, ... , β_{17} las comunas de la 1 a la comuna 16, β_{18} Sexo (Hombre), β_{19} Edad, β_{20}, ... , β_{24} los estratos del 1 al 6

Figura 3. Traceplot del Modelo 1



V-2. Modelo 2: Para este modelo, en vista a las comunas no significativas que tiene la variable Comuna se decide ignorar dicha varible. Por tanto, el modelo queda de la siguiente manera:

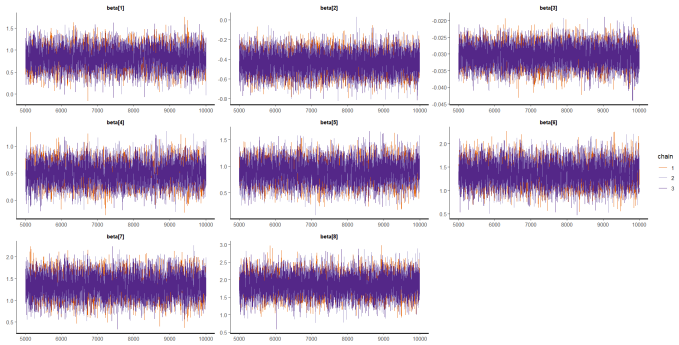
- Y: Vulnerabilidad
- x_2**:** Sexo
- x_3**:** Edad
- x_4**:** Estrato

De igual forma que en el modelo anterior, se hace uso del software R para obtener los resultados:

Tabla III
RESUMEN DEL AJUSTE MODELO 2

β_i	Mean	se mean	sd	2.5 %	97.5 %	n eff	$\hat{R}$
β_1	-0.82	0.00	0.23	-1.28	-0.36	4605	1
β_2	0.44	0.00	0.11	0.22	0.65	8946	1
β_3	0.03	0.00	0.00	0.02	0.04	8753	1
β_4	-0.50	0.00	0.20	-0.89	-0.11	4726	1
β_5	-0.89	0.00	0.20	-1.29	-0.49	4833	1
β_6	-1.36	0.00	0.24	-1.83	-0.89	5391	1
β_7	-1.33	0.00	0.25	-1.83	-0.84	5811	1
β_8	-1.85	0.00	0.29	-2.42	-1.29	6406	1

Figura 4. Traceplot del Modelo 2



Realizando los métodos anteriormente definidos.
Con el factor de Bayes se tienen en cuenta los 2 modelos:

- M_1 : Y $\sim$ x_1 + x_2 + x_3 + x_4
- M_2 : Y $\sim$ x_2 + x_3 + x_4

Donde su prueba de hipótesis esta sería:

$$BF_{1,2} = \frac{\int p(D|M_1)\,p(M_1)}{\int p(D|M_2)\,p(M_2)} = \frac{1{,}128534e^{-219}}{1{,}671865e^{-220}} \approx 6{,}750151$$

Con el resultado obtenido del Factor de Bayes, hay una fuerte evidencia que respalda el hecho de que M_1 explica mejor los datos a comparacion de M_2

Con el método de LOOCV y la ayuda de la libreria "loo."[en] R se comparan los modelos M_1 y M_2, en donde se obtuvo:

Tabla IV
COMPARACION DE MODELOS LOOVC

Modelo	elpd diff	se diff
M_1	0.00	0.00
M_2	-0.2	5.70

Los resultados obtenidos de la tabla IV hay una pequeña diferencia entre M_1 y M_2 en cuanto a la diferencia de la densidad puntual predictiva (elpd diff) lo cual indica que M_1 tiene un ligero mejor rendimiento en cuando al M_2. Por otro lado la diferencia del error cuadrado (se diff) de M_2 es mucho mayor en comparación a M_1, lo que suguiere que M_2 tiene mayor variabilidad en termino a las predicciones.

Los resultados obtenidos por el metodo LOOVC son consistentes a los obtenidos por el Factor de Bayes.

En la curva ROC obtenemos lo siguiente:

Figura 5. Curva ROC del Modelo 1

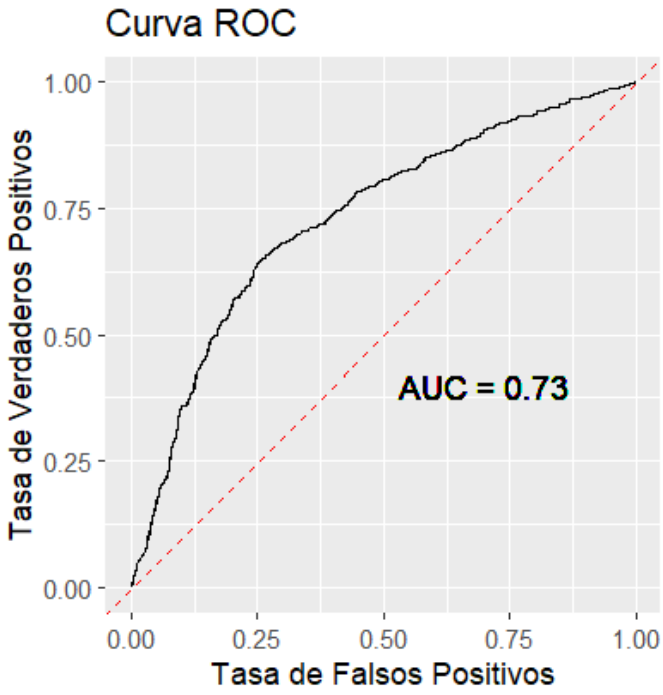
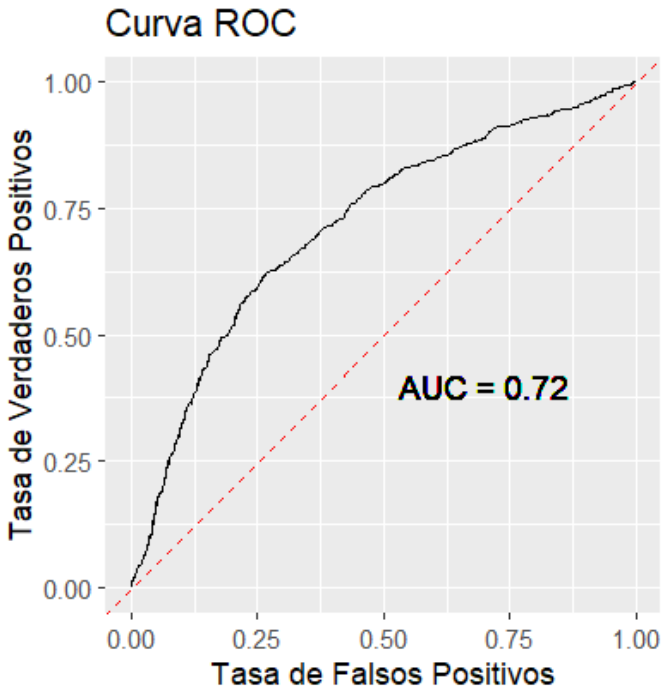


Figura 6. Curva ROC del Modelo 2



En este caso, la diferencia entre los valores de 0.73 y 0.72 es mínima. Por lo tanto, no se puede afirmar con certeza que uno de los modelos sea significativamente mejor que el otro en función únicamente de estos valores.

VI. Conclusiones

- Todas las variables del modelo full, Sexo, Edad, Comuna y Estrato, son significativas al momento de explicar la vulnerabilidad de un individuo.
- El modelo full es una buena herramienta para tomar decisiones acerca de mejoras para la ciudad, ya que tiene una buena capacidad predictiva y según los métodos utilizados, es mejor que el modelo reducido.
- Podemos ver con el método de MCMC que cumple con los requisitos, ya que no se puede concluir que no converge y además sus cadenas son estacionarias.
- Según el análisis descriptivo se ha observado que los grupos de mayor edad suelen enfrentar mayores riesgos y desafíos, lo que los hace más propensos a la vulnerabilidad.
- Las condiciones socioambientales de una comuna o área geográfica pueden tener un impacto significativo en la vulnerabilidad de las personas. Factores como la infraestructura, la calidad de vivienda, la disponibilidad de servicios básicos y la exposición a riesgos pueden influir en la vulnerabilidad. Es por esto que se observa que la Comuna 14 (Poblado) tiene más porcentaje de No vulnerabilidad, mientras que otras comunas si tienen más proporción de vulnerabilidad, según el análisis descriptivo. Lo cual tiene sentido ya que Poblado se conoce por ser una zona de riqueza en la ciudad.
- Dado nuestro modelo, para futuras investigaciones se pueden interpretar los parámetros dados en la tabla II para identificar poblaciones más vulnerables y así tomar mejores decisiones para la ciudad.
- Teniendo en cuenta que el nivel de referencia para cada variable es el primero, en el caso de la comuna, solo las comunas 8, 13, 14, 15 y 16 representan un cambio significativo comparando con la comuna 1

Referencias

[1] J. Hernandez, D. Cardona, A. Segura ,*Construcción y análisis de un índice de vulnerabilidad social en la población joven*, Medellín, Colombia: Revista Latinoamericana de Ciencias Sociales, 2017. https://www.redalyc.org/journal/773/77355376026/html/#B4.
[2] J. Rodriguez, *Vulnerabilidad y grupos vulnerables: un marco de referencia conceptual mirando a los jóvenes*, Santiago de Chile, Chile: 2001. http://repositorio.cepal.org/bitstream/handle/11362/7150/S018659_es.pdf.
[3] L. Castro, R. Cano ,*POBREZA Y VULNERABILIDAD: FACTORES DE RIESGO EN ELPROCESO EDUCATIVO*, Chile: Contextos Educativos, 2013. https://publicaciones.unirioja.es/ojs/index.php/contextos/article/view/1290/1209.
[4] *Medellín Cómo Vamos* Medellín, Colombia, 2022. https://www.medellincomovamos.org/calidad-de-vida/encuesta-de-percepcion-ciudadana
[5] J. Uribe, D. Puentes ,*Código de proyecto en GitHub*, Colombia: Universidad Nacional de Colombia, 2023. https://github.com/jiurgi/Bayesiana.
[6] V. Pascual de Olmo, *Métodos avanzados de muestreo : MCMC*, Madrid, España: Universidad Carlos III de Madrid, 2011. http://hdl.handle.net/10016/13735.
[7] H. Chitarroni, *La regresión Logística*, Buenos Aires, DIC/2002 https://racimo.usal.edu.ar/83/1/Chitarroni17.pdf.
[8] A. Vehtari, A. Gelman, J. Gabry.

Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017. https://arxiv.org/abs/1507.04544.

[9] J.A Martinez, P.S Pérez*La Curva ROC*, Madrid, España, 2022. https://www.sciencedirect.com/science/article/abs/pii/S1138359322001952?via%3Dihub.

[10] Kass, R. E., & Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90(430), 773-795.