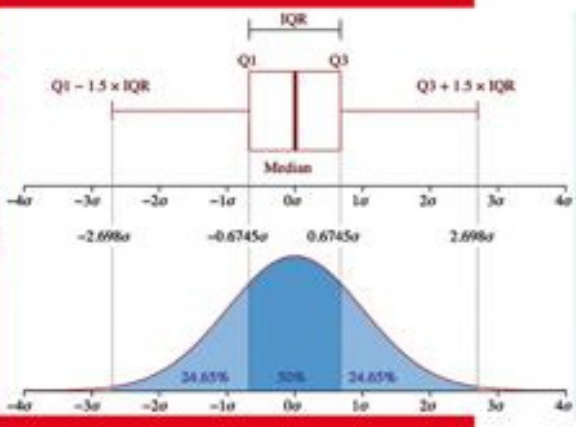
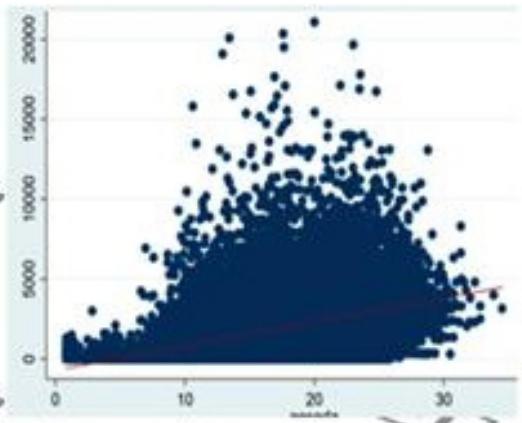
Universidad Mayor de San Andrés
Facultad de Ciencias Puras y Naturales
Carrera de Estadística
Instituto de Estadística Teórica y Aplicada

Varianza Nº 26

Revista del Instituto de Estadística Teórica y Aplicada

ISSN 2789-3510





$$=E[(x-\mu)^2]$$
$$=E[x^2-2x\mu+\mu^2]$$
$$=E[x^2]-2\mu E[x]+\mu^2$$
$$=E[x^2]-2\mu^2+\mu^2$$
$$=E[x^2]-\mu^2$$
$$=E[x^2]-E[x]^2$$

Varianza

Revista de Investigación Científica del
Instituto de Estadística Teórica y Aplicada

Número 26
Octubre, 2025
La Paz - Bolivia

Universidad Mayor de San Andrés
Facultad de Ciencias Puras y Naturales
Carrera de Estadística
Instituto de Estadística Teórica y Aplicada (I.E.T.A.)

ISSN 2789-3510 VERSIÓN IMPRESA
ISSN 2789-3529 VERSIÓN EN LÍNEA

DEPÓSITO LEGAL
4-1-285-2021 P.O**.**

REVISTA VARIANZA
Nº 26 - Octubre, 2025

DIRECTOR CARRERA DE ESTADÍSTICA
Ph. D. Juan Carlos Flores López

DIRECTOR a.i. INSTITUTO DE ESTADÍSTICA TEÓRICA Y APLICADA
Lic. Esp. Raúl León Delgado Álvarez

VISIBILIDAD: REVISTAS BOLIVIANAS

La Paz - Bolivia
Edificio Bloque FCPN - Campus Cota Cota
Teléfonos: 2612824 -2612844
Email: *ieta@umsa.bo*
Página web: https://ojs.umsa.bo/ojs/index.php/revistavarianza

COMITÉ EDITORIAL

COMITÉ CIENTÍFICO INTERNACIONAL

Martha Patricia Bohorquez Castañeda, Ph.D.

(Estadístico)

Universidad Nacional de Colombia

Bogotá, Colombia

E-mail: *mpbohorquezc@unal.edu.co*

Adriana D'Amelio, Mg.

(Estadístico)

Universidad Nacional de Cuyo

Mendoza, Argentina

E-mail: *estat06@hotmail.com*

COMITÉ CIENTÍFICO NACIONAL

Franz Cuevas Quiroz, Ph.D.

(Informático)

Universidad Mayor de San Andrés

La Paz, Bolivia

E-mail: *franzcq@gmail.com*

Ernesto Eusebio Cupe Clemente, M.Sc.

(Matemático)

Universidad Mayor de San Andrés

La Paz, Bolivia

E-mail: *ecupe@fcpn.edu.bo*

PRESENTACIÓN

El Instituto de Investigación Teórica y Aplicada de la Carrera de Estadística, de la Facultad de Ciencias Puras y Naturales de la Universidad Mayor de San Andrés, se complace en presentar a la comunidad en general y al ámbito académico-científico en particular, la edición N° 26 de su revista científica "Varianza".

Esta edición reafirma nuestro compromiso con la divulgación del conocimiento estadístico mediante cuatro artículos originales que abordan problemáticas de relevancia actual con solvencia metodológica. Cabe resaltar que, en un significativo avance hacia la equidad de género, tres de los cuatro trabajos son autoría de mujeres que contribuyen al enriquecimiento de la ciencia estadística, reflejando así el creciente y valioso liderazgo femenino en la disciplina.

Los artículos incluidos en esta edición son:

1. *Análisis e Impacto de los Datos Atípicos para la Confiabilidad de los Resultados. Este estudio subraya la importancia crítica de identificar y analizar datos atípicos para garantizar la fiabilidad de los resultados y evitar interpretaciones sesgadas en la investigación.*

2. *Relación entre Educación e Ingresos en Bolivia: Un Análisis de la Encuesta de Hogares 2021. A partir de estos datos, la investigación analiza la correlación entre años de estudio e ingresos personales, encontrando una cifra inferior al 50%. Empleando análisis de componentes principales y regresión, se concluye que en Bolivia los años de estudio no se traducen directamente en mayores ingresos.*

3. *Predicción del Rendimiento Académico mediante Machine Learning: Regresión Logística vs. Árbol de Decisión. Este artículo evalúa y compara la eficacia de estos dos algoritmos para predecir el rendimiento académico en estudiantes de la Escuela Militar de Ingeniería (2022). Demuestra que el modelo de Regresión Logística, con ajuste de hiperparámetros y validación cruzada, alcanza una precisión del 90.38%, superando al árbol de decisión.*

4. *Factores Determinantes de la Fecundidad en Mujeres Adultas en Bolivia: Un enfoque con Regresión de Poisson, utilizando datos de la Encuesta de Demografía y Salud (2023), este trabajo identifica los factores sociodemográficos y económicos asociados a la fecundidad. El análisis provee evidencia empírica de que la baja escolaridad, el nivel de*

riqueza y el acceso a la planificación familiar continúan determinando los patrones reproductivos, ofreciendo insumos valiosos para el diseño de políticas públicas.

El reconocimiento a las y los autores de los artículos, por el esfuerzo y dedicación a la investigación.

Finalmente, invito a visitar en la web la página de la revista Varianza que se encuentra suscrita a las revistas UMSA a través de la siguiente dirección electrónica y código QR.

https://ojs.umsa.bo/index.php/revistavarianza

Lic. Esp. Raúl León Delgado Álvarez
DIRECTOR a.i. INSTITUTO DE ESTADÍSTICA TEÓRICA Y APLICADA

Nuestro más sincero agradecimiento al Comité de revisores, nacionales e internacionales, cuya desinteresada dedicación y expertise fueron pilares esenciales para materializar la edición N° 26 de "Varianza". Su compromiso con el avance del conocimiento estadístico, reflejado en cada evaluación minuciosa, ha enriquecido incuestionablemente el valor de esta publicación.

ÍNDICE

ARTÍCULOS ORIGINALES

Artículo Original
Revista Varianza Nro. 26, Octubre de 2025
ISSN 2789-3529 versión en línea
DOI: *https://doi.org/10.53287/qhvv8628oh49s*
Páginas: 1-8

MÉTODOS EN LA IDENTIFICACIÓN DE DATOS ATÍPICOS

METHODS IN THE IDENTIFICATION OF ATYPICAL DATA

Veronica Cuenca Ramallo[1]
Universidad Mayor de San Andrés, La Paz-Bolivia
✉ *vcramallo12@gmail.com*

Artículo recibido: 10/09/2025
Artículo aceptado: 16/10/2025

RESUMEN

Los datos atípicos son en ocasiones una cuestión subjetiva (significa que depende de la percepción, interpretación o criterio personal de quien analiza), y no de una regla totalmente fija u objetiva, ya que su identificación depende del contexto de la investigación y de los criterios utilizados. No obstante, existen diversos métodos que permiten clasificar y analizar estos valores para obtener resultados más confiables y evitar interpretaciones erróneas.

El objetivo es conocer los métodos para determinar los datos atípicos en diferentes investigaciones recomendando su identificación y analizando el alcance de sus efectos sobre la serie de datos, con el fin de evitar errores en los resultados finales, especialmente dentro la aplicación de modelos de regresión lineal o múltiple.

Se implementa el método descriptivo, para conocer el objeto de estudio, así como el método analítico y sintético (Analizar los distintos tipos de valores atípicos, Integra los distintos métodos de identificación para proponer un enfoque más completo), que permite determinar los puntos de datos observados que se alejan de la línea de mínimos cuadrados (errores o residuales). La principal aportación de este trabajo consiste en destacar la importancia de identificar oportunamente los valores atípicos y proponer un enfoque combinado de métodos que contribuye a mejorar la precisión de los análisis estadísticos y la fiabilidad de los resultados obtenidos.

En conclusión, es necesario determinar los valores atípicos para alcanzar resultados más precisos y evitar que estos influyan negativamente en las conclusiones de la investigación y en la correcta interpretación de los valores.

Palabras clave: Datos atípicos, Modelos de regresion lineal, Modelos de regresión multiple, Investigación social.

ABSTRACT

Outliers are sometimes a subjective matter (meaning they depend on the personal perception, interpretation, or judgment of the analyzer) and not a completely fixed or objective rule, as their identification depends on the context of the research and the criteria used. However, there are various methods that allow these values to be classified and analyzed to obtain more reliable results and avoid misinterpretations.

The objective is to understand the methods for determining outliers in different research projects, recommending their identification and analyzing the extent of their effects on the data series, in order to avoid errors in the final results, especially when applying linear or multiple regression models.

[1] Docente de la Carrera de Estadística de la Facultad de Ciencias Puras y Naturales de la Universidad Mayor de San Andrés. ORCID: 0009-0003-0196-7862

The descriptive method is implemented to understand the object of study, as well as the analytical and synthetic method (analyzing the different types of outliers, integrating the different identification methods to propose a more comprehensive approach), which allows for the identification of observed data points that deviate from the least squares line (errors or residuals). The main contribution of this work is to highlight the importance of identifying outliers in a timely manner and to propose a combined approach to methods that contributes to improving the accuracy of statistical analyses and the reliability of the results obtained.

In conclusion, it is necessary to identify outliers to achieve more accurate results and prevent them from negatively influencing research conclusions and the correct interpretation of values.

Keywords: Atypical data, Linear regression models, Multiple regression models, Social research.

1. INTRODUCCIÓN

En estadística, se cuenta con muestreo estratificado, siendo una de ellas el valor atípico considerado como una observación que es numéricamente distante del resto de los datos. Las estadísticas derivadas de los conjuntos de datos que incluyen valores atípicos serán frecuentemente engañosas.

Los valores atípicos pueden ser indicativos de datos que pertenecen a una población diferente del resto de las muestras establecidas.

Los valores atípicos son en ocasiones una cuestión subjetiva, y existen algunos métodos para clasificar y obtener resultados. El método más utilizado por su sencillez y resultados es el test de Tukey, que toma como referencia la diferencia entre el primer cuartil Q1 y el tercer cuartil Q3, o rango intercuartílico.

En un diagrama de caja se considera un valor atípico el que se encuentra 1,5 veces esa distancia de uno de esos cuartiles (atípico leve) o a 3 veces esa distancia (atípico extremo). (Greene, 2016)

Las series económicas pueden estar influidas por una serie de procesos no determinista, ni conocidos para el analista, podrían incidir en que estas observaciones presenten estructura distinta al del resto de estas series, teniendo la capacidad de sesgar los resultados obtenidos y de afectar la capacidad de estimaciones de los modelos. (Gujarati,2017)

Al iniciar el análisis estadístico, necesariamente se debe observar si existen datos atípicos que puedan incidir sobre los resultados. La inclusión de valores atípicos influye en la estimación de los parámetros en los modelos de series temporales estacionarios, debido a que la posición de los atípicos será siempre desconocida.

Se recomienda identificarlos y estimar sus efectos, con el objetivo de eliminar sus efectos sobre la serie, ya que los datos atípicos pueden incidir en el análisis de regresión. (Lindsey, 2017)

Los métodos a disposición para la identificación de valores atípicos son los siguientes: el de identificación a partir de gráficos. El método de la normalización de variables: identificación a partir de la distancia de la media. El método de la identificación a partir de la prueba de Tukey. El método de la identificación a partir de la prueba de Tukey ajustada.

2. MATERIAL Y MÉTODOS

En este artículo se implementa el método descriptivo, con la finalidad de conocer el objeto de estudio, en este caso, el conocimiento de los Datos Atípicos, son

fundamentales en el estudio de la Estadística. Se utiliza el método analítico, sintético, ya que permite determinar los puntos de datos observados que se alejan de la línea de mínimos cuadrados, tomando en cuenta el error o residuo que llega a ser la distancia vertical de la línea al punto.

Posteriormente, se realiza la síntesis para recomponer el objeto de investigación articulándolo con los elementos descompuestos, generando así las conclusiones y recomendaciones.

El método utilizado para el desarrollo de la investigación, fue el de acción participativa, cuyo objetivo es producir conocimiento y sistematizar las experiencias, a fin de consolidar la implementación del proceso de obtención de los datos atípicos. Se utilizó la técnica de revisión bibliográfica, ya que se recurrió a fuentes de información, que permitieron establecer directrices, doctrina y conceptos; referentes a la aplicación de los métodos de obtención de datos atípicos.

El muestreo utilizado fue no probabilístico, empleando para ello juicio u opinión; esto en atención a la complejidad del proceso, ya que la población al contactar generalmente asume con mucha precaución, exponer sus experiencias.

En cuanto al concepto de Datos Atípicos, se refiere a la serie de programas, políticas, iniciativas y actividades que se distribuyen en forma ordenada, una vez que se ha identificado la presencia de una observación atípica, se debe investigar su procedencia y si se concluye que se ha generado por errores en el muestreo se debe eliminar. Es conveniente repetir el análisis estadístico sin la observación atípica y examinar las nuevas conclusiones.

Estas transformaciones teórico-doctrinarias, organizacionales están orientadas a incrementar la mejor aproximación a la obtención de un resultado exacto y real; para que responda con efectividad las necesidades en la observación atípica, que se denomina influyente. En este punto, el investigador debe enjuiciar si es posible su aparición por un error experimental (eliminarlo) o si tal observación podría volver a buscar modelos más complejos.

La investigación es fundamentalmente cualitativa, con una intencionalidad descriptiva y exploratoria; con esta metodología se extraen las implicaciones observables, particularmente con referencia a los fundamentos, la identificación de valores atípicos en la predicción de series temporales es importante porque los valores atípicos influyen en el modelo de predicción que se utiliza para predecir valores futuros

Los criterios sirven para sintetizar los fundamentos de cómo las variables independientes causan un cambio en el proceso; la teoría, la investigación parte del supuesto de la experiencia.

El objetivo final de esta investigación es proponer recomendaciones basados en los aspectos positivos en la aplicación de los métodos para la identificación de datos atípicos en el análisis estadístico.

3. RESULTADOS

Un valor atípico es una observación extrañamente grande o pequeña. Los valores atípicos pueden tener un efecto desproporcionado en los resultados estadísticos, como la media, lo que puede conducir a interpretaciones engañosas..

Cuadro N° 1.
Causas Comunes de los valores atípicos (outliers)

CAUSAS	ACCIONES POSIBLES
Error de entrada de datos	Corregir el error y volver a analizar los datos
Problema del proceso	Investigar el proceso para determinar la causa del valor atípico.
Factor faltante	Determinar si no se consideró un factor que afecta el proceso.
Probabilidad aleatoria	Investigar el proceso y el valor atípico para determinar si este se produjo en virtud de las probabilidades; realice el análisis con y sin el valor atípico para ver su impacto en el resultado.

Fuente: Soporte de Minitab (R) 18

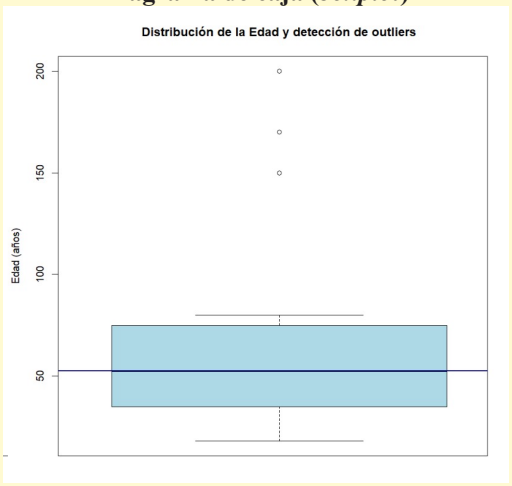
1. Identificación a partir de gráficos

El gráfico de caja (boxplot) constituye una primera opción al momento de analizar e identificar datos atípicos, el mismo presenta la mediana, el primer y tercer cuartil, además del 1.5 o rango intercuartílico. En el caso de R, se puede verificar que la opción determinada, permite identificar los valores considerados como atípico y los valores utilizados para representar el *boxplot*. (Gujarati,2017).

A modo de ejemplo, se puede utilizar la base de datos del Censo 2012, considerando la variable edad, para ilustrar la identificación de valores atípicos mediante un diagrama de caja

```
boxplot (datos EDAD,
        main = "Distribución de la Edad y detección de outliers",
      ylab = "Edad (años)",
      col = "lightblue")
```

Figura 1.
Diagrama de caja (*boxplot*)



Fuente: Elaboración propia, Censo Nacional de Población y Vivienda 2012 de Bolivia, disponibles públicamente a través del INE (https://anda4.ine.gob.bo).

2. Normalización de variables: identificación a partir de la distancia de la media

Otra estrategia para la identificación de valores atípicos, consiste en normalizar la variable de interés de la forma tradicional, lo que permite obtener una nueva variable xz, que se interpreta como el número de unidades (positivas o negativas, dependiendo del signo) en que se una observación se encuentre alejada de la media de la serie. (White, 2019)

3. Identificación a partir de la prueba de Tukey

Una alternativa es utilizar una medida de dispersión robusta a valores atípicos, y posteriormente establecer los rangos que permitan la identificación de los datos atípicos. (Tobin, 2018)

4. Identificación a partir de la prueba de Tukey ajustada

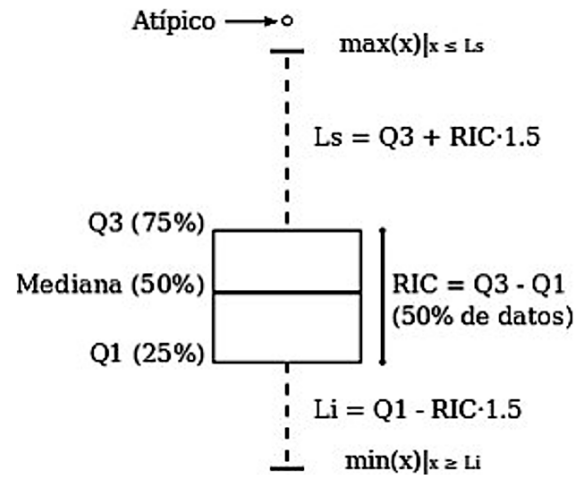
Una alternativa práctica para aumentar la sensibilidad de la prueba, es obtener los quintiles a partir del vector de variables

originales recortado a partir de los valores extremos. (Rundel, 2017)

5. Identificación a partir de la distancia de Cook

A partir del enfoque de regresión se puede utilizar el concepto de distancia de Cook. Esta realizar la estimación del cambio de cada valor ajustado, con la i-ésima observación. Por lo que, mide la influencia de cada observación. (*j*) es el valor de la respuesta ajustada *j*, cuando se incluyen todas las observaciones; es el valor de la observación *j* cuando no se incluye la observación *i*, *P* es el número de coeficientes en el modelo de regresión; por último, MSE es el error cuadrático medio del modelo con todas las observaciones. (Tobin, 2018) basado en Tukey (1977).

Figura. 2.
Diagrama del Valor Atípico



Fuente: Tukey, J. W. (1977). *Exploratory data analysis. Addison-Wesley.*

4. DISCUSIÓN

4.1 RESULTADOS DE LOS VALORES ATÍPICOS

En algunos conjuntos de datos, hay valores (puntos de datos observados), llamados valores atípicos. Los valores atípicos son puntos de datos observados que se alejan de la línea de mínimos cuadrados. Tienen grandes "errores", donde el "error" o residual es la distancia vertical de la línea al punto.

Los valores atípicos deben examinarse de cerca. A veces, por una u otra razón, no deben incluirse en el análisis de los datos. Es posible que un valor atípico sea el resultado de datos erróneos. Otras veces, un valor atípico puede contener información valiosa sobre la población estudiada y debe seguir incluyéndose en los datos.

La clave está en examinar cuidadosamente las causas de que un punto de datos sea un valor atípico. (Faraway, 2017)

El método más habitual por su sencillez y resultados es el test de Tukey, que toma como referencia la diferencia entre el primer cuartil (Q1) y el tercer cuartil (Q3), o rango intercuartílico. En un diagrama de caja se considera un valor atípico el que se encuentra 1,5 veces esa distancia de uno de esos cuartiles (atípico leve) o a 3 veces esa distancia (atípico extremo). Se trata de un método paramétrico que supone que la población es normal (Figura 3). No obstante, también existen métodos no paramétricos cuando la muestra no supere la prueba de normalidad correspondiente.

Además de los valores atípicos, una muestra puede presentar ciertos puntos, denominados puntos influyentes, que tienen un efecto significativo sobre los resultados del análisis.

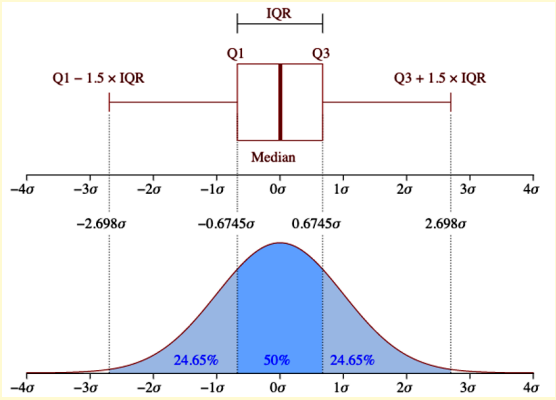
Los datos atípicos distorsionan los resultados de los análisis y por esta razón hay que identificarlos y tratarlos de manera adecuada, generalmente excluyéndolos del análisis (Rundel, 2017)

Se trata de puntos de datos observados que están alejados de los demás en la dirección horizontal. Estos puntos pueden tener un

gran efecto en la pendiente de la línea de regresión.

Para empezar a identificar un punto influyente, puede eliminarlo del conjunto de datos y ver si la pendiente de la línea de regresión cambia significativamente.

Figura 3.
Detección paramétrica de valores atípicos, basado en la curva de distribución normal



Fuente: Valores Atípicos Universidad Politécnica de Madrid (Yepes, 2018)

Es posible identificar visualmente posibles valores atípicos mediante las observaciones del diagrama de dispersión y la línea de mejor ajuste. Sin embargo, sería importante contar con alguna directriz sobre la distancia que debe tener un punto para considerarse un valor atípico. Los datos atípicos son ocasionados por:

a). Errores de procedimiento

b). Acontecimientos extraordinarios

c). Valores extremos. Por ejemplo, una muestra de datos del número de cigarrillos consumidos a diario contiene el valor porque hay un fumador que fuma sesenta cigarrillos al día.

d). Causas no conocidas

Como regla general, podemos señalar como valor atípico cualquier punto que esté situado más de dos desviaciones típicas por encima o por debajo de la línea de mejor ajuste. La desviación típica utilizada es la de los residuales o errores. (Springer, 2016)

Podemos hacerlo visualmente en el diagrama de dispersión al dibujar un par de líneas adicionales que estén dos desviaciones típicas por encima y por debajo de la línea de mejor ajuste.

Todos los puntos de datos que se encuentren fuera de este par de líneas adicionales se marcan como posibles valores atípicos. Alternativamente, podemos hacerlo numéricamente, al calcular cada residual y compararlo con el doble de la desviación típica.

En el enfoque gráfico es más fácil. En primer lugar, se muestra el procedimiento gráfico, seguido de los cálculos numéricos. Por lo general, solo tendrá que utilizar uno de estos métodos.

En el ejemplo de la variable edad del Censo 2012, se calcularon los cuartiles en R, lo que facilitó el análisis de la distribución de los datos y la detección de valores atípicos.

```
Calcular cuartiles
Q1 <- quantile(edad_censo, 0.25)
Q3 <- quantile(edad_censo, 0.75)
IQR_val <- Q3 - Q1
# Calcular límites para detectar outliers
limite_inferior <- Q1 - 1.5 * IQR_val
limite_superior <- Q3 + 1.5 * IQR_val
# Identificar outliers
outliers <- edad_censo[edad_censo < limite_inferior | edad_censo > limite_superior]
# Resultados
cat("Primer cuartil (Q1):", Q1, "\n")
cat("Tercer cuartil (Q3):", Q3, "\n")
cat("Rango intercuartílico (IQR):", IQR_val, "\n")
cat("Límite inferior:", limite_inferior, "\n")
cat("Límite superior:", limite_superior, "\n")
```

cat("Valores atípicos detectados en Edad:", outliers, "\n")

Resultados

Primer cuartil (Q1): 23

Tercer cuartil (Q3): 68

Rango intercuartílico (IQR): 45

Límite inferior: -44.5

Límite superior: 135.5

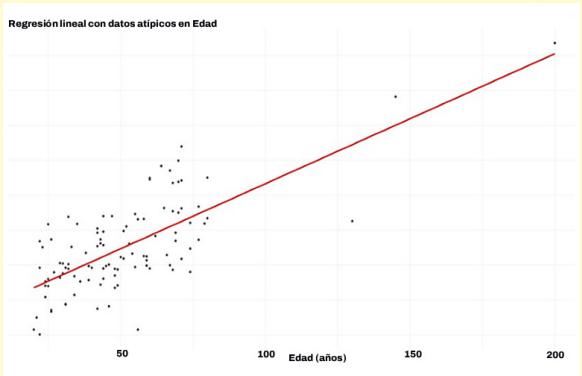
Valores atípicos detectados en Edad: 150

En estudios como el análisis de la variable edad del Censo 2012, la regresión lineal permite identificar outliers, analizar su efecto y garantizar que el modelo represente fielmente la relación entre variables.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- β_0 = intercepto (valor esperado de Y cuando $X = 0$)
- β_1 = pendiente (cambio esperado en Y por unidad de cambio en X)
- ε = término de error aleatorio (residuo)

Figura 4.
Regresión Lineal



Fuente: Elaboración propia, Censo Nacional de Población y Vivienda 2012 de Bolivia, disponibles públicamente a través del INE (https://anda4.ine.gob.bo).

5. CONCLUSIONES

1. Es fundamental conocer los métodos para identificar los datos atípicos en distintas investigaciones, así como evaluar sus efectos sobre la serie, con el fin de evitar errores en los resultados finales.

2. Estos ajustes teórico-práctico y organizacionales están orientadas a incrementar la mejor aproximación a la obtención de un resultado exacto y real; para que responda con efectividad las necesidades en la observación atípica, que se denomina influyente. En este punto, el investigador debe enjuiciar si es posible su aparición por un error experimental (y por tanto eliminarlo) o si es necesario considerarlo para desarrollar modelos más complejos.

3. Es necesario investigar tanto el proceso como el valor atípico para determinar si su aparición se debe a fenómenos probabilísticos, se recomienda realizar el análisis con y sin el valor atípico para evaluar su impacto en los resultados.

4. Los valores atípicos son observaciones que se alejan significativamente de la línea de mínimos cuadrados. Presentan grandes “errores”, es decir, la distancia vertical de la línea al punto, por lo que deben examinarse cuidadosamente. Dependiendo del caso, un valor atípico puede ser resultado de un error en los datos o contener información relevante sobre la población estudiada, por lo que se debe decidir si se excluye o se mantiene en el análisis.

CONFLICTO DE INTERESES

La autora declara que no hay conflicto de intereses con respecto a la publicación de este documento.

REFERENCIAS BIBLIOGRÁFICAS

Badhan, A., & Ganpati, A. (2024). *Overview of outlier detection methods and evaluation metrics. In Challenges in Information, Communication and Computing Technology* (pp. 736-741). CRC Press.

Durdağ, U. M., & Yılmaz, S. (2024). *Minimum - variance-based outlier detection method using geodetic networks. Geodesy and Geodynamics*, 17(4), 2187-2199.

Faraway. (2017). *Linear Models*. Boston - USA: McGraw-Hill.

Fisher, R.A. (2010). *Limiting forms of the frequency distribution of the largest or smallest member of a sample*. . London, England.

Fréchet, M. (2010). *Sur la loi de probabilité de l'écart maximum. Annales de la Société Mathématique Polonaise*, Paris, France.

Geoffroy, J. (1990). *Contributions a la théorie des valeurs extrêmes*. Paris, France: Publications de l'Institut de Statistique de l'Université de Paris.

Gnedenko, B. (1980). *Sur la distribution limite du terme maximum d'une serie aléatoire.* . Paris, France: McGraw-Hill.

Greene, W.H. (2016). *Analisis Econométrico*. Madrid España: Prentice-Hall Editions.

Gujarati, D. N. (2017). *Econometría*. Ciudad de Mexico, Mexico: McGraw-Hill.

Kale, B.K. (2000). *Estimation of expected life in the presence of an outlier observation. Technometrics*, Madrid, España.

Karamata, J. (2000). *Sur un mode de croissance régulière des fonctions. Mathematica*. Madrid, España: McGraw-Hill.

Neyman, J.(2013). *Outlier proneness of phenomena and of related distributions, In Optimizing Methods in Statistics* . New York, EE.UU.: Academic Press.

Rundel, C, (2017). *Openintro Statistics*. Barcelona, España: McGraw-Hill.

Tobin, J. (2018). *Estimation of relationships for limited dependent variables. Econometrica*. Madrid, España: McGraw-Hill.

Von Mises, R. (1998). *La distribution de la plus gran de n valeurs*. Math Union Interbalkan, Paris, France.

White, H. (2019). *A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity*. Brussels, Belgium: McGraw-Hill.

Yepes, R. (2018). *¿Qué hacemos con los valores atipicos ?* Madrid, España: Universidad Politécnica de Madrid.

Artículo Original
Revista Varianza Nro. 26, Octubre de 2025
ISSN 2789-3529 versión en línea
DOI: *https://doi.org/10.53287/dddi7166kc43g*
Páginas: 9-22

ANÁLISIS CORRELACIONAL ENTRE LOS INGRESOS PERSONALES Y LOS AÑOS DE ESTUDIO EN BOLIVIA

CORRELATIONAL ANALYSIS BETWEEN PERSONAL INCOME AND YEARS OF STUDY IN BOLIVIA

Raúl León Delgado Álvarez[1]
Universidad Mayor de San Andrés, La Paz - Bolivia
✉ *rldelgado3@umsa.bo*

Jaime Tito Pinto Ajhuacho[2]
Universidad Mayor de San Andrés, La Paz - Bolivia
✉ *titojaime_pinto@yahoo.com*

Artículo recibido: 09/09/2025
Artículo aceptado: 16/10/2025

RESUMEN

En Bolivia con la información de la Encuesta de Hogares año 2021, se realizó un análisis de correlación entre ingresos personales y años de estudio, la influencia de esta última muestra una correlación con coeficiente menor a 0.5, indicando que no hay una buena correlación, lo que en otros países muestra una tendencia positiva.

En la búsqueda de explicar una correlación buena que explique el comportamiento de los ingresos se encontró otras variables independientes que explican la variable ingresos personales mostrando un Coeficiente de Determinación con tendencia positiva.

En el presente estudio se realiza un análisis correlacional para mostrar que el coeficiente de correlación entre el ingreso personal y los años de estudio no es significativo, lo que no ocurre en otros países, destacándose que en Bolivia a mayores años de estudio no se tienen mejores ingresos, esta situación tiene factores socioeconómicos que son de mucho interés, en vista de que la educación superior es una inversión para obtener mejores ingresos en el futuro.

Palabras clave: Correlación, ingresos personales, años de estudio.

ABSTRACT

In Bolivia, with information from the 2021 Household Survey, a correlation analysis was carried out between personal income and years of study. The influence of the latter shows a correlation with a coefficient less than 0.5, indicating that there is no good correlation, which in other countries shows a positive trend.

In the search for a strong correlation that explains the behavior of income, other independent variables were found that explain the personal income variable, showing a Coefficient of Determination with a positive trend.

In the present study, a correlational analysis is conducted to show that the correlation coefficient between personal income and years of education is not significant, which is not the case in other countries. It is worth noting that in Bolivia, higher years of education do not lead to higher incomes. This situation has socioeconomic factors that are of great interest, given that higher education is an investment in obtaining better incomes in the future.

[1] Docente de la Carrera de Estadística de la Facultad de Ciencias Puras y Naturales de la Universidad Mayor de San Andrés. ORCID: 0000-0001-7886-5264

[2] Docente de la Carrera de Estadística de la Facultad de Ciencias Puras y Naturales de la Universidad Mayor de San Andrés. ORCID: 0000-0002-1157-2165

Keywords: Correlation, personal income, years of education.

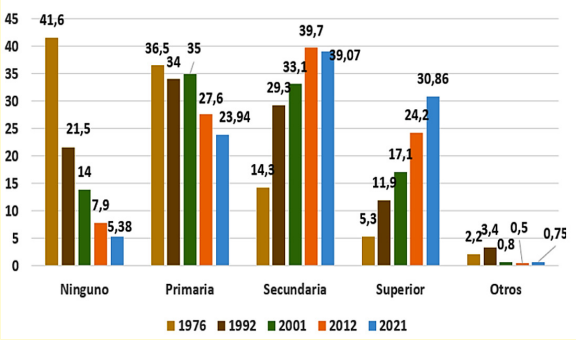
1. INTRODUCCIÓN

La educación en Bolivia es un derecho universal y responsabilidad del Estado, está constituida según niveles (Primaria, Secundaria y Superior) a los cuales se tiene acceso en función de la edad, pues al establecer los años de estudio de las personas se conocen los niveles máximos alcanzados.

La disparidad entre la cobertura tecnológica y el acceso a materiales educativos entre zonas urbanas y rurales es un desafío.

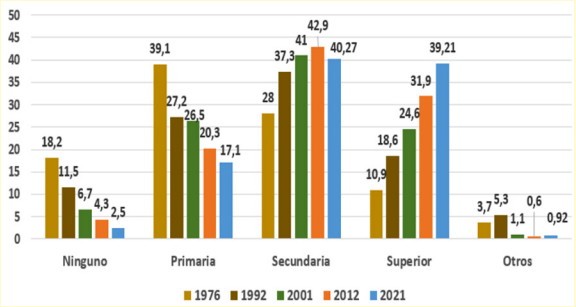
En referencia al nivel educativo general, según información del Instituto Nacional de Estadísticas se tiene lo siguiente:

GRÁFICO 1.
BOLIVIA: Nivel de instrucción alcanzado de la población de 19 años o más de edad, Censos 1976, 1992, 2001, 2013 y EH 2021
(En porcentaje)



Fuente: Instituto Nacional de Estadística, EH 2021.

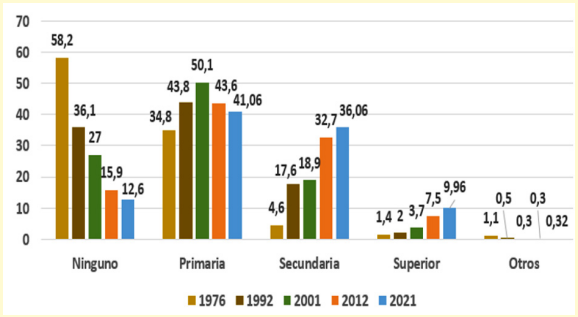
GRÁFICO 2.
BOLIVIA AREA URBANA: Nivel de instruccion más alto alcanzado de la poblacion de 19 años o más de edad, Censos 1976, 1992, 2001, 2012 y EH 2021
(En Porcentaje)



Fuente: Instituto Nacional de Estadística, EH 2021

Observando los datos del nivel de instrucción más alto alcanzado de la población de 19 años o más de edad, se puede ver que hasta el año 2021 el 30,86% alcanzó el nivel Superior.

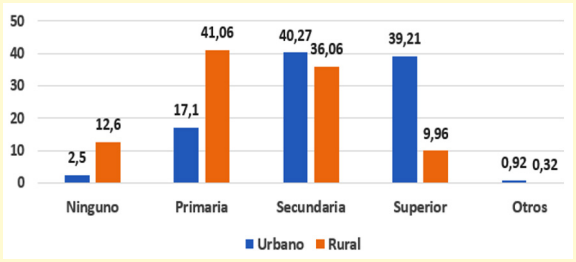
GRÁFICO 3.
BOLIVIA AREA RURAL: Nivel de instruccion mas alto alcanzado de la poblacion de 19 años o mas de edad, Censos 1976, 1992, 2001, 2012 y EH 2021
(En Porcentaje)



Fuente: Instituto Nacional de Estadística, EH 2021

En la desagregación por áreas urbana y rural, se observa que en el área rural desciende en el nivel superior al 9,96% en el año 2021.

GRÁFICO 4.
BOLIVIA: Comparación de Nivel de instrucción más alto alcanzado de la población, Encuesta de Hogares 2021
(En Porcentaje)



Fuente: Elaboración propia.

El Gráfico 4, muestra el nivel educativo "Secundaria" alcanzado por la población urbana igual a 40,27% y en "Superior" se tiene un porcentaje similar (39,21%), en cambio en el área rural el nivel "Secundaria" indica 36,06% y a nivel "Superior" se tiene un descenso que apenas alcanza al 10% de la población, esto evidencia que hay posibles factores que influyen en este descenso.

El ingreso personal es todo el dinero o los beneficios económicos que una persona o un hogar recibe de diversas fuentes, este ingreso es la remuneración principal que se recibe por el desempeño de un trabajo o actividad laboral, incluye sueldos, salarios, comisiones, bonificaciones, horas extras, rentas de propiedades, pensiones, intereses de inversiones, y otras transferencias o prestaciones, que pueden ser de naturaleza ordinaria o extraordinaria.

Se puede indicar que los ingresos personales en la mayoría de los casos se relacionan con la actividad laboral en Bolivia, y existen variadas formas de contratos laborales y la principal problemática del país es la alta tasa de informalidad, que afecta a la mayoría de la población ocupada.

En los últimos años la tasa de informalidad subió, más del 80%, al mismo tiempo, el país enfrenta desafíos como la precariedad laboral, especialmente en el sector informal, y la necesidad de garantizar empleos dignos y salarios suficientes, concentrando la mayor parte de los empleos en el sector terciario (comercio, servicios de alojamiento y comida, transporte) en el área urbana, mientras que en el sector rural domina el sector primario (agricultura).

La informalidad laboral en Bolivia ha crecido de 62,4% a 84,2% entre 2005 y 2024, según datos del Instituto Nacional de Estadística.

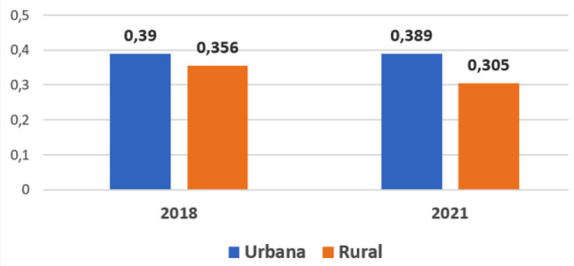
Tomando en cuenta que el acceso a la educación es uno de los derechos humanos más importantes para todas las personas, es la educación, la base para mejorar la calidad de vida de las personas.

En Bolivia, los profesionales solo representan el 8,17% de la población con empleo. (INE/2011), una encuesta realizada por el Banco Interamericano de Desarrollo (BID), en Bolivia indica para el año 2015, el 25% de las empresas no pueden contratar a los profesionales de reciente titulación porque no tienen experiencia. En el año 2019 de 19.000 titulados universitarios, las empresas sólo requirieron aproximadamente el 27% de profesionales, y el 73% restante está desempleado.

El Centro de Estudios de la Realidad Económica y Social en Bolivia, indica que una persona demora 7 meses en conseguir un empleo, tiempo bastante largo, tomando en cuenta que en otros países como Perú y Chile demoran solo 3 meses.

Se conoce por referencia que, en otros países, un mayor grado de educación permite obtener un mayor ingreso económico; la evidencia de esta relación llama a la investigación, especialmente en países en vías de desarrollo.

GRÁFICO 5.
BOLIVIA: CORRELACION DE INGRESOS PERSONALES Y AÑOS DE ESTUDIO POR URBANA/RURAL, SEGÚN AÑOS 2018-2021 (EH-2018, EH-2021)



Fuente: Instituto Nacional de Estadística, EH 2021

La correlación de ingresos personales y años de estudio en los años 2018 al 2021 disminuyó tanto en el área urbana como rural.

1.1 OBJETIVO DE LA INVESTIGACIÓN

Objetivo General

Análisis de la correlación de ingresos personales y nivel educativo alcanzado en Bolivia en las áreas urbana y rural.

Objetivos Específicos

1. Elaborar estadísticas de indicadores de ingreso personal vs. años de estudio en Bolivia.

2. Evaluar los ingresos personales con indicadores del modelo lineal a nivel nacional (urbana y rural) de Bolivia.

3. Analizar indicadores del ingreso personal y nivel educativo de la población de Bolivia.

4. Examinar y analizar los ingresos personales en función de otros factores a nivel nacional, urbana y rural de Bolivia.

2. ANTECEDENTES

La educación de una sociedad tiene como objetivo mejorar las condiciones de vida de una comunidad, para conocer la realidad del acontecer educativo se puede trabajar en la construcción de indicadores globales para considerar las políticas educativas.

Un país debe estar comprometido con la calidad educativa, conocer a fondo los indicadores educativos más relevantes que contribuyan a la implementación de los objetivos de desarrollo sostenible.

Estos indicadores deben medir el comportamiento y desempeño dentro de un proceso educativo, estos deben ser cuantificables, medibles y, dentro de lo posible, estar bajo control y seguimiento.

En el proceso de conocer indicadores respecto a los niveles educativos, se debe garantizar una recolección eficiente de datos y contar con herramientas de medición y análisis poderosos que provean una visión clara y objetiva de las medidas a tomar para cumplir los objetivos en especial de políticas públicas.

Datos de la Oficina de Estadísticas Laborales de EE. UU. muestran que los individuos con títulos universitarios o superiores tienen ingresos medios significativamente más altos que aquellos con niveles educativos más bajos.

Investigaciones en países como Bolivia y Paraguay confirman que la educación está fuertemente correlacionada con los ingresos, y que cada año adicional de estudio puede resultar en un aumento proporcional en el salario.

Autores y teóricos abordan la correlación entre los años de estudio e ingresos personales, Gary Becker, es un economista pionero en la teoría del capital humano, sus trabajos explican cómo la inversión en educación (aumento de años de estudio) y otras habilidades incrementa la productividad individual, y por ende, el potencial de ingresos. Un articulo al respecto cita “Al ser más productivo y valioso, el ingeniero se vuelve más competitivo en el mercado laboral, justificando un salario más alto y oportunidades de ascenso más rápidas”.

Jacob Mincer, es conocido por su "ecuación de Mincer" (1974), que formaliza la relación entre el salario, la educación y la experiencia laboral; esta ecuación es fundamental en el estudio del capital humano y la determinación de ingresos.

La relación entre años de estudio e ingresos personales suele ser positiva, pero su magnitud varía significativamente entre países; en Estados Unidos, Japón y China muestran un alto gasto público en educación.

En América Latina, como en Chile y Argentina, la relación también es evidente, aunque con cifras de ingresos distintas. Estados Unidos, por su parte, ha cuantificado que las personas con estudios universitarios (licenciatura) tienden a ganar sumas significativamente mayores a lo largo de su vida que aquellos con estudios de secundaria.

3. MÉTODO

La metodología que se utilizó tiene un enfoque cuantitativo con un alcance correlacional. Inicialmente, se determinó la correlación entre la variable dependiente (ingreso de las personas) y la variable independiente (años de estudio) usando el instrumento de medición (Coeficiente de Pearson) siendo que ambas variables son cuantitativas.

Los instrumentos que se utilizaron:

3.1 INSTRUMENTO DE ANÁLISIS CORRELACIONAL

El instrumento que permite ver la correlación que se obtiene entre las variables de estudio para explicar la realidad de ingresos personales y años de estudio es el Coeficiente de correlación de Pearson (r), medida estadística que cuantifica la fuerza y la dirección de la relación lineal entre dos variables. Este coeficiente nos ayuda a entender si existe una relación lineal entre dos conjuntos de datos y, en caso afirmativo, qué tan fuerte y en qué dirección es esa relación, su valor varía entre -1 y +1, donde:

- **1:** Indica una correlación positiva perfecta (a mayor valor de una variable, mayor valor de la otra).
- **-1:** Indica una correlación negativa perfecta (a mayor valor de una variable, menor valor de la otra).
- **0:** Indica que no hay correlación lineal entre las variables.

La interpretación de la fuerza de la correlación también puede ser más específica recurriendo al siguiente criterio: Entre 0.00 y 0.29: Correlación débil, entre 0.30 y 0.59: Correlación moderada y entre 0.60 y 1.00: Correlación fuerte.

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

3.2 MODELO LINEAL GENERAL

Un modelo lineal general es un modelo estadístico en el que la variable dependiente es una combinación lineal de variables independientes, y se usa para unificar varios métodos estadísticos, como ANOVA y regresión.

La ecuación del Modelo Lineal General, puede expresar en forma matricial como $Y = X\beta + U$, donde Y es la matriz de variables dependientes, X es la matriz de variables independientes, β es el vector de coeficientes a estimar y U es el vector de errores. Esta es una generalización de la regresión lineal que permite manejar múltiples variables dependientes simultáneamente.

Este modelo es un marco flexible que permite analizar variables de respuesta con distribuciones diferentes a la normal, ya que se relaciona la media de la variable con el predictor lineal a través de una función de enlace y permite que la varianza sea una función de su valor predicho.

Un modelo lineal generalizado (MLG) tiene tres componentes principales:

Componente aleatorio: Identifica la variable de respuesta y su distribución de probabilidad.

Predictor lineal: Especifica las variables explicativas (predictores) mediante una ecuación de predicción lineal.

Función de enlace: Vincula la esperanza (media) de la variable de respuesta con el predictor lineal, modelando la relación entre ellos.

Se adapta a diversos tipos de datos, incluyendo variables categóricas y no normales, lo que lo hace muy versátil.

3.3 ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

El Análisis de Componentes Principales (ACP), o PCA por sus siglas en inglés, es una técnica multivariante que reduce la dimensionalidad de los datos creando nuevas variables (componentes principales) que son combinaciones lineales de las originales y capturan la mayor parte de la variabilidad del conjunto de datos.

Su objetivo es simplificar datos complejos, identificar patrones subyacentes y facilitar la visualización de datos de alta dimensión, así como la reducción de la redundancia en los datos.

En su aplicación realiza:

Reducción de la dimensionalidad. PCA encuentra un nuevo conjunto de ejes (los componentes principales) que son ortogonales entre sí.

Maximización de la varianza. El primer componente principal (PC1) se elige como el eje que maximiza la varianza de los datos proyectados.

Creación de nuevas variables. El segundo componente principal (PC2) es perpendicular a PC1 y maximiza la varianza restante, y así sucesivamente para los componentes siguientes.

Selección de componentes. Se seleccionan los componentes que explican la mayor parte de la variabilidad total, generalmente utilizando un criterio como autovalores mayores que uno, o también observando un gráfico de sedimentación para ver dónde la pendiente disminuye.

Se recurre a componentes principales para generar una regresión lineal, y luego calcular el coeficiente de determinación R2 para evaluar que también se ajusta el modelo. El enfoque conocido como regresión sobre componentes principales, es utilizado en lugar de las variables originales para estimar el modelo de regresión y calcular su R^2.

R^2 mide la proporción de la varianza de la variable dependiente que es explicada por los componentes principales en el modelo de regresión. Un R^2 más cercano a 1 indica que el modelo se ajusta bien a los datos.

Los datos que se utilizaron en el presente análisis son de la Encuesta Nacional de Hogares realizada por el INE (2021).

4. RESULTADOS

Inicialmente, se realizó un primer análisis exploratorio de los datos, para revisar su consistencia y tener una visión general del comportamiento de éstos, la revisión, muestra lo siguiente:

4.1. ANÁLISIS DE CORRELACIÓN INGRESOS PERSONALES Y AÑOS DE ESTUDIO

El Análisis de datos ha tomado variables que tienen que ver con la correlación entre años de estudio e ingresos personales en Bolivia, ésta relación de datos estadísticos ha develado el estado del nivel educativo respecto a los ingresos.

CUADRO 1.
BOLIVIA: CORRELACION DE PEARSON DE INGRESOS PERSONALES Y AÑOS DE ESTUDIO (EH-2018-2021).

COEFICIENTE	AÑO	
	2018	2021
r	0,4049	0,3930

Fuente: Elaboración propia

El Cuadro 1, indica que la fuerza de la relación lineal entre las variables es muy baja o inexistente (menor al 50%) y que esta relación ha tendido a debilitarse con el tiempo, aunque con algunas fluctuaciones. Esto implica que hay poca o nula asociación lineal entre las variables, es decir, no se puede predecir el comportamiento de una variable

basándose en la otra de manera consistente, lo cual es la interpretación de correlación débil o inexistente.

CUADRO 2.
BOLIVIA: CORRELACIÓN DE INGRESOS PERSONALES Y AÑOS DE ESTUDIO POR DEPARTAMENTOS, SEGÚN AÑOS 2018-2021 (EH-2018, EH-2021).

DEPARTAMENTO	AÑO	
	2018	2021
Chuquisaca	0,417	0,420
La Paz	0,398	0,385
Cochabamba	0,378	0,369
Oruro	0,440	0,351
Potosí	0,430	0,392
Tarija	0,374	0,387
Santa Cruz	0.392	0,409
Beni	0,479	0,424
Pando	0,497	0,435

Fuente: Elaboración propia

A nivel departamental los coeficientes de correlación disminuyen o en algunos casos hay un leve ascenso, pero todos son menores al 50% en el periodo 2018-2021.

CUADRO 3.
BOLIVIA: CORRELACIÓN DE INGRESOS PERSONALES Y AÑOS DE ESTUDIO POR URBANA/RURAL, SEGÚN AÑOS 2018-2021 (EH-2018, EH-2021)

ÁREA	AÑO	
	2018	2021
Urbana	0,390	0,389
Rural	0,356	0,305

Fuente: Elaboración propia

Las características entre las zonas, urbana y rural, son similares, lo que implicaría una relación fuerte y una tendencia conjunta, pero los valores de correlación que disminuyen indican que esta relación se está reduciendo, volviéndose más dispersa, y esa pérdida de coherencia es preocupante porque indica un cambio o una desconexión entre las variables.

CUADRO 4.
BOLIVIA: INGRESOS PROMEDIOS PERSONALES (Bs./mes) POR DEPARTAMENTO, SEGÚN AÑOS 2018-2021. (EH-2018, EH-2021).

DEPARTAMENTO	AÑO	
	2018	2021
Chuquisaca	1.097	989
La Paz	1.355	1.251
Cochabamba	1.452	1.304
Oruro	1.244	1.161
Potosí	962	968
Tarija	1.478	1.290
Santa Cruz	1.557	1.443
Beni	1.108	1.254
Pando	1.234	1.076

Fuente: Elaboración propia.

La interpretación estadística de que los ingresos en Potosí (entre 962 y 968 Bs/mes) son bajos es un indicador de un reducido nivel de vida y posible precariedad económica, pues los ingresos son significativamente inferiores al umbral que apenas alcanza los 1.000 Bs/mes en otros departamentos. Este dato sugiere una desigualdad económica considerable, dado que Potosí experimenta una situación económica menos favorable que el resto del país, limitando el acceso a bienes, servicios y oportunidades para sus habitantes.

CUADRO 5.
BOLIVIA: PORCENTAJE DE PERSONAS POR NIVEL EDUCATIVO GENERAL, SEGÚN AÑOS 2018-2021.
(EH-2018, EH-2021).

NIVEL EDUCATIVO GENERAL	AÑO	
	2018	2021
Ninguno	11,2	10,6
Primaria	30,9	29,9
Secundaria	37	37,8
Superior	20,6	21,2
Otros	0,3	0,5
Total	**100**	**100**

Fuente: Elaboración propia.

Según los datos recopilados entre 2018 y 2021, el 37% de la población encuestada ha completado la educación secundaria,

mientras que solo el 21% ha llegado a completar la educación superior.

Esto sugiere una brecha significativa en la finalización de los estudios superiores en comparación con la educación secundaria, donde el porcentaje es notablemente mayor.

4.2. APLICACIÓN DEL MODELO DE REGRESIÓN LINEAL

Los procedimientos estadísticos como la elaboración de indicadores y el modelo de regresión lineal ayudan a conocer y explicar el proceso educativo.

El P-valor F es 0,0000 y nos da indicios de confiabilidad del modelo.

CUADRO 6.
BOLIVIA: ANALISIS DE REGRESION LINEAL VARIABLES INGRESO PERSONAL, AÑOS DE ESTUDIO Y NIVEL EDUCATIVO GENERAL, (EH-2021)

MODELO	COEFICIENTES
Constante	306,227
Años de estudio (aestudio)	209,216
Nivel Educativo general (niv_ed_g)	-432,598
Coeficiente de Determinación	0,1592
Coeficiente de Pearson	0,393

Variable dependiente: Ingreso personal (Bs./Mes).
Fuente: Elaboración propia en base eh2021.

4.3. ANÁLISIS VARIABLES INGRESO PERSONAL, AÑOS DE ESTUDIO, NIVEL EDUCATIVO Y EDAD

Aplicando componentes principales podemos crear nuevas variables para el análisis de regresión y poder explicar el comportamiento de ingreso personal y otras variables independientes:

Componentes principales.

aeseda=0,6937*aestudio+0,6902*niv_ed_g+0,2061*s01a_03

Análisis de regresión. En el cuadro ANOVA se tiene el coeficiente de Determinación igual a 0,2570 y su coeficiente de correlación 0,5070, la variable de respuesta "yper" se puede explicar por la variable (aeseda).

Planteando la regresión lineal para la estimación del modelo propuesto se tiene: $yper = \hat{\beta}_o + \hat{\beta}_1 * aeseda.$ El modelo es

$\widehat{yper}$ = -707,18+150,23*aeseda, donde:

y_i = Ingreso personal (Bs/Mes) (**yper**).
aeseda = la variable combinación de:
x_{1i} = Años de estudio (**aestudio**)
x_{2i} = Nivel educativo general (**niv_ed_g**).
x_{3i} = Edad (**s01a_03**)

CUADRO 7.
CUADRO DE ANALISIS DE LA VARIANZA (ANOVA) VARIABLE YPER VS, AESEDA

Source	SS	df	MS	Number of obs	=	37,624
				F(1, 37622)	=	13014.76
Model	3.5336e+10	1	3.5336e+10	Prob > F	=	0.0000
Residual	1.0215e+11	37,622	2715078.59	R-squared	=	0.2570
				Adj R-squared	=	0.2570
Total	1.3748e+11	37,623	3654222	Root MSE	=	1647.7

yper	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
aeseda	150.2331	1.316884	114.08	0.000	147.6519	152.8142
_cons	-707.1813	19.72723	-35.85	0.000	-745.8472	-668.5154

Fuente: Elaboración propia.

Según muestra el Cuadro 7 (ANOVA) se aprecia que el P-valor F es 0,000 esto indica que hay confiabilidad en el modelo.

Un intervalo de confianza de nivel (1 - α)% para el parámetro β_1 (pendiente de la recta de regresión poblacional) está dada por:

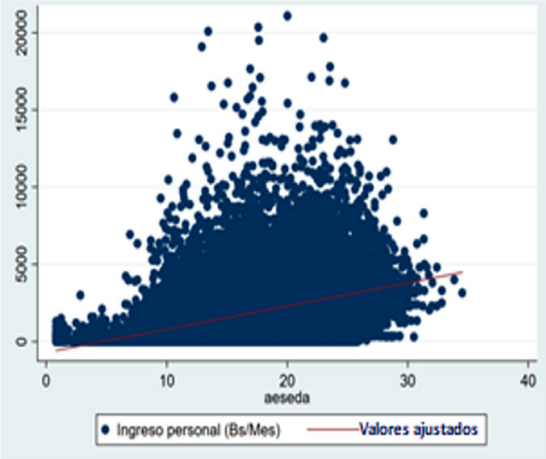
$\hat{\beta}_1 \pm z_{\frac{\alpha}{2}} SE(\hat{\beta}_1)$, donde $z_{\alpha/2}$ es el valor asignado para un nivel de confianza y $SE(\hat{\beta}_1)$ es el error estándar del estimador de la pendiente, el intervalo para un nivel de confianza de 95% da: [147,6522; 152,814]

Si aeseda = 3,03, multiplicando los extremos del IC por la constante obtendremos: [447,38 ; 463,02].

De igual forma el intervalo de confianza para

el parámetro $\hat{\beta}_0 \pm z_{\frac{\alpha}{2}} SE(\hat{\beta}_0)$, donde $z_{\alpha/2}$ es el valor asignado para un nivel de confianza y es el error estándar del estimador $\hat{\beta}_0$, el intervalo para un nivel de confianza de 95% da:[-745,83 ; -668,52]

CUADRO 6.
BOLIVIA: DISPERSION Y TENDENCIA DEL ÍNDICE AESEDA E INGRESO PERSONAL



Fuente: Elaboración propia.

La grafica de los datos, muestra la relación que existe entre "aeseda" y "yper". su relación tiene una tendencia a ser lineal, pero se nota una dispersión y correlación que apenas supera el 50,70% al ajustar una línea de regresión por el método de los mínimos cuadrados ordinarios.

CUADRO 8.
CORRELACIÓN DE PEARSON

	yper	aeseda
yper	1.0000	
	40294	
aeseda	0.5070	1.0000
	0.0000	
	37624	37624

Fuente: Elaboración propia con datos de EH_2021

De acuerdo al Coeficiente de correlación de Pearson; el resultado de la prueba establece una relación significativa entre las variables ingresos de las personas (yper) con la variable aeseda de 0,5070 es un alto nivel de significancia de P-valor de 0,000 menor a 0,05 establecido

4.4. ANÁLISIS VARIABLES INGRESO PERSONAL, AÑOS DE ESTUDIO, NIVEL EDUCATIVO GENERAL, EDAD E INGRESO LABORAL

En el análisis se plantea resumir cuatro factores en un índice total y encontrar su tendencia y la relación existente entre este índice y el ingreso personal (yper).

Inicialmente para la resolución del problema, revisamos el diagrama de dispersión entre **Y (yper)** y las variables $\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}$ y $\mathbf{X_4}$ **(aestudio, niv_ed_g, s01a_03, ylab)**, respectivamente, la evaluación es si resulta razonable pensar la existencia de una relación lineal entre la variable dependiente y las distintas variables independientes. Recurriendo al método de componentes principales tenemos:

La primera componente principal que contribuye con un 58,67% de la **variación** es:

aesylab = 0,6185*aestudio + 0,6126*niv_ed_g - 0,3666*s01a_03 + 0,3281*ylab

Planteando y aplicando la viabilidad del método de regresión lineal para la estimación del modelo propuesto se tiene:

$$yper = \hat{\beta}_o + \hat{\beta}_1 * aesylab.$$

Recurriendo a la Tabla de Análisis de la Varianza (ANOVA), tenemos:

CUADRO 9.
CUADRO DE ANALISIS DE LA VARIANZA (ANOVA) VARIABLE YPER VS, AESYLAB

Source	SS	df	MS		
				Number of obs	= 15,205
				F(1, 15203)	> 99999.00
Model	5.8013e+10	1	5.8013e+10	Prob > F	= 0.0000
Residual	2.8550e+09	15,203	187792.755	R-squared	= 0.9531
				Adj R-squared	= 0.9531
Total	6.0868e+10	15,204	4003432.23	Root MSE	= 433.35

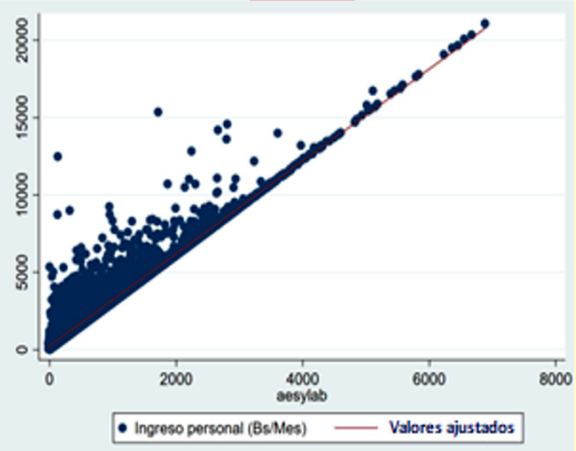
yper	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
aesylab	2.981195	.0053637	555.81	0.000	2.970682	2.991709
_cons	253.5764	5.932695	42.74	0.000	241.9476	265.2052

Fuente: Elaboración propia.

Se puede también apreciar que el P-valor F es 0,000 que nos indica que hay confiabilidad del modelo.

Los intervalos de confianza para el parámetro β_1 y β_0 de la recta de regresión poblacional) para un nivel de confianza de 95% es: [2,970682; 2,991] y [241,9503; 265,2025], respectivamente.

GRÁFICO 7.
BOLIVIA: DISPERSIÓN Y TENDENCIA DEL ÍNDICE AESYLAB E INGRESO PERSONAL.



Fuente: Elaboración propia.

La grafica de los datos, muestra la relación que existe entre "aesylab" y "yper". su relación tiene una tendencia a ser lineal, lo cual nos dice que se puede ajustar una línea de regresión.

El análisis de regresión obtenido con la incorporación de las variables independientes, indica que el coeficiente de determinación se incrementa; es decir, las variables independientes explican en un 95,31% la varianza de la variable dependiente ingresos de las personas por mes.

La prueba establece una relación significativa entre las variables yper y aesylab de 0,9763

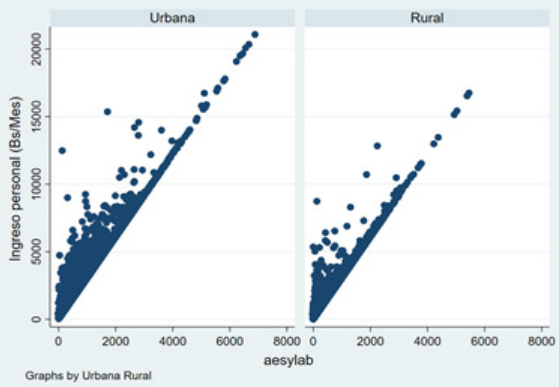
CUADRO 10
CORRELACIÓN DE PEARSON

	yper	aesylab
yper	1.0000	
	40294	
aesylab	0.9763	1.0000
	0.0000	
	15205	15205

Fuente: Elaboración propia con datos de EH_2021.

Sobre la dispersión y tendencia a nivel urbana y rural se tiene el Gráfico 8:

GRÁFICO 8.
BOLIVIA: DISPERSIÓN Y TENDENCIA DEL ÍNDICE AESYLAB E INGRESO PERSONAL A NIVEL URBANA Y RURAL.



Fuente: Elaboración propia.

Se puede ver en ambos casos que la tendencia es similar, con más dispersión en la urbana que la rural.

Los coeficientes de correlación y de determinación, corroboran lo indicado, con una aproximación al 100% de tendencia lineal.

CUADRO 11.
BOLIVIA: CORRELACIÓN DE INGRESOS PERSONALES Y VARIABLE AESYLAB POR URBANA/RURAL (EH-2021)

ÁREA	COEFICIENTE	
	PEARSON	DETERMINACIÓN
Urbana	0,975	0,951
Rural	0,973	0,946

Fuente: Elaboración propia

Una correlación de 0.975 entre los ingresos personales y un grupo de variables (años de estudio, nivel educativo, edad e ingreso laboral) indica una relación lineal positiva y muy fuerte, casi perfecta.

Esto significa que a medida que los años de estudio, el nivel educativo, la edad y el ingreso laboral tienden a aumentar, los ingresos personales también tienden a aumentar de manera muy predecible.

5. CONCLUSIONES

En referencia a los niveles educativos, en las zonas urbanas la concurrencia a los niveles educativos es mayor que en las rurales.

A nivel nacional la distribución de los niveles educativos en la población, de 19 años o más de edad, la participación a nivel primario y secundario es del 22,6% y 38,4% respectivamente y se percibe que en el área rural tienen una participación del 5,9 %, lo cual es bajo respecto al área urbana.

La interpretación económica de la baja participación educativa en el área rural de Bolivia se traduce en una menor productividad, menores ingresos y una mayor desigualdad económica, ya que la educación es un motor clave para el desarrollo individual y colectivo, permitiendo acceder a empleos mejor remunerados y fomentando la innovación y crecimiento económico.

Inicialmente los ingresos personales y años de estudio no muestran una correlación buena, apenas llegan al 50%, dato similar se ve en el área urbana y rural.

Una correlación baja o nula entre los años de estudio y los ingresos personales en las áreas urbanas y rurales sugieren que, aunque la educación suele ser un factor de mejora, existen otros elementos económicos, sociales o del mercado laboral que influyen mucho más en los niveles de ingreso, o que el sistema educativo no está alineado con las demandas del mercado, lo que limita la traducción de años de estudio en mayores ingresos.

Los ingresos personales y los años de estudio debieran estar correlacionadas porque la educación es una inversión en capital humano, que incrementa la productividad, las habilidades y las oportunidades laborales de una persona, resultando en salarios más altos y, en general, mayores ingresos a lo largo de su vida.

La correlación positiva entre los años de estudio e ingresos se interpreta económicamente bajo la teoría del capital humano. Esta teoría establece que la educación es una inversión que aumenta la productividad y habilidades de un individuo, lo que a su vez eleva su potencial de ingresos al mejorar sus oportunidades laborales y su capacidad para obtener ingresos más altos.

Incorporando otras variables independientes se puede explicar en un 95,31% en un modelo lineal se mejora la correlación de ingreso personal.

Amartya Sen (1999) sostiene que la mejor educación es fundamental porque expande las capacidades humanas, que son las oportunidades y libertades reales que una persona tiene para vivir como desea.

La educación no solo, incrementa el ingreso, sino que también es una capacidad valorable en sí misma, que permite a las personas ser y hacer cosas significativas par su vida y que, a su vez, mejora su potencial para obtener un ingreso y un trabajo digno.

6. DISCUSIÓN

La educación es un instrumento importante para que el individuo mejore sus condiciones de vida a partir del incremento de sus ingresos, en ella se puede percibir que hay diversos factores que influyen en su desarrollo.

Si analizamos el nivel educativo en general se observa que afecta al ingreso personal y es más notorio en al área rural que el urbano, y se ve disminución en el nivel educativo superior.

Las personas pobres por lo general provienen de hogares pobres y se los puede identificar en el área rural, el nivel secundario lo logran, pero su situación de pobreza es una dificultad en el que si se quiere lograr un nivel educativo superior.

Los argumentos a favor de la educación parecen mostrar una realidad, la educación por sí misma parece que no es suficiente para que las personas superen las condiciones de pobreza en que viven en el área rural.

Los ingresos personales son indicadores de la salud financiera de una persona y son clave para entender el comportamiento de consumo de los hogares y la economía en general.

Los ingresos personales y los años de estudio están relacionados porque la educación es una inversión, que aumenta la productividad, las habilidades y las oportunidades laborales de una persona, resultando en salarios más altos y, en general, mayores ingresos a lo largo de su vida.

Esta relación esta explicada por modelos estadísticos que consideran tanto la escolaridad como la experiencia laboral para estimar los ingresos de las personas.

La educación proporciona conocimientos y habilidades que son valorados por el mercado laboral, permitiendo a los individuos desempeñar trabajos más complejos y productivos, lo que se traduce en salarios más altos.

Los años dedicados a estudiar implican dejar de trabajar y ganar ingresos en ese período, pero la recompensa futura por esta inversión educativa suele ser mayor que el ingreso que se habría obtenido de no estudiar.

El análisis realizado indica que es posible una correlación entre los ingresos personales y los años de estudio, y existe evidencia empírica que demuestra una relación positiva y significativa: a mayor nivel educativo, generalmente mayores son los ingresos.

En la búsqueda de respuesta a la correlación de ingreso personal con otras variables independientes se encontró que la actividad laboral logra una correlación buena, en la cual se ve también que una explicación de los ingresos no está directamente en función de años de estudio.

Las personas de clase social alta tienen la oportunidad de acceder a mejores escuelas, colegios y universidades donde la calidad educativa es de mejor calidad; logrando de esta manera una ventaja en la formación de habilidades y capacidades, por lo tanto, también mejoras considerables en sus ingresos.

Las personas de clase media, al no poseer suficientes ingresos, trabajan y estudian asignando sus recursos en otras prioridades, como la alimentación y otros servicios básicos, reduciendo su perspectiva de crecimiento y acceso a establecimientos educativos superiores.

En cuanto a las personas de menores ingresos en el área urbana junto a los del área rural, su principal prioridad es la alimentación, de ellos existen grupos minoritarios que acceden a la formación primaria, pero en su gran mayoría no la culminan y menos la secundaria, empiezan a trabajar desde niños para lograr ingresos y poder sopesar los gastos del hogar.

El análisis refleja que existe una gran diferencia entre ingresos de las personas y años de estudio, siendo en el área rural muy baja la relación con los años de estudio y por ende bajos ingresos del hogar.

Parece inalcanzable el acceso para las personas del área rural a niveles superiores de educación, posiblemente por el alto costo de la educación superior, lo que sugiere que una verdadera política orientada a favorecer a las personas de menores ingresos, será aquella que permita acceder a estos a la educación superior, en consecuencia mejorar su economía familiar, también se hace necesario desarrollar nuevas habilidades en los estudiantes, principalmente de educación superior, favoreciendo a nivel personal y también como país.

Se debe poner mucha atención al análisis de que la correlación entre ingresos personales y años de estudios, porque con ella se valora el mercado laboral, lo cual se traduce en mejores oportunidades de empleo y salarios más altos.

La existencia de personas con educación superior plantea la necesidad de llevar a cabo investigaciones que permitan explicar más claramente los factores que explican esta situación.

La pobreza es más notoria en el área rural, en la que los años de estudio llegan hasta 19 años, analizando el área rural de Potosí vemos que la tendencia del nivel educativo general si tomamos en cuenta indicadores de pobreza, los niveles educativos en especial "Superior" tiende a disminuir.

Es importante tomar en cuenta estos hechos observables que se dan, y deben dar lugar a revisión de las políticas públicas, si queremos salir de la pobreza y viabilizar un país en vías de desarrollo tecnológico.

Se deben tomar en cuenta estos factores que influyen en la educación, los cuales pueden ser superables permitiendo que todas las personas pobres accedan a altos niveles educativos y la cual permita acceso a un mercado laboral digno y de superación, caso contrario las personas seguirán en estado de pobreza (por ingresos) pues no tendría ingresos que le permitieran superar la línea de pobreza o indigencia.

CONFLICTO DE INTERESES

Los autores declaran que no hay conflicto de interés con respecto a la publicación de este documento.

REFERENCIAS BIBLIOGRÁFICAS

Cao y Fernández Casa, (2020), *Técnicas de Simulación y Re muestreo,* Capítulo 9 , Métodos de remuestreo, rubenfcasal. github.io/simbook/bootstrap.html.

Carlos Gamero Buron (2015) José L. Iranzo Acosta, *Modelos Probabilísticos Variables Aleatorias Continuas*, Departamento de economía Aplicada, Universidad de Málaga.

Censo de Población y vivienda 2012, *Características de la Población, Estado Plurinacional de Bolivia,* Instituto Nacional de Estadística, febrero 2015.

Cochran, William G., (1996), *Técnicas de muestreo*, Compañía editorial Continental S.A. México.

Federico Stezano (2021), *Enfoques, definiciones y estimaciones de pobreza y desigualdad en América Latina y el Caribe*, Documentos de Proyectos NACIONES UNIDAS, CEPAL, FIDA-Invertir en la Población.

Gujarati Damodar, (2010), *Econometría*, McGRAW-HILL/INTERAMERICANA EDITORES, S.A. DE C.V.

Informe Nacional Voluntario 2021, *Estado Plurinacional de Bolivia*, año 2021.

Instituto Nacional de Estadística Bolivia, *Nivel de instrucción alcanzado por la población de 19 años o más edad por sexo*, *según área, 2011 – 2023*, INE Estado Plurinacional de Bolivia, año 2020.

Jorge Muro Guerrero, (2019), *Trabajo académico de Análisis de datos censurados: técnicas de estimación e inferencia no paramétricas y paramétricas*, Facultad de Ciencias, Universidad de Zaragoza.

Kunio Takezawa, K. (2006). *Introduction to nonparametric regression, National Agricultural Research Center*; Tsukuba -shi Ibaraki-ken, japan; Jhon Wiley & Sons.

Landa Fernando, (2023), *Curso Medición de la pobreza y desigualdad con STATA*.

Minor Mora Salas, Juan Pablo Pérez Saenz, CSO Facultad Latino Americana de Ciencias Sociales, Costa Rica. *http://bibliotecavirtual.clacso.org.ar/ar/libros/costar/flacso/cuad131.pdf*.

Reichmann, W. J., (1965), *Uso y abuso de las estadísticas*, Deusto, Bilbao.

Robín Cavagnoud, Sophia Lewandowski y Cecilia Salazar, (2015), *Pobreza, desigualdades y educación en Bolivia*, Boletín del Institut Français d' Etudes Andines.

Secretaria Municipal de Planificación para el desarrollo, Dirección de Investigación e información municipal, (2014), *Medición de la calidad educativa en el municipio de La Paz*, Gobierno autónomo Municipio de La Paz.

Siegel, Sidney, Castellan, N. John. (1998). *Estadística no Paramétrica aplicada a las ciencias de la conducta*. Trillas. México.

UDAPE, (2005), *Unidad de Análisis de Políticas Sociales y Económicas*, Análisis Económico, Volumen 20.

Walpole Ronald E., Raymond H. Myers, Sharon L. Myers y Keying Ye;(2012) *Probabilidad y estadística para ingeniería y ciencias*; Novena edición, Pearson Educación, México.

Artículo Original
Revista Varianza Nro. 26, Octubre de 2025
ISSN 2789-3529 versión en línea
DOI: *https://doi.org/10.53287/knfb3373zs29o*
Páginas: 23-34

PREDICCIÓN Y CLASIFICACIÓN DEL RENDIMIENTO ACADÉMICO A TRAVÉS DE MÉTODOS DE *MACHINE LEARNING*: REGRESIÓN LOGÍSTICA Y ÁRBOL DE DECISIONES

PREDICTION AND CLASSIFICATION OF ACADEMIC PERFORMANCE USING MACHINE LEARNING METHODS: LOGISTIC REGRESSION AND DECISION TREE

Benito Oscar Siñani Beltran[1]
Universidad Mayor de San Andrés, La Paz - Bolivia
✉ *benitoscar.sb@gmail.com*
Lizeth Mendoza Pinto[2]
Instituto Nacional de Estadística, La Paz - Bolivia
✉ *lmendoza.ps@gmail.com*

Artículo recibido: 15/09/2025
Artículo aceptado: 16/10/2025

RESUMEN

El objetivo principal de este trabajo de investigación es evaluar métodos de Machine Learning: regresión logística y árbol de decisiones para la predicción del rendimiento académico considerando diferentes variables sociodemográficas, académicas, económicas y otras, con base en registros de la gestión 2022 de la Escuela Militar de Ingeniería (E.M.I.), departamento de Ciencias Básicas. Los resultados obtenidos muestran que el método más preciso respecto a las métricas obtenidas (precisión, recall, especificidad, F1-score y otros) es el método de regresión logística con ajuste de hiperparámetros con validación cruzada con una precisión del 90.38%. El método del árbol de decisiones mostró tener una baja precisión con 47.81%. Para mejorar estas métricas se analizaron: el árbol de decisiones con ajuste de hiperparámetros, el Random Forest y el Random Forest con ajuste de hiperparámetros, siendo este último el que mostró las métricas mejoradas (54.6% de precisión). Los resultados obtenidos a partir del análisis de los métodos propuestos, servirán para identificar a los estudiantes en riesgo de bajos rendimientos y la toma de decisiones para evitar la posible deserción estudiantil.

Palabras clave: Machine learning, regresión logística, árbol de decisiones, rendimiento académico.

ABSTRACT

The main objective of this research work is to evaluate Machine Learning methods: logistic regression and decision tree for the prediction of academic performance considering different sociodemographic, academic, economic and other variables, based on records from the 2022 management of the Military School of Engineering (E.M.I.), Department of Basic Sciences. The results obtained show that the most accurate method with respect to the metrics obtained (precision, recall, specificity, F1-score and others) is the logistic regression method with hyperparameter tuning with cross-validation with an accuracy of 90.38%. The decision tree method showed low accuracy with 47.81%. To improve these metrics, the following were analyzed: the decision tree with hyperparameter tuning, the Random Forest and the Random Forest with hyperparameter tuning, the latter being the one that showed the improved metrics (54.6% accuracy). The results obtained from the analysis of the proposed methods will be used to identify students at risk of low performance and to make decisions to prevent possible student dropouts.

[1] Docente de la carrera de Estadística, Facultad de Ciencias Puras y Naturales de la UMSA. Docente de la Escuela Militar de Ingeniería. Maestría en Innovación Educativa. ORCID: 0000-0001-9562-5983

[2] Instituto Nacional de Estadística (INE). ORCID: 0009-0009-6142-8530

Keywords: Machine learning, logistic regression, decision tree, academic performance.

1. INTRODUCCIÓN

El rendimiento académico es un tema muy complejo y preocupante a nivel mundial, la UNESCO ha confirmado rendimientos deficientes en las pruebas PISA 2022 (Programa para la Evaluación Internacional de Estudiantes en América Latina y el Caribe), especialmente en el área de matemáticas, ciencias y lenguaje, en este informe, se muestra que un alto porcentaje de jóvenes de 15 años no tienen las competencias mínimas frente a los nuevos desafíos (UNESCO 2024). Esta población es la base para estudios universitarios, a la falta de competencias mínimas en estas áreas, se espera bajos rendimientos y/o el fracaso de los estudiantes en los primeros años de universidad.

En cuanto al uso de las tecnologías para estudiar el fenómeno de la deserción, el uso de tecnologías avanzadas, como la inteligencia artificial, el aprendizaje automático y el big data, también fueron destacados como un enfoque emergente para predecir y mitigar el abandono. Estos modelos de predicción, como se describió en varios artículos, lograron identificar con un alto grado de precisión a los estudiantes en riesgo, lo que permitió la implementación de intervenciones tempranas (Herreño M. Martha et al. 2024).

Frente a este problema, las instituciones de educación superior requieren identificar oportunamente a los estudiantes en riesgo y para eso es necesario recurrir a las herramientas que proporcionan las nuevas tecnologías, para diseñar e implementar diversas políticas que eviten bajos rendimientos y posterior abandono y/o fracaso de los estudiantes en instituciones de educación superior.

Bolivia no es la excepción respecto a esta problemática, algunos autores, hacen referencia a la deserción universitaria que es una consecuencia de los bajos rendimientos académicos.

En Bolivia el Sistema de la Universidad Boliviana SUB está conformado por 15 universidades públicas, 11 gratuitas y 4 privadas. El SUB en diez años ha tenido un crecimiento de casi el 100% en la matrícula estudiantil, pasando de 256.834 en el año 2004 a 440.918 en el año 2015. Las 11 universidades públicas gratuitas representan algo más del 75% de la matrícula universitaria total en Bolivia, registrando para el mismo periodo de tiempo una deserción definitiva promedio del 10,66% (Poveda V. Iván 2019).

Este tema refleja no solo el desempeño de los estudiantes en el aula, sino también lo efectivas que son las estrategias de enseñanza utilizadas, factores que influyen en el bajo rendimiento académico, sociales, económicos, académicos, psicológicos y otros, considerando que estamos en una etapa donde la tecnología es crucial y debe ser incorporada en el proceso de enseñanza aprendizaje. Conocer y/o determinar los factores que se relacionan directa o indirectamente con el rendimiento académico coadyuvaría en la implementación de políticas y/o programas para mejorar o incrementar el rendimiento académico en los estudiantes.

El *Machine Learning* responde a esta necesidad, proporcionando diferentes métodos y/o técnicas con las que podemos aprender diferentes tareas, reconocer patrones, hacer predicciones, etc., a partir de un conjunto de datos. En este trabajo se evaluarán dos métodos del Machine Learning: regresión logística y árboles de decisiones para predecir el rendimiento académico en estudiantes de la EMI gestión 2022 utilizando como software base Python.

2. MATERIALES Y MÉTODOS

El análisis se realizará sobre una base de datos que contiene información relevante de variables sociodemográficas, académicas y otras, de estudiantes de pregrado de Cs. Básicas de la EMI unidad académica La Paz gestión 2022, que considera información de 650 estudiantes.

A partir de la información se analizarán los métodos del Machine Learning (regresión logística y árbol de decisiones), para luego realizar la evaluación y mostrar las diferentes métricas de cada uno de los métodos.

2.1. DEFINICIÓN DE VARIABLES

En el presente estudio las variables a considerar son:

- Regresión logística: se han considerado todas las variables independentes X1,…,X19 y la variable dependiente Y: estado de aprobación (Reprobado, aprobado)

- Árbol de decisiones: se han considerado todas las variables independientes X1,…,X19 y la variable dependiente Y: rendimiento categórico (deficiente (0 a 51), regular (51 a 70), bueno (70 a 85), excelente (85 a 100)).

Las variables independientes a considerar en ambos métodos son:

V. Ind	Descripción	Codificación
X1	Edad del estudiante	Numérica
X2	Género del estudiante	Categórica binaria
X3	Nivel de ingreso familiar	Numérica
X4	Nivel educativo más alto de los padres o tutores	Categórica
X5	Lugar de residencia del estudiante	Categórica
X6	Hábitos de estudio del estudiante	Categórica
X7	Calidad percibida de la enseñanza en la institución	Categórica
X8	Estado de la infraestructura de la institución de educación superior	Categórica
X9	Recibe apoyo de los docentes	Categórica
X10	Horas dedicadas al estudio por semana	Numérica
X11	Asistencia regular a clases	Categórica
X12	Participación en actividades extracurriculares	Categórica
X13	Colegio de procedencia	Categórica
X14	Utilización de tecnologías de información	Categórica
X15	Promedio de notas obtenidos en el colegio	Numérica
X16	Número de materias reprobadas	Numérica
X17	Tiempo de traslado a la universidad en minutos	Numérica
X18	Acceso a internet en casa	Categórica
X19	Situación laboral actual del estudiante	Categórica

Fuente: Elaboración propia

2.2. RENDIMIENTO ACADÉMICO

El rendimiento académico hace referencia al nivel de aprendizaje alcanzado por un estudiante, evaluado a partir de calificaciones, exámenes y otras medidas de desempeño. En este trabajo, se utilizó el promedio de calificaciones como indicador del rendimiento académico.

En la práctica, la mayoría de investigaciones destinadas a explicar el éxito o el fracaso en los estudios, analizan el rendimiento académico a través de las calificaciones o la certificación académica de un estudiante (Ocaña Fernández 2011).

2.3. *MACHINE LEARNING*

El Machine Learning (ML) o aprendizaje automático es parte de la inteligencia artificial, cuya característica es dotar a las computadoras la capacidad de aprender utilizando diversos algoritmos, estos se caracterizan por el manejo de grandes volúmenes de información, análisis y la toma de decisiones.

El aprendizaje automático, comúnmente abreviado como ML, es un tipo de inteligencia artificial (IA) que "aprende" o se adapta con el tiempo. En lugar de seguir reglas estáticas codificadas en un programa, esta tecnología identifica patrones de entrada y contiene algoritmos que evolucionan con el tiempo (Damián et al. 2021).

El *Machine Learning*, presenta una serie de métodos supervisados y no supervisados, con la ayuda de estos métodos es posible predecir el rendimiento académico considerando diversas variables explicativas y una fuente de datos confiable. Para lograr este propósito es necesario entrenar a los modelos para obtener modelos más precisos y confiables.

El objetivo del *Machine Learning* es crear modelos que permitan resolver de forma automática una tarea dada, basándose en los algoritmos de *Machine Learning*, una vez creado el modelo, este debe ser entrenado con datos para que así el sistema sea capaz de predecir una respuesta o salida, para una entrada de datos de los cuales se desconozca la salida o resultado (Vargas 2021).

2.4. MÉTODO DE REGRESIÓN LOGÍSTICA

Es un método del Machine Learning de clasificación, que se utiliza principalmente para predecir la probabilidad de una variable dependiente, esta variable es dicotómica del tipo binario (0,1).

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 explicativas o independientes, que pueden ser cuantitativas (que se denominan covariables o covariadas) o categóricas.

Logistic Regression, es un modelo matemático que determina la existencia o ausencia de relaciones entre una variable dependiente y n variables independientes. La regresión logística analiza datos distribuidos binomialmente (Zabarte 2022).

El objetivo de este método es calcular la probabilidad de que ocurra un determinado evento, determinar que variables son las más significativas en un conjunto de variables explicativas, considerando la probabilidad de aquellas que explican más respecto a las otras.

El objetivo de la regresión logística binaria no es, como en la regresión lineal, predecir el valor de la variable dependiente (Y) a partir de una o varias variables independientes (X), sino predecir la probabilidad de que ocurra el evento que caracteriza la variable dependiente (éxito, enfermedad, etc.), conocidos los valores de las variables independientes (Ochoa Sangrador, Molina Arias, y Ortega Páez 2023).

El modelo de regresión logística es:

$$Y = \left(\frac{p}{1-p}\right) = b_0 + b_1X_1 + \cdots + b_nX_n$$

Donde los X_i ,i = 1,2,…,n son variables independientes y la variable Y es la variable dependiente dicotómica.

Las probabilidades son:

$$P(Y) = \frac{1}{1 + e^{-(b_0+b_1X_1+\cdots+b_nX_n)}}$$

Donde:

$P(Y)$, es la probabilidad de que un estudiante apruebe.

b_0: es el intercepto, (constante del modelo)

b_i, i = 1,…,n: son los parámetros del modelo.

Para estimar los parámetros del modelo de regresión logística, utilizando las observaciones que se tienen de las variables

independientes o explicativas X_i, utilizamos el método de máxima verosimilitud.

Si consideramos una muestra de n observaciones independientes (X_i, Y_i), $i = 1,\ldots, n$, con $X = (X_1,\ldots, X_k)$, con:

$$Y_i = \begin{cases} 1, & evento\ ha\ ocurrido \\ 0, & evento\ no\ ha\ ocurrido \end{cases}$$

Estimamos los parámetros del modelo $\beta = (b_0,\ldots, b_k)$ con el método de máxima verosimilitud cuyo objetivo es maximizar la probabilidad de obtener el conjunto de datos observados. La función de verosimilitud es:

$$L(X/\theta) = \prod f(x_i, \theta)$$

Derivando para encontrar el vector β que maximice la función de verosimilitud, se tiene:

$$\sum (Y_i - P_i) = 0;$$

$$\sum X_{ij}(Y_i - P_i) = 0, \quad j = 1, \ldots, k$$

Resolviendo el sistema se obtienen los parámetros del modelo.

Curva ROC

La curva ROC representa gráficamente la sensibilidad (probabilidad de que un individuo con el evento de interés se clasifique cuando ocurre el evento de interés) y especificidad (probabilidad de que una observación que no tiene el evento de interés se clasifique correctamente) para los diferentes puntos del valor de corte "c" (0,1). La curva ROC muestra la capacidad del modelo de regresión logística para separar las observaciones que tienen el evento de interés frente a los que no tienen el evento de interés

Para decidir si el evento ocurre o no a partir de las probabilidades estimadas de ocurrencia se considera el valor de corte "c", donde:

$P(Y = 1 / \ X = X_i) >= c$

=> el evento ocurre

$P(Y = 1 / \ X = X_i) < c$

=> el evento no ocurre

El área de la curva ROC varía entre 0.5 y 1, si los valores son cercanos a 0.5 existe clasificación con determinado error, próximos a 1, la clasificación es excelente.

Matriz de confusión

La matriz de confusión permite mostrar el desempeño de los métodos del *Machine Learning* en cuanto a la precisión, sensibilidad (*recall*), especificidad y la exactitud del método.

Figura 1. Matriz de confusión para dos clases (positivos y negativos)

		Valores predecidos	
		Positivo	**Negativo**
Valores reales	**Positivos**	VP	FN
	Negativos	FP	VN

Donde:

VP: valores positivos predichos correctos

VN: valores negativos predichos correctos

FN: falsos negativos, cuando en realidad son valores positivos

FP: falsos positivos, cuando en realidad son negativos

De la Figura 1, la diagonal principal muestra los valores reales de cada una de las clases.

- Precisión (accuracy): es una medida que muestra el desempeño del modelo en general, considera todas las clases (positivas y negativas)

$$Precision = \frac{VP + VN}{VP + VN + FN + FP}$$

- Sensibilidad (*recall*): es la tasa de verdaderos positivos, esta medida

muestra el desempeño del modelo para detectar los casos positivos

$$Sensibilidad = \frac{VP}{VP + FN}$$

- Especificidad: esta medida identifica correctamente los casos negativos.

$$Especifidad = \frac{VN}{VN + FP}$$

- F1-*score*: es la media armónica entre la precisión y la sensibilidad, valores próximos a 1 indican un desempeño del modelo excelente.

$$F1 - score = 2 * \frac{Precision * Sensibilidad}{Precision + Sensibilidad}$$

2.5. ÁRBOL DE DECISIONES

Los árboles de decisión buscan reglas de clasificación en base a un conjunto de variables, generalmente la variable dependiente es dicotómica del tipo binario (0,1), la selección de los mejores atributos se hace mediante la entropía (medida de la incertidumbre de una variable aleatoria, cuanto mayor es el valor de la entropía, mayor será el contenido de la información)

Classification Tree: es apropiado para resolver problemas de clasificación y de regresión. Utilizan la estructura de un árbol, formado por nodos y ramas. Los nodos contienen una condición sobre una columna de datos. Según los valores de la columna sobre la que se aplica la condición, las observaciones se separan por distintas ramas. Los nodos hoja, nodos que no tienen descendentes, proporcionan las clasificaciones de las instancias (Zabarte 2022).

Los árboles de decisión son modelos de clasificación, estos consideran las variables más influyentes para realizar la clasificación considerando diversos criterios, hasta llegar a un punto (hoja) donde ya no es posible clasificar.

Los elementos de un árbol de decisiones son:

- Nodos: Son las que tienen a los atributos
- Arcos: Son las que contienen valores posibles del nodo padre
- Hojas: Son los nodos que clasifican como 0 o 1

Estos modelos de clasificación surgen a partir de particiones, sea (x_1 ,…, x_N) vector de mediciones correspondiente a un caso, que contiene a todas las posibles mediciones, sea *C* el conjunto de las J clases posibles, "*y*" la salida del modelo, con y ϵ C, asignamos a cada caso una clase de C (clasificador), la muestra considerada tiene N casos con sus respectivas características (variables explicativas), sea la muestra L.

Sea h(x) el clasificador y sea $R(h)^*$ la tasa de error de la clasificación verdadera que es el cociente entre la cantidad de valores de *X* a los que $h(x)$ le asigna una clase incorrecta sobre el total de valores de *X*. Para calcular la tasa de error de clasificación utilizamos una muestra L denominado estimador interno.

$$R(h)^* = P(h(x) \neq Y)$$

Estimador de sustitución (error de entrenamiento)

Para determinar el error de entrenamiento hacemos correr todos los casos de la muestra L en el clasificador h(x):

$$R(h) = \frac{1}{N} \sum \mathrm{P}(\mathrm{h}(\mathrm{x_n}) \neq j_n)$$

Con j_n: la clase del caso *n*

Para estimar el verdadero error de clasificación se utiliza la estimación por conjunto de prueba denominado error de prueba, dividiendo el conjunto de aprendizaje L en dos subconjuntos L1 y L2. El modelo h(x) se entrena utilizando L1 para luego

verificar los casos mal clasificados utilizando L2, por lo tanto, el error de prueba es:

$$R^{ep}(h) = \frac{1}{N_n}\sum P(h(x_n) \neq j_n)\ en\ L_2$$

La condición es que L1 y L2 estén idénticamente distribuidos, donde L1+L2=L, el inconveniente para estimar el verdadero error de clasificación es la cantidad de casos, lo que implica en la reducción de la eficiencia del modelo.

Para resolver este problema utilizamos la estimación denominada validación cruzada de k dobleces (*k_fold cross validation*). El proceso de validación cruzada consiste en dividir L2 en k subconjuntos de tamaño similar, se construye o entrena el clasificador h(x) con cada uno de los k subconjuntos de L, denominados $h^{(k)}(x)$, luego el error de prueba es:

$$R^{ep}\left(h^{(k)}\right) = \frac{1}{N}\sum P\left(h^{(k)}(x_n) \neq j_n\right)$$

Se tienen k estimaciones, el error por validación de k dobleces es:

$$R^{cv}(h) = \frac{1}{K}\sum R^{ep}(h^{(k)})$$

El clasificador por árbol binario predice la clase para un caso de la siguiente manera:

1. La selección de las condiciones
2. La decisión de seguir dividiendo a cada nodo
3. La asignación de una clase especifica a cada nodo terminal

La idea principal es que la pureza de los subconjuntos generados a partir de una división sea mayor a la pureza del conjunto padre.

La impureza hace referencia a la entropía, que es una medida de desorden (impuro).

La entropía se define como:

$$i(t) = \sum -p(j\ /\ t\ )\log\left(p(j/t)\right)$$

Donde p(j / t) es la probabilidad de la clase j dentro del subconjunto o nodo t que varía de 1 a 0 (1 cuando pertenece a la clase y 0 cuando no existen miembros que pertenecen a la clase) (C. Arana s. f.)

La entropía no es la única medida que mide la impureza, también se tiene a el índice de Gini.

El índice de diversidad de Gini, trata de minimizar la impureza existente en los subconjuntos de casos de entrenamiento generados al ramificar por un atributo determinado, su función es la siguiente (Medina Merino y Ñique Chacón 2017):

$$G(C_i) = 1 - \sum p(j\ /\ t\ )\text{^}2$$

Donde C_i es la clase

Donde p(j/t) es la proporción de observaciones de la clase j en el nodo t.

En el proceso de construcción se selecciona en cada nodo la variable que minimiza la impureza (entropía), luego se realiza una poda para evitar sobreajustes.

Random forest

El *Random Forest* se basa en los árboles de decisión es un conjunto de árboles de decisión entrenados, para realizar la clasificación, cada uno de los árboles de decisión da una clasificación y la decisión con mayor cantidad de votos es la predicción del algoritmo,

Los modelos de Bosque de árboles aleatorios o Random Forest (RF) es un algoritmo de aprendizaje supervisado que utiliza un método de tipo ensemble. Como los modelos ensemble, RF está compuesto por un conjunto de modelos de árboles de decisión que se consideran como submodelos o modelos

sencillos, relativamente no correlacionados que operan como un comité. (Vakaruk 2023)

El *Random Forest* se utiliza para mejorar las métricas del árbol de decisiones, ya que este método reduce la varianza debido al uso de múltiples árboles.

3. RESULTADOS

En este trabajo se evaluaron los métodos Regresión Logística y Árbol de Decisiones para predecir el rendimiento académico de los estudiantes de la E.M.I. Unidad Académica La Paz, a partir de registros académicos de la gestión 2022 que corresponde a 650 registros de estudiantes de primer y segundo semestre del departamento de Cs. Básicas, considerando estado de aprobación (0 reprobado, 1 aprobado), rendimiento categórico (1 deficiente, 2 regular, 3 bueno, 4 excelente) para cada uno de los métodos respectivamente.

3.1. REGRESIÓN LOGÍSTICA

Se ha considerado la regresión logística, regresión logística con ajuste de hiperparámetros y regresión logística con validación cruzada.

La Tabla 1 muestra las métricas de la regresión logística y se puede apreciar que la regresión logística con validación cruzada muestra un desempeño efectivo para poder predecir a los estudiantes aprobados y reprobados.

Exactitud: El modelo de regresión logística clasifica correctamente el 83.5% de los datos.

Precisión: El 85.7% de las predicciones realizadas para la clase 1 (aprobado) son correctas.

Recall: El 96.2% de las predicciones realizadas para la clase 1 fueron identificadas correctamente.

ROC-AUC: El modelo se desempeña bien ya que el valor obtenido 0.83 es próximo a 1.

Según estas métricas, el modelo tiene un buen rendimiento y es capaz de identificar los verdaderos positivos.

En la Figura 2, el modelo predijo 7 muestras de la clase 0 (reprobado) que son correctas;156 muestras que fueron predichas de la clase 1 (aprobado) que son correctas; 26 muestras que fueron predichas de la clase 1 (aprobado) pero son incorrectas, estas pertenecen a la clase 0 (reprobado) y 6 muestras que fueron predichas de la clase 0 (reprobado) pero son incorrectas, estas pertenecen a la clase 1 (aprobado)

En la Figura 3, el área bajo la curva es de 0.84 lo cual indica un buen rendimiento del modelo de regresión logística capaz de distinguir correctamente entre las clases (0 reprobado y 1 aprobado) en un 84%.

3.2. ÁRBOL DE DECISIONES

Para el análisis y para la mejora de las métricas se han considerado: árbol de decisiones, árbol de decisiones con ajuste de hiperparámetros, *Random Forest*, *Random Forest* con ajuste de hiperparámetros, estos métodos son de clasificación, ya que la variable dependiente

Tabla 1.
Comparación entre las métricas regresión logística, regresión logística con ajuste de hiperparámetros y regresión logística con validación cruzada.

Métrica	Regresión logística	Regresión logística con ajuste de hiperparámetros	Regresión logística con validación cruzada
Exactitud	0.8358974358974359	0.8358974358974359	0.8769230769230768
Precisión	0.8571428571428571	0.8571428571428571	0.9038508671421749
Recall	0.9629629629629629	0.9629629629629629	0.9561301084236863
ROC-AUC	0.838383838383838	0.8393191170968949	0.8728575797291395

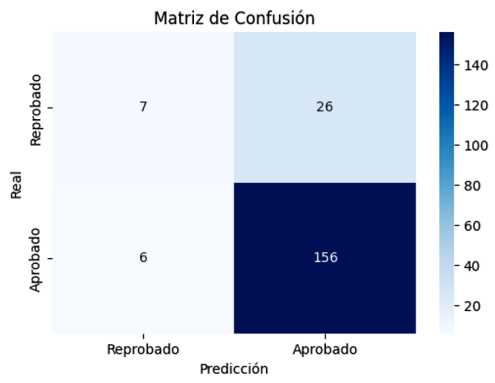
Fuente: Elaboración propia

es rendimiento categórico (deficiente (0 a 51), regular (51 a 70), bueno (70 a 85), excelente (85 a 100)).

Se puede observar en la tabla 2 mejoras en las métricas del árbol de decisiones. Los resultados del *Random Forest* con ajuste de hiperparámetros muestra un modelo con desempeño moderado cuya exactitud es de 56.41%, existe un buen rendimiento del modelo, especialmente para la categoría regular, y hay dificultades en la última categoría, esto porque existen pocas observaciones para esta categoría.

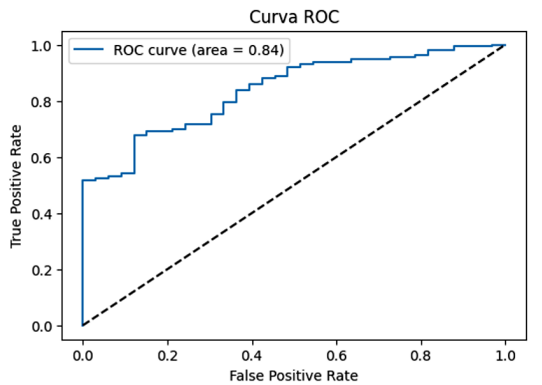
En la tabla 3, se puede observar la comparación de las métricas de los métodos utilizados, por lo tanto, el método más efectivo es el método de regresión logística con validación cruzada.

Figura 2.
Matriz de confusión



Fuente: Elaboración propia

Figura 3.
Curva ROC



Fuente: Elaboración propia

Tabla 2
Comparación de los métodos: árbol de decisiones, árbol de decisiones con ajuste de hiperparámetros, Random Forest y Random Forest con ajuste de hiperparámetros

Métrica	Árbol de decisiones	Árbol de decisiones con ajuste de hiperparámetros	Random Forest	Random Forest con ajuste de hiperparámetros
Exactitud	0.4615	0.5025	0.5641	0.5641
Precisión	0.4781	0.4980	0.5237	0.5464
Recall	0.4615	0.5025	0.5641	0.5641

Fuente: Elaboración propia

Tabla 3.
Comparación de las métricas de todos los métodos analizados

Medida	Regr. Logística	Regr. logística con ajuste de hiperparámetros	Regresión logística con validación cruzada	Árbol de decisiones	Árbol de decisiones con ajuste de hiperparámetros	Random Forest	Random Forest con ajuste de hiperparámetros
Exactitud	0,8359	0,8359	0,8769	0,4615	0,5025	0,5641	0,5641
Precisión	0,8571	0,8571	0,9038	0,4781	0,498	0,5237	0,5464
Recall	0,9629	0,9629	0,9561	0,4615	0,5025	0,5641	0,5641

Fuente: Elaboración propia

Para evaluar el modelo utilizamos las métricas exactitud (*accuracy*), reporte de clasificación y la matriz de confusión.

El modelo de regresión logística considerando las variables más significativas es:

$$P(Y) = \frac{1}{1 + e^{-(2.9881+0.5607x_3+0.5184x_4+1.0124x_6+0.5194x_9+0.9008X_{10}+0.6987X_{13}+1.3360X_{15}-0.5811X_{19})}}$$

Donde las variables más significativas son: X3 Nivel de ingreso familiar; X4 Nivel educativo más alto de los padres; X6 Hábitos de estudio del estudiante; X9 Apoyo de los docentes; X10 Horas dedicadas al estudio; X13 Colegio de procedencia; X15 Promedio de notas del colegio; X19 Situación laboral del estudiante.

El método de regresión logística muestra que las variables académicas y hábitos de estudio son las que más influyen en el rendimiento académico, además de las variables socioeconómicas, pero la situación laboral del estudiante afecta de manera negativa en el rendimiento académico.

La Figura 4, muestra el análisis de las variables predictoras en el método del *Random Forest* con ajuste de hiperparámetros el cual fue aplicado como un modelo de clasificación considerando diferentes niveles de clasificación del rendimiento académico, los resultados muestran que las variables que influyen con mayor fuerza en la clasificación son: promedio_colegio, horas_estudio_semanal, nivel_educativo_padres, ingreso_familiar, tiempo_traslado; variables que son determinantes en el rendimiento académico.

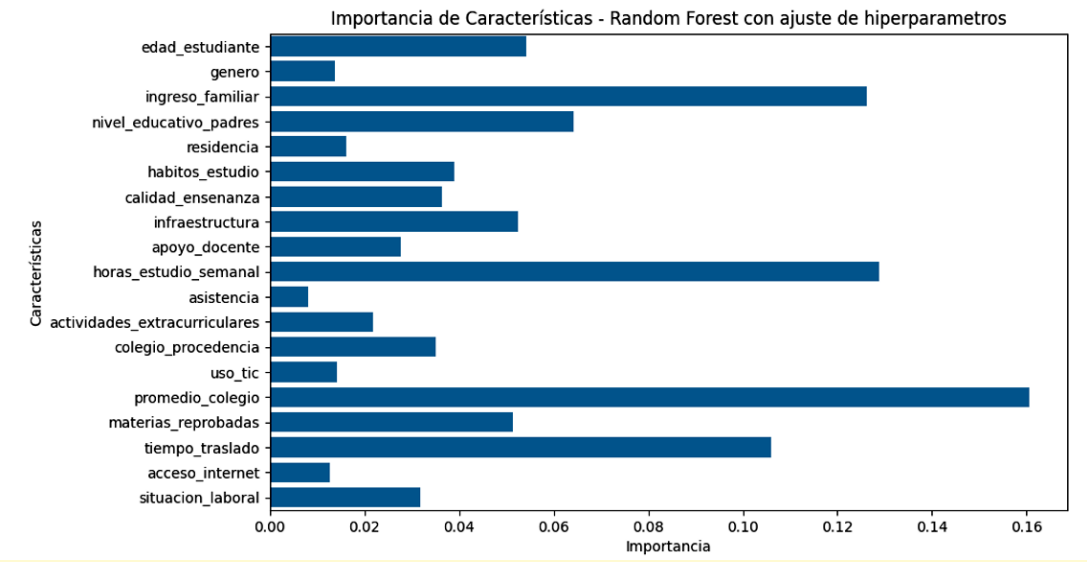
Según los resultados obtenidos en el *Random Forest* con ajuste de hiperparámetros, el método este subajustado de acuerdo a las medidas obtenidas: exactitud (0,5641) es un valor que no capta de manera correcta los patrones generales de los datos originales; recall (0,5641) indica un rendimiento equilibrado, pero bajo y F1-score muestra un buen rendimiento solo en una clase y dificultad para predecir las clases restantes.

4. CONCLUSIONES

El método que destaca considerando las diferentes métricas es el modelo de regresión logística que además de obtener las mejores métricas, identifica a los estudiantes en riesgo.

Figura 4.

Importancia de las variables en el Random Forest con ajuste de hiperparámetros



Fuente: Elaboración propia

Regresión logística con ajuste de hiperparámetros con validación cruzada, método más efectivo (90.38% de precisión y 95.61% de *recall*) para la predicción del rendimiento académico, considerando a estudiantes aprobados y/o reprobados.

Árbol de decisiones, este método sirvió para mostrar la jerarquía (método de clasificación) de las diferentes variables, si bien las métricas no son tan favorables (respecto a la precisión 47.81%) se ha realizado mejoras, como ser: árbol de decisiones con ajuste de hiperparámetros (con una mejora en la precisión del 49.80%); *Random Forest* (con una mejora en la precisión del 52.37%) y un *Random Forest* con ajuste de hiperparámetros (con una mejora en la precisión del 54.6%), donde se ha evidenciado que el mejor método es el *Random Forest* con ajuste de hiperparámetros.

Comparando las métricas de los diferentes métodos, se concluye que el método de regresión logística con ajuste de hiperparámetros es el método más efectivo con una exactitud del 87.69%, precisión del 90.38%, *recall* del 95.61%, ROC-AUC del 87.86% para la predicción del rendimiento académico, considerando todas las variables explicativas. Si bien existen diferencias no significativas entre las métricas del método de regresión logística y regresión logística con ajuste de hiperparámetros, se concluye que este último distingue correctamente entre las clases (aprobado y reprobado).

5. DISCUSIÓN

Los hallazgos de este trabajo muestran que la regresión logística con validación cruzada es el método más efectivo para predecir el rendimiento académico en estudiantes de la EMI del departamento de Cs. Básicas.

Este resultado se alinea con estudios previos que han destacado la capacidad de este algoritmo para manejar variables de distinta naturaleza y proporcionar modelos estables con buen balance entre Precisión y *Recall*. La alta sensibilidad encontrada (95.61%) indica que el modelo es particularmente eficiente para identificar a los estudiantes en riesgo de reprobar, información relevante para diseñar e implementar diversas políticas de intervención temprana.

Respecto al método de los árboles de decisión, estos mostraron ciertas limitaciones en cuanto a Precisión y Exactitud. Métricas que fueron mejoradas utilizando el *Random Forest* evidenciando que los métodos de ensamblado son una alternativa viable y útil ya que permiten interpretar la jerarquía y la importancia de las variables identificando los factores que más influyen en el rendimiento académico.

Un aspecto relevante del análisis es la influencia de factores académicos y sociodemográficos en la predicción. La interacción de estas variables confirma que el rendimiento académico no depende de un único elemento, sino de la combinación de condiciones personales, sociales y del entorno institucional. Esto sugiere la necesidad de que las instituciones de educación superior implementen diversos sistemas que se basan en métodos del *Machine Learning* para detectar estudiantes en riesgo y posterior deserción.

CONFLICTO DE INTERESES

Los autores declaran que no hay conflicto de intereses respecto a la publicación de este documento.

REFERENCIAS BIBLIOGRÁFICAS

Damián, Anderson, Jiménez Alfaro, José Vicente, y Díaz Ospina. 2021. *Revisión sistemática de literatura: Técnicas de aprendizaje automático (Machine Learning)*. Activa: 113-21.

Herreño M. Martha, Romero T. Jose, Mejía R. Jennifer, y Roman S. Wanda. 2024. *Deserción estudiantil en educación superior. Tendencias y oportunidades en la era post pandemia*. Koinonia X.

Medina Merino, Fátima Rosa, y Carmen Ñique Chacón. 2017. *Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python*. Interfaces. *https://www.kaggle.com/primaryobjects*

Ocaña Fernández, Yolvi. 2011. *Variables académicas que influyen en el rendimiento académico de los estudiantes universitarios*. Investigación educativa 15: 165-79.

Ochoa Sangrador, Carlos, Manuel Molina Arias, y Eduardo Ortega Páez. 2023. *Regresión logística múltiple*. Evid Pediatr. *http://www.evidenciasenpediatria.es*

Poveda V. Iván. 2019. *Los factores que influyen sobre la deserción universitaria. Estudio en la UMRPSFXCh – Bolivia, análisis con ecuaciones estructurales*. REV.INV.&NEG 12: 61-77.

UNESCO. 2024. PISA 2022 El panorama de los países de América Latina y el Caribe. Chile. *https://www.unesco.org/es/open-access/cc-sa*.

Vargas, Peña Dennis. 2021. *Modelo De Detección De Botnets En El Tráfico Del Sistema De Nombres De Dominio De La Red, Basado En Aprendizaje De Máquina Para Banca Central En Bolivia*. Universidad Mayor de San Andrés.

Vakaruk, Stanislav. 2023. *Contribuciones a la Aplicación de Machine Learning en Escenarios Novedosos de Tiempo Real*. Universidad Politécnica de Madrid

Zabarte, Moreno Gorka. 2022. *Utilización de técnicas de aprendizaje automático para la predicción del rendimiento de los jugadores de futbol*. Universidad Politécnica de Madrid.

Artículo Original
Revista Varianza Nro. 26, Octubre de 2025
ISSN 2789-3529 versión en línea
DOI: https://doi.org/10.53287/scos9764py57q
Páginas: 35-45

ANÁLISIS DE LOS DETERMINANTES EN LA TASA DE FECUNDIDAD DE LAS MUJERES EN BOLIVIA

ANALYSIS OF THE DETERMINANTS OF THE FERTILITY RATE AMONG BOLIVIAN WOMEN

Valentina Valdez Vega[1]
Universidad Mayor de San Andrés, La Paz-Bolivia
✉ *vvaldezv@fcpn.edu.bo*

Artículo recibido: 01/09/2025
Artículo aceptado: 23/09/2025

RESUMEN

El estudio analiza los factores que determinan la fecundidad en mujeres adultas en Bolivia, utilizando los datos de la Encuesta de Demografía y Salud de 2023, a partir de un modelo de regresión Poisson. Se identifican variables sociodemográficas y económicas asociadas a la ocurrencia y cantidad de hijos nacidos vivos. Para esto, se consideran variables como edad, educación, estado conyugal, ingreso del hogar, pertenencia indígena y uso de métodos anticonceptivos, que siguen el enfoque teórico de T.P. Schultz (2005). Los resultados muestran que factores como la baja escolaridad, quintiles de riqueza y la planificación familiar siguen condicionando los patrones reproductivos en el país. Estos hallazgos ofrecen evidencia empírica útil para el diseño de políticas públicas orientadas al desarrollo y la planificación reproductiva, con enfoque de género y territorial en Bolivia.

Palabras clave: Baja escolaridad, Fertilidad, Planificación familiar, Regresión Poisson.

ABSTRACT

The study analyzes the factors determining fertility among adult women in Bolivia, using data from the Demographic and Health Survey of 2023, and a Poisson regression model. Sociodemographic and economic variables associated with the occurrence and number of live births are identified. Variables considered include age, education, marital status, household income, indigenous identity, and contraceptive use, following the theoretical framework of T.P. Schultz (2005). The results show that factors such as low educational attainment, wealth quintiles, and family planning continue to influence reproductive patterns in the country. These findings provide useful empirical evidence for the design of public policies aimed at development and reproductive planning, with a gender and territorial perspective in Bolivia.

Keywords: Family Planning, Fertility, Low educational attainment, Poisson Regression.

1. INTRODUCCIÓN

Según los resultados de la última Encuesta de Demografía y Salud (EDSA) realizada por el Instituto Nacional de Estadística (INE, 2023), se observó una disminución en la tasa de fecundidad[2] en Bolivia. La tasa global de fecundidad en 2023 llegó a 2,1 en comparación a la tasa de 2,9 que se obtuvo en la pasada encuesta EDSA de 2016. Esta caída en la fecundidad no es una sorpresa para países dentro de la región latinoamericana y puede ser relacionada con cambios culturales, nuevas expectativas de vida por parte de la población joven y la creciente inserción al mercado laboral de la población femenina (Enríquez-Canto et al., 2018). Por otra parte, se ha visto una fuerte asociación respecto a

[1] Estudiante de la carrera de Estadística – Universidad Mayor de San Andrés (UMSA). https://orcid.org/0009-0007-4335-5441
[2] Según el INE, la tasa global de fecundidad se refiere al "Número promedio de hijos/as nacidos vivos por mujer en edad fértil (15 y 49 años)."

elevados niveles de pobreza, especialmente en hogares monoparentales o rurales.

La EDSA constituye una fuente de información relevante al momento de obtener indicadores de salud y demografía a nivel nacional, que posteriormente ayudarán al diseño de políticas públicas implementadas en Bolivia. Esta encuesta ofrece información importante para analizar la fecundidad en el país, ya que permite vincular los patrones reproductivos con variables sociodemográficas, culturales y de salud, además de gozar de una cobertura a nivel nacional. De esta forma, esta encuesta actúa como fuente de información esencial para la construcción de modelos predictivos que identifican los factores más significativos asociados a la fecundidad de mujeres adultas en edad fértil[3] (INE, 2023).

En este sentido, resulta necesario analizar los factores determinantes en la fecundidad de las mujeres en Bolivia, para así poder entender de mejor forma por que existe una tendencia a la baja respecto a la tasa de fecundidad y el número de hijos por mujer. Para tal efecto, es posible utilizar modelos estadísticos de regresión como ser la regresión *Poisson*, ya que nos permite examinar cómo varía la fecundidad en cuanto al número de hijos según características individuales como la edad, nivel educativo, actividad económica o pertenencia étnica (Schultz, 2005).

La fecundidad constituye uno de los indicadores demográficos más relevantes dentro de la dinámica de un país, ya que influye de forma directa en el crecimiento de una población, la estructura por edades y, como consecuencia, permite diseñar políticas sociales y económicas como sean requeridas. Como menciona Schultz (2005), el estudio de la fecundidad no solo se reserva a una decisión propia de una familia o persona, sino que tiene fuertes implicaciones económicas y sociales dentro de una sociedad. A partir de esto, se refuerza la influencia de la fecundidad como un factor esencial del desarrollo de un país y la importancia de comprender los factores que determinan la fecundidad dentro de una economía, para poder entender su comportamiento.

En Bolivia, tomando en cuenta los resultados presentados en la EDSA, se ha observado que persisten diferencias geográficas, étnicas y socioeconómicas que inciden en los patrones de fecundidad de las mujeres. Como se observa en la Figura 1, los resultados posteriores a la recolección de información para la EDSA (en sus dos versiones 2016 y 2023) muestran una tendencia a la baja en lo que compete a la tasa global de fecundidad a nivel Bolivia. En ambos casos, se observa una disminución de 3,5 a 2,9 para el año 2016, y una reducción a 2,1 para el año 2023. Sin embargo, comparando con otros países de América Latina, se observa que desde el 2008 la tasa global de fecundidad boliviana se mantiene mayor respecto a países como Ecuador, Brasil y Chile. (INE, 2025) Aunque, como región, en todos los países se observa una tendencia decreciente. Esta reducción en la tasa global de fecundidad conduce a una población con una proporción menor de jóvenes, lo cual genera desafíos para la fuerza laboral, la sostenibilidad de los sistemas sociales y afecta directamente el crecimiento poblacional.

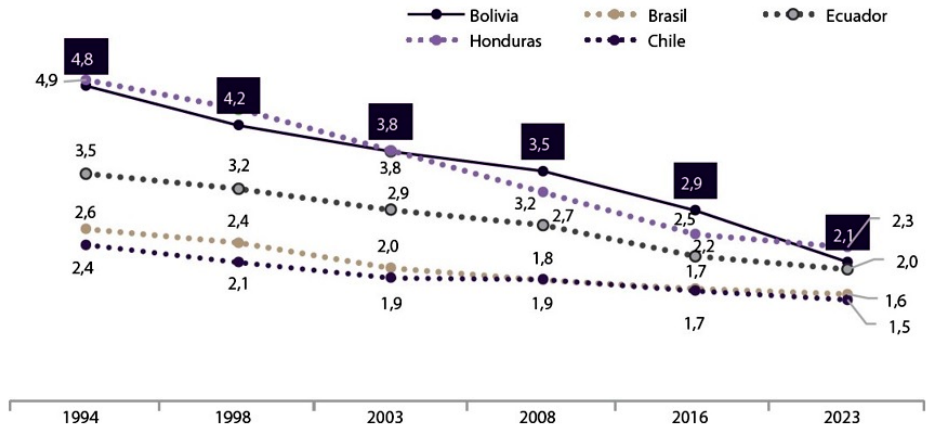
De esta forma, la investigación surge como un aporte dentro de la necesidad de generar evidencia empírica que contribuya a entender los determinantes sociales, económicos y culturales de la fecundidad en Bolivia, utilizando la EDSA, que además se enfoque en analizar componentes actuales dentro de esta temática. Para que así, se pueda entender de mejor forma la disminución en la tasa global de fecundidad observada en los

[3] Según la definición de la EDSA (2023) considera una mujer adulta en edad fértil a aquellas mujeres entre 20 y 49 años.

últimos años, y se puedan plantear políticas o programas que respondan a estos resultados. Igualmente, la investigación emplea el modelo de regresión *Poisson*, el cual en base a su connotación teórica y práctica permite llevar a cabo el estudio.

Figura 1.

Evolución de la Tasa Global de Fecundidad en Bolivia y países de la región



Fuente: (INE, 2025) - en base a la EDSA 1994, 1998, 2003 y 2008, EDSA 2016 y 2023, y Banco Mundial 2022.

2. MATERIALES Y MÉTODOS

2.1 METODOLOGÍA

El enfoque de la presente investigación es cuantitativo, debido a que se basa en la recolección y análisis de datos numéricos, a partir de la EDSA 2023, para examinar los determinantes de la fecundidad de manera objetiva. El tipo de estudio corresponde a una investigación correlacional, cuyo propósito es identificar y analizar la relación existente entre las variables que determinan la fecundidad, sin embargo, no se establece causalidad directa. Igualmente, esta investigación se enmarca dentro del paradigma positivista, que sostiene que el conocimiento se obtiene a través de la observación empírica y la medición objetiva. Esto sucede dado que la investigación busca patrones generales que expliquen los fenómenos sociales y económicos de la fecundidad en Bolivia.

Fecundidad y Fertilidad

La fecundidad y la fertilidad son conceptos centrales en la demografía debido a ser indicadores relevantes al analizar la dinámica demográfica y el crecimiento de una población. Aunque frecuentemente son usados como sinónimos, tienen significados distintos. Mientras que la fertilidad alude a la capacidad biológica de tener hijos, la fecundidad se refiere al número efectivo de nacimientos (Rodríguez Vignoli, 2014). De esta forma, la fecundidad se convierte en una variable clave en la planificación de políticas públicas debido a su impacto en la estructura poblacional y en la demanda de servicios educativos, sanitarios, laborales y sistemas de seguridad social (CEPAL, 2020).

Un aporte relevante al estudio de la fertilidad es realizado por T.P. Schultz (2005), quien establece una relación inversa entre el ingreso y la fecundidad. Según su análisis, a medida que aumenta el ingreso y la educación de las mujeres, la fecundidad tiende a disminuir, lo que responde a mayor participación femenina

en el mercado laboral como a un cambio en las preferencias familiares respecto a la planificación reproductiva. Este enfoque ha sido retomado en estudios posteriores para explicar patrones reproductivos en países en desarrollo. En el caso de América Latina, ha atravesado una transición demográfica significativa donde se observó una reducción en las tasas de fecundidad en la región. Sin embargo, esta disminución no ha sido homogénea, ya que persisten niveles elevados en ciertos grupos poblacionales, especialmente en zonas rurales y entre mujeres con menor acceso a educación y/o servicios (Schultz, 2005).

Por otro lado, como Schultz (2005) menciona en su capítulo "Fertility and Income", las variables determinantes de la fecundidad femenina pueden ser fundamentadas bajo un enfoque económico y provenir de la evidencia y la teoría microeconómica. Ante esto, las principales variables que afectan la fecundidad según Schultz son: el ingreso, la educación, el uso de anticonceptivos y la planificación familiar, la mortalidad infantil, el entorno social y cultural y otras variables exógenas que pueden tomar en cuenta la edad de la mujer y el estado conyugal.

Regresión Poisson

La Regresión *Poisson* es un modelo no lineal que toma en cuenta a una variable cualitativa discreta como la variable dependiente, en general modela y predice variables de conteo. Esta regresión parte de la distribución Poisson que precisamente se utiliza para analizar conteos o números de eventos en un intervalo dado. Así, en estos casos, la variable dependiente toma más de dos valores discretos no negativos (números del 0, 1, 2, en adelante). La distribución Poisson es utilizada comúnmente para modelar el número de ocurrencias de cierto evento en un medio continuo (generalmente hace referencia al tiempo), por ejemplo, es útil para modelar conteos como el número de hijos que tiene una mujer adulta en edad fértil. De esta forma, esta regresión permite examinar cómo varía la fecundidad según características individuales como la edad, nivel educativo, actividad económica o pertenencia étnica (Long & Freese, 2014). Siendo de esta forma un instrumento correcto para el presente estudio.

Partiendo de la distribución Poisson, sea una variable aleatoria Y que representa (en el caso de la investigación) el número de hijos de una mujer en edad fértil. Esta variable sigue una distribución Poisson con parámetro μ. Como lo indica la siguiente ecuación.

Sea Y~*Poisson*(μ), su distribución de probabilidad está dada por:

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}; \quad y = 0,1,\dots,\infty \; ; \mu > 0$$

Además, teniendo como característica que su esperanza matemática y varianza son iguales al parámetro.

$$\mathrm{E}[Y] = \mathrm{V}[Y] = \mu$$

De esta forma, la regresión Poisson (Agresti, 2010) utiliza la función de enlace log-lineal obtenida del modelo lineal generalizado (MLG) con k variables predictoras. La regresión Poisson está dada por la siguiente ecuación:

$$\log \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

Operando sobre el parámetro:

$$\mu = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k}$$

Para una muestra aleatoria de n observaciones, se tienen las estimaciones:

$$\hat{\mu}_i = \hat{E}(y_i) = e^{b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}}; \quad i = 1,2,\dots n$$

Descripción del Dataset

Los datos que se usaron para trabajar con el modelado de la regresión *Poisson*, fueron obtenidos a través de la última versión de la Encuesta de Demografía y Salud (EDSA), realizada por el INE en 2023. La encuesta EDSA forma parte de las investigaciones que se realizan de forma periódica y a nivel nacional, con el fin de proporcionar información para el cálculo referente a los principales indicadores de salud y demografía como ser la fecundidad, salud materna e infantil, mortalidad infantil y de niñez, vacunación, estado nutricional de los menores de seis años y anticoncepción. Posteriormente, esta información se convierte en el pilar fundamental para formular y evaluar el diseño de políticas públicas y programas que se implementen en Bolivia bajo el Plan de Desarrollo Económico y Social (PDES).

La EDSA 2023 fue realizada bajo cuatro cuestionarios temáticos: cuestionario hogar, cuestionario mujer, cuestionar hombre y el cuestionario de primera infancia, donde cada uno de ellos aborda preguntas diferentes enfocados en la población a la que se encuesta. Para motivos del análisis de datos de la investigación, se utilizaron los datos provenientes del cuestionario mujer, donde se obtiene información respectiva a los datos personales y preferencias de reproducción de las mujeres, y también la información del cuestionario hogar para obtener el ingreso de la familia (esta variable será aproximada mediante el quintil de riqueza).

Así, las variables[4] seleccionadas de forma inicial para realizar la regresión *Poisson* son detalladas en la Tabla 1. El motivo por el cual se seleccionaron estas variables va en línea con la teoría descrita por Schultz (2006) sobre los factores económicos y sociales que influyen en la tasa de fecundidad. En el caso de determinante de mortalidad infantil, no existe una variable que recoja esta información proveniente de la EDSA 2023, por lo que no será tomada en cuenta.

Tabla 1.
Descripción de las variables usadas en ambos modelos de forma inicial.

Factor (según Schultz)	Código variable	Descripción
Variable dependiente	ms02_0208	Número de hijos nacidos vivos
Ingreso	qriqueza	Quintil de riqueza del hogar
	ms08_0809	Si trabajó de forma remunerada durante la semana pasada
Educación	niv_ed	Nivel de educación alcanzado general
	aestudio	Años de estudio alcanzados por la encuestada
Métodos anticonceptivos y planificación familiar	ms02_0276	Si recibió información sobre educación sexual
	cono_algmet	Conocimiento de métodos anticonceptivos
	actmaconcep_cme_m	Si actualmente usa algún método anticonceptivo
Entorno social y cultural	pertenece	Pertenencia a una nación o pueblo indígena (hecho en base a la variable ms01_0108)
	idioma_orig	Si el primer idioma aprendido en la niñez es un idioma originario (elaborado en base a la variable ms01_0106)
	departamento	Departamento de residencia
	region	Región geográfica de residencia (altiplano, valle o llano)
	area	Área de residencia (urbana o rural)
Edad	ms01_0101a	Edad cumplida en años
Estado conyugal	actuni_m	Si actualmente se encuentra en unión con una pareja

Fuente: Elaboración propia

[4] El diccionario de variables de la EDSA 2023 puede ser encontrado en el catálogo ANDA.

Respecto a las variables idioma orig y pertenece se hizo una transformación de las variables ms01_0106 (idioma o lengua que se aprendió a hablar en la niñez) y ms01_0108 (nación o pueblo originario al que pertenece). Así, la variable idioma orig toma en cuenta si la mujer aprendió a hablar un idioma originario en su niñez como el valor de éxito, para de esta forma conocer si la mujer nació en una familia originaria. En el caso de la variable pertenece se toma en cuenta si la mujer pertenece a un pueblo o nación originaria como el valor de éxito, para conocer si se considera parte de un pueblo indígena.

Regresión Poisson

Para la regresión *Poisson* se toma en cuenta a las mujeres mayores a 19 años (mujeres adultas en edad fértil), donde la variable dependiente en este caso corresponde a la variable ms02_0208, que representa el número de hijos vivos de una mujer.

3. RESULTADOS Y ANÁLISIS

En principio se realizó la prueba de sobredispersión de Dean-Lawrence el valor del estadístico z (-12.386) excede considerablemente el valor crítico de ±1.96 para un nivel de significancia del 5%, indicando una desviación altamente significativa del supuesto de equidispersión. El parámetro de dispersión de 0.852 demuestra que la varianza observada es 14.8% menor que la media, configurando un caso de subdispersión.

Se empleó el método de eliminación hacia atrás (backward reduction) al momento de trabajar con la regresión y la selección de variables significativas. Este procedimiento consiste en iniciar con todas las variables independientes disponibles y, posteriormente, eliminar secuencialmente aquellas que no resultan estadísticamente significativas, de acuerdo con un nivel de significancia definido. En cada etapa se vuelve a estimar el modelo hasta obtener una especificación que mantenga únicamente las variables con efecto significativo sobre la variable dependiente.

Este método es una práctica habitual en el análisis de regresión, en especial debido a que permite simplificar el modelo sin comprometer su capacidad explicativa, conduciendo a menudo a un modelo más preciso. Además, facilita la identificación de los factores más relevantes según los criterios estadísticos. De esta forma, se seleccionaron las variables más relevantes, equilibrando el criterio estadístico con fundamentos la teoría económica señalada por Schultz sobre la fecundidad. Así, se buscó construir un modelo parsimonioso y explicativo, sin perder variables importantes.

En la Tabla 2, se observa los resultados de la regresión Poisson donde se obtuvo un modelo con variables significativas y un criterio de Akaike[5] igual a 32073, que corresponde al menor valor encontrado El modelo Poisson ajustado muestra que variables socioeconómicas como nivel de riqueza, educación, condición laboral y área de residencia influyen significativamente en el número esperado de hijos. Las mujeres del quintil más bajo tienen en promedio 15% más en la tasa de fecundidad que las del quintil de referencia (quintil superior) y pertenecer al cuarto quintil de riqueza disminuye la tasa en aproximadamente un 12%, mientras que tener solo primaria completa se asocia con 16% más hijos que el grupo de referencia (secundaria o más). Cada año adicional de estudio reduce en promedio el número de hijos en 3%. Asimismo, no conocer métodos anticonceptivos reduce notablemente la incidencia. A nivel territorial, los hogares

[5] El criterio de Akaike es una medida de la calidad relativa de un modelo estadístico. Cuanto más pequeño sea su valor, indica que el modelo tiene un mejor ajuste.

de La Paz, Cochabamba y Tarija presentan menor incidencia, en tanto que en Potosí es mayor la incidencia del número de hijos respecto al departamento de Pando. La interacción indica que, en áreas urbanas, los hogares del segundo quintil de ingresos menores tienen mayor incidencia en el número de hijos que en áreas rurales.

Tabla 2.
Resultados de la regresión Poisson

Variable	Referencia	Categorías	Coeficiente	Error estándar	Valor P[6]	Sig.	exp(b)
Constante			-0,817	0,061	< 2e-16	***	0,44
Quintil de riqueza	Quintil superior	Quintil inferior	0,142	0,031	3,4E-06	***	1,15
		Segundo quintil	0,026	0,027	0,3452		1,03
		Quintil intermedio	-0,018	0,032	0,5695		0,98
		Cuarto quintil	-0,127	0,040	0,0013	**	0,88
Nivel de educación	Secundaria	Ninguno	-0,029	0,042	0,4882		0,97
		Primaria incompleta	0,023	0,021	0,2756		1,02
		Primaria completa	0,152	0,019	1,4E-15	***	1,16
Años de estudio			-0,030	0,004	5,0E-12	***	0,97
Educación sexual	No sabe	Si recibe información	-0,002	0,020	0,9130		1,00
		No recibe información	0,065	0,020	0,0011	**	1,07
Conoce métodos anticonceptivos	Si conoce	No conoce	-0,052	0,023	0,0240	*	0,95
Usa métodos anticonceptivos	Si usa	No usa	-0,118	0,008	< 2e-16	***	0,89
Idioma originario	Habla idoma originario	No habla idioma originario	-0,026	0,009	0,0038	**	0,97
Departamento	Pando	Chuquisaca	-0,023	0,028	0,4170		0,98
		La Paz	-0,076	0,016	0,0000	***	0,93
		Cochabamba	-0,059	0,018	0,0010	**	0,94
		Oruro	-0,036	0,031	0,2530		0,96
		Potosi	0,034	0,024	0,1531		1,03
		Tarija	-0,111	0,030	0,0002	***	0,90
		Santa Cruz	0,001	0,016	0,9681		1,00
		Beni	0,119	0,030	0,0001	***	1,13
Área	Rural	Urbana	-0,033	0,025	0,1794		0,97
Edad			0,047	0,001	< 2e-16	***	1,05
Estado conyugal	Está en pareja	No está en pareja	-0,270	0,009	< 2e-16	***	0,76
Quintil de riqueza * Área	Quintil superior * Área urbana	Quintil inferior * Área urbana	0,071	0,030	0,0166	*	1,07
		Segundo quintil * Área urbana	0,072	0,027	0,0079	**	1,07
		Tercer quintil * Área urbana	0,019	0,032	0,5455		1,02
		Cuarto quintil * Área urbana	0,010	0,039	0,7955		1,01
		Chi-cuadrado de Pearson	9163,9		< 2e-16	***	
		Devianza	9607,6				
		Pseudo R2 (McFadden)	0,223				

Fuente: Elaboración propia.
Nota: En la tabla ***, ** y * indican que el estimado es significativo al 0.1%, 1% y 5% respectivamente.

[6] El Valor P representa el riesgo de rechazar la hipótesis nula dado que es verdadera, es decir rechazar que la variable independiente no tenga efecto sobre la variable dependiente cuando si es significativa.

Finalmente, en áreas urbanas, las mujeres del quintil más bajo tienen 7% más hijos que sus equivalentes rurales. El ajuste global es bueno (devianza residual/grados de libertad ≈ 0.89), lo que sugiere ausencia de sobredispersión.

Comparando con la teoría de Schultz, se observa que las variables que explican el número de hijos están asociadas con los determinantes de ingreso familiar, educación, conocimiento y uso de métodos anticonceptivos, el entorno social y cultural de la mujer, la edad y el estado conyugal. Asimismo, los resultados indican que área de residencia (es decir el área rural o urbana) no presenta una asociación significativa con la fecundidad, en contraste con la hipótesis propuesta por Schultz (2007).

4. DISCUSIÓN

El ingreso puede entenderse de dos formas, según lo que menciona Schultz. En primer lugar, se refiere al ingreso derivado del capital físico (como ser la tierra) que tiende a aumentar la fecundidad, pues incrementa el consumo familiar y el valor financiero inmediato de tener más hijos. En este sentido, el capital físico se entiende como aquellos activos que generan un retorno económico para la mujer y, consecuentemente, su familia. Siendo así que estos ingresos al producir un flujo de ingresos o bienes para el hogar, reducen el costo de tener más hijos. La segunda forma de ingreso se refiere a aquel derivado del capital humano de la mujer (como ser los ingresos propios o la educación de la mujer), los cuales tienden a reducir la fecundidad. Esto sucede debido al mayor costo de oportunidad de la maternidad, dado que el valor del tiempo disponible para trabajar de la mujer es mayor al valor de su tiempo invertido en ser madre (Schultz, 2007).

Con respecto al nivel de educación de la madre, al igual que observaba Schultz, resulta en una variable relevante, ya que mayores años de escolaridad femenino están asociados de forma negativa con la tasa de fecundidad. Este hecho también es señalado por Enríquez-Canto, Ortiz-Romaní y Ortiz-Montalvo (2018), quienes mencionan que la educación incrementa el costo de oportunidad de la maternidad y repercute de manera positiva en el status y la autonomía de la mujer.

En relación con los métodos anticonceptivos y la planificación familiar, Schultz denota que el acceso a métodos anticonceptivos reduce de forma efectiva la fecundidad para una pareja. Así mismo, Schultz indica que dicho aspecto está íntimamente ligado a la educación y la salud reproductiva dentro de una pareja. Por otro lado, al hablar de la mortalidad infantil, una reducción en este indicador genera que la esperanza por tener más hijos disminuya para una familia, lo cual reduce la fecundidad. De esta forma, se enfatiza la influencia de la supervivencia infantil en las decisiones reproductivas (Schultz, 2005).

Las normas sociales y culturales sobre familia y fecundidad, así como el entorno social, condicionan las expectativas sobre el tamaño ideal del hogar; aunque Schultz se centra más en factores económicos, su análisis reconoce esta dimensión. Adicionalmente, es necesario mencionar la relevancia de esta última variable en el estudio de la fecundidad, ya que se asocia la reducción de las tasas de fecundidad a nivel de país e incluso a nivel mundial con el cambio dentro de las expectativas personales y/o profesionales en los jóvenes y los nuevos ideales del tamaño familiar. Mismas variables están intrínsecamente relacionadas con la sociedad y cultura moderna (Enríquez-Canto et al., 2018).

No obstante, y con los resultados obtenidos, es importante recalcar el papel que juegan las políticas públicas en la fecundidad. Como explica Castro Torres (2021), estas políticas deben estar dirigidas a reducir las desigualdades estructurales, como ser el acceso a la educación, salud o empleo, que reproducen patrones divergentes en la fecundidad. Al igual que se observa en los resultados obtenidos mediante la regresión Poisson en el presente trabajo, estas variables son relevantes para determinar la tasa de fecundidad en el caso boliviano. Sin embargo, de la misma forma que lo plantea Castro Torres, resulta ser necesario estudiar la fecundidad comparando grupos sociales si se busca estudiar patrones y detectar soluciones para grupos específicos.

Finalmente, pese a que Schultz no menciona a la edad y el estado conyugal de una mujer como determinantes causal principales en su modelo, estas variables pueden ser cruciales en términos del horizonte fértil y el momento de decisión de reproducción para una familia. Como tal, Schultz analiza la fecundidad acumulada por edades específicas, añadiendo así de forma implícita a la edad como parte de la estructura de oportunidades de fecundidad. Respecto al estado conyugal, el hecho de estar en pareja formal (ya sea en matrimonio o convivencia) facilita o incrementa la probabilidad de tener hijos. A su vez, la decisión de formar pareja también puede verse influida por variables económicas, como el ingreso esperado o la educación.

En el caso de América Latina, estudios recientes muestran que la región ha experimentado una transición demográfica acelerada, con una notable reducción en los niveles de fecundidad y un acercamiento hacia tasas de reemplazo. Rosero-Bixby (2018) destaca que el acceso creciente a la educación secundaria y superior, especialmente en mujeres, constituye uno de los factores más influyentes en este cambio. De manera complementaria, Cavenaghi y Alves (2019) subrayan el rol de las políticas públicas en la ampliación del acceso a métodos anticonceptivos modernos y en la disminución de las desigualdades reproductivas. Investigaciones como la de Echarri-Cánovas y Juárez (2020) en México evidencian que los cambios en la participación laboral femenina y en la estructura social han sido determinantes en la reducción de la fecundidad.

Estos hallazgos son coherentes con los resultados obtenidos en el presente trabajo para el caso boliviano y permiten situarlos dentro de una dinámica regional, donde los factores socioeconómicos, educativos y culturales interactúan de forma significativa en la configuración de los patrones de fecundidad.

Los resultados también sugieren la necesidad de políticas públicas focalizadas. Por un lado, la asociación inversa entre educación femenina y fecundidad puede interpretarse como evidencia del efecto causal que ejerce la expansión educativa en la postergación de la maternidad y en la reducción del número de hijos. Por otro lado, el acceso desigual a servicios de salud reproductiva y la heterogeneidad en los niveles de ingreso del hogar evidencia la persistencia de brechas socioeconómicas en la fecundidad.

5. LIMITACIONES

El estudio presenta algunas limitaciones que deben ser consideradas al interpretar los resultados. En primer lugar, los datos utilizados provienen de encuestas de hogares que, si bien son representativas a nivel nacional, pueden estar sujetos a sesgos por parte de las personas encuestadas. En segundo

lugar, el modelo aplicado se basa en supuestos estadísticos que no siempre capturan toda la complejidad de los determinantes sociales y culturales de la fecundidad, lo que puede limitar la generalización de los hallazgos. En tercer lugar, la naturaleza transversal de la información impide establecer relaciones de causalidad estrictas, de modo que las asociaciones identificadas deben interpretarse como correlaciones.

Pese a estas limitaciones, los resultados aportan evidencia valiosa sobre los factores que influyen en la fecundidad en Bolivia y pueden servir de base para el diseño de políticas públicas más inclusivas y que promuevan una mayor equidad en las oportunidades de desarrollo social y económico del país.

6. CONCLUSIONES Y RECOMENDACIONES

Finalmente, se logró realizar y analizar el modelo Poisson para determinar los factores relevantes en la fecundidad en Bolivia. Siendo este aspecto de gran importancia para la población boliviana y la creación de políticas sociales, vista la tendencia de la tasa global de fecundidad durante los últimos años y habiendo destacado la relevancia de estos indicadores dentro de la dinámica demográfica y el crecimiento poblacional. Igualmente, se buscó analizar estos factores a través de la teoría económica y social propuesta por Schultz, y los estudios de Rosero-Bixby, Cavenaghi y Alves y Echarri-Cánovas y Juárez, la cual aporta sobre las variables que juegan un rol importante al momento de analizar la fecundidad en un país.

De los resultados obtenidos se observó que los factores relevantes que van en concordancia con la teoría de revisión bibliográfica son el ingreso, educación, métodos anticonceptivos y planificación familiar, quintiles de riqueza * área, entorno social y cultural. Sin embargo, como variables exógenas se añadieron la edad y el estado conyugal las cuales resultaron significativas en el modelo final.

Como recomendaciones se resalta la importancia de llevar a cabo un estudio con datos recientes y posiblemente tomando en cuenta la variable mortalidad infantil para así analizar el efecto que tiene sobre la fecundidad en Bolivia. Igualmente, es aconsejable realizar estudios adicionales considerando otras variables como ser el ingreso de la familia o variables sobre la pareja de la mujer.

REFERENCIAS BIBLIOGRÁFICAS

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). John Wiley & Sons.

Batyra, E. (2020). Increasing educational disparities in the timing of motherhood in the Andean region: A cohort perspective. Population Research and Policy Review, 39(3), 283–309. *https://doi.org/10.1007/s11113-019-09535-0*

Castro Torres, A. F. (2021). Analysis of Latin American fertility in terms of probable social classes. European Journal of Population, 37(2), 297–339. *https://doi.org/10.1007/s10680-020-09569-7*

Cavenaghi, S., & Alves, J. E. D. (2019). *La transición de la fecundidad en América Latina: diversidad y retos*. CEPAL.

Comisión Económica para América Latina y el Caribe (CEPAL). (2020). L*a matriz de la desigualdad social en América Latina*. Naciones Unidas.

Echarri-Cánovas, C. J., & Juárez, F. (2020). *Determinantes socioeconómicos de la fecundidad en México: una visión desde la transición demográfica*. Estudios Demográficos y Urbanos, 35(1), 45–72.

Enríquez-Canto, Y., Ortiz-Romaní, K., & Ortiz-Montalvo, Y. (2018). Análisis de los determinantes próximos e impacto de la ocupación en la fertilidad de mujeres peruanas. Revista Panamericana de Salud Pública, *42*, e160. *https://doi.org/10.26633/RPSP.2018.160*

Instituto Nacional de Estadística (INE). (2023). *Encuesta de Demografía y Salud (EDSA) 2023*. *https://anda.ine.gob.bo/index.php/catalog/119*

Instituto Nacional de Estadística (INE). (2025). *Fecundidad y maternidad: Una mirada a la transición demográfica en Bolivia EDSA 2023*. *https://anda.ine.gob.bo/index.php/catalog/119/download/1240*

Long, J. S., & Freese, J. (2014). *Regression models for categorical dependent variables using Stata* (3rd ed.). College Station, TX: Stata Press.

Rodríguez Vignoli, J. (2014). *Fecundidad y maternidad adolescente en América Latina: desigualdades socioeconómicas y espaciales. Comisión Económica para América Latina y el Caribe (CEPAL)*. Naciones Unidas.

Rosero-Bixby, L. (2018). La fecundidad en América Latina: tendencias recientes y perspectivas futuras. Revista Latinoamericana de Población, 12(22), 7–33.

Schultz, T.P. (2005) *Fertility and income. Working Paper 925, Economic Growth Center, Yale University*. URL *http://www.econ.yale.edu/growth_pdf/cdp925.pdf*. Publicado en 2006.

Schultz, T. P. (2007). *Population policies, fertility, women's human capital, and child quality.* En T. P. Schultz & J. Strauss (Eds.), Handbook of Development Economics (pp. 3249–3303). Elsevier. *https://doi.org/10.1016/S1573-4471(07)04052-1*

INSTRUCCIONES PARA AUTORES

REVISTA VARIANZA
Revista Científica del Instituto de Estadística Teórica y Aplicada (IETA),
Carrera de Estadística, Facultad de Ciencias Puras y Naturales
Universidad Mayor de San Andrés
La Paz, Bolivia

ISSN 2789-3510, versión impresa
ISSN 2789-3529, versión en linea
https://ojs.umsa.bo/index.php/revistavarianza/
https://ieta.umsa.bo/

MISIÓN

Difundir principalmente artículos originales de investigación científica en diferentes ámbitos de la vida, basados en el uso de métodos y técnicas estadísticas. También difundir artículos de naturaleza teórica en el campo de la Estadística. Todo ello con el propósito de contribuir al desarrollo de nuestra sociedad.

VISIÓN

Llegar a ser la revista científica nacional de mayor calidad e impacto en el campo de la estadística aplicada y teórica, así como ser el principal referente para el contexto internacional.

TIPOS DE MANUSCRITOS

En la Revista Varianza se publican principalmente **artículos originales**, aquéllos que resultan de una investigación científica y que contribuyen, en alguna medida, al conocimiento científico y/o solución de alguna problemática. Los artículos originales pueden ser de naturaleza teórica o práctica. Los de naturaleza práctica se enfocan en dar respuesta, con base en el uso de métodos y/o técnicas estadísticas apropiadas, a problemas o preguntas de investigación en distintos campos de la vida; mientras los de naturaleza teórica presentan un nuevo método o técnica estadística, o pueden ofrecer una versión mejorada de uno existente.

Además, si bien no son artículos originales de investigación, en la Revista Varianza también se publican otro tipo de manuscritos, como ser:

De revisión, que constituyen básicamente informes sobre avances o estado del arte de un tema particular, con base en la recopilación y selección de artículos científicos originales,

Comunicaciones breves, manuscritos que comunican de manera breve algunos datos de una investigación original que el editor cree que serán interesantes para muchos investigadores y que probablemente estimularán más la investigación en esa área,

Estudios de caso, informan los resultados sobre casos específicos de fenómenos interesantes. Su propósito es hacer que otros investigadores conozcan la posibilidad de que un fenómeno específico pueda ocurrir,

Reseñas, consisten en resúmenes concisos generalmente sobre libros recientemente publicados en el campo de la Estadística,

Notas científicas, presentan observaciones y descripciones científicas breves de métodos o resultados, comunican resultados de estudios pequeños, avances de trabajos de investigación o noticias de interés científico,

De enseñanza, son manuscritos sobre temas relacionados a la enseñanza de la Estadística, por ejemplo la comprensión de un teorema o un método de estimación. Tiene el propósito de clarificar y complementar los conocimientos estadísticos de los estudiantes y los docentes.

PROCESO DE REVISIÓN DE MANUSCRITOS

Luego de haber recibido el manuscrito, se inicia el proceso de su revisión, el cual tiene tres etapas:

Primera etapa: Revisión por el editor

En esta etapa, el editor revisa si el tema del manuscrito es apropiado para la Revista y si cumple con las instrucciones para los autores. Se revisa la pertinencia del manuscrito para la Revista, los aspectos de forma del manuscrito y el cumplimiento de requisitos básicos exigidos en las normas de la Revista Varianza. El autor será contactado para informarle si su manuscrito es apto para pasar a la revisión por pares a doble ciego o si requiere mejorar algunos detalles del manuscrito sugeridos por el editor o si es rechazado (por no presentarse en el formato exigido en las normas, por tener errores metodológicos importantes, porque el manuscrito ha sido publicado previamente o porque el aporte no es nuevo, entre otros). En caso de no existir faltas o errores, el manuscrito pasa a la segunda etapa.

Segunda etapa: Revisión por evaluadores externos

Cada manuscrito que llega a esta etapa es sometido al proceso de revisión por pares a “doble ciego”. Esta modalidad significa que cada manuscrito es revisado por dos evaluadores externos a nuestra institución, ambos miembros del comité científico, con la restricción de que ni el evaluador sabe el nombre del autor del manuscrito y ni el autor sabe quiénes son sus evaluadores. Para la asignación del manuscrito a los dos evaluadores se toma en cuenta el vínculo entre el tema del manuscrito y la especialidad o experiencia de los evaluadores.

En esta etapa se evalúa rigurosamente el contenido del manuscrito, poniendo énfasis en los aspectos metodológicos. A través de una ficha de evaluación, el dictamen de la revisión es una de las siguientes cuatro alternativas: (i) aceptado sin modificaciones, (ii) aceptado con observaciones leves, (iii) aceptado con observaciones profundas o (iv) rechazado. Se comunica al evaluador tanto el dictamen como las observaciones, en caso de existir. Una vez que el autor subsana las observaciones, nuevamente el manuscrito corregido es derivado

al revisor para su evaluación, y así sucesivamente hasta que el manuscrito es aceptado o rechazado. En caso de no corregirse las observaciones, el manuscrito es rechazado.

Algunos aspectos que se toman en cuenta en la evaluación, son:

a) Claridad en el planteamiento del objetivo principal y/o la hipótesis a probar.
b) Pertinencia del método estadístico para el cumplimiento del objetivo y/o prueba de hipótesis.
c) Datos apropiados para el cumplimiento del objetivo y/o prueba de hipótesis.
d) Grado de profundidad de la investigación.
e) Coherencia en el análisis cuantitativo, incluido la correcta interpretación de los resultados.
f) Generación de nuevo conocimiento científico y aporte para la formulación de políticas, programas y proyectos.
g) Pertinencia de la referencia bibliográfica usada.
h) Calidad de la redacción, incluido el resumen (síntesis del manuscrito en forma clara y adecuada).
i) Ajuste del formato a las normas APA.

Tercera etapa: Diagramación

Los manuscritos aprobados por el comité científico pasan a diagramación, a una última revisión de redacción y a maquetación. Esta tarea es realizada por un profesional en diagramación y el editor, en coordinación con el autor. Se trata de una última revisión del manuscrito, sólo de forma. No se acepta ninguna modificación con relación al contenido del manuscrito ya aceptado, sea de texto, tabla o gráfico, como tampoco se acepta la inclusión de un texto adicional, una tabla o un gráfico.

FORMATO PARA ARTÍCULOS ORIGINALES DE INVESTIGACIÓN

Título

El título debe estar en español y en inglés, ambos en mayúscula, en no más de 15 palabras cada uno, por lo que debe ser lo más corto posible y debe reflejar plenamente el contenido del manuscrito.

Autor(es)

Debajo del título deben colocarse el nombre completo del autor, su afiliación institucional durante la realización del manuscrito y su correo electrónico. Si son dos o más autores, colocar el nombre completo, la afiliación institucional y el correo electrónico de cada uno. Para el caso de dos o más autores, el principal autor es el que más ha contribuido a la investigación, y deberá encabezar la lista. En pie de página, y con la numeración correspondiente para cada autor, se debe informar la profesión, un breve curriculum vitae si se desea, y el número ORCID.

Resumen

El resumen debe estar en español y en inglés, con títulos en mayúscula (RESUMEN y ABSTRACT, respectivamente), y muy bien redactado en no más de 250 palabras. Debe incluir con absoluta claridad y precisión el objetivo, la metodología, los datos usados, el principal resultado, alguna idea de discusión y la principal conclusión.

Palabras claves

Debajo del resumen en español deben incluirse las palabras clave (Palabras clave), las cuales sirven para identificar el artículo por parte de usuarios nacionales e internacionales. Incluir de cuatro a siete palabras clave que no formen parte del titulo del artículo, ordenadas alfabéticamente y separadas por comas. Igualmente, debajo del resumen en inglés incluir las mismas palabras clave, pero en inglés (Key words).

Introducción

Debe presentar el problema dentro de un marco teórico y/o revisión bibliográfica que acompañe a la hipótesis y/o objetivo principal del trabajo. El problema debe describir claramente lo que se resolverá con la investigación; la revisión bibliográfica expone el marco de referencia que da sustento al trabajo de investigación, resalta, a través de citas, estudios previos que se relacionan específicamente con la investigación; el objetivo principal debe indicar claramente, en forma precisa y sin ambigüedad, la finalidad de la investigación; mientras la hipótesis debe plantear lo que trata de probar la investigación. El objetivo y la hipótesis deben estar al final de esta sección.

Materiales y métodos

En esta sección se describe minuciosamente la forma en que se realizó la investigación, de modo que un lector que acceda al artículo pueda comprenderlo plenamente y reproducirlo a fin de determinar la confiabilidad y la validez de los resultados. Esta sección debe describir claramente los aspectos principales respecto de los datos, los instrumentos, y los métodos y técnicas estadísticas usados en la investigación.

Resultados

En esta sección se expone el principal o los principales hallazgos obtenidos con la investigación, todos ellos en estricta consonancia con el objetivo principal y/o con la hipótesis de la investigación. La correcta interpretación de los resultados es de suma importancia en esta sección. Para transmitir los principales hallazgos pueden incluirse, si es necesario, cuadros, gráficos y diagramas, evitando la redundancia, evitando un número excesivo de datos y manteniendo la objetividad (imparcialidad y honestidad). Todos los cuadros, gráficos y diagramas deben enumerarse en el orden que aparecen en el texto.

Discusión

Esta sección está orientada a interpretar los resultados de la investigación en relación con el objetivo principal, la hipótesis y el estado de conocimiento actual del tema de la investigación, esto es, se debe indicar qué significan los hallazgos encontrados y cómo estos se relacionan con el conocimiento actual sobre el tema.

Además de compararlos y contrastarlos con los resultados de otros estudios relevantes, resaltando sus limitaciones y ventajas tanto conceptuales como metodológicas, argumentar las implicaciones de los resultados para la formulación de políticas, programas y/o proyectos, y argumentar las implicaciones para futuras investigaciones.

Conclusiones

Las conclusiones son derivadas de los resultados y de la discusión, y responden al objetivo y/o la hipótesis de la investigación. Constituyen los aportes y las innovaciones de la investigación

Agradecimiento

En esta sección no se incluye ningún elemento científico, sólo se trata de ser cortés con quienes colaboraron en la investigación. Se puede reconocer la contribución de personas o instituciones que ayudaron realmente en la investigación, pero no se las puede considerar como coautores.

Conflicto de intereses

Acá, el autor o autores deben declarar no tener ningún conflicto de intereses con su artículo científico.

Referencias bibliográficas

Esta sección contiene la referencia de libros y artículos citados en las diferentes secciones del manuscrito, en formato APA. Debe existir siempre una correspondencia entre las citas que se hace en el trabajo y las que se lista en las referencias bibliográficas, ya que normalmente los lectores estarán interesados en verificar los datos que efectivamente se utilizaron para la investigación.

Ejemplos de cómo enunciar las referencias bibliográficas en formato APA se pueden encontrar en la siguiente dirección (URL): https://normas-apa.org/referencias/

ESPECIFICACIONES PARA LA PRESENTACIÓN DE MANUSCRITOS

Los manuscritos que se presentan deben estar escritos en Word, hoja tamaño carta, doble columna, letra times new román de tamaño 11, espacio simple, margen izquierdo de 2.5 cm. y los demás márgenes de 2.0 cm. Todo el manuscrito, incluido texto, gráficos, cuadros, diagramas y otros, debe contener entre 10 y 20 páginas, con títulos y subtitulos enumerados.

Los gráficos, cuadros y diagramas no deben exceder el 30 por ciento del manuscrito. Adicionalmente, todas las notas y referencias deben ir acorde al formato APA.

Para la presentación del manuscrito debe acompañarse una carta en la que se indique el tipo de manuscrito que se está enviando (artículo original, revisión, estudio de caso, reseña, nota científica o manuscrito de enseñanza) a la siguiente dirección: ieta@umsa.bo. Los autores pueden enviar sus manuscritos en cualquier momento del año.

PERIODICIDAD DE LA PUBLICACIÓN

La versión impresa de la Revista Varianza se publicó por primera vez el año 2001, desde ese año hasta el 2020 se publicó anualmente, si bien no se pudo editar en algunos años. Sin embargo, a partir del segundo semestre de 2021 la publicación es semestral, en los meses de abril y octubre.

En cambio, la versión digital (on line) de la Revista se publica desde el segundo semestre del año 2021, también en los meses de abril y octubre.

Con el propósito de incrementar la visibilidad de la Revista Varianza y facilitar la búsqueda de artículos por parte de los lectores, desde el año 2023 la Revista Varianza también se publica junto a las revistas científicas de otras unidades de la Universidad Mayor de San Andrés. Se puede acceder a la página a través de la dirección *https://ojs.umsa.bo*.

CONFLICTO DE INTERESES

La Revista Varianza tiene la política de evitar cualquier conflicto de interés de los autores, del comité editorial y del propio editor responsable. Se recomienda a cada autor evitar cualquier conflicto de interés relacionado con su artículo, debiendo comunicar oportunamente al editor responsable, como también se pide al comité editorial impedir cualquier conflicto de interés en el proceso editorial.

ÉTICA DE PUBLICACIÓN

La revista Varianza tiene compromiso con la ética de la investigación, por ello, promueve los siguientes aspectos:

a) Evitar conflictos de intereses
b) Evaluar objetivamente los manuscritos
c) Respetar los criterios de evaluación de los evaluadores
d) Conservar la confidencialidad de los autores y evaluadores, durante todo el proceso de revisión.

FINANCIAMIENTO DE LA REVISTA

La Universidad Mayor de San Andrés (UMSA) asigna anualmente recursos financieros al Instituto de Estadística Teórica y Aplicada (IETA) para impresión y difusión de la Revista Varianza. La publicación en la revista es gratuita bajo la modalidad Open Access.

PROPIEDAD INTELECTUAL

LICENCIAMIENTO

Dirección: Calle 27 de Cota Cota
Bloque FCPN - Primer Piso
Email: ieta@umsa.bo
Pagina web: https://ojs.umsa.bo/index.php/revistavarianza

LA PAZ - BOLIVIA