

Análisis de la influencia geográfica en la evaluación Ser Estudiante a través de estadística multivariante

Analysis of the geographical influence on the assessment Ser Estudiante using multivariate statistics

Mario Orlando Suárez-Ibujés¹[0000-0002-3962-5433], Nadia Nathaly Sánchez-Pozo¹[0000-0003-0509-2082]

¹ Universidad Politécnica Estatal del Carchi, Posgrado. Tulcán, 040101, Ecuador.

¹{mario.suarez, nadia.sanchez}@upec.edu.ec

CITA EN APA:

Suárez-Ibujés, M. O., & Sánchez-Pozo, N. N. (2023). Análisis de la influencia geográfica en la evaluación Ser Estudiante a través de estadística multivariante. *Prometeo Conocimiento Científico*, 3(1), e27. <https://doi.org/10.55204/pcc.v3i1.e27>

Recibido: 2023-05-13

Revisado: 2023-05-15 al 2023-06-01

Corregido: 2023-06-04

Aceptado: 2023-06-06

Publicado: 2023-06-09

Prometeo
Conocimiento Científico
ISSN: 2953-4275



Los contenidos de este artículo están bajo una licencia de Creative Commons Attribution 4.0 International (CC BY 4.0)

Los autores conservan los derechos morales y patrimoniales de sus obras.

Resumen.

Introducción: Existen múltiples elementos que inciden en los resultados de las evaluaciones Ser Estudiante, encontrándose la zona geográfica rural o urbana de la institución educativa entre uno de los factores que influyen en el rendimiento escolar logrado.

Objetivo: El objetivo de este estudio es analizar la influencia de la zona geográfica de la institución educativa en la evaluación Ser Estudiante de tercero de bachillerato 2021-2022 mediante descriptivos estadísticos, test de normalidad a través de la prueba de Kolmogórov-Smirnov, test de homocedasticidad con la prueba de Fligner-Killeen, test de igual de medianas utilizando la prueba de Wilcoxon-Mann-Whitney, y modelación con el método de Naive Bayes.

Métodos: Para ello se ha utilizado una metodología de enfoque cuantitativo, no experimental, transversal y correlacional. La muestra la integran 4466 estudiantes, 2610 del área urbana y 1856 del área rural, que rindieron la evaluación Ser Estudiante de tercero de bachillerato 2021-2022.

Resultados: Los resultados revelan que los estudiantes del área urbana obtienen mejores desempeños que los estudiantes del área rural en la mayoría de los campos evaluados.

Conclusiones: Se concluye que la zona geográfica de los establecimientos educativos influye para que en su mayoría los estudiantes del área urbana obtengan mejores resultados que los estudiantes del área rural en la evaluación Ser Estudiante de tercero de bachillerato 2021-2022.

Palabras Clave: influencia geográfica; evaluación del estudiante; rendimiento escolar; análisis cuantitativo.

Abstract:

Introduction: There are multiple elements that influence the results of the Ser Estudiante evaluations, and the rural or urban geographical area of the educational institution is one of the factors that influence the school performance achieved.

Objective: The objective of this study is to analyze the influence of the geographic area of the educational institution on the Ser Estudiante evaluation of third year of high school 2021-2022 by means of statistical descriptions, normality test through the Kolmogorov-Smirnov test, homoscedasticity test with the Fligner-Killeen test, test of equal medians using the Wilcoxon-Mann-Whitney test, and modeling with the Naive Bayes method.

Methods: A quantitative, non-experimental, cross-sectional and correlational methodology was used. The sample was composed of 4466 students, 2610 from urban areas and 1856 from rural areas, who took the Ser Estudiante 2021-2022 third year high school evaluation.

Results: The results reveal that students from urban areas perform better than students from rural areas in most of the fields evaluated.

Conclusions: It is concluded that the geographic area of the educational establishments influences the majority of students from urban area to obtain better results than students from rural areas in the Ser Estudiante assessment for the third year of high school 2021-2022

Keywords: geographical influence; student evaluation; school performance; quantitative analysis.

1. INTRODUCCIÓN

Varios estudios han mencionado la existencia de diversos factores externos e internos que inciden en el rendimiento escolar como la condición económica, el nivel de educación de los padres de familia y el entorno institucional (Borja et al., 2021). De la misma manera, según Ainsworth (como se citó en Santillán y Vargas, 2020) presentó que el nivel de pobreza del vecindario de la institución educativa está asociado a los resultados del rendimiento escolar, incluso posterior a controlar los antecedentes familiares de los estudiantes. Aunque se han hallado varios elementos que pueden afectar el rendimiento estudiantil, varias investigaciones han asociado el rendimiento estudiantil con variables de carácter socioeconómico, entre las más importantes destaca la zona geográfica donde se encuentra la escuela (Pozo et al., 2022).

En relación a lo anterior, investigaciones sobre las pruebas PISA han indicado que uno de los factores que más condicionan el desempeño estudiantil es el tipo de institución educativa, siendo las instituciones privadas las que tienen mejores desempeños que las públicas (Castro Aristizabal et al., 2018). Así mismo, estudios regionales han mencionado que “entre un 40% y un 50% de las diferencias de aprendizaje entre los estudiantes depende de la escuela, esto indica que en nuestra región es muy relevante el establecimiento educativo al que asisten los estudiantes” (UNESCO, 2021,p.36). De manera similar, “la investigación social y educativa ha mostrado que las escuelas situadas en zonas de niveles económicos más desfavorecidos obtienen peores resultados en el rendimiento académico de sus estudiantes, en comparación con las situadas en contextos económicos más favorecidos”(Rodríguez et al., 2019,p.124).

Dentro de este marco, el artículo 26 de la Constitución de la República del Ecuador establece que “la educación es un derecho de las personas a largo de su vida y un deber ineludible e inexcusable del Estado” (Constitutivos y Estado, 2021, p.18). Sin embargo, en las zonas geográficas rurales las políticas públicas no han llegado de la forma esperada, las instituciones en área rurales disponen de menos recursos educativos y tecnológicos, los estudiantes y docentes del sector rural luchan en medio de situaciones adversas, los estudiantes de las zonas urbanas tienen más acceso a museos, bibliotecas y otros recursos culturales y educativos en comparación con los estudiantes de las zonas rurales, lo cual ha ocasionado que los resultados de aprendizaje no sean los que el sector rural aspira y se merece (Garofalo y Villao, 2018).

En relación a la problemática expuesta, en las evaluaciones Ser Estudiante (SEST) que son aplicadas por el Instituto Nacional de Evaluación Educativa (Ineval) en Ecuador desde el 2014 y constituyen un conjunto de iniciativas y políticas públicas utilizadas para obtener mejores resultados de calidad educativa en el país, históricamente los estudiantes de las instituciones del área rural han obtenido en promedio resultados más bajos que aquellos estudiantes de las instituciones educativas ubicadas en el área urbana (INEVAL, 2019).

También, al analizar los resultados en la evaluación Ser Bachiller 2016, evaluación similar al SEST, se observó un promedio mayor en las instituciones educativas del área urbana en comparación con las instituciones educativas del área rural, concluyendo que el contexto de la institución educativa influye en los mencionados resultados (Toscano y Valencia, 2020).

En este contexto surge la necesidad de seguir analizando las variables que inciden en las evaluaciones Ser Estudiante para tomar decisiones de mejora educativa. Al respecto, el análisis de los resultados en las evaluaciones han sido un instrumento útil para proponer innovadas estrategias que favorezcan al desarrollo integral del estudiantado (INEVAL, 2019). También, el análisis de los resultados han sido de utilidad para “proveer indicadores nacionales e internacionales de aprendizaje que contribuyan a la consecución de los Objetivos de Desarrollo sostenible y contribuir a la política pública para la mejora de la calidad educativa” (INEVAL, 2022, p.4). Igualmente, al identificar estas posibles causas se ha creado la base para mejorar los procesos educativos en el corto y mediano plazo (Medina y Villarreal, 2018).

En definitiva, el análisis de la influencia de la zona geográfica ha sido un tema motivo de estudio por los diversos factores influyen en los logros escolares que determinan la calidad educativa que se imparte en los centros educativos (Pérez y Romero, 2018). Además, diversos trabajos de investigación han empleado la estadística inferencial para estudiar la correlación entre diversas variables que influyen en las evaluaciones escolares (Aliu et al., 2021). Sin embargo, estos factores de influencia en los resultados de las evaluaciones Ser Estudiante en Ecuador es un tema aun poco investigado desde la estadística multivariante.

La más reciente evaluación Ser Estudiante de tercero de bachillerato fue aplicada en el año lectivo 2021-2022 a una muestra representativa a nivel nacional con base al currículo, los estándares de aprendizaje, régimen (Costa-Galápagos y Sierra-Amazonía), área (rural y urbana) y financiamiento (público y privado) (INEVAL, 2022). En esta evaluación se consideraron los campos de “Matemática, Lengua y Literatura, Biología, Física, Química, Historia, Educación para la Ciudadanía y Filosofía”(INEVAL, 2022, p. 8).

El presente artículo tuvo por objetivo analizar la influencia de la zona geográfica rural o urbana de la institución educativa en la evaluación Ser Estudiante de tercero de bachillerato 2021-2022 mediante descriptivos estadísticos (media aritmética, desviación estándar, percentil 25, mediana, percentil 75), comprobación del supuesto de normalidad utilizando la prueba el test de Kolmogórov-Smirnov, comprobación del supuesto de homocedasticidad empleando el test no paramétrico Fligner-Killeen, comprobación de la existencia de diferencia significativa de los resultados obtenidos por los estudiantes del área rural y urbana mediante el test no paramétrico de Wilcoxon-Mann-Whitney, y modelación de la influencia del zona geográfica urbana y rural a través el método de Naive Bayes (Frank et al., 2000).

2. METODOLOGÍA O MATERIALES Y MÉTODOS

2.1. Diseño de la investigación

La investigación se realizó mediante el enfoque cuantitativo de investigación con diseño no experimental de tipo transversal (Hernández y Mendoza, 2018) a través de la cálculos estadísticos y prueba de hipótesis sin la manipulación directa y deliberada en un contexto natural y momento único. Así mismo, cumplió el criterio de investigación correlacional (González et al., 2020), debido a que se evaluó la

correlación de las variables presentes en la evaluación Ser Estudiante de tercero de bachillerato 2020-2021 empleando el método estadístico de Naive Bayes.

2.2. Población y muestra

La muestra estuvo conformada por 4466 estudiantes (2610 del área urbana y 1856 del área rural) del sistema educativo nacional ecuatoriano que rindieron la evaluación Ser Estudiante de tercero de bachillerato del año lectivo 2021-2022. La muestra se obtiene de la base de datos del Ineval (Bases de Datos Ser Estudiante, 2022).

Ahora bien, la base de datos empleada se conformó por 15 variables correspondientes al estudiantado cuyo estado de evaluación del sustentante se registró como evaluado: área de asentamiento de la institución educativa rural o urbana (tp_area), región natural del ecuador (nm_regi), sostenimiento de la institución educativa (sostenimiento), sexo del sustentante (tp_sexo), autoidentificación étnica (etnibee), clasificación según el índice socioeconómico (quintil), promedio global obtenido por el estudiante (inev), promedio obtenido en la asignatura de Matemática (imat), promedio obtenido en la asignatura de Lengua y Literatura (ilyl), promedio obtenido en la asignatura de Física (ifis), promedio obtenido en la asignatura de Química (iqui), promedio obtenido en la asignatura de Biología (ibio), promedio obtenido en la asignatura de Historia (ihis), promedio obtenido en la asignatura de Filosofía (ifil) y promedio obtenido en la asignatura de Educación para la Ciudadanía (ied) (INEVAL, 2022b).

2.3. Procedimiento de recogida y análisis de datos

Con la base de datos de la evaluación Ser Estudiante de tercero de bachillerato del año lectivo 2021-2022 que cumple con el supuesto de independencia (Amat, 2020) debido a que se aplicó a una muestra representativa de estudiantes con particularidades distintas que pertenecen a unidades educativas de diferente zona geográfica (INEVAL, 2019) se realizó el tratamiento estadístico empleando RStudio versión 4.1.3.

Se comprobó el supuesto de normalidad de las variables (Lugo y Pino, 2021) a través del test de Kolmogórov-Smirnov (con la correlación de Lilliefors), la misma es una herramienta estadística para realizar el contraste de prueba de normalidad de una muestra cuyo conjunto de datos sea mayor a 50 (Mata, 2016). Además, esta prueba es una herramienta útil en muchos análisis estadísticos (Flores, 2021), sin embargo, es importante recordar que esta prueba solo puede detectar desviaciones significativas de la normal en muestras grandes, ya que en el caso de tener muestras de tamaño menor a 50 el test de Shapiro-Wilk es el más apropiado. En esta prueba el contraste de hipótesis de normalidad para este estudio fue:

H_0 : Los datos de la evaluación Ser Estudiante siguen una distribución normal

H_1 : Los datos de la evaluación Ser Estudiante no siguen una distribución normal

También se comprobó el criterio de homocedasticidad o igualdad de varianzas empleando el test no paramétrico de Fligner-Killeen para evaluar si la varianza de los datos es constante a lo largo de todas las variables independientes garantizando que los errores de las predicciones sean uniformes y no estén influenciados por la magnitud de los valores de las variables independientes (Ortiz, 2017). Esta prueba no

paramétrica es aplicada sin requerir suposición alguna de la distribución de los datos. Al respecto, la homocedasticidad implica “que la varianza es la misma para cualquier valor de la media”(Oddi et al., 2020, p.2). En esta prueba el contraste de hipótesis de igual de varianzas fue:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Donde:

σ_1^2 = varianza de la población en la zona geográfica urbana

σ_2^2 = varianza de la población en la zona geográfica rural

Para comprobar si existe diferencia significativa de los resultados obtenidos por los alumnos de la zona geográfica urbana y de la zona geográfica rural se empleó el test de Wilcoxon-Mann-Whitney o prueba U de Mann-Whitney. Este test “es una prueba no paramétrica aplicada a dos muestras independientes en donde no se tiene parámetros a estimar sino distribuciones a comparar”(Quispe et al., 2019, p.15). “Históricamente la comparación de medias basada en rangos tuvo dos desarrollos separados.

En 1945 Wilcoxon propuso el estadístico W (suma de rangos), y en 1947 Mann y Whitney propusieron el estadístico U”(Campos, 2018, p. 3). Además, “se demuestra que el estadístico U está relacionado algebraicamente con el estadístico W, siendo la prueba de Wilcoxon-Mann-Whitney un procedimiento que combina el uso de uso de ambos estadísticos” (Campos, 2018, p. 4). Este test es útil cuando los datos no cumplen los supuestos de normalidad o homocedasticidad, supuestos que si son necesarios cumplir para el empleo del test t de Student, por lo que en el presente artículo se empleó el test de Wilcoxon-Mann-Whitney para contrastar la igual de las medianas entre la zona geográfica urbana y la zona geográfica rural (Iglesias et al., 2023). La formulación de hipótesis para esta prueba fue:

$$H_0: \tilde{u}_1 = \tilde{u}_2$$

$$H_1: \tilde{u}_1 \neq \tilde{u}_2$$

Donde:

$\tilde{u}_1$ = mediana en la población en la zona geográfica urbana

$\tilde{u}_2$ = mediana en la población en la zona geográfica rural

Se acepta H_0 si $p\text{-valor} \geq \alpha$; se rechaza H_0 si $p\text{-valor} < \alpha$

El nivel de significación del contraste α , cuyo valor empleado en este estudio es de 0.05 , representa la más alta posible probabilidad de no aceptar H_0 cuando es cierta (Arias y Molina, 2017). El p-valor es la menor probabilidad con la cual se rechaza la hipótesis nula. Si el p-valor es mayor o igual que α no existe evidencia significativa para rechazar la hipótesis nula; y si el p-valor es menor a α se rechaza la hipótesis nula. Sin embargo, es importante tener en cuenta que el p-valor no indica la magnitud del efecto observado, sino simplemente la probabilidad de lograr un resultado similar o más extremo bajo la hipótesis nula, por lo tanto, un resultado estadísticamente significativo no siempre implica una diferencia significativa o relevante en términos prácticos (Martínez et al., 2017).

Con la finalidad de fundamentar la interpretación en el test anterior considerando el efecto de la diferencia entre los grupos de comparación (área urbana y área rural) se emplea el “estimador del tamaño

del efecto de la d de Cohen” (Iglesias et al., 2023, p.33). La magnitud del efecto (Rendón et al., 2021) se clasifica: no efecto o diferencia intrascendente (d =inferior a 0.19), diferencia pequeña (d =0.2-0.49), moderada (d =0.5-0.79), grande (d =0.8-1.29), y muy grande (d =superior a 1.29).

Para modelar la influencia del área de asentamiento de la institución educativa rural o urbana (zona geográfica) se empleó el método de Naive Bayes. Este método es “un algoritmo que se basa en el Teorema de Bayes que emplea una técnica de clasificación probabilística para asumir si alguna característica de la clase está relacionada o no respecto a la otra característica” (Shah, 2020, p.5). Naive Bayes “es un método probabilístico en el que una nueva observación es clasificada según estimaciones de las probabilidades condicionales asociadas a que pertenezca a una clase o a otra”(De La Parra et al., 2020, p.6).

Además, “el modelo Naive Bayes es un clasificador muy representativo y de gran simplicidad bajo el supuesto de independencia condicional entre cada par de atributos dado el valor de la clase” (Lorenzo et al., 2022, p. 74). El método de Naive Bayes “ha sido ampliamente investigado debido a su precisión sorprendentemente alta, en comparación con métodos de aprendizaje más sofisticados”(Frank et al., 2000, p. 6). Este método “se utiliza para suponer si alguna característica de la clase está relacionada o no, respectivamente, con la otra característica basándose en el Teorema de Bayes, por lo que también se conoce como clasificador probabilístico”(Shah, 2020, p.5).

Para estimar el desempeño método de Naive Bayes suele emplearse las métricas de precisión, sensibilidad, especificidad (De La Parra et al., 2020). Los elementos que intervienen en este modelo son “Verdaderos positivos (VP) = Casos positivos correctamente clasificados, Verdaderos negativos (VN) = Casos negativos correctamente clasificados, Falsos positivos (FP) = Casos negativos incorrectamente clasificados, Falsos negativos (FN) = Casos positivos incorrectamente clasificados” (Mosquera et al., 2018, p.157).

La precisión se obtiene de la siguiente manera: $\text{precisión} = \text{VP}/(\text{VP}+\text{FP})$. La sensibilidad o cobertura (fracción positiva verdadera) es la probabilidad de identificar correctamente a los estudiantes en una determinada área (rural o urbana) dado que el estudiante pertenece a dicha área (Mosquera et al., 2018) y se obtiene de la siguiente manera: $\text{sensibilidad} = \text{VP}/(\text{VP}+\text{FN})$.

La especificidad (fracción negativa verdadera) se define como la capacidad del modelo para identificar los casos negativos, y se calcula dividiendo el doble de la multiplicación de la precisión y cobertura para la suma de la precisión y cobertura ($\text{especificidad} = 2 * (\text{precisión} * \text{cobertura}) / (\text{precisión} + \text{cobertura})$) (Adolfo, 2020). En este estudio en particular se analizó las métricas de sensibilidad y especificidad.

3. RESULTADOS Y DISCUSIÓN

Los resultados de los descriptivos estadísticos de las variables cuantitativas de los alumnos que pertenecen al área geográfica urbana y al área geográfica rural fueron los siguientes (Tabla 1):

Tabla 1*Estadísticos descriptivos de la evaluación Ser Estudiante de tercero de bachillerato 2021-2022*

<i>Variable</i>	<i>Tipo de área</i>	<i>Media</i>	<i>Desviación estándar</i>	<i>Percentil 25</i>	<i>Mediana</i>	<i>Percentil 75</i>
<i>inev</i>	urbana	691.0	28.73	674.0	687.0	706.0
	rural	688.3	30.72	672.0	685.0	703.0
<i>imat</i>	urbana	694.0	23.77	684.0	694.0	700.0
	rural	697.7	29.76	687.0	694.0	706.0
<i>ilyl</i>	urbana	665.2	51.05	625.0	660.0	694.0
	rural	657.5	53.66	616.0	652.0	694.0
<i>ifis</i>	urbana	694.0	17.19	692.0	694.0	702.0
	rural	692.7	23.66	684.0	692.0	702.0
<i>iqui</i>	urbana	696.9	35.97	691.0	694.0	711.0
	rural	694.6	38.69	683.0	691.0	704.0
<i>ibio</i>	urbana	676.7	84.27	623.0	654.0	718.0
	rural	674.1	84.08	610.0	654.0	713.0
<i>ihis</i>	urbana	694.4	38.77	675.0	688.0	713.0
	rural	691.2	41.81	675.0	688.0	710.0
<i>ifil</i>	urbana	699.6	58.13	680.0	705.0	729.0
	rural	695.7	56.77	680.0	692.0	717.0
<i>ied</i>	urbana	706.5	46.63	683.0	709.0	730.0
	rural	703.0	48.61	683.0	703.0	730.0

Nota: *Elaboración propia con RStudio*

Análisis: Los resultados del promedio global (*inev*) con la información que proporciona la media se identifica que la media de la zona geográfica urbana fue mayor que la media de la zona geográfica rural. Así mismo, con la información de la desviación estándar de *inev* se evidencia que los datos del área urbana estuvieron menos dispersos que los valores del área rural. Además, con la información de los percentiles de *inev* se verificó que el 25% de los datos en el área urbana son menores de 674, mientras que en el área rural son menores de 672. Estos primeros resultados ratifican la incidencia de la zona geográfica, en general, se puede mencionar que el estudiantado de la zona rural obtiene resultados más bajos en comparación con los estudiantes de la zona urbana (Meza et al., 2021).

Los resultados de la aplicación del test de Kolmogórov-Smirnov (con la correlación de Lilliefors) empleado para verificar el criterio de normalidad fueron los siguientes (Tabla 2):

Tabla 2*Resultados del test de Kolmogorov-Smirnov*

<i>Variables</i>	<i>Tipo de área</i>	<i>p-valor</i>
<i>inev</i>	urbana	2.2×10^{-16}
	rural	2.2×10^{-16}
<i>imat</i>	urbana	2.2×10^{-16}
	rural	2.2×10^{-16}
<i>ilyl</i>	urbana	2.2×10^{-16}
	rural	2.2×10^{-16}
<i>ifis</i>	urbana	2.2×10^{-16}
	rural	2.2×10^{-16}
<i>iqui</i>	urbana	2.2×10^{-16}
	rural	2.2×10^{-16}
<i>ibio</i>	urbana	2.2×10^{-16}
	rural	2.2×10^{-16}
<i>ihis</i>	urbana	2.2×10^{-16}
	rural	2.2×10^{-16}
<i>ifil</i>	urbana	2.2×10^{-16}
	rural	2.2×10^{-16}
<i>ied</i>	urbana	2.2×10^{-16}
	rural	2.2×10^{-16}

Nota: *Elaboración propia con RStudio*

Análisis: Como el $p\text{-valor} = 2.2 \times 10^{-16}$ es menor a 0.05 se rechaza la hipótesis nula, por lo tanto, existe evidencia significativa de que los datos de la zona geográfica urbana y los datos de la zona geográfica rural no siguieron una distribución normal al 95% de confianza. Como los datos no siguen una distribución normal es un primer criterio para emplear posteriormente el test de Wilcoxon-Mann-Whitney con la finalidad de comparar si las medianas poblacionales del área urbana y del área rural son iguales (Iglesias et al., 2023).

A continuación, se verificó el criterio de homocedasticidad, el cual es un segundo criterio para emplear el test de Wilcoxon-Mann-Whitney. Los resultados que se obtuvieron al aplicar el test no paramétrico Fligner-Killeen para verificar el criterio de homocedasticidad se muestra a continuación (Tabla 3):

Tabla 3

Resultados del test de Fligner-Killeen

<i>Variables</i>	<i>Tipo de área</i>	<i>Fligner-Killeen</i>	<i>p-valor</i>
<i>inev</i>	urbana	0.068	0.794
	rural		
<i>imat</i>	urbana	19.428	0.000
	rural		
<i>ilyl</i>	urbana	0.584	0.445
	rural		
<i>ifis</i>	urbana	4.976	0.026
	rural		
<i>iqui</i>	urbana	2.359	0.125
	rural		
<i>ibio</i>	urbana	0.017	0.897
	rural		
<i>ihis</i>	urbana	4.155	0.042
	rural		
<i>ifil</i>	urbana	4.752	0.029
	rural		
<i>ied</i>	urbana	2.284	0.131
	rural		

Nota: *Elaboración propia con RStudio*

Análisis: Como $p\text{-valor} = 0.000$ (en imat), $p\text{-valor} = 0.026$ (en ifis), $p\text{-valor} = 0.042$ (en ihis) y $p\text{-valor} = 0.029$ (en ifil) son menores a 0.05 se rechaza la hipótesis nula, por lo tanto, existe evidencia significativa de que las varianzas del área geográfica urbana en imat, ifis, ihis e ifil son diferentes a sus respectivas varianzas del área rural al 95% de confianza. También, debido a que $p\text{-valor} = 0.794$ (en inev), $p\text{-valor} = 0.445$ (en ilyl), $p\text{-valor} = 0.125$ (en iqui), $p\text{-valor} = 0.897$ (en ibio) y $p\text{-valor} = 0.131$ (en ied) son mayores a 0.05 no existe evidencia significativa para rechazar la hipótesis nula, por lo tanto, las varianzas del área geográfica urbana en inev, ilyl, iqui, ibio e ied son iguales a sus respectivas varianzas del área geográfica rural al 95% de confianza.

Con base a los resultados obtenidos en las 2 pruebas anteriores, con la finalidad de comparar las medianas poblacionales de dos muestras independientes (área geográfica urbana y área geográfica rural) se aplicó el test de Wilcoxon-Mann-Whitney (W). Además, se presentan los resultados que se obtuvieron de los estimadores “del tamaño del efecto d de Cohen” (Iglesias et al., 2023, p.33). (Tabla 4).

Tabla 4*Resultados del test de Wilcoxon-Mann-Whitney y efecto d de Cohen*

<i>Variables</i>	<i>Tipo de área</i>	<i>Mediana</i>	<i>W</i>	<i>p-valor</i>	<i>d de Cohen</i>
<i>Inev</i>	urbana	687.0	2556776	0.002	0.091
	rural	685.0			
<i>Imat</i>	urbana	694.0	2225983	0.000	-0.109
	rural	694.0			
<i>Ilyl</i>	urbana	660.0	2646807	0.000	0.148
	rural	652.0			
<i>Ifis</i>	urbana	694.0	2472837	0.230	0.066
	rural	692.0			
<i>Iqui</i>	urbana	694.0	2456914	0.410	0.061
	rural	691.0			
<i>Ibio</i>	urbana	654.0	2447096	0.555	0.030
	rural	654.0			
<i>Ihis</i>	urbana	688.0	2529134	0.012	0.078
	rural	688.0			
<i>Ifil</i>	urbana	705.0	2571220	0.000	0.067
	rural	692.0			
<i>Ied</i>	urbana	709.0	2535573	0.007	0.073
	rural	703.0			

Nota: *Elaboración propia con RStudio*

Análisis: Como p-valor = 0.002 (en inev), p-valor = 0.000 (en imat), p-valor = 0.000 (en ilyl), p-valor = 0.012 (en ihis), p-valor = 0.000 (en ifil), y p-valor = 0.007 (en ied) son menores a 0.05 se rechaza la hipótesis nula, por lo tanto, existe evidencia significativa de que las medianas de inev, imat, ilyl, ihis, ifil e ied en el área urbana son diferentes a sus respectivas medianas en el área rural al 95% de confianza. Así mismo, debido a que p-valor = 0.23 (en ifis), p-valor = 0.4102 (en iqui) y p-valor = 0.5551 (en ibio) son mayores a 0.05 no existe evidencia significativa para rechazar la hipótesis nula, por lo tanto, las medianas de ifis, iqui e ibio en el área urbana son iguales a sus respectivas medianas el área rural al 95% de confianza. Además, como todos los valores de d de Cohen son inferiores a 0.19 se concluye que no existe efecto, es decir, la diferencia entre los grupos de comparación (área urbana y área rural) es intrascendente, lo que implica que no existen errores de interpretación de los resultados en base al p-valor de la prueba no paramétrica W.

Estos hallazgos de igualdad de medianas en Física, Química y Biología del área urbana y rural contradice en parte a lo que se ha mencionado históricamente que los estudiantes del área rural obtienen resultados inferiores que los estudiantes del área urbana (INEVAL, 2019), lo que representa una esperanza de mejora de la calidad educativa en la zona rural en estas 3 asignaturas, sin embargo, en el promedio global, Matemática, Lengua y Literatura, Historia, Filosofía y Educación para la Ciudadanía aún persiste mejores resultados en el área urbana. En este sentido, investigaciones han mencionado que el área de residencia rural o urbana es un predictor que puede acentuar la desigualdad en el aprendizaje de los estudiantes y limitar el acceso a oportunidades (Santillán y Vargas, 2020).

La aplicación del modelo de Naive Bayes para la decisión de clasificación de si pertenece al área urbana o al área rural se obtuvo los siguientes 6 primeros resultados (Tabla 5).

Tabla 5*Resultados del modelo de Naive Bayes para una decisión de clasificación*

Umbral	Precisión	Sensibilidad	Especificidad
0.05	42.38560	100.00000	59.536354
0.10	42.85714	99.292035	59.871932
0.15	44.18403	90.088496	59.289458
0.20	47.70241	77.168142	58.958756
0.25	59.23710	63.539823	56.848773
0.30	55.57621	52.920354	54.215775

Nota: *Elaboración propia con RStudio*

Análisis: De estos resultados se seleccionó el umbral 0.10 debido a que este umbral maximiza la especificidad, es decir, la especificidad crece a medida que los umbrales aumentan pero llega un punto (umbral 0.10) en que empieza a decaer (Adolfo, 2020). Con la selección del umbral 0.10 se realizó la matriz de confusión (Tabla 6).

Tabla 6*Matriz de confusión del modelo Naive Bayes*

Tipo de área	Falso	Verdadero
urbana	24	748
rural	4	561

Nota: *Elaboración propia con RStudio*

Análisis: De esta matriz se obtienen los datos que permiten clasificar los estudiantes que pertenecen al área urbana: VP = 561; VN = 24; FP = 748; FN = 4. Con estos resultados se obtiene una sensibilidad de 99.29 y una especificidad de 59.87. Con la sensibilidad de 99.29 y una especificidad de 59.87 indica que el modelo de Naive Bayes tiene un desempeño aceptable para modelar la influencia del área de asentamiento rural o urbana de la institución educativa (Adolfo, 2020).

Con base a los resultados obtenidos se concluye que mediante los estadísticos descriptivos (Tabla 1) se evidencia los primeros indicios de influencia de la zona geográfica en favor de los estudiantes del área urbana. Por otra parte, con la prueba de Kolmogórov-Smirnov se demuestra que existe evidencia significativa de que los datos del área geográfica urbana y los datos del área geográfica rural no cumplen con el criterio de normalidad.

Además, a través de la prueba de Fligner-Killeen se demuestra que existe evidencia significativa de que los promedios obtenidos en Matemática, Física, Historia y Filosofía no cumplen con el criterio de homocedasticidad. También, con esta misma prueba se concluye que existe evidencia significativa de que el promedio global obtenido por el estudiante y los promedios obtenidos en Lengua y Literatura, Química y Biología cumplen con el criterio de homocedasticidad.

Asimismo, con la prueba de Wilcoxon-Mann-Whitney se evidencia que las medianas del promedio global obtenido por el estudiante y las medianas de los promedios obtenidos en las asignaturas Matemática, Lengua y Literatura, Historia, Filosofía y Educación para la Ciudadanía en el área urbana son diferentes a sus respectivas medianas en el área rural. Igualmente, con esta prueba se revela que existe evidencia significativa de que las medianas de los promedios obtenidos en las asignaturas Física, Química y Biología del área urbana son iguales a sus respectivas medianas del área rural.

Finalmente, con los resultados del modelo de Naive Bayes se concluye que este modelo puede ser empleando en investigaciones similares con efectos positivos de modelamiento que complementa los cálculos estadísticos descriptivos y las pruebas de hipótesis empleadas en el presente estudio para evidenciar la influencia de la zona geográfica en el rendimiento escolar.

4. CONCLUSIONES

Existe evidencia significativa de que las medianas del promedio global obtenido por el estudiante y las medianas de los promedios obtenidos en las asignaturas Matemática, Lengua y Literatura, Historia, Filosofía y Educación para la Ciudadanía en el área urbana son diferentes a sus respectivas medianas en el área rural al 95% de confianza empleando la prueba de Wilcoxon-Mann-Whitney. También, con esta prueba se revela que existe evidencia significativa de que las medianas de los promedios obtenidos en las asignaturas Física, Química y Biología del área urbana son iguales a sus respectivas medianas del área rural al 95% de confianza. Además, utilizando d de Cohen se concluye que no existen errores de interpretación de los resultados en base al p -valor en la prueba de Wilcoxon-Mann-Whitney.

Se evidencia que Naive Bayes tiene un desempeño aceptable de sensibilidad y especificidad para modelar y verificar que el área geográfica del centro educativo influye en los resultados de la evaluación Ser Estudiante de tercero de bachillerato 2021-2022. Sin embargo, cabe mencionar que la presente investigación tiene una limitación debido a que la zona geográfica es uno más de los factores que influyen en el rendimiento escolar, ya que la educación es un proceso complejo que está relacionado con una variedad de factores influyentes como los antecedentes familiares, la motivación del estudiante, la calidad de los maestros, entre otros, y, además, cada estudiante es único y puede responder de manera diversa a diferentes entornos educativos.

En definitiva, la zona geográfica es uno de los factores que puede tener un impacto significativo en el rendimiento escolar de un estudiante, por lo que es trascendental que los educadores, los padres de familia y demás actores educativos trabajen juntos para garantizar el acceso a recursos educativos y a una educación de calidad, independientemente de la zona geográfica de donde vivan.

FINANCIACIÓN

Los autores no recibieron financiación para el desarrollo de la presente investigación.

CONFLICTO DE INTERESES

Los Autores declaran que no existe conflicto de intereses

RECONOCIMIENTO A REVISORES:

La revista reconoce el tiempo y esfuerzo del editor Raúl Lozada Yanez, y de revisores anónimos que dedicaron su tiempo y esfuerzo en la evaluación y mejoramiento del presente artículo.

CONTRIBUCIÓN DE AUTORÍA

En concordancia con la taxonomía establecida internacionalmente para la asignación de créditos a autores de artículos científicos (<https://credit.niso.org/>). Los autores declaran sus contribuciones en la siguiente matriz:

Participar activamente en:	Suárez M.	Sánchez N.
Conceptualización	X	X
Análisis formal	X	X
Adquisición de fondos	X	X
Investigación	X	X
Metodología	X	X
Administración del proyecto	X	X
Recursos	X	X
Redacción –borrador original	X	X
Redacción –revisión y edición	X	X
La discusión de los resultados	X	X
Revisión y aprobación de la versión final del trabajo.	X	X

REFERENCIAS

- Adolfo, S. B. (2020). *Regresión logística y Nāive Bayes. Detección de churn bancario*. [https://rpubs.com/AdSan-R/Rlog_NB_ChurnBank#:~:text=AUC \(Area Under The Curve,0.5%2C no tiene capacidad discriminativa.](https://rpubs.com/AdSan-R/Rlog_NB_ChurnBank#:~:text=AUC (Area Under The Curve,0.5%2C no tiene capacidad discriminativa.)
- Aliu, A., Rexhepi, S., & Iseni, E. (2021). Analysis and comparison of commitment, homework, extra hours, preliminary grades and testing of students in Mathematics using linear regression model. *Mathematics Teaching-Research Journal*, 13(3), 21–52.
- Arias, M. M., & Molina, M. (2017). Lectura crítica en pequeñas dosis ¿Qué significa realmente el valor de p? *Rev Pediatr Aten Primaria*, 19, 377–381.
- Bases de datos Ser Estudiante – Banco de Información*. (n.d.). Retrieved April 6, 2023, from <http://evaluaciones.evaluacion.gob.ec/BI/bases-de-datos-ser-estudiante/>
- Borja, M., Martínez, J., Barreno, S., & Haro, O. (2021). Factores asociados al rendimiento académico: Un estudio de caso factors associated with academic performance: a case study. *Revista Educare*, 25(3), 54–68. <https://revistas.investigacion-upelipb.com/index.php/educare/article/view/1509/1516>
- Campos, B. (2018). Noparametrica-apuntes1819. *Bioestadística Básica*, 1, 1–12. <http://diposit.ub.edu/dspace/bitstream/2445/127108/1/Noparametrica-apuntes1819.pdf>
- Castro Aristizabal, G., Castillo Caicedo, M., & Mendoza Parra, J. C. (2018). Factores asociados a la adquisición de competencias en América Latina. *Revista de Ciencias Sociales*, 23(4), 33–52. <https://doi.org/10.31876/rcs.v23i4.25136>
- Constitutivos, E., & Estado, D. E. L. (2021). *Página 1 de 228*. 1–228.
- De La Parra-González, I., Luna-Rosas, F. J., Rodríguez-Martínez, L. C., & Frausto-Reyes, C. (2020). Evaluación de la regresión logística como clasificador de espectros Raman en el diagnóstico automático de cáncer de mama. *Revista de Ingeniería Biomédica y Biotecnología*, 4(12), 1–12. <https://doi.org/10.35429/jbeb.2020.12.4.1.12>
- Estadística-con-R/12_T-test.pdf en master · JoaquinAmatRodrigo/Estadística-con-R · GitHub*. (n.d.). Retrieved March 31, 2023, from https://github.com/JoaquinAmatRodrigo/Estadística-con-R/blob/master/PDF_format/12_T-test.pdf
- Flores, C., & Flores, K. (2021). Tests To Verify the Normality of Data in Production Processes : Anderson-Darling , Ryan-Joiner , Shapiro-Wilk and. *Societas Revistas*, 23(2), 83–97. <http://portal.amelica.org/ameli/jatsRepo/341/3412237018/index.html>
- Frank, E., Trigg, L., Holmes, G., & Witten, I. H. (2000a). Naive Bayes for Regression. *Machine Learning*, 41(1), 5–25. <http://www.cs.waikato.ac.nz/~eibe/pubs/nbr.pdf>

- Frank, E., Trigg, L., Holmes, G., & Witten, I. H. (2000b). Technical note: Naive Bayes for regression. *Machine Learning*, 41(1), 5–25. <https://doi.org/10.1023/A:1007670802811>
- Garofalo García, R., & Villao Villacrés, F. (2018). Crisis De La Escuela Rural, Una Realidad Silenciada Y Su Lucha Para Seguir Adelante [Rural School Crisis, A Silenced Reality, And Its Struggle To Move On]. *Conrado*, 14(62), 152–157. <http://scielo.sld.cu/pdf/rc/v14n62/rc266218.pdf>
- Hernández, R., & Mendoza, C. (2018). Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta. In *Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta*. http://www.biblioteca.cij.gob.mx/Archivos/Materiales_de_consulta/Drogas_de_Abuso/Articulos/SampieriLasRutas.pdf
- Iglesias Rodríguez, A., Martín González, Y., & Hernández Martín, A. (2023). Evaluación de la competencia digital del alumnado de Educación Primaria. *Revista de Investigación Educativa*, 41(1), 33–50. <https://doi.org/10.6018/rie.520091>
- INEVAL. (2019a). *Informe de resultados Ser Bachiller lectivo 2018-2019*. <https://cloud.evaluacion.gob.ec/dagireportes/sbciclo19/distrito/03D02.pdf>
- INEVAL. (2019b). La educación en Ecuador: Logros alcanzados y nuevos desafíos Resultados educativos 2017-2018. *Ministerio de Educación*, 212. <https://www.evaluacion.gob.ec>
- INEVAL. (2022a). *Dirección de Análisis, Aplicación y Cobertura Territorial SEST 2022*. http://evaluaciones.evaluacion.gob.ec/archivosPD/SEST/SEST22_diseno_muestral_20220716.pdf
- INEVAL. (2022b). *Ser Estudiante 2022*. 1–23. <http://evaluaciones.evaluacion.gob.ec/BI/folleto-ser-estudiante-2022/>
- INEVAL. (2023). *Informe Nacional de Resultados Ser Estudiante del nivel de Bachillerato. Año lectivo 2021-2022*. 86. www.evaluacion.gob.ec
- Lorenzo, S., Lorenzo, S., Lorenzo, S., & Lorenzo, S. (2022). *Evaluación de exactitud de Naive Bayes y Regresión Logística para clasificación con atributos y clases binarios*. 13(1), 73–84.
- Lugo-Armenta, J. G., & Pino-Fan, L. R. (2021). Inferential Reasoning Levels for t-Student Statistical. *Bolema - Mathematics Education Bulletin*, 35(71), 1776–1802. <https://doi.org/10.1590/1980-4415V35N71A25>
- Luis, J., & Gonzáles, A. (2020). Formulación de los objetivos específicos desde el alcance correlacional en trabajos de investigación. *Ciencia Latina Revista Científica Multidisciplinar*, 2215(2), 237–247. https://doi.org/10.37811/cl_rcm.v4i2.73
- Martínez-Ezquerro, J. D., Riojas-Garza, A., & Rendón-Macías, M. E. (2017). Clinical significance vs statistical significance. How to interpret the confidence interval at 95 % [Significancia clínica sobre significancia estadística. Cómo interpretar los intervalos de confianza a 95 %]. *Revista Alergia Mexico*, 64(4), 477–486. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054126316&doi=10.29262%2Fram.v64i4.334&partnerID=40&md5=22afeaf39a1b2774be1ec57985ab9496>
- Mata, M. (2016). *Kolmogorov-Smirnov en R (Lilliefors) para testear la normalidad*. Viva El Software Libre. <https://vivaelssoftwarelibre.com/test-de-kolmogorov-smirnov-en-r/>
- Medina, M. A. G., & Villarreal, D. C. T. (2018). Logro educativo y factores asociados en estudiantes de sexto grado de educación primaria en el estado de Nuevo León, México. *Perfiles Educativos*, 40(159), 107–125. <https://doi.org/10.22201/issue.24486167e.2018.159.58143>

- Meza-Cascante, L. G., Suárez-Valdés-Ayala, Z., Agüero-Calvo, E., Jiménez-Céspedes, R., Calderón-Ferrey, M., Sancho-Martínez, L., Pérez-Tyteca, P., & Monje-Parrilla, J. (2021). La matemática como dominio masculino: Un estudio de la percepción en la educación media costarricense. *Revista Electrónica Educare*, 25(3), 1–15. <https://doi.org/10.15359/ree.25-3.35>
- Mosquera, R., Castrillón, O. D., & Parra, L. (2018). Máquinas de Soporte Vectorial, Clasificador Naïve Bayes y Algoritmos Genéticos para la Predicción de Riesgos Psicosociales en Docentes de Colegios Públicos Colombianos. *Información Tecnológica*, 29(6), 153–162. <https://doi.org/10.4067/s0718-07642018000600153>
- Oddi, F. J., Miguez, F. E., Benedetti, G. G., & Garibaldi, L. A. (2020). Cuando la variabilidad varía: Heterocedasticidad y funciones de varianza. *Ecología Austral*, 30(3), 438–453. <https://doi.org/10.25260/ea.20.30.3.0.1131>
- Paola, A., Palomo, T., Roberto, E., & Núñez, V. (2020). Análisis de resultados del examen ser bachiller en el dominio matemático. *Revista Cognosis*. ISSN 2588-0578, 5(2), 13–32. <https://doi.org/10.33936/COGNOSIS.V5I2.2282>
- Pérez-Obregón, J. M., & Romero-Díaz, T. (2018). Análisis del rendimiento académico mediante regresión logística y múltiple. *Revista Electrónica de Conocimientos, Saberes y Prácticas*, 1(2), 33–42. <https://doi.org/10.30698/recsp.v1i2.10>
- Pozo-Burgos, E. J., Burbano-Pulles, M. R., Vidal-Chica, J. I., & Revelo-Salgado, G. E. (2022). Sociocultural and Demographic Factors That Influence Academic Performance: the Pre-University Case of the Universidad Politécnica Estatal Del Carchi. *Journal of Technology and Science Education*, 12(1), 147–156. <https://doi.org/10.3926/JOTSE.1359>
- Quispe Andía, A., Calla Vasquez, K. M., Yangali Vicente, J. S., Rodríguez López, J. L., & Pumacayo Palomino, I. I. (2019). Estadística no paramétrica aplicada a la investigación científica. In *Resultado de Investigación*.
- Rendón-Macías, M. E., Zarco-Villavicencio, I. S., & Villasís-Keever, M. Á. (2021). Statistical methods for effect size analysis. *Revista Alergia Mexico*, 68(2), 128–136. <https://doi.org/10.29262/ram.v65i2.949>
- Rodríguez Rodríguez, D., Remedios, |, & Rosquete, G. (2019). Rendimiento académico y factores sociofamiliares de riesgo Variables personales que moderan su influencia. *Perfiles Educativos* |, XLI. <https://doi.org/10.22201/iisue.24486167e.2019.164.58925>
- RPubs - EBA 14 Homogeneidad de varianzas u homocedasticidad. (n.d.). Retrieved March 31, 2023, from <https://rpubs.com/Rortizdu/333014>
- Santillán Hernández, A. S., & Vargas Sánchez, J. R. (2020). Diferencia regional de factores que afectan el desempeño escolar para la misma modalidad educativa. *Paradigma Económico*, 12(1), 5. <https://doi.org/10.36677/paradigmaeconomico.v12i1.14788>
- Shah, H. B. (2020). Comparing Machine Learning Algorithms for Credit Card Fraud Detection. 2020. <http://broncoscholar.library.cpp.edu/handle/10211.3/216678>
- UNESCO. (2021). *Los Aprendizajes fundamentales en América Latina y el Caribe. Evaluación de logros de los estudiantes. Estudio Regional Comparativo y Explicativo (ERCE 2019)*. Erce, 77. <https://unesdoc.unesco.org/ark:/48223/pf0000380257>