PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS SOCIALES

TRABAJO FINAL – GRUPO 8

Determinantes de la deserción escolar en la Educación Básica Regular en el Perú durante el 2021

Milagros Alejandra Tolentino Curo (Código: 20196150)

Jesús Alberto Nicho Rosado (Código: 20176035)

Regina Victoria Huerta Rojas (Código: 20197116)

Diego Arturo Camargo Pacheco (Código: 20160256)

Curso

Laboratorio de programación: R y Python

Docente

Mauricio Vallejos

Lima, 2023

Objetivos

La investigación busca abordar de manera analítica la problemática de la deserción escolar en la Educación Básica Regular (EBR) en Perú, en los años 2020-2021. El objetivo principal es analizar cómo las características sociodemográficas y socioeconómicas están asociadas con la deserción. Es importante hacer énfasis en los años 2020-2021, ya que podrían proporcionarnos resultados más diferenciados debido a las medidas tomadas por el gobierno, como la cuarentena. Hemos definido tres objetivos específicos: realizar un análisis de las diferencias existentes entre diversos grupos de estudiantes, determinar la probabilidad de deserción, centrándose en ciertas características específicas de los estudiantes en la EBR durante 2020-2021. Es importante abordar estos efectos en la deserción en la EBR para implementar políticas que ayuden a reducir las brechas de acceso, promoviendo así la calidad educativa y, en consecuencia, mejorando la calidad de vida a largo plazo.

Formulación del problema

La deserción escolar es uno de los problemas más grandes que enfrenta el Perú no solo por el acceso limitado a los centros educativos sino que también porque ello se relaciona con otros problemas ya enraizados como la pobreza y la desigualdad. En el año 2020, la pandemia golpeó fuertemente al Perú, para contrarrestar la propagación del virus se tomaron en consideración la implementación de ciertas medidas como la cuarentena obligatoria. Ante esas medidas, el retorno a clases fue de manera virtual, lo que acaba acentuando más el problema la decisión de asistir a un centro educativo ya que no todos los centros educativos y las familias estaban tecnológicamente preparadas.

Por otra parte, si bien el gobierno tomó medidas para ampliar el acceso tecnológico de aquellas familias o estudiantes que no contaban con los medios para acceder a sus clases virtuales, estas no fueron suficientes para incluir a todos. “En el Perú solo el 34 % de los hogares contaba con computadora y solo el 28 % tenía acceso a internet” (INEI, 2022). En consecuencia, el impedimento del acceso ya sea por distancia o por falta de recursos tecnológicos o económicos, se convierte en una problemática crucial , dado que la educación es un factor importante para el desarrollo profesional, lo que se convierte a largo plazo en formación de capital humano lo que es fundamental para la obtención de mejores oportunidades laborales y con ello, mejores ingresos .

Revisión de literatura

La deserción escolar no solo se atribuye a las características individuales y familiares, como la edad, el número de hijos o los ingresos; también se ve influida por factores externos, como la ubicación geográfica y

el tipo de institución educativa a la que asisten los estudiantes. En esta revisión de literatura, se explorarán los elementos que han tenido un impacto significativo en la deserción escolar, abarcando tanto las características individuales y familiares como los factores externos.

En el contexto colombiano, se ha identificado una correlación entre la distancia entre el hogar y la escuela y la tasa de deserción escolar (Drysdale, 1972). Este hallazgo concuerda con investigaciones que señalan que la deserción es más pronunciada en áreas rurales, donde la lejanía entre el lugar de residencia y la institución educativa puede aumentar la probabilidad de que los estudiantes abandonen la escuela. Santos (2009) sugiere que pertenecer a un área rural disminuye la probabilidad de asistir a una institución educativa, aunque los efectos no son altamente significativos, lo que sugiere que la tasa de deserción puede estar asociada a otras variables explicativas, como los ingresos familiares.

Por otro lado, la literatura respalda la noción de que existe una disyuntiva entre trabajar y estudiar, influida por las condiciones socioeconómicas del hogar. En Ecuador, se ha observado que los hogares con bajos ingresos pueden estar impulsando a sus hijos en edad de estudiar a trabajar, lo que resulta en la renuncia a la asistencia a las escuelas (Carrión, 2014). Sin embargo, otros estudios argumentan que el trabajo y el estudio no son necesariamente causas directas de la deserción escolar, ya que los adolescentes tienden a gestionar horarios compatibles. Este efecto podría variar según la edad del estudiante (Peña y Soto, 2016).

Adicionalmente, la literatura sugiere la existencia de variables como el número de hermanos, la presencia de hermanos menores en el hogar y el rendimiento académico como posibles influencias en la deserción escolar (Cerrutti, 2004). También se han identificado efectos asociados a las regiones donde predomina el habla vernácula, como en Perú, donde la escasez de colegios en estas áreas dificulta la asistencia de los estudiantes a las escuelas (Rodriguez, 2011).

Metodología y bases de datos

En este trabajo se utilizará la Encuesta Nacional de Hogares (ENAHO) con corte transversal para el año 2021. Esta encuesta se encuentra disponible en las bases de datos del INEI. Se emplea principalmente la integración del módulo 300 de Educación y el módulo 500 de Empleo e Ingresos. Del módulo 300 de Educación, se tomará en consideración los ingresos, edad, lengua materna, área geográfica, programas de financiamiento, recursos disponibles, conectividad a internet y el tipo de colegio al que asisten los encuestados. Resulta importante incorporar el año 2020 para el análisis ya que es un contexto educativo de virtualidad y por ello nos ayudará a comprender su impacto sobre la deserción a través de las variables ya mencionadas, especialmente

el de acceso a conectividad.

Asimismo, del módulo 500 de Empleo e Ingresos, se extrajo datos sobre el tipo de empleo que los padres tienen y cuál sería su situación laboral. Estas variables ayudan a complementar las características socioeconómicas.Para completar la base de datos también las bases de datos de Sumaria 2021, el módulo de Salud y el módulo del Hogar de la Enaho que nos aportarán variables las cuales incluiremos en el análisis para poder hacer una mejor estimación.

El presente estudio llevará un análisis econométrico para examinar los determinantes que pueden afectar la decisión de matricularse en un centro educativo en Perú durante el año 2019. Para explorar la relación entre ambas variables, se implementará el modelo logístico. Este modelo tomará como variable dicotómica, dependiente y nominal si el estudiante desertor, es decir, si se matriculó en el año 2020 pero no en el año 2021, o si el estudiante se matriculó en el año 2020 y 2021. Este modelo determina cómo es que la probabilidad de decisión de matrícula cambia, en función de las características socioeconómicas y sociodemográficas.Adicionalmente, se añadieron otras variables de control, como conectividad, programas de financiación, entre otros ,con el propósito de añadir validez al modelo. La ecuación planteada para el modelo econométrico es la siguiente:

$$P(Matrícula = 1/X) = \beta_0 + \beta_1 ingreso\ familiar + \beta_2 edad + \beta_3 sexo + \beta_4 área\ geo +$$
$$\beta_5 tipo\ de\ colegio + \beta_6 programas + \beta_7 internet + \beta_8 tv + \beta_9 celular + \beta_{10} situación\ laboral +$$
$$\beta_{11} vacunas\ covid + \beta_{12} educ\ madre + \beta_{13} tamaño\ familiar + \beta_{14} lengua\ materna + \mu$$

Figura 1: Ecuación del modelo

Como se mencionó anteriormente, se tomará en consideración sólo aquellos alumnos que pueden tomar la decisión de matricularse al siguiente año, es decir se está descartando a aquellos estudiantes que cursan el quinto año de secundaria ya que estarían culminando su etapa escolar. Para poder identificar al grupo descartado, se tomaron en cuenta las 3 variables siguientes: primero, estudiantes que estuvieron en secundaria en el año 2020 (p304a_21=3), segundo, alumnos que estuvieron en quinto de secundaria en el 2020 (p304b_21=5) y, tercero, de este grupo, aquellos que aprobaron en el 2020 (p305_21=1).

Conclusiones/Resultados

Antes de interpretar la relación entre las variables independientes sobre la deserción escolar, es importante establecer la correlación que puede haber entre variables, es por ello que realizamos un análisis de multicolinealidad sujeto a la matriz de correlación. Se encontró correlaciones positivas entre las variables de situación

laboral-edad con 0.47 , situación laboral-edad con 0.47, años de educación de la madre-años de educación del padre con 0.63, área geográfica-internet con 0.38. Por otro lado, encontramos correlaciones negativas relativamente altas como tipo de colegio-ingresos con -0.34, programas sociales-ingreso familiar con -0.21 y tipo de colegio-ingreso familiar con -0.26. Consideramos que la correlación entre educación del padre y de la madre es muy alta. Con el fin de prevenir posibles complicaciones relacionadas con la multicolinealidad, optamos por retener únicamente la variable que representa la educación de la madre, descartando la variable correspondiente al padre.

Después, realizamos una regresión logit simple para observar el efecto positivo o negativo que podrían tener las variables independientes sobre la deserción escolar. Asimismo, buscamos observar la significancia de cada uno de los efectos.En este contexto, se puede observar de primera instancia que el pseudo $R\hat{2}$, que indica la la capacidad explicativa de las variables independientes sobre Y es bastante pequeño, lo que podría significar que no se está logrando capturar todo el efecto a través de las variables independientes. Encontramos que las variables edad, tipo de colegio y situación laboral, tienen un impacto positivo y significativo sobre la probabilidad de matricularse el siguiente año.

Asimismo, las variables lengua materna, si tiene TV y área geográfica muestran un efecto positivo en la probabilidad de matricularse; sin embargo, no son estadísticamente significativas. Por otro lado, identificamos que las variables de si la persona está incluida en algún programa social, el estar vacunado contra el covid y el acceso a internet tienen un efecto negativo y significativo sobre la probabilidad de matricularse el año 2021. Del mismo modo, se identifico que el sexo, los años de educacion de la madre, los ingresos y si tiene celular influyen negativamente pero no son estadisticamente significativos.

1. Análisis de gráficos

La figura 1 muestra la relación entre la deserción escolar y la edad. Así, la deserción escolar es igual a cero si el estudiante se matriculó en el año t+1, luego de haber estudiado en el año t. Es necesario que el estudiante haya cursado al menos un año de estudios para ser considerado en deserción escolar. Por otro lado, si la deserción escolar es igual a 1, consideramos que el estudiante, con las características anteriores, no se matriculó. En esa medida, el gráfico 1 revela que hay un mayor porcentaje de estudiantes de alrededor de 12 años que no se matriculó en el nivel educativo que le corresponde. Esto se puede deber a que los estudiantes que están por terminar la secundaria optan por trabajar, influenciado por su situación socioeconómica o sus preferencias. En esa medida, como la edad en promedio de deserción es convergente con el inicio de la el nivel secundario de educación básica, es necesario una base de apoyo emocional u orientaciones para reducir

la deserción en este grupo de edad.

Por otro lado, en la Figura 2, se puede observar que el efecto de las vacunas por COVID-19 fue positivo en la tasa de matrícula de los estudiantes de educación básica regular con respecto a no haber recibido dicha vacuna. Esto se puede deber a que la vacunación aumenta la confianza de los padres por lo que se sentirán más seguros enviando a los estudiantes a entornos educativos presenciales. Asimismo, la vacunación puede contribuir a estabilizar la situación, permitiendo una continuidad más efectiva de la educación presencial y reduciendo la posibilidad de interrupciones académicas significativas.

Asimismo, la Figura 3 revela que el área geográfica que presenta una mayor tasa de deserción es el área rural sobre el área urbano. Sin embargo, las diferencias no son tan pronunciadas, lo que se puede deber a la baja población en las zonas rurales con respecto a las zonas urbanas. Por otro lado, la alta tasa de deserción se puede deber a la distancia entre las instituciones educativas y los hogares de los estudiantes. Además, la enseñanza en los centros educativos, según la literatura, no se logran adaptar a las lenguas nativas en el método de enseñanza, lo cual desmotiva a los estudiantes a asistir a las clases.

Finalmente, la Figura 4 muestra que mientras más nivel educativo tenga la madre, es posible que haya menos probabilidad de deserción escolar. Esto se puede deber a una mayor influencia de la madre en el estudiante, quien es tomada como modelo de referencia por la cercanía que tiene con sus hijos y su poder de decisión en las actividades diarias del hogar.

2. Limitaciones

Como mencionamos anteriormente, una de las limitaciones que encontramos fue el valor del pseudo $R\hat{2}$ ya que es 0.1313 lo que según la teoría, es relativamente pequeño lo cual puede sugerir que el modelo puede no estar capturando el efecto para la variable independiente. Asimismo otra observación es que la variable *edad* no se limita solo a aquellos menores a 18 años sino que incluye valores atípicos como personas hasta mayores de 30 años lo que puede significar que hayan otro tipo de caracteristicas o variables que los esten afectando.Por otro lado, encontramos que, a pesar de que el modelo predice correctamente la mayoria de variables, podria no estar capturando correctamente el efecto de otras variables explicativas. Otro de los principales desafíos de este modelo es su poder predictivo. A través de la matriz de confusión del cuadro 3 se puede verificar que el modelo tiene una precisión del 85.44 %, lo que indica la proporción de predicciones correctas en el conjunto de datos. Pero su rendimiento en la clasificación de la clase positiva (deserción escolar) es limitado, con bajo recall y F1-score según el cuadro 4. Esto sugiere que el modelo puede estar presentando dificultades al momento de identificar adecuadamente a los casos de deserción escolar.

Anexos

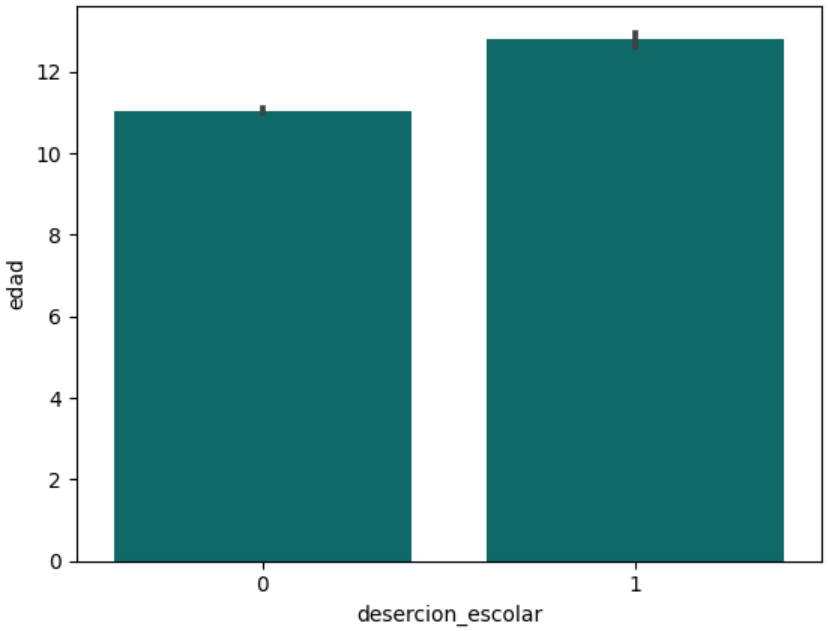

Figura 2: **Distribución de la deserción escolar segun la edad** Fuente: *ENAHO (2021)*

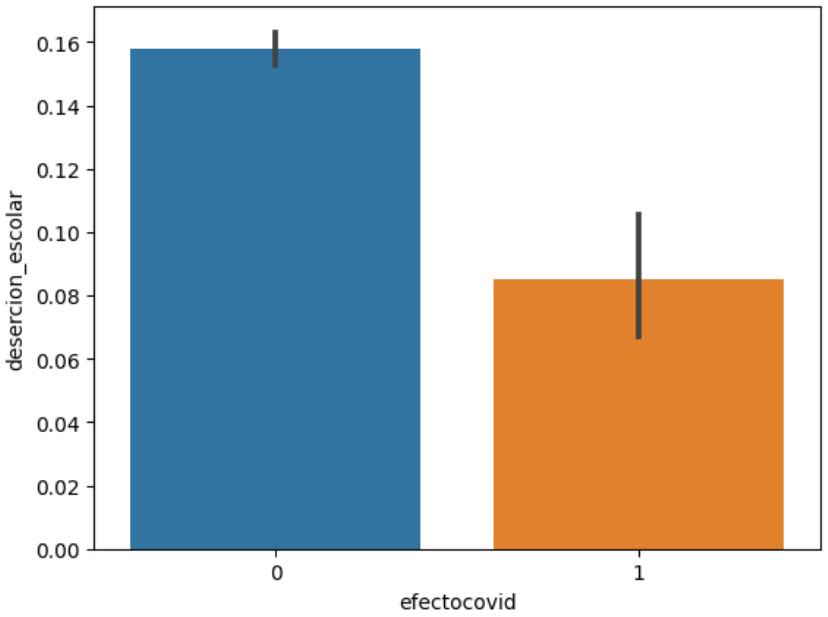

Figura 3: **Distribución de la vacunación covid sobre la deserción escolar** Fuente: *ENAHO (2021)*

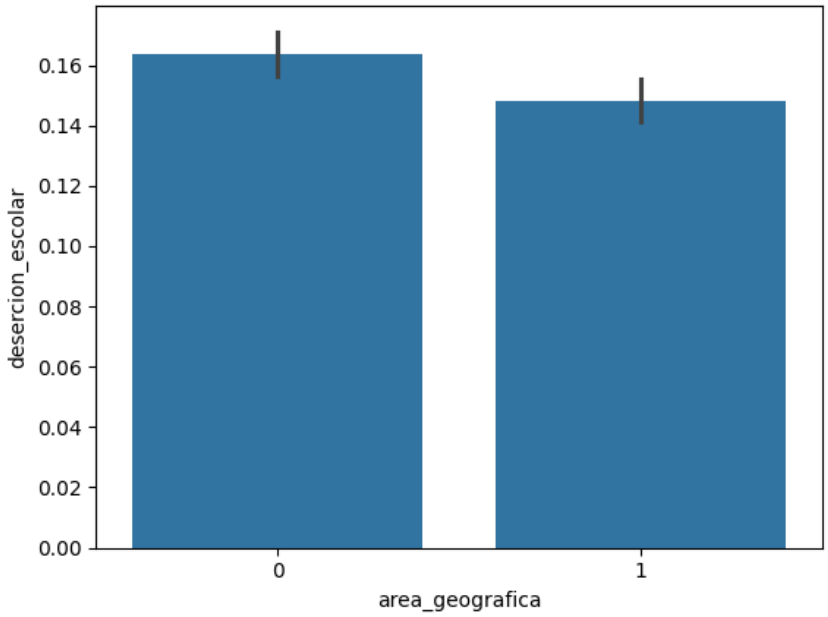

Figura 4: **Deserción escolar segun área geografica:rural o urbano** Fuente: *ENAHO (2021)*

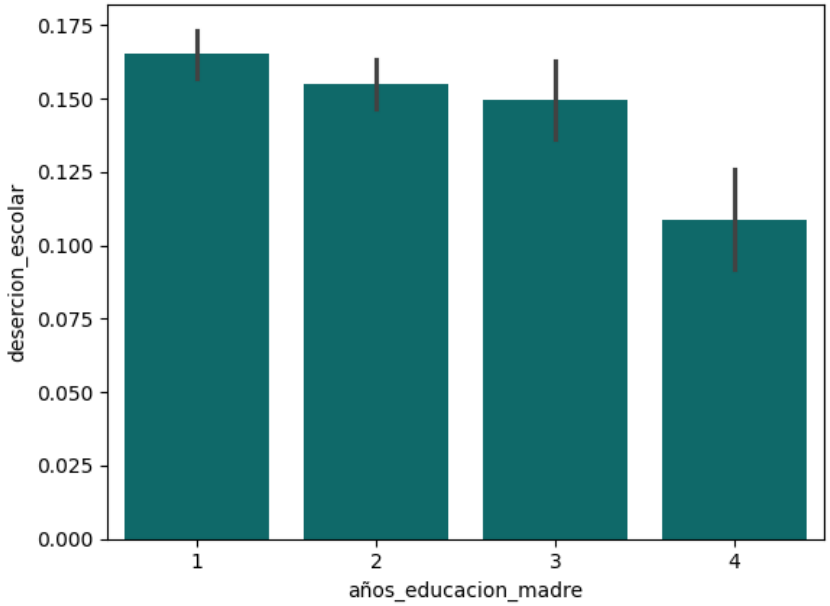

Figura 5: **Deserción escolar segun el nivel educativo de la madre** Fuente: *ENAHO (2021)*

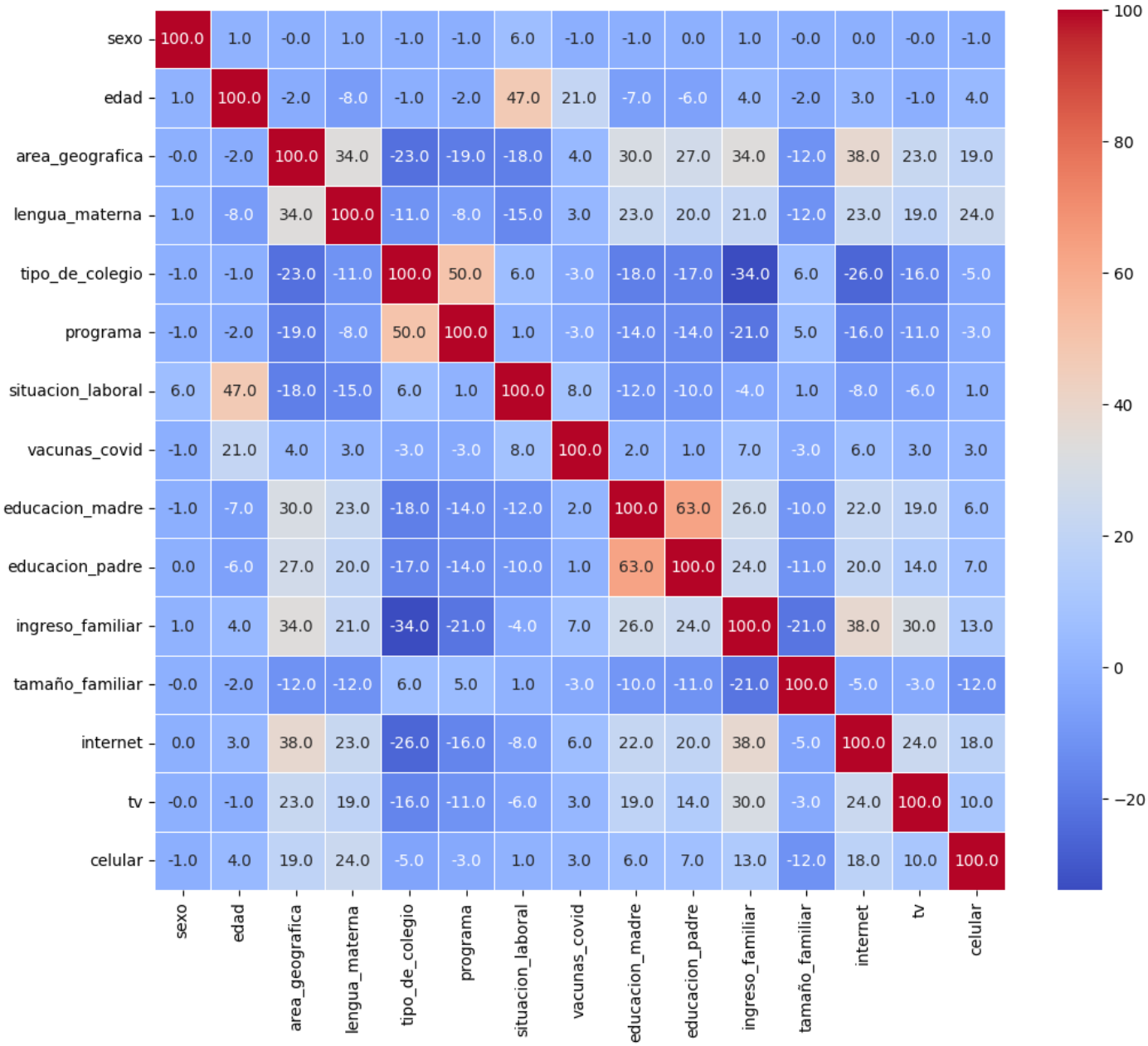

Figura 6: **Matriz de correlación** Fuente: *ENAHO (2021)*

Variables	Definición	Fuente
Deserción escolar	Variable dicotomica que toma el valor de 1 en el caso de que el individuo haya estado matriculado en el año precedente (2020) y no lo esté en el año presente (2021), y toma el valor de 0 si continúa matriculado en el presente año	ENAHO (2021)
Sexo	Variable dicotomica que toma el valor de 1 en el caso de que el individuo sea hombre, y toma el valor de 0 si el individuo es mujer	ENAHO (2021)
Edad	Variable continua que representa la cantidad de años transcurridos desde el nacimiento de un individuo hasta el momento de la observación	ENAHO (2021)
Área geográfica	Variable dicotomica que toma el valor de 1 en el caso de que el individuo reside en una zona urbana, y asume el valor de 0 si reside en una zona rural	ENAHO (2021)
Vacunas COVID	Variable dicotomica que toma el valor de 1 en el caso de que el individuo haya recibido al menos una dosis de la vacuna contra el COVID, y adquiere el valor de 0 en caso contrario	ENAHO (2021)
Lengua Materna	Variable dicotomica que toma el valor de 1 si la lengua materna del individuo es el castellano, y toma el valor de 0 si se trata de alguna lengua nativa	ENAHO (2021)
Educación del padre	Variable categórica que representa el nivel educativo más alto alcanzado por el padre del individuo	ENAHO (2021)
Eduación de la madre	Variable categórica que representa el nivel educativo más alto alcanzado por la madre del individuo	ENAHO (2021)
Ingresos familiares	Variable continua que representa el ingresos mensual per capita del hogar donde reside el individuo (en soles)	ENAHO (2021)
Situación laboral	Variable dicotomica que toma el valor de 1 si el individuo está empleada en el año presente (2021), y toma el valor de 0 en el caso contrario	ENAHO (2021)
Tipo de colegio	Variable dicotomica que toma el valor de 1 si es que el individuo estudió en un colegio público, y toma el valor de 0 si estudió en un colegio privado	ENAHO (2021)
Programas sociales	Variable dicotomica que toma el valor de 1 si es que el individuo está vinculado a un programa social que les brinde recursos educativos para sus estudios, y toma el valor de 0 en el caso contrario	ENAHO (2021)
Internet	Variable dicotomica que toma el valor de 1 si el hogar del individuo cuenta con internet, y toma el valor de 0 en el caso contrario	ENAHO (2021)
TV	Variable dicotomica que toma el valor de 1 si el hogar del individuo cuenta con un televisor, y toma el valor de 0 en el caso contrario	ENAHO (2021)
Celular	Variable dicotomica que toma el valor de 1 si el hogar del individuo dispone de al menos un teléfono celular, y toma el valor de 0 en caso contrario	ENAHO (2021)

Cuadro 1: **Descripción de las variables de interés**

Cuadro 2: **Coeficientes y valores adicionales**

	Coef	Std. err	P-value	dy/dx
const.	-16.1344	0.185	0.000	-
sexo	-0.007377	0.043	0.865	-0.0003
edad	0.089454	0.007	0.000	0.004
area_geografica	0.136171	0.050	0.006	0.0061
lengua_materna	0.022717	0.070	0.745	0.0010
tipo_de_colegio	14.688583	0.074	0.000	0.6655
programa	-1.339402	0.046	0.000	-0.0603
situacion_laboral	0.632032	0.063	0.000	0.0285
vacunas_covid	-1.324343	0.141	0.000	-0.0597
educacion_madre	-0.017734	0.026	0.495	-0.0008
ingreso_familiar	-0.000013	5.5e-05	0.817	-5.744e-07
tamano_familiar	-0.042803	0.013	0.001	-0.0019
internet	-0.335450	0.048	0.000	-0.0151
tv	0.113199	0.051	0.013	0.0051
celular	-0.003572	0.113	0.975	-0.0002
Observaciones:				18514
Pseudo R2:				0.1313
LLR p-value:				0.0000

Clase real \ Clase predicha	No	Sí
No	15469	171
Sí	2525	349

Cuadro 3: **Matriz de confusión**

	Precision	Recall	F1-Score	Support
0	0.86	0.99	0.92	15640
1	0.67	0.12	0.21	2874
Accuracy	-	-	0.56	18514

Cuadro 4: **Informe de Clasificación**

Códigos

```
#Importacion de las principales librerias
import statsmodels.api as sm
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
from sklearn.preprocessing import LabelEncoder

#Importacion de la base de datos DATA1 donde est n las variables de interes
data = pd.read_csv("../data/final/DATA1.csv", index_col=0)
"../data/graphs"

#Verificar la base de datos: el numero de observaciones (18514) y el numero de
   variables (23)
data.shape

#Informacion de las variables
data.info()

#Graficos de barra de la variable dependiente (deserci n escolar) clasificandolo por
   Edad, vacunacion contra el COVID,  educacion de la madre y area geografica
sns.barplot(x='desercion_escolar',y='edad',data=data,color='#007878',lw=4,ls='solid')
sns.barplot(data=data, x = 'vacunas_covid', y='desercion_escolar')
sns.barplot(data=data, x = 'educacion_madre',
   y='desercion_escolar',color='#007878',lw=4,ls='solid')
sns.barplot(data=data, x = 'area_geografica', y='desercion_escolar')

#Matriz de correlacion de las variables independientes para el analisis de
   multicolinealidad
matriz_correlacion= data[['sexo', 'edad', 'area_geografica', 'lengua_materna',
   'tipo_de_colegio',
       'programa', 'situacion_laboral', 'vacunas_covid', 'educacion_madre',
          'educacion_padre',
       'ingreso_familiar', 'tama o_familiar', 'internet',
```

```
        'tv', 'celular']].corr().round(2)
matriz_correlacion

#Mapa de calor de la matriz de correlacion
plt.figure(figsize=(12, 10))
annot_kws = {"size": 10}
sns.heatmap(matriz_correlacion * 100, annot=True, cmap="coolwarm", fmt=".1f",
    linewidths=.5, annot_kws=annot_kws)

#Creacion de la matriz "X" con las variables independientes y de la matriz "y" con la
    variable independiente. Se incluye un vector de constantes a la matriz X.
X = data[['sexo', 'edad' ,'area_geografica', 'lengua_materna', 'tipo_de_colegio',
        'programa', 'situacion_laboral', 'vacunas_covid', 'educacion_madre',
        'ingreso_familiar', 'tama o_familiar', 'internet',
        'tv', 'celular']]
y = data['desercion_escolar']
X = sm.add_constant(X)

#Ajuste del modelo de regresion logistica a datos representados por las variables
    independientes (X) y la variable dependiente (y). Se utiliza el metodo HC3 para
    corregir la heterocedasticidad en la estimaci n de la covarianza de los
    coeficientes del modelo.
logit_model = sm.Logit(y, X)
result = logit_model.fit(cov_type='HC3')

#presentacion de la tabla de resultados del modelo de regresi n logistica
print(result.summary())

#Presentar los efectos marginales del modelo de regresion logistico. Se calculan
    mediante el valor medio de las variables independientes (at mean).
margins = result.get_margeff(at='mean', method='dydx')
print(margins.summary())

#Valores predichos de las variables independientes
predictions = result.predict(X)

```

```
#Evaluacion de las predicciones del modelo de regresi n con un umbral (threshold) de
    0.5. Se clasifican las predicciones comparando con el umbral y, de acuerdo a
    ello,se calculan metricas de evaluacion mediante la Matriz de Confusion.
threshold = 0.5
predicted_classes = (predictions > threshold).astype(int)

from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
accuracy = accuracy_score(y, predicted_classes)
conf_matrix = confusion_matrix(y, predicted_classes)
report = classification_report(y, predicted_classes)
print(f'Precisi n del modelo: {accuracy}')
print('Matriz de Confusi n:')
print(conf_matrix)
print('Informe de Clasificaci n:')
print(report)
```

Listing 1: **Codigos utilizados de la regresion logistica (Python)**

Referencias

[1] Carrión, J. E. V. (2014). Características y Determinantes del Trabajo Infantil y su Influencia en la Deserción Escolar en el Ecuador, 2012 (Doctoral dissertation, Universidad de Cuenca).

[2] Cerrutti, M., & Binstock, G. (2004, April). Camino a la exclusión: determinantes del abandono escolar en el nivel medio en la Argentina. In Trabajo presentado en el I Congreso ALAP (Asociación Latinoamericana de Población).

[3] Drysdale, R. (1972). Factores determinantes de la deserción escolar en Colombia. *Revista del Centro de Estudios Educativos*, 2(3). https://www.cee.edu.mx/rlee/revista/r1971_1980/rtexto/t1972_302.pdf

[4] INEI. (2022). Pandemia y deserción educativa en la Educación Básica Regular: factores asociados y posibles efectos, 2017–2021. Instituto Nacional de Estadística e Informática. https://www.inei.gob.pe/media/MenuRecursivo/investigaciones/desercion-escolar.pdf

[5] Peña Axt, J. C., Soto Figueroa, V. E., & Calderón Aliante, U. A. (2016). La influencia de la familia en la deserción escolar: estudio de caso en estudiantes de secundaria de dos instituciones de las comunas de Padre las Casas y Villarrica, Región de la Araucanía, Chile. Revista mexicana de investigación educativa, 21(70), 881-899. https://www.redalyc.org/articulo.oa?id=14046162011

[6] Rodríguez Lozano, E. (2011). ¿Barreras lingüísticas en la educación?: la influencia de la lengua materna en la deserción escolar. Enlace: https://revistas.pucp.edu.pe/index.php/economia/article/download/2711/2655/

[7] Santos, H. (2009). Dinámica de la deserción escolar en Chile. Santiago de Chile: Centro de Políticas Comparadas de Educación, CPCE. https://www.desarrollosocialyfamilia.gob.cl/btca/txtcompleto/mideplan/ser.estsoc-dinamdeserc.escolar.pdf