METODOLOGÍA EMPLEADA EN EL REPORTE 2:

1. ANÁLISIS A NIVEL DISTRITAL:

Para el análisis a nivel distrital se usaron las siguiente variables:

Tabla 1. Variables usadas en el análisis distrital.

Variable	Indicador	Identificador	Tipo de variable
Caducidad	% de población del distrito con DNI caduco	R_CAD	Continua
Nivel de vida	Índice de desarrollo humano	IDH_2024	Continua
Pobreza	% de la población con pobreza monetaria	POB_MON	Continua
Presencia del Estado	Índice de densidad del Estado	IDE_2024	Continua
Accesibilidad	Distancia en kilómetros a la oficina Reniec más cercana	DIST_KM	Continua

Fuente: Elaboración del equipo de investigación de la SDPEG/DRE.

Respecto la operacionalización de las variables del estudio:

- **Caducidad:** Se operacionalizó mediante el porcentaje de electoras y electores que, al corte de información del V simulacro de cierre del padrón, tenían el DNI caducado, por distrito.
- **Nivel de vida:** Se tomó el IDH 2024 calculado por el PNUD con base en las proyecciones hechas a partir del censo 2017 y las ENAHO realizadas por el INEI.
- **Pobreza:** Se tomaron las estimaciones hechas por el INEI con base en el censo del 2017 y la ENAHO del mismo año.
- **Presencia del Estado:** Se usó el IDE 2024 desarrollado por el PNUD.
- **Accesibilidad:** Se operacionalizó como la distancia euclidiana (lineal) en kilómetros entre el centroide poblacional de cada distrito y la agencia del Reniec más cercana con capacidad de realizar trámites de renovación del DNI de manera permanente. El centroide poblacional se construyó a partir de la información georreferenciada de centros poblados del Censo 2017 (INEI), ponderando las coordenadas de cada centro poblado según el tamaño de su población, con lo que se obtuvo un único punto representativo del área de mayor concentración poblacional en cada distrito. La relación de agencias del Reniec habilitadas para la renovación del DNI fue proporcionada por la Dirección de Servicios Registrales, y sus coordenadas fueron generadas por

el equipo de georreferenciación de las SDPEG de las DRE a partir de sus direcciones oficiales.

Respecto a los distritos incluidos en el análisis:

De los 1892 distritos existentes al momento de redactar este reporte, solo se incluyeron 1874 en el análisis. Esta decisión se tomó debido a que no se cuenta con información de la pobreza monetaria, IDH e IDE para los distritos creados posteriores al censo del 2017. Asimismo, que en distritos de reciente creación los porcentajes de DNIs caducados serían engañosamente bajos.

Tabla 2. Distritos excluidos del análisis.

Ubigeo Reniec	Departamento	Provincia	Distrito	Fecha de creación
070916	CUSCO	LA CONVENCIÓN	KUMPIRUSHIATO	17 de marzo del 2021
050412	AYACUCHO	LA MAR	UNIÓN PROGRESO	05 de marzo del 2021
080534	HUANCAVELICA	TAYACAJA	COCHABAMBA	09 de marzo del 2021
080533	HUANCAVELICA	TAYACAJA	LAMBRAS	19 de diciembre del 2020
050415	AYACUCHO	LA MAR	RIO MAGDALENA	09 de diciembre del 2020
050413	AYACUCHO	LA MAR	PATIBAMBA	16 de marzo del 2021
070919	CUSCO	LA CONVENCIÓN	UNIÓN ASHANINKA	18 de mayo del 2021
050314	AYACUCHO	HUANTA	PUTIS	09 de marzo del 2021
250206	UCAYALI	PADRE ABAD	BOQUERÓN	18 de marzo del 2021
050414	AYACUCHO	LA MAR	NINABAMBA	16 de marzo del 2021
070918	CUSCO	LA CONVENCIÓN	MANITEA	07 de abril del 2021
070917	CUSCO	LA CONVENCIÓN	CIELO PUNCO	07 de abril del 2021
210806	SAN MARTIN	TOCACHE	SANTA LUCÍA	04 de marzo del 2021
250207	UCAYALI	PADRE ABAD	HUIPOCA	09 de marzo del 2021
030712	APURIMAC	CHINCHEROS	AHUAYRO	02 de mayo del 2021
170107	MOQUEGUA	MARISCAL NIETO	SAN ANTONIO	15 de junio del 2021
120113	LA LIBERTAD	TRUJILLO	ALTO TRUJILLO	15 de diciembre del 2022
Aún no se ha creado	LORETO	MARISCAL RAMÓN CASTILLA	SANTA ROSA DE LORETO	03 de julio de 2025

Fuente: Elaboración del equipo de investigación de la SDPEG/DRE.

Análisis descriptivo:

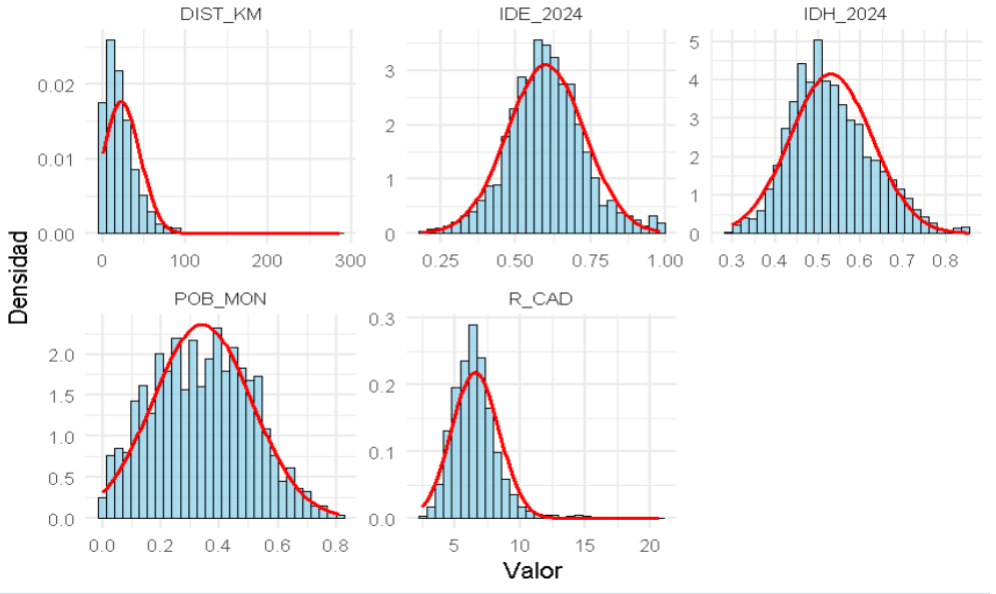
Cómo primera etapa del análisis, se presentan algunas características descriptivas de las variables del estudio. Salvo IDE_2024, todas las variables muestran distribuciones no normales, lo cual se evidencia tanto en las pruebas de normalidad Kolmogorov-Smirnov (K-S), como de manera gráfica en los histogramas de las variables. Este resultado se debe principalmente a la asimetría positiva de la distribución de las variables, especialmente marcada en R_CAD y DIST_KM, donde se observan valores que se alejan de manera considerable de sus promedios. Lo anterior refleja que los escenarios en los que un distrito presenta porcentajes muy elevados de población electoral con el DNI caduco o se ubica a grandes distancias de una agencia Reniec constituyen patrones atípicos, dado que la mayoría de las observaciones se concentran en valores bajos.

Tabla 3. Estadísticos descriptivos de las variables.

Variable	Media	Mediana	Asimetría	Normalidad (K-S) p-value
R_CAD	6.5342	6.3545	1.8452	0.0000
IDH_2024	0.5305	0.5185	0.4666	0.0000
POB_MON	0.3414	0.3430	0.1194	0.0033
IDE_2024	0.6011	0.5965	0.1736	0.0603
DIST_KM	22.8128	17.7333	3.2381	0.0000

Fuente: Elaboración del equipo de investigación de la SDPEG/DRE con base en Reniec (2025), INEI (2017) y PNUD (2025).

Gráfico 1. Histogramas de las variables.

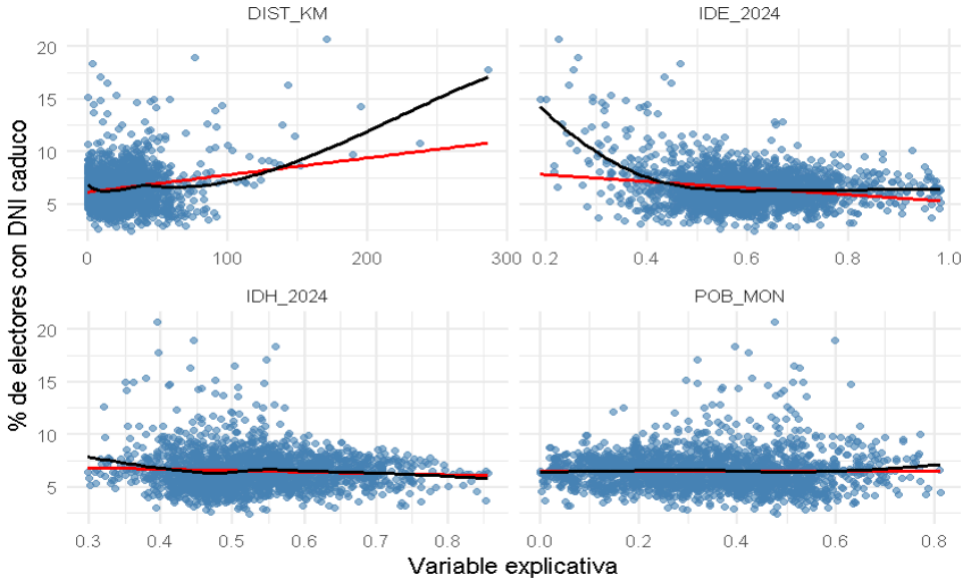


Fuente: Elaboración del equipo de investigación de la SDPEG/DRE con base en Reniec (2025), INEI (2017) y PNUD (2025).
Nota: La línea roja indica la distribución normal teórica de las observaciones.

Análisis bivariado:

Continuando con el análisis, se exploran las tendencias en la relación uno a uno de las variables explicativas y R_CAD. Como se observa en el Gráfico 2, IDH_2024 y POB_MON presentan una relación predominantemente lineales, respecto a R_CAD en todo su rango, salvo ligeras desviaciones en los valores más bajos en el caso de IDH_2024, y en los valores más altos en el caso de POB_MON. En contraste, DIST_KM e IDE_2024 presenta desviaciones más marcadas respecto a la tendencia lineal global. En el caso de DIST_KM, la relación se mantiene cercana a la lineal global en la primera mitad del rango, pero posteriormente incrementa su pendiente de manera considerable. En el caso de IDE_2024, la pendiente resulta mucho más pronunciada en el rango bajo del índice, para luego atenuarse hasta converger con lo relación lineal global en la parte media y superior del rango.

Gráfico 2. Diagramas de dispersión de las variables explicativas vs R_CAD



Fuente: Elaboración del equipo de investigación de la SDPEG/DRE con base en Reniec (2025), INEI (2017) y PNUD (2025).
Nota: La línea roja indica la tendencia global de las variables asumiendo una función lineal constante, obtenida mediante el comando lm (). La línea negra refleja la relación de las variables ajustándose a las variaciones locales, sin asumir una función específica, obtenida mediante el comando low ess().

Tabla 4. Pruebas de correlación entre las variables explicativas y R_CAD

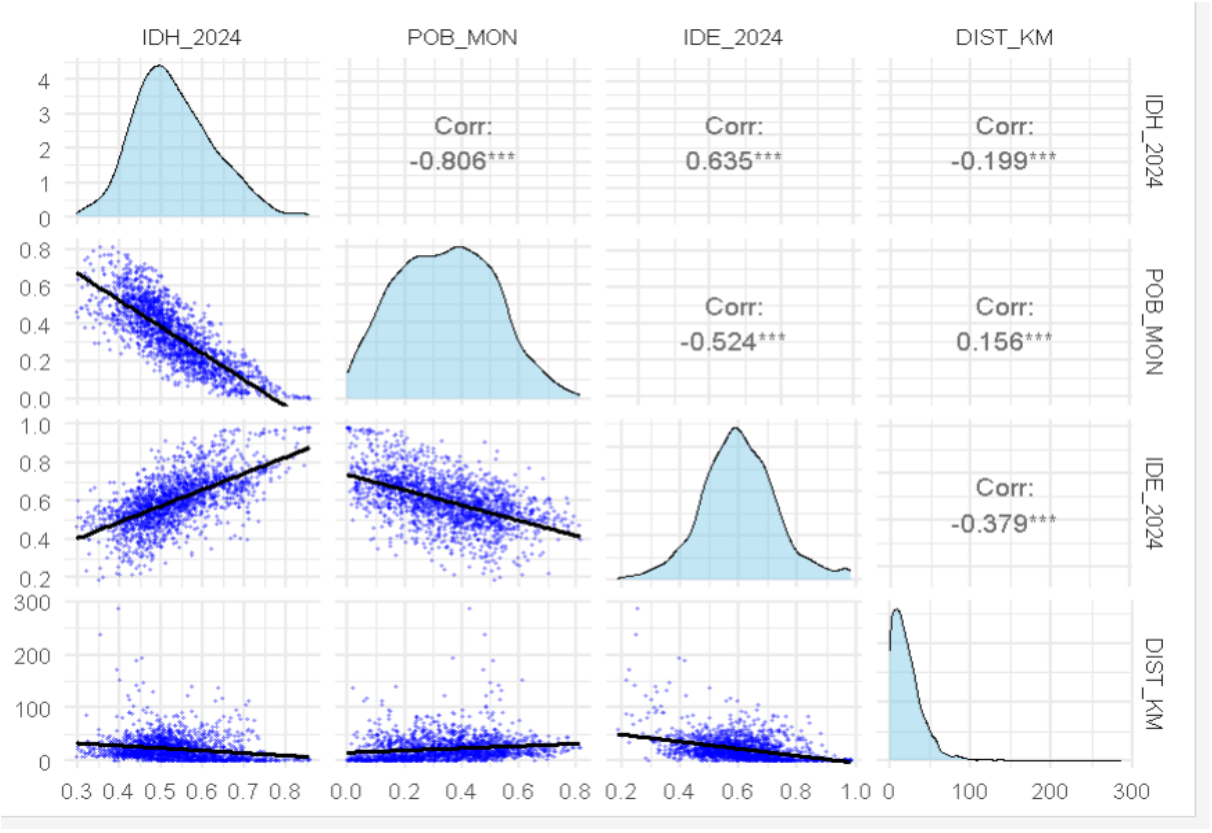
Variable	R de Pearson	Significancia	Rho Spearman	Significancia
IDH_2024	-0.071	0.0022	0.003	0.8809
POB_MON	-0.002	0.9335	-0.076	0.0010
IDE_2024	-0.221	0.0000	-0.094	0.0000
DIST_KM	0.200	0.0000	0.019	0.4121

Fuente: Elaboración del equipo de investigación de la SDPEG/DRE con base en Reniec (2025), INEI (2017) y PNUD (2025).

A continuación, se presentan los resultados de las pruebas de correlación R de Pearson y Rho de Spearman entre las variables explicativas y R_CAD. Los coeficientes y p_values de Pearson muestran la existencia de una relación lineal global y estadísticamente significante en todas las variables, salvo en POB_MON donde la relación es tan tenue, que es posible que sea nula. Por su parte, los resultados de Spearman evidencian que, en el caso de IDH_2024 y DIST_KM, la dirección de la relación no es uniforme a lo largo del rango de valores, pudiendo invertirse en determinados tramos. Siendo esto último algo que también se puede visualizar en el Gráfico 2.

Finalizando esta sección, se muestran las pruebas de correlación entre las variables explicativas a fin de identificar posibles problemas de multicolinealidad en las pruebas de regresión múltiple. En el Gráfico 3 observamos de manera compacta tanto los resultados de las pruebas de Pearson como los diagramas de dispersión correspondientes. Primero, tenemos que hacer notar que todas las variables presentan algún grado de correlación entre ellas, siendo en todos los casos estadísticamente significativa. En segundo lugar, destaca la fuerte correlación negativa entre IDH_2024 y POB_MON (r = -0,8), un valor considerablemente alto que sugiere una relación muy estrecha entre las variables y que deberemos tener en cuenta en los análisis posteriores para prevenir problemas de multicolinealidad.

Gráfico 3. Correlación de las variables explicativas



Fuente: Elaboración del equipo de investigación de la SDPEG/DRE con base en Reniec (2025), INEI (2017) y PNUD (2025).

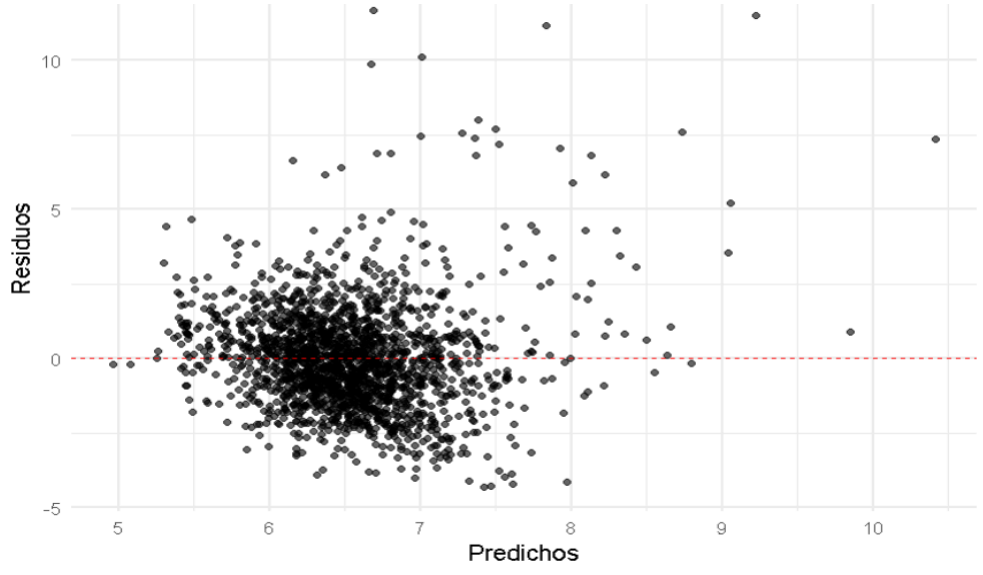
Modelo lineal base:

El primer modelo que hemos desarrollado es una regresión lineal múltiple usando el comando `lm()`. A continuación se presentan los resultados.

```
Call:
lm(formula = R_CAD ~ IDH_2024 + POB_MON + IDE_2024 + DIST_KM,
    data = DATA)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3163 -1.0907 -0.1103  0.8505 11.6904

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  9.355623   0.518797  18.033 < 0.0000000000000002 ***
IDH_2024    -0.586352   0.788025  -0.744                0.457
POB_MON     -1.897511   0.405387  -4.681  0.00000306399384885 ***
IDE_2024    -3.493894   0.434420  -8.043  0.00000000000000155 ***
DIST_KM      0.010419   0.001947   5.352  0.00000009779440469 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.756 on 1869 degrees of freedom
Multiple R-squared:  0.08241, Adjusted R-squared:  0.08045
F-statistic: 41.96 on 4 and 1869 DF,  p-value: < 0.00000000000000022
```

Lo primero a destacar es que el modelo lineal muestra que, en conjunto las variables explicativas ejercen un efecto significativo sobre R_CAD ($F = 41.96$; $p < 0.001$). Sin embargo, la capacidad del modelo es limitada, puesto que sólo logra explicar cerca del 8% de la varianza de R_CAD ($R^2 = 0.08241$). Respecto a los coeficientes, POB_MON e IDE_2024 muestran efectos negativos y estadísticamente significativos, mientras que DIST_KM presenta efectos positivos. Por el contrario, IDH_2024 si bien presenta efectos negativos, estos no resultan estadísticamente significativos.

Gráfico 4. Diagrama de dispersión de los residuos del modelo lineal base.

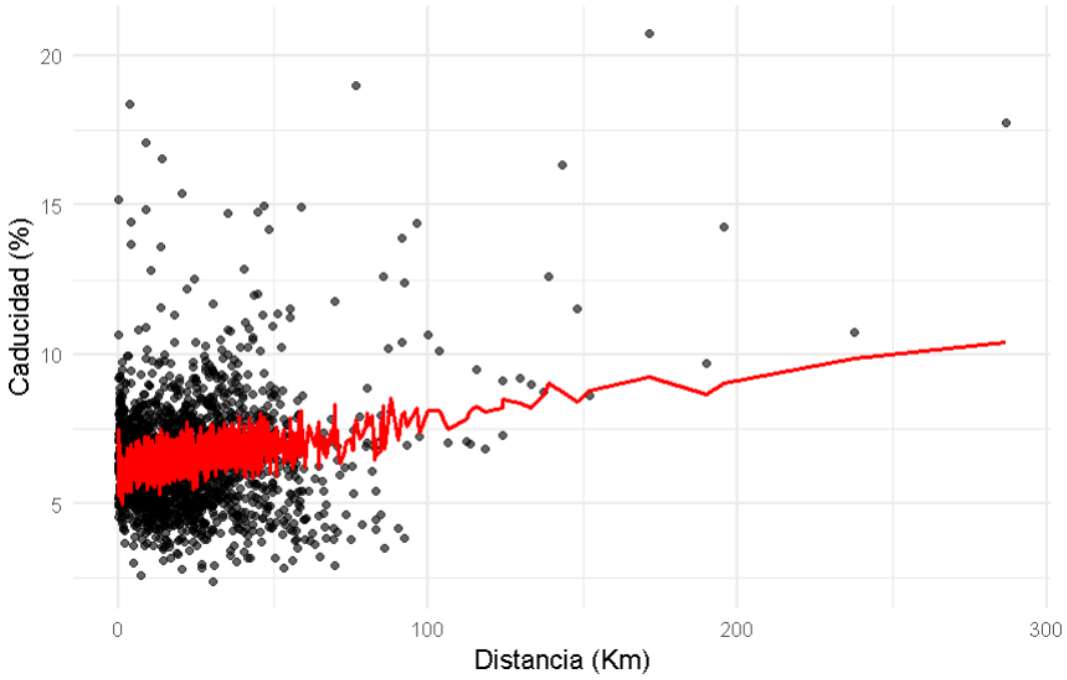


Fuente: Elaboración del equipo de investigación de la SDPEG/DRE con base en Reniec (2025), INEI (2017) y PNUD (2025).

Pasando al análisis de residuos (ver Gráfico 4), vemos que en los valores ajustados más altos los residuos se concentran en valores positivos, indicando que el modelo tiende a subestimar los valores más altos de R_CAD. Asimismo, observamos en general una ligera asimetría de los residuos, con residuos positivos más grandes que los negativos. Finalmente, aunque la mayoría de los residuos se encuentran alrededor de cero, la dispersión aumenta en los valores ajustados más altos, sugiriendo posibles problemas de heterocedasticidad.

La distribución antes vista, recuerda a los resultados del Gráfico 2, donde DIST_KM presentaba un aumento considerable de su pendiente en sus rangos más elevados, alejándose de la linealidad global asumida. Para comprobar que esta no linealidad se relaciona con la asimetría de los residuos, se ha realizado un diagrama de dispersión entre DIST_KM y R_CAD agregando los valores predichos por el modelo lineal múltiple (línea roja). Como se observa en el Gráfico 5, efectivamente el cambio en la relación entre ambas variables en los niveles más elevados de la distancia es una de las razones por las cuales el modelo no logra ajustarse adecuadamente.

Gráfico 5. Diagrama de dispersión DIST_KM vs R_CAD



Fuente: Elaboración del equipo de investigación de la SDPEG/DRE con base en Reniec (2025), INEI (2017) y PNUD (2025).
Nota: La línea roja indica los valores ajustados según el modelo lineal múltiple.

Modelo lineal segmentado:

Con el propósito de captar el cambio en la pendiente de la variable DIST_KM, se optó por implementar un modelo lineal segmentado a partir del modelo lineal base antes desarrollado. Para ello, se utilizó la función `segmented()`, incorporando sólo dos puntos de quiebre a fin de evitar una complejización innecesaria del modelo.

Los resultados del modelo segmentado han ajustado dos puntos de quiebre óptimos y estadísticamente significativos aproximadamente en los 7,22 km y los 58,12 km, donde la relación entre R_CAD y DIST_KM cambia. La inclusión de estos quiebre mejora la capacidad explicativa del modelo el cual pasa de explicar casi el 8% de la varianza de R_CAD en el modelo lineal base a explicar aproximadamente el 14% en el modelo segmentado.

```
      ***Regression Model with Segmented Relationship(s)***

Call:
segmented.lm(obj = modelo_base, seg.Z = ~DIST_KM, psi = c(5,
    100))

Estimated Break-Point(s):
                Est. St.Err
psi1.DIST_KM  7.217  1.170
psi2.DIST_KM 58.124  7.658

Coefficients of the linear terms:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 11.50644    0.56419  20.395 < 0.0000000000000002 ***
IDH_2024    -1.75811    0.77738  -2.262               0.0238 *
POB_MON     -1.87890    0.39295  -4.782           0.00000188 ***
IDE_2024    -4.15164    0.42848  -9.689 < 0.0000000000000002 ***
DIST_KM     -0.16156    0.03636  -4.443           0.00000938 ***
U1.DIST_KM   0.16444    0.03649   4.507                   NA
U2.DIST_KM   0.03393    0.00581   5.841                   NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.698 on 1865 degrees of freedom
Multiple R-Squared: 0.1434,  Adjusted R-squared: 0.1397

Boot restarting based on 6 samples. Last fit:
Convergence attained in 3 iterations (rel. change -0.00000000004859)
```

Los quiebres antes detectados (U1 y U2) generan aumentos en el coeficiente de DIST_KM, 0.164 y 0.034 respectivamente. De forma que el efecto del aumento de un kilómetro extra sobre R_CAD por cada rango de la distancia queda de la siguiente forma:

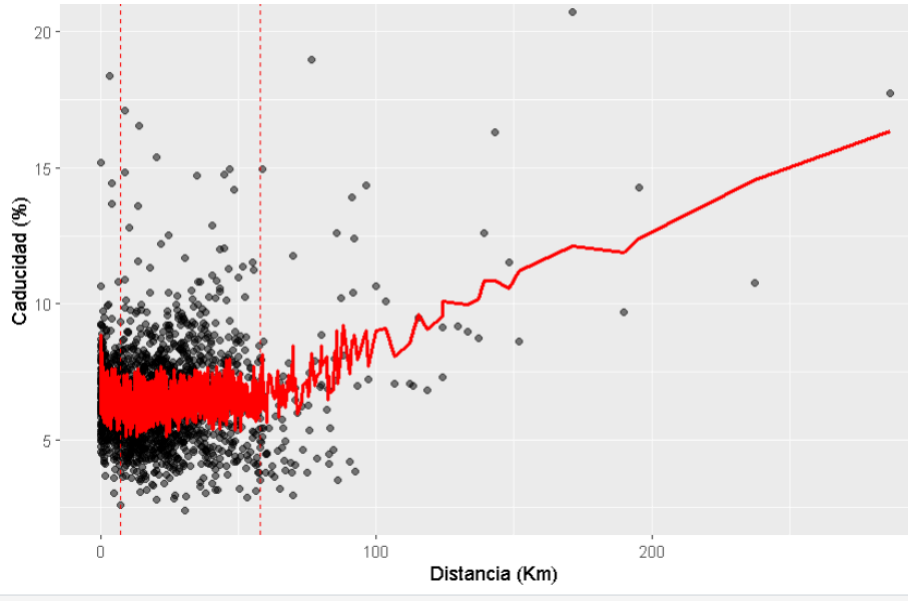
- Menor a 7,22 km: efecto = -0.162
- Entre 7,22 km y 58,12 km: efecto = -0.162 + 0.164 = +0.002
- Mayor a 58,12 km: efecto = -0.162 + 0.164 + 0.034 = +0.036

La mejora del modelo producto de la segmentación se observa en el Gráfico 6, donde se observa como la línea producto de la regresión se ajusta a los cambios en la relación entre R_CAD y DIST_KM. Asimismo, como se aprecia en el Gráfico 7, el modelo segmentado ha conseguido mejorar la distribución de los residuos en el último rango de los valores fijados. Sin embargo, a pesar de la mejora, aún persiste la asimetría en la distribución de los residuos, teniendo en general, los positivos mayor magnitud que los negativos. Esto debido a la presencia de distritos con un porcentaje atípicamente alto de porcentajes de caducidad (R_CAD) que el modelo no consigue ajustar.

En resumen, el análisis apoya la idea de la relación no lineal entre DIST_KM y R_CAD, donde el efecto de la misma depende del rango de distancia observado. El modelo

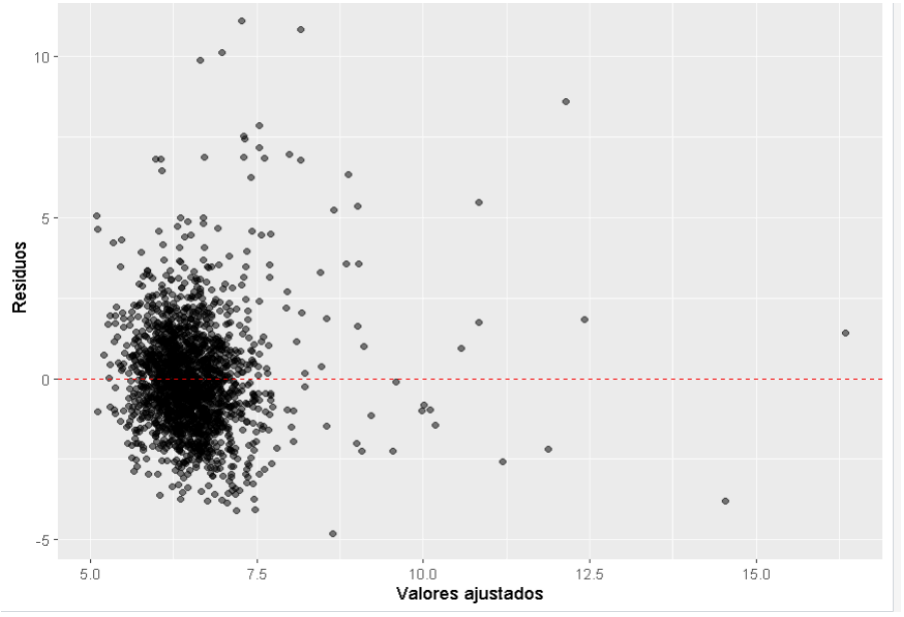
segmentado permite identificar tres efectos distintos: un efecto negativo en los entornos más cercanos, un efecto casi nulo en las distancias intermedias y un efecto positivo en las distancias más alejadas.

Gráfico 6. Diagrama de dispersión DIST_KM vs R_CAD, en el modelo segmentado



Fuente: Elaboración del equipo de investigación de la SDPEG/DRE con base en Reniec (2025), INEI (2017) y PNUD (2025).
Nota: La línea roja indica los valores ajustados según el modelo lineal segmentado.

Gráfico 7. Diagrama de dispersión de los residuos del modelo lineal segmentado.



Fuente: Elaboración del equipo de investigación de la SDPEG/DRE con base en Reniec (2025), INEI (2017) y PNUD (2025).

Para confirmar formalmente la dispersión asimétrica de los residuos, así como la aparente heterocedasticidad de residuos observable en el Gráfico 7, se han realizado las pruebas Shapiro-Wilk y Breusch-Pagan, respectivamente. Los resultados de ambas (p value menor

a 0,05) confirman las sospechas de residuos asimétricos (no normales) y con heterocedasticidad.

```
Shapiro-Wilk normality test
data:  resid_seg2
W = 0.93623, p-value < 0.00000000000000022

studentized Breusch-Pagan test
data:  modelo_seg2
BP = 134.39, df = 8, p-value < 0.00000000000000022
```

A fin de robustecer los resultados del modelo segmentado se ha aplicado la corrección de White-HC2 mediante la función coeftest(). Esta corrección no afecta a los coeficientes estimados del modelo, pero sí al cálculo de los p value a fin de hacerlos robustos ante problemas de heterocedasticidad en los residuos.

```
t test of coefficients:

               Estimate Std. Error t value             Pr(>|t|)
(Intercept)  11.5064432  0.6084886 18.9099 < 0.00000000000000022 ***
IDH_2024     -1.7581092  0.7816056 -2.2494            0.0246061 *
POB_MON      -1.8789021  0.3924661 -4.7874     0.000001822473651 ***
IDE_2024     -4.1516420  0.6045634 -6.8672     0.000000000008889 ***
DIST_KM      -0.1615562  0.0290971 -5.5523     0.000000032235999 ***
U1.DIST_KM    0.1644414  0.0293793  5.5972     0.000000025018468 ***
U2.DIST_KM    0.0339319  0.0093314  3.6363            0.0002841 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabla 5. Comparación de los coeficientes del modelo lineal base y el modelo lineal segmentado

Variable	Coeficiente (modelo base)	Coeficiente (modelo segmentado)
IDH_2024	-0.59	-1.76*
POB_MON	-1.89***	-1.88***
IDE_2024	-3.49***	-4.15***
DIST_KM	0.01***	-0.16***
		0.002***
		0.04***

Fuente: Elaboración del equipo de investigación de la SDPEG/DRE con base en Reniec (2025), INEI (2017) y PNUD (2025).

Finalizamos esta sección comparando los resultados de ambos modelos (Tabla 5). Como se observa, además de los cambios en la pendiente de DIST_KM producto de la segmentación, también se evidencia un cambio en el efecto estimado de las demás variables explicativas. En el caso del IDH_2024, este no resultaba estadísticamente

significativo en el modelo base, mientras que en el modelo segmentado sí alcanza significancia ($p < 0.05$), con un efecto negativo mayor en magnitud (-1.76). En cuanto a la variable POB_MON, su efecto negativo se mantiene estable entre ambos modelos (aprox. -1.89). Por su parte, el IDE_2024 también refuerza su efecto negativo al pasar de -3.49 en el modelo base a -4.15 en el modelo segmentado, confirmando su fuerte peso explicativo.

Limitaciones del análisis a nivel distrital:

En primer lugar, la información sobre la caducidad de los DNI fue extraída del RUIPN casi tres meses antes del cierre del padrón electoral y nueves meses antes del día de la votación. Es esperable que para dichas fechas el porcentaje de DNIs caducados disminuya de manera desigual entre distritos, lo cual podría intensificar o disminuir los efectos de los factores propuestos.

En segundo lugar, las mediciones hechas para la construcción de los indicadores algunas variables explicativas se encuentran desactualizadas. La pobreza monetaria, por ejemplo, se construyó a partir de datos del censo de 2017, mientras que el IDH y el IDE provienen de estimaciones del PNUD con base en proyecciones del censo de 2017 ajustadas con las ENAHO.

Finalmente, para la accesibilidad a la agencia Reniec se recurrió a una variable proxy: la distancia lineal al centroide poblacional. Este cálculo no considera la presencia, o no, de vías de acceso, ni la presencia de obstáculos naturales como ríos o montañas. Además, no fue posible georreferenciar todas las agencias, por lo que en ciertos casos se emplearon coordenadas referenciales. En consecuencia, la estimación del efecto estimado de la distancia es perfectible.

2. ANÁLISIS A NIVEL INDIVIDUAL:

Para el análisis a nivel individual se usaron las siguiente variables:

Tabla 1. Variables usadas en el análisis individual.

Variable	Indicador	Identificador	Tipo de variable
Caducidad	Tener el DNI caduco	CAD	Dicotómica
Edad	Edad en años	YEAR	Continua
Discapacidad	Declarar discapacidad en el DNI	DISC	Dicotómica
Residencia	Lugar de residencia (nacional o extranjero) declarada en el DNI	EXT	Dicotómica
Sexo	Sexo que figura en el DNI	SEX	Dicotómica
Nivel educativo	Nivel educativo declarado en el DNI	EDUC	Categórica

Fuente: Elaboración del equipo de investigación de la SDPEG/DRE.

Respecto la operacionalización de las variables del estudio:

- **Caducidad:** Se utilizó la información registrada en el RUIPN al 31 de julio de 2025. La variable se operacionalizó de manera dicotómica, asignando el valor 1 cuando la persona presentaba el DNI vencido y 0 cuando contaba con el documento vigente.
- **Edad:** Se utilizó la información registrada en el RUIPN al 31 de julio de 2025. La variable se definió como continua calculando la edad en años que cada ciudadano tendrá al 12 de abril del 2025, día de las Elecciones Generales 2026.
- **Discapacidad:** Se utilizó la información registrada en el RUIPN al 31 de julio de 2025. Se operacionalizó como variable dicotómica, 1 si la persona declaró tener una discapacidad y 0 si no declaró tener una discapacidad.
- **Residencia:** Se utilizó la información registrada en el RUIPN al 31 de julio de 2025. Se operacionalizó como variable dicotómica, 1 si la persona declaró residir en el extranjero y 0 si declaró residir en territorio peruano.
- **Sexo:** Se operacionalizó con base en los datos declarados del DNI que figuraba en el RUIPN al 31 de julio de 2025. Se operacionalizó como variable dicotómica, 1 si figuraba como mujer y 0 si figuraba como hombre.
- **Nivel educativo:** Se operacionalizó con base en los datos declarados del DNI que figuraba en el RUIPN al 31 de julio de 2025. Se operacionalizó como variable categórica usando los siguientes criterios:

    - Iletrado: Si en la declaración figuraba como iletrado. Además esta se usó como categoría base para el análisis.

- Educación inicial: Si en la declaración figuraba educación inicial como máximo nivel educativo alcanzado.
- Educación primaria: Si declaró haber cursado algún grado o haber completado la educación primaria.
- Educación secundaria: Si declaró haber cursado algún grado o haber completado la educación secundaria.
- Educación superior: Si declaró haber cursado estudios superiores, ya sean técnicos o universitarios, completos o incompletos.
- Educación especial: Si declaró haber recibido educación especial.

Respecto a los distritos incluidos en el análisis:

De los 1892 distritos existentes al momento de redactar este reporte, solo se incluyeron 1874 en el análisis. Esta decisión se tomó debido a que en distritos de reciente creación, aún no ha transcurrido suficiente tiempo como para que los ciudadanos que se registraron tengan el DNI caducado. Por lo que considerar estos distritos en el modelo explicativo introduciría un sesgo en los resultados.

Tabla 2. Distritos excluidos del análisis.

Ubigeo Reniec	Departamento	Provincia	Distrito	Fecha de creación
070916	CUSCO	LA CONVENCIÓN	KUMPIRUSHIATO	17 de marzo del 2021
050412	AYACUCHO	LA MAR	UNIÓN PROGRESO	05 de marzo del 2021
080534	HUANCAVELICA	TAYACAJA	COCHABAMBA	09 de marzo del 2021
080533	HUANCAVELICA	TAYACAJA	LAMBRAS	19 de diciembre del 2020
050415	AYACUCHO	LA MAR	RIO MAGDALENA	09 de diciembre del 2020
050413	AYACUCHO	LA MAR	PATIBAMBA	16 de marzo del 2021
070919	CUSCO	LA CONVENCIÓN	UNIÓN ASHANINKA	18 de mayo del 2021
050314	AYACUCHO	HUANTA	PUTIS	09 de marzo del 2021
250206	UCAYALI	PADRE ABAD	BOQUERÓN	18 de marzo del 2021
050414	AYACUCHO	LA MAR	NINABAMBA	16 de marzo del 2021
070918	CUSCO	LA CONVENCIÓN	MANITEA	07 de abril del 2021
070917	CUSCO	LA CONVENCIÓN	CIELO PUNCO	07 de abril del 2021
210806	SAN MARTIN	TOCACHE	SANTA LUCÍA	04 de marzo del 2021
250207	UCAYALI	PADRE ABAD	HUIPOCA	09 de marzo del 2021
030712	APURIMAC	CHINCHEROS	AHUAYRO	02 de mayo del 2021

170107	MOQUEGUA	MARISCAL NIETO	SAN ANTONIO	15 de junio del 2021
120113	LA LIBERTAD	TRUJILLO	ALTO TRUJILLO	15 de diciembre del 2022
Aún no se ha creado	LORETO	MARISCAL RAMÓN CASTILLA	SANTA ROSA DE LORETO	03 de julio de 2025

Fuente: Elaboración del equipo de investigación de la SDPEG/DRE.

Modelo logístico:

Dada la naturaleza dicotómica de la variable dependiente, R_CAD, se ha optado por un modelo binario logístico. Considerando los altos costos computacionales que implica trabajar con toda la población electoral, se optó por usar una muestra representativa. Para ello, se realizó un muestreo aleatorio y estratificado del 4% a nivel provincial en el territorio nacional y a nivel de país para la población en el extranjero. Finalmente, el modelo se ha implementado mediante la función `glm()` y la especificación `family = binomial(link = "logit")`. Se adjuntan los resultados:

```
Call:
glm(formula = DNI_CAD ~ SEX + YEARS + YEARS_2 + GRUP_ED2 + EXT +
    DISC, family = binomial(link = "logit"), data = DATA_small)

Coefficients:
                Estimate  Std. Error z value             Pr(>|z|)
(Intercept)  -2.75513175  0.03103328 -88.780 < 0.0000000000000002 ***
Mujer        -0.25194491  0.00733230 -34.361 < 0.0000000000000002 ***
YEARS         0.07672780  0.00131755  58.235 < 0.0000000000000002 ***
YEARS_2      -0.00101041  0.00001454 -69.488 < 0.0000000000000002 ***
INICIAL       0.28302143  0.04043717   6.999     0.00000000000258 ***
PRIMARIA     -0.58989527  0.01746509 -33.776 < 0.0000000000000002 ***
SECUNDARIA   -1.06131077  0.01670393 -63.537 < 0.0000000000000002 ***
SUPERIOR     -1.27390325  0.02059310 -61.861 < 0.0000000000000002 ***
ESPECIAL      0.03189062  0.10845193   0.294               0.7687
Extranjero    1.91156041  0.01136951 168.130 < 0.0000000000000002 ***
Con Discapacidad 0.08805838  0.03830368   2.299            0.0215 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 603194  on 1095722  degrees of freedom
Residual deviance: 568045  on 1095712  degrees of freedom
AIC: 568067

Number of Fisher Scoring iterations: 6
```

Dado que se trata de un modelo logístico, no se pueden cuantificar directamente los efectos de cada variable, pero los signos y significancias de los coeficientes permiten identificar las tendencias. En general, observamos que la probabilidad de tener el DNI caducado está

asociada significativamente con el sexo, la edad, la residencia en el extranjero, tener alguna discapacidad y, de manera muy marcada, con el nivel educativo. Asimismo, destaca que el efecto de la edad no es lineal, sino que aumenta y luego se atenúa, como lo indica el signo negativo en el coeficiente cuadrático.

A fin de cuantificar el efecto de cada variables sobre la probabilidad de tener el DNI caduco, se calcularon los efectos marginales promedio derivados del modelo logístico mediante la función `margins()`. Se adjuntan los resultados:

factor	AME	SE	z	p	lower	upper
Con Discapacidad	0.0063	0.0028	2.2236	0.0262	0.0007	0.0119
Extranjero	0.2528	0.0022	114.0455	0.0000	0.2484	0.2571
ESPECIAL	0.0043	0.0148	0.2913	0.7708	-0.0247	0.0333
INICIAL	0.0412	0.0062	6.6150	0.0000	0.0290	0.0533
PRIMARIA	-0.0656	0.0022	-29.7647	0.0000	-0.0699	-0.0612
SECUNDARIA	-0.1013	0.0021	-47.3713	0.0000	-0.1055	-0.0971
SUPERIOR	-0.1136	0.0022	-51.5408	0.0000	-0.1179	-0.1093
Mujer	-0.0174	0.0005	-34.3835	0.0000	-0.0184	-0.0164
YEARS	0.0053	0.0001	57.7807	0.0000	0.0051	0.0055
YEARS_2	-0.0001	0.0000	-68.7142	0.0000	-0.0001	-0.0001

Los efectos marginales obtenidos muestran que, manteniendo constantes las demás variables, la probabilidad de tener el DNI vencido se ve influida de manera significativa por casi todas las variables incluidas en el modelo. Destaca como variable de mayor efecto el residir en el extranjero, donde la probabilidad de tener el DNI caduco aumenta en 25,3 puntos porcentuales respecto a los que residen en territorio nacional. En segundo lugar, tener mayor nivel educativo disminuye la probabilidad de tener el DNI caduco. En comparación con los iletrados, haber cursado primaria reduce la probabilidad en 6,6 puntos porcentuales, secundaria en 10,1 puntos, y educación superior en 11,4 puntos. Siendo el haber recibido educación especial, el único nivel educativo que no muestra un efecto estadísticamente significativo.

Respecto a las demás variables, estas tienen un efecto más pequeño, por ejemplo ser mujer reduce la probabilidad de tener el DNI caduco en 1,7 puntos porcentuales respecto a ser hombres. Asimismo, tener alguna discapacidad incrementa ligeramente la probabilidad de tener el DNI caduco en 0,6 puntos.

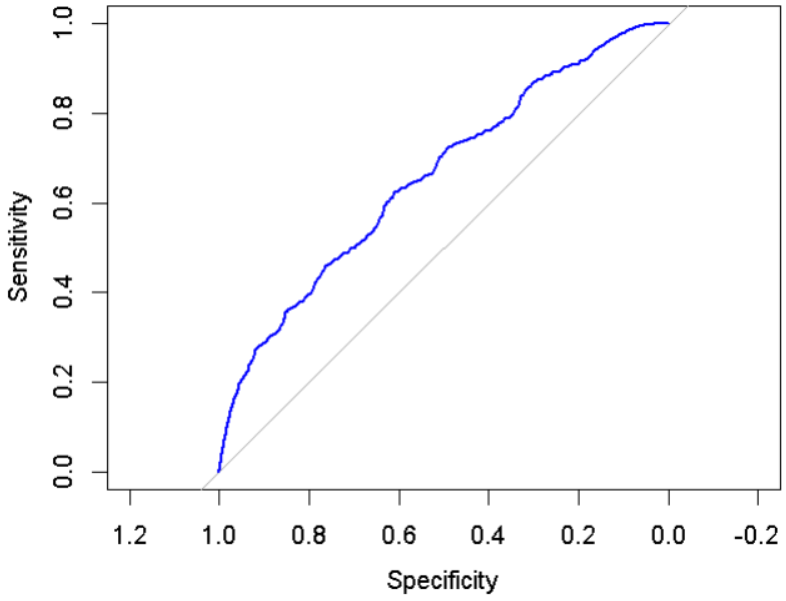
Sobre la edad, esta muestra un efecto no lineal: cada año adicional incrementa la probabilidad en 0,53 puntos porcentuales, pero este efecto disminuye progresivamente en edades más avanzadas, como lo refleja el coeficiente cuadrático negativo. Finalmente, mediante el uso del método delta, se estimó que la máxima probabilidad de caducidad se alcanza alrededor de los 38 años, a partir de los cuales la probabilidad comienza a descender.

```
$edad_max
   YEARS
37.96848

$IC95
[1] 37.65432 38.28263
```

Sobre las métricas de bondad de ajuste, se adjuntan los pseudo R cuadrados más usados, así como el valor AUC. Como vemos, hemos obtenido valores relativamente bajos, algo que nos señala la necesidad de incluir más variables explicativas en futuros modelos.

```
fitting null model for pseudo-r2
McFadden             r2ML             r2CU
0.05827227       0.03156976       0.07457341
```

Gráfico 1. Curva ROC del modelo.




```
Area under the curve (AUC): 0.6593
```

Limitaciones del análisis a nivel distrital:

En primer lugar, la información sobre la caducidad de los DNI fue extraída del RUIPN casi tres meses antes del cierre del padrón electoral y nueves meses antes del día de la votación. Es esperable que para dichas fechas muchos ciudadanos actualicen sus DNIs, lo cual podría intensificar o disminuir los efectos detectados.

En segundo lugar, debe considerarse que la información contenida en el RUIPN proviene de datos declarados por los propios ciudadanos en su última actualización del DNI. En muchos casos, estas variables pueden no haberse actualizado en periodos largos de tiempo, lo que introduce un sesgo potencial en el análisis. Este riesgo es especialmente relevante en el caso de los ciudadanos con el DNI caducado pues los mismos llevarían más de 8 años (tiempo que dura vigente un DNI) sin actualizar su información. En conjunto, estas limitaciones deben ser tenidas en cuenta al interpretar los resultados, ya que podrían influir en la magnitud de algunos de los efectos estimados.

3. ANÁLISIS DE TWEETS: DESINFORMACIÓN Y SENTIMIENTOS.

Para la elaboración de los gráficos y el análisis, se realizó una recolección de datos mediante el uso de APIs y web scraping obteniendo aquellas publicaciones que hicieran referencia a Reniec en la red social “X” (antes Twitter). El proceso de recolección permitió extraer una muestra de 1037 tweets del año 2021 entre enero y diciembre. Si bien esta estrategia fue una muestra aleatoria, la mayoría de observaciones (tweets) pertenecen a la segunda parte del año, es decir, de manera posterior al proceso electoral vigente.

Se realizó un procesamiento y limpieza de la base de datos construida a partir de los tweets, se filtró aquellos tuits provenientes de autoría del twitter oficial del RENIEC. De manera exploratoria, se realizó el procesamiento del texto para la construcción de gráficos estilo “wordcloud”.

Para analizar la información de los tweets, se realizó un procesamiento a través de la librería “Spanish - NLP” , un modelo de inteligencia artificial entrenado con texto en español para procesar y argumentar texto-data en español[1]. El modelo interpreta el texto ofrecido y lo evalúa a partir de ciertos clasificadores de comunicación en un rango entre 0 y 1. Aquellos valores más cercanos al 1 indican mayor presencia de dicha categoría.

Categorías:

a. Discurso de Odio (hate speech)
b. Incivilidad (incivility)

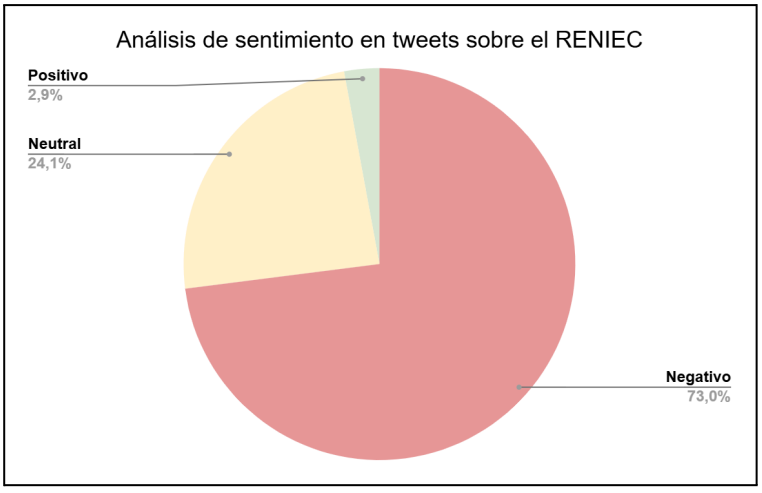
[1] En el siguiente enlace se puede acceder al repositorio oficial del modelo: https://github.com/jorgeortizfuentes/spanish_nlp

c. Análisis de Sentimiento (sentiment analysis)
d. Análisis de emociones (emotion analysis)
e. Ironía (irony)
f. Sexismo (sexist)
g. Racismo (racismo)

Esta metodología, aunque innovadora, ya tiene algunos precedentes en el estudio de las redes sociales en los contextos electorales latinoamericanos (Ponte Torrel, J.,2022 ; Rodriguez, 2025 ; Blanco-Fernández et al. 2024). Existe amplia evidencia sobre la efectividad de los modelos pre entrenados para poder clasificar y categorizar corpus de texto en idiomas diferentes al inglés de base sobre el que se construyen y entrenan los Large Language Models (LLMs) cómo Bert, RoBERTo, RoBERTuito, etc. Aunque el modelo seleccionado, *Spanish - NLP*, seguramente ha sido "afinado" en su capa final con textos provenientes del contexto chileno, es bastante poderoso para poder predecir y efectuar análisis de sentimiento.

De hecho, recientemente se han dado diferentes iniciativas similares en el campo de la observación electoral y la comunicación política. Por ejemplo, desde hace unos años PNUD lidera el uso de estas herramientas con una aplicación propia especializada en el combate de la desinformación y la prevención de conflictos electorales. Con el uso de eMonitor+ , PNUD ha contribuido con una poderosa herramienta en la automatización del análisis de redes sociales. Recientemente, en el informe "La institucionalidad bajo ataque electoral" (PNUD, 2024) presentaron algunos de los resultados de su aplicación en el contexto electoral peruano del 2021.

Resultados:



El modelo identifica altos niveles de sentimiento negativo. Encontramos que un 73% de tweets contienen contenido clasificado con sentimiento negativo. Esto confirma la existencia de contenido posiblemente desinformador, bajo la estrategia de la deslegitimación de las instituciones electorales. Si bien es cierto que el modelo no identifica o realiza verificación de tipo “fact-checking”, en diferentes manuales sobre la lucha contra la desinformación se considera la difusión de contenido tóxico sobre los organismos electorales cómo parte de las estrategias de los sujetos desinformadores y/o individuos antidemocráticos para socavar la confianza electoral y deslegitimar la elección.

Bibliografía:

Ponte Torrel, J. (2022). La campaña peruana en Twitter. Análisis de la polarización afectiva durante la segunda vuelta de las elecciones generales 2021. Cuadernos.info, (53), 138-161. https://doi. org/10.7764/cdi.53.49539

Blanco-Fernández, Y., Otero-Vizoso, J., Gil-Solla, A., & García-Duque, J. (2024). Enhancing Misinformation Detection in Spanish Language with Deep Learning: BERT and RoBERTa Transformer Models. Applied Sciences, 14(21), 9729. https://doi.org/10.3390/app14219729

Rodriguez (2025) Radar Político by FK Economics © 2025. ¿Qué dicen los candidatos presidenciales? ¿Qué temas priorizan?. https://www.radarpolitico-fk.com/

Programa de las Naciones Unidas para el Desarrollo. (2 de abril de 2024). Emonitor+: IA contra la desinformación. *PNUD Perú*. https://www.undp.org/es/peru/noticias/emonitor-ia-contra-la-desinformacion

La institucionalidad electoral bajo ataque (2024) https://www.transparencia.org.pe/publicaciones