

Reporte de resultados 2

Asistente de investigación: Erick Gabriel Fajardo Martínez
Investigador: Dr. Gabriel Purón Cid

2022-06-29

Descripción de los modelos

En este segundo intento se compara el desempeño de tres modelos entrenados. El primer modelo (percepción sin normalizar) fue creado utilizando la pregunta de percepción de la ENCIG para construir la clasificación de los municipios en corruptos y no corruptos. El segundo modelo (percepción) es similar al primero ya que también se utilizó la pregunta de percepción para construir la clasificación de corrupción, sin embargo, este modelo se entrenó con los datos normalizados y se añadieron dos variables explicativas más (grado promedio de escolaridad y población total). El tercer modelo (incidencia) es igual al segundo, es decir, también fue entrenado con los datos normalizados y variables explicativas adicionales; la diferencia está en la construcción de la clasificación de corrupción la cual fue hecha con la pregunta de incidencia de la ENCIG.

Igual que en el primer reporte, para el primer modelo se cuenta con información de 233 municipios y 234 variables de egresos, las cuales están al nivel de partidas. Mientras que para los dos modelos nuevos de percepción e incidencia, se cuenta con información del mismo número de municipios y 236 variables.

Normalización de los datos de egresos

Atendiendo a lo comentado en la reunión con la Dra. Daniela Moctezuma, esta vez se procedió a la normalización de los datos. Los modelos de percepción e incidencia cuentan con esta normalización, la cual fue hecha de la siguiente manera:

$$Z_i = \frac{X_i}{PEA_i}$$

donde : X = Partida de egresos, i = Municipio

Clasificación de corrupción

A diferencia del primer reporte donde la clasificación fue hecha de la siguiente manera:

$$Corrup = \begin{cases} 1 & \text{si proporción que contestó muy frecuente} > 0 \\ 0 & \text{si proporción que contestó muy frecuente} = 0 \end{cases}$$

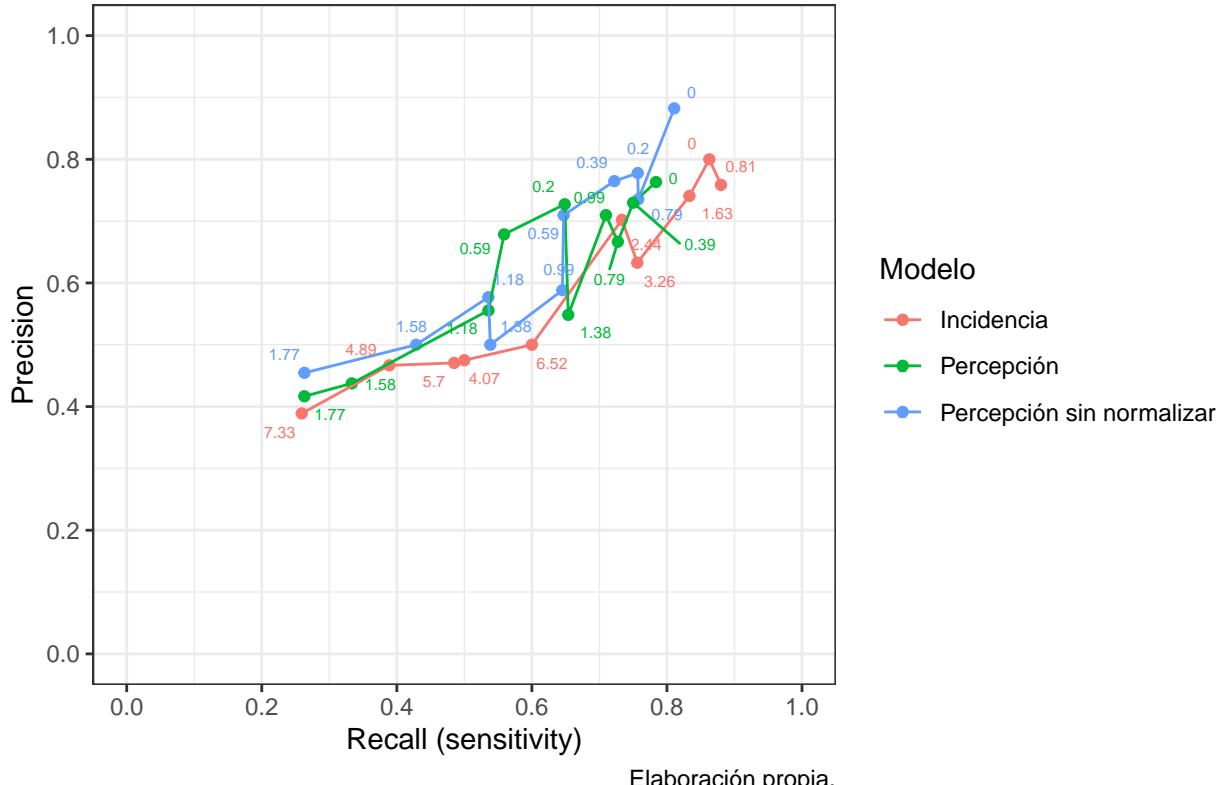
En los modelos de percepción e incidencia se optó por probar con varios umbrales con el objetivo de encontrar aquel que **maximizara la sensibilidad del modelo**. Recordemos que la sensibilidad es la capacidad del modelo para clasificar un resultado positivo cuando este resultado es positivo en la realidad.

En el modelo de percepción los umbrales corresponden a la proporción de encuestados que contestó que el fenómeno de corrupción es algo muy frecuente en su lugar de residencia, mientras que para el modelo de incidencia los umbrales corresponden a la proporción de encuestados que contestó que sí han estado involucrados en un acto de corrupción.

Curvas de precision y recall

Para elegir el umbral óptimo en ambos modelos se calcularon y entrenaron varios modelos para capturar y graficar sus valores de precision y recall. Ambas métricas permiten conocer la capacidad del modelo para clasificar correctamente los valores positivos, en este caso que los municipios sean corruptos. Ambos valores se encuentran entre 0 y 1, y se busca que ambos valores sean lo más cercano a 1.

Precision–Recall



Comparación de los modelos

Con base en las métricas de la siguiente tabla es posible percibir que cada modelo tiene sus ventajas y desventajas.

El primer modelo de percepción sin normalizar es bueno clasificando los municipios que no son corruptos (alta especificidad) y tiene un desempeño regular al clasificar los municipios corruptos (sensibilidad).

El segundo modelo de percepción tiene un buen desempeño al clasificar las dos categorías, es el modelo más balanceado. Sin embargo, el umbral óptimo para alcanzar este balance es 0, es decir, clasificar a los municipios como corruptos cuando la proporción de encuestados que respondió que la corrupción es un fenómeno muy frecuente sea mayor a 0.

El tercer modelo, el de incidencia, es muy bueno para clasificar los municipios corruptos, sin embargo, también asigna municipios que en realidad no son corruptos a esta categoría. Por lo tanto, es el modelo menos balanceado.

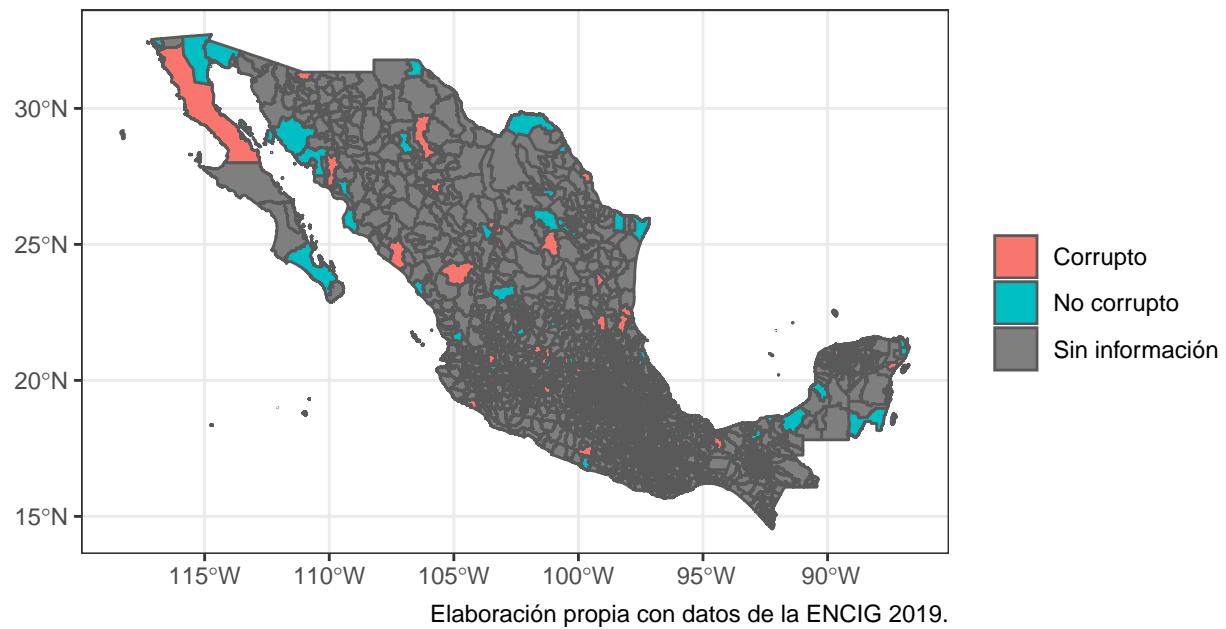
Modelo	Umbral	Sensibilidad	Especificidad	F1_score	Precisión_balanceada
Percepción sin normalizar	0.79	0.62	0.84	0.71	0.73
Percepción	0.00	0.78	0.72	0.77	0.75
Incidencia	0.81	0.88	0.26	0.81	0.57

Mapas modelados

Mapa de valores reales

Clasificación **real** de corrupción de los municipios

Total de municipios encuestados en la ENCIG: 233

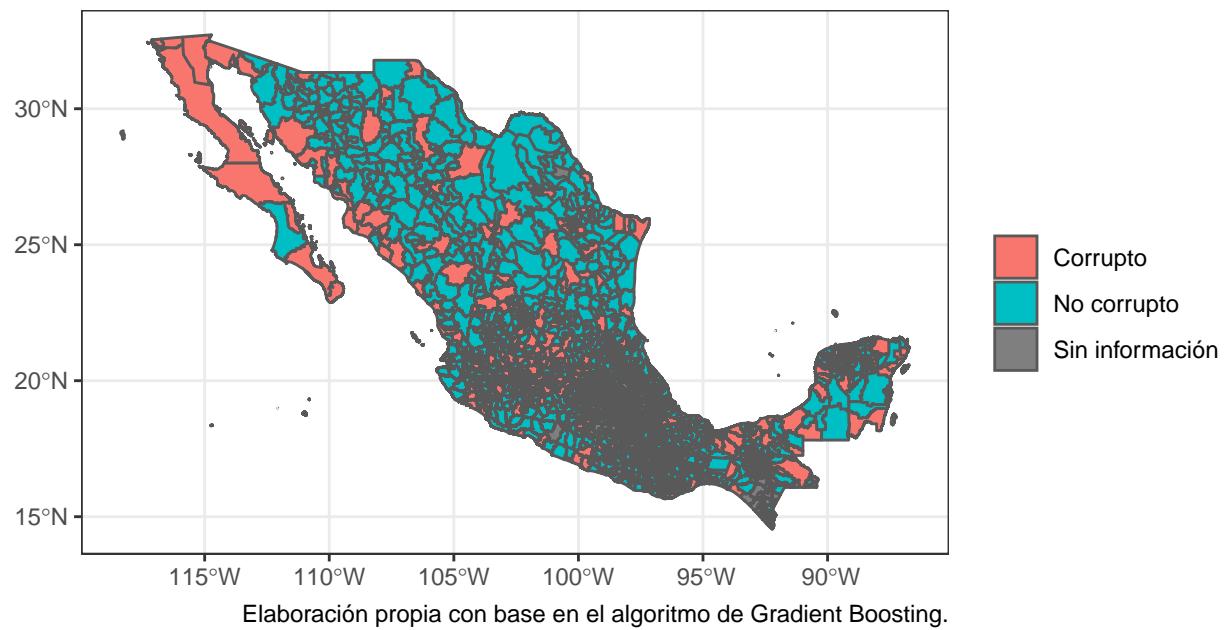


Modelo de percepción sin normalizar

Clasificación **predicha** de corrupción de los municipios

Total de municipios modelados: 2121

Sensibilidad del modelo: 0.62 | Especificidad: 0.84



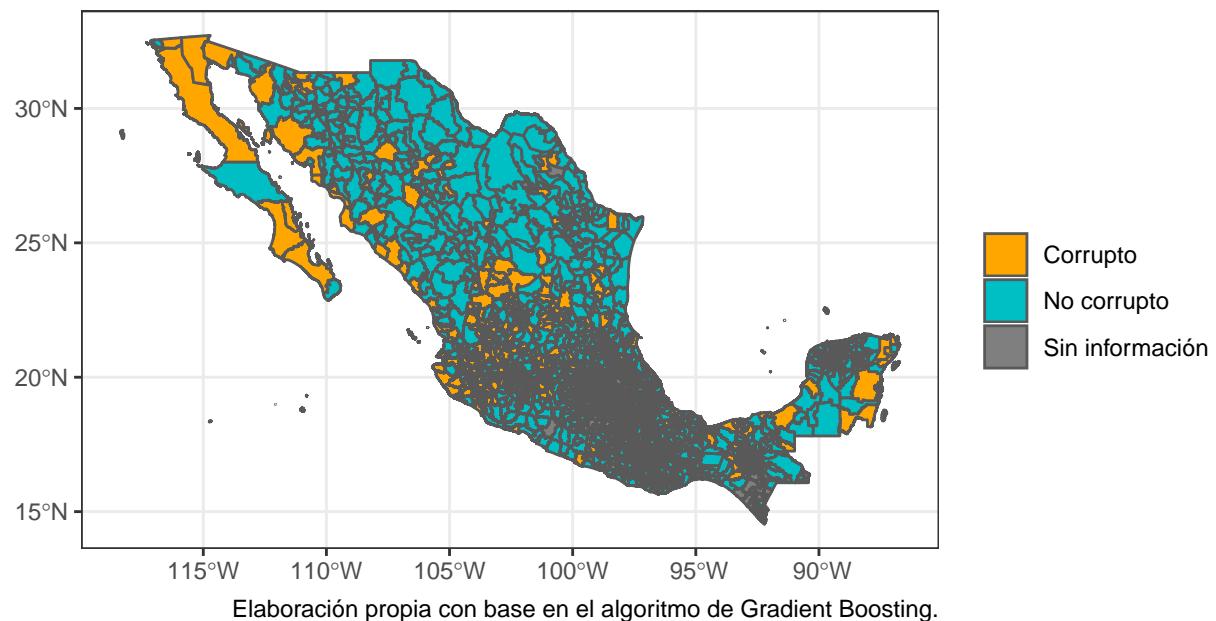
Modelo de percepción

Clasificación **predicha** de corrupción de los municipios

Modelo de percepción

Total de municipios modelados: 2121

Sensibilidad del modelo: 0.78 | Especificidad: 0.72



Modelo de incidencia

Clasificación **predicha** de corrupción de los municipios

Modelo de incidencia

Total de municipios modelados: 2121

Sensibilidad del modelo: 0.88 | Especificidad: 0.26

