Reporte de resultados 3

Ingresos

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

2022-07-14

Descripción de los modelos

En este tercer reporte se realiza la modelación con las partidas de **ingresos** de cada municipio, y de igual manera se compara el desempeño de los modelos utilizando el reactivo de percepción y de incidencia de corrupción. Ambos modelos cuentan con datos normalizados por la población económicamente activa y cuentan con variables de contexto como la población total del municipio y grado promedio de escolaridad.

Posteriormente se pretende agregar más variables de contexto que correspondan a indicadores económicos y políticos de los municipios.

En este caso se cuenta con información de 233 municipios y 141 variables de ingresos, las cuales son menos que las 234 variables de egresos (partidas de egresos).

La clasificación de corrupción de los municipios se realizó igual que en los reportes anteriores: 1) Primero se calcularon las tasas de percepción e incidencia de cada municipio con base en los datos de la ENCIG 2019; 2) Se asignó la etiqueta de corrupción a los municipios que cuyas tasas de percepción o incidencia fueran mayores a un umbral arbitrario de dichas tasas. Este umbral fue elegido a través de un método de optimización con la ayuda de curvas ROC (Receiver Operating Characteristic), donde se selecciona el umbral que maximice la sensibilidad del modelo.

Curvas ROC

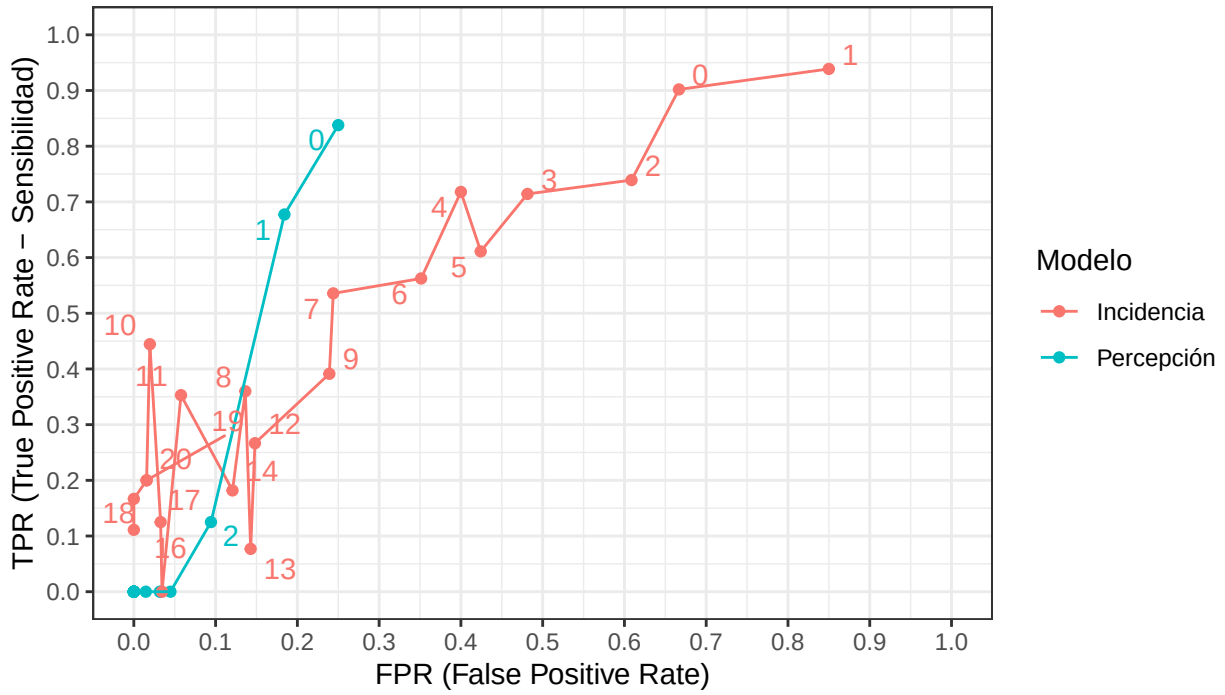
Para elegir el umbral óptimo, se entrenaron varios modelos para capturar y graficar sus tasas de positivos verdaderos (True Positive Rate - TPR/sensibilidad) y falsos negativos (False Positive Rate - FPR). Ambas métricas permiten conocer la capacidad del modelo para clasificar correctamente los valores positivos, en este caso que los municipios sean corruptos. Ambos valores se encuentran entre 0 y 1, y se busca que la razón de positivos verdaderos sea alta (eje Y).

Cada punto de la gráfica 1 corresponde a los valores obtenidos de TPR y FPR de cada modelo y los números que acompañan a cada punto son los valores de las respectivas tasas de percepción e incidencia (umbrales) con las cuales se realizó la clasificación de corrupción, por ejemplo, el punto más alto de la línea roja está etiquetado con el valor de 1, por lo tanto, en ese caso si el municipio cuenta con una tasa de incidencia mayor al umbral 1 es clasificado como corrupto.

En la gráfica 1 es posible observar que el modelo de percepción tiene una buena sensibilidad (eje Y) únicamente cuando se clasifican como corruptos a los municipios con tasas de percepción de entre 0 y 1. Esta sensibilidad se ubica aproximadamente entre los valores de 0.68 y 0.85. En cambio, el modelo de incidencia tiene un mejor desempeño; en primer lugar, es posible clasificar a los municipios como corruptos en un rango más grande en términos de tasa de incidencia, dicho rango va desde 4 a 0; en segundo lugar, la sensibilidad para este rango se ubica entre 0.70 y 0.95, lo cual supera al modelo de percepción. Sin embargo, es más probable que el modelo de incidencia clasifique municipios como corruptos cuando en realidad no lo son. Esto se mide con el eje X, que es la tasa de clasificación de falsos positivos.

En resumen, el modelo de percepción es confiable en un pequeño rango del umbral y no es tan probable que clasifique como corruptos a municipios que no lo son. El modelo de incidencia tiene un mejor desempeño en un rango más grande del umbral y logra obtener una sensibilidad mayor, pero es más probable que clasifique más municipios como corruptos cuando no lo son.

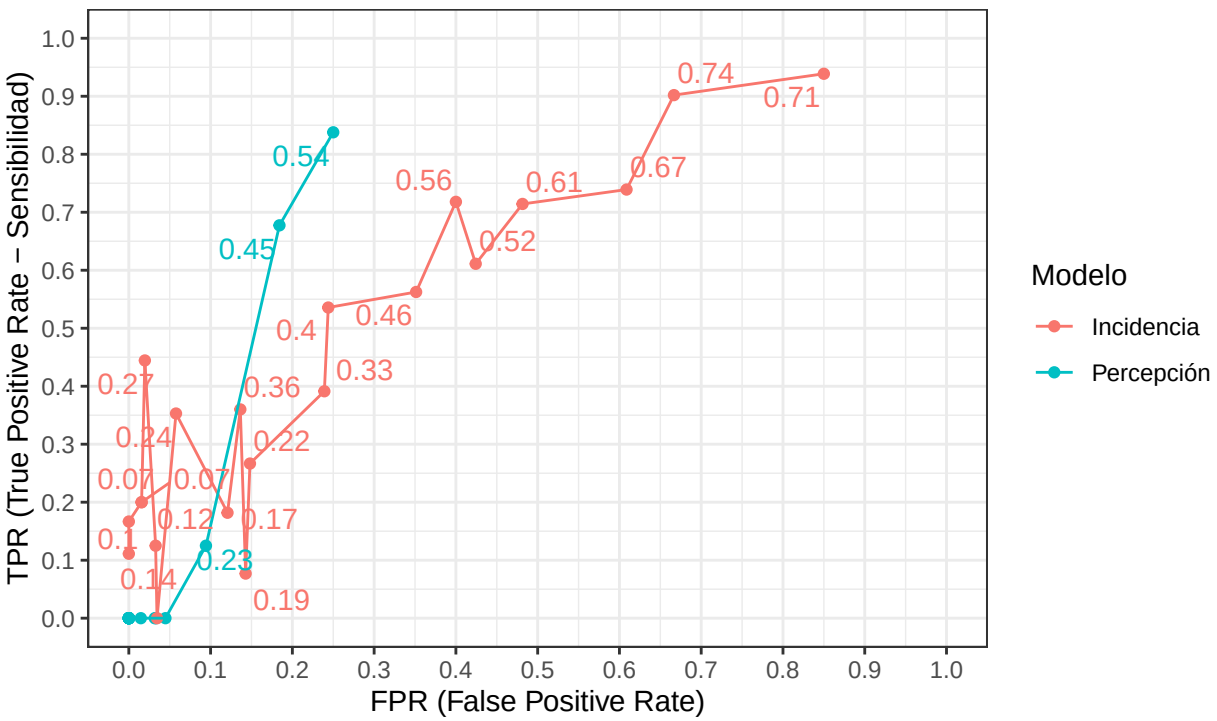
Gráfica 1: Curvas ROC − Umbrales



Elaboración propia.
Nota: El rango de los umbrales va desde 0 hasta 20, aumentando en una unidad.

Algo que tienen en común ambos modelos es que su sensibilidad aumenta cuando la prevalencia de municipios corruptos es balanceada, es decir, cuando se ubica cercana al 50%. Esto se puede apreciar en la gráfica 2, donde las etiquetas de los puntos ahora representan la prevalencia asociada a cada clasificación. Otro aspecto importante, es que existe una relación negativa entre la magnitud del umbral y la prevalencia, lo cual se traduce en una menor sensibilidad.

Gráfica 2: Curvas ROC − Prevalencia



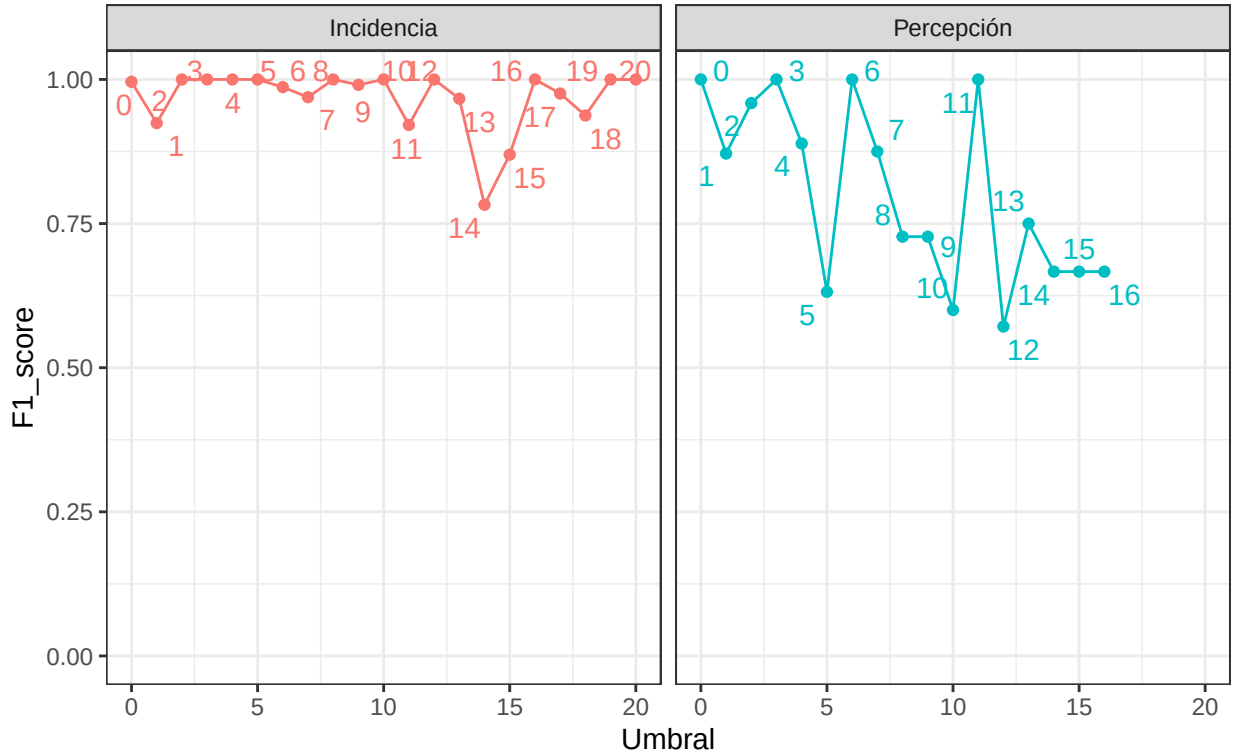
Elaboración propia.

F1-score

Otra manera de optimizar el desempeño de los modelos es compararlos en términos de la métrica F1-score, la cual es una ponderación entre la sensibilidad y especificidad del modelo. Entre más cercano a uno, el modelo contará con un mejor balance de sensibilidad y especificidad, es decir, será mejor en clasificar correctamente a los municipios como corruptos y no corruptos.

En la gráfico 3 es posible observar que el modelo de incidencia es mucho más consistente en casi todos los valores del umbral.

Gráfica 3: F1−score



Elaboración propia.

Comparación de los modelos

Con base en lo mostrado en la gráfica 1 de las curvas ROC, se seleccionó el rango de umbrales de 0 a 4 para entrenar 8 modelos: 4 de percepción y 4 de incidencia. La comparación se muestra en la tabla 1:

Table 1: Comparación.

Modelo	Umbral	Prevalencia	Sensibilidad	Especificidad	F1_score	Precisión_Balanceada
Percepción	0	0.54	0.84	0.75	0.82	0.79
Percepción	1	0.45	0.68	0.82	0.71	0.75
Percepción	2	0.23	0.12	0.91	0.17	0.52
Percepción	3	0.13	0.00	0.97	NA	0.48
Percepción	4	0.09	0.00	0.97	NA	0.48
Incidencia	0	0.74	0.90	0.33	0.84	0.62
Incidencia	1	0.71	0.94	0.15	0.82	0.54
Incidencia	2	0.67	0.74	0.39	0.72	0.57
Incidencia	3	0.61	0.71	0.52	0.71	0.62
Incidencia	4	0.56	0.72	0.60	0.71	0.66

Es posible apreciar que, como se mencionó anteriormente, los modelo de percepción solo son

confiables en el rango de umbrales de 0 a 1, puesto que sus niveles de sensibilidad disminuyen drásticamente a partir del umbral con valor de 2. El umbral que maximiza la precisión de este modelo es el que toma el valor de 0.

Por otro lado, los modelos de incidencia tienen un desempeño constante el cual destaca por sus buenos niveles de sensibilidad pero es débil en cuanto a sus niveles de especificidad, esto quiere decir que puede clasificar más municipios como corruptos de los que en realidad hay, no obstante, es muy seguro que no se equivoque con los que sí son corruptos. En este caso el umbral que nos otorga la máxima precisión es el que toma el valor de 4.