Métodos de aprendizaje de ḿaquina para inferir el nivel de cobertura de banda ancha fija en municipios de México

César Zamora Martínez[a]

[a]Alumno de Maestría en Ciencias de Datos (ITAM)

Aunque en fechas recientes se reconoce el impacto benéfico que la banda ancha tienen sobre en entorno económico y social, la penetración de tales servicios en los municipios obedece a múltiples factores que inciden en el despliegue de la infraestructura que permite su prestación. Motivado por ello, en este trabajo se plantea el uso de métodos basados en aprendizaje de máquina que permitan clasificar los municipios conforme a su nivel de cobertura a través de indicadores de penetración y establecer factores que propician o desincentivan los despliegues de banda ancha fija.

Durante las últimas tres décadas las telecomunicaciones han tenido un avance sin precedentes en el mundo, posicionándose como herramientas que potencian el desarrollo económico y social, pues, como ha sido ampliamente documentado en la literatura (Pradhan et al. (2014)), permiten crear oportunidades, reducir la pobreza e impulsar el progreso económico y social para el bienestar de la población[*]. Uno de los ejes que permiten explicar lo anterior es el impacto benéfico de los servicios de banda ancha en los procesos productivos, financieros y en general el bienestar de la población (Katz (2012)).

En México, a cerca de cuatro años de la reforma a de telecomunicaciones en 2013, que llevo a la promulgación de la Ley Federal de Telecomunicaciones y Radiodifusión (LFTyR) junto con la creación del Instituto Federal de Telecomunicaciones (IFT), se estimó un crecimiento superior al 37% en las conexiones de banda ancha fija (BAF), traduciéndose a que para entonces casi la mitad los hogares contaban con servicios de Internet (IFT (2017) e IFT (2018a)).

Aunque este fenómeno revela una tendencia favorable con respecto al entorno internacional[†], la Encuesta sobre Disponibilidad y Uso de Tecnologías de la Información y la Comunicación en los Hogares 2018 (ENDUTIH 2018, INEGI (2018)) dejó en claro la existencia de una brecha en la adopción de estos servicios para la población mexicana (por tanto en sus conducentes beneficios), pues sólo cerca de 65.8% de la población con seis años o más es usuario de servicios de Internet en los hogares del país, además de que mostró que este es un fenómeno urbano, puesto el 73.1% del total de la población urbana son usuarios de este servicio en contraste con la población conectada en zonas rurales que es cercana a 40.6%.

A efecto de explicar el entorno de la penetración de servicios de Internet a nivel municipal, en adelante nos centraremos en los servicios de banda ancha fija[‡], los cuales son servicios de acceso a Internet y transmisión de datos orientados a usuarios finales (personas físicas o empresas), que se brindan a través de equipos terminales (módems, terminales ópticas y demás) que tienen una ubicación geográfica determinada y fija (IFT (2018d)). Ello obliga a los operadores de telecomunicaciones interesados a realizar inversiones que les permitan alcanzar los puntos geográficos en donde se localizan los clientes potenciales, esto es, cerca de hogares y edificios de empresas, aprovechando las capacidades de las tecnologías en las que se basan sus redes.

Dicho contexto les condiciona a establecer un circuito físico o virtual a través del cual se pueda conectar la ubicación del usuario a la red del operador y a través del que se prestarán los servicios ("Acceso de datos" o simplemente como "acceso", IFT (2018d)). Por ende, dado que afrontan costos considerables en infraestructura, equipos, permisos y recursos humanos para poder brindar servicios[§], típicamente los operadores concentran su oferta en zonas densamente pobladas donde existe suficiente capacidad económica para asegurar no solo que recuperarán sus inversiones sino que serán rentables desde la visión de negocio.

Además de los aspectos socio-económicos, también se destacan otros factores que pueden ser tomados en cuentan por un operador para evaluar una zona como idónea para brindar servicios: 1) Viabilidad de permisos para desarrollar los despliegues (e.g. concesiones para operar, medio ambiente), 2) viabilidad tecnológica (e.g. limitadas técnicas por la distancia que limitan la velocidad, calidad, entre otras), 3) existencia de infraestructura cercana a la zona de la que puedan disponer para proveer servicios (por ejemplo, propia o arrendada); y 4) existencia de competencia en el área; es decir de proveedores de servicios de telecomunicaciones.

Para cuantificar la cobertura de banda ancha fija la OCDE define una medida de penetración en una zona como la cantidad de accesos en ella por cada 100 habitantes, el cual es un proxy del indicador de suscriptores por cada 100 habitantes[¶]:

$$PenBAFHabitantes = \frac{Accesos}{Habitantes} \times 100 \qquad [1]$$

En este sentido, de acuerdo a la información publicada por este organismo a diciembre de 2018, sus países miembros contaban con una penetración media de 30.92 accesos por

[*] Katz (2018) muestra que un avance del 1% en un índice sobre digitalización, genera un incremento de la productividad que redunda en un crecimiento económico de un 0.3% del PIB.

[†] A finales de 2018 México fue el cuarto país con mayor crecimiento de penetración de banda ancha fija entre los países de la Organización para la Cooperación y Desarrollo Económicos (OCDE); y mostró un crecimiento de 17.9% en la penetración de accesos por medio de fibra óptica (IFT (2019))

[‡] Ello dado que desafortunadamente, el Banco de Información de Telecomunicaciones (BIT) del IFT sólo posee el detalle desagregado de servicios de Internet de banda ancha móvil para nivel estado, sin que se hayan podido localizar fuentes con datos precisos al respecto.

[§] En línea con Escobar (2017), no sólo se enfrentan costos directos, sino oportunidad y de transacción; así como el riesgo de afrontar costos hundidos.

[¶] http://www.oecd.org/internet/broadband/broadband-faqs.htm

[2]E-mail: czamora5email.itam.mx

cada 100 habitantes, ocupando México el penúltimo lugar (con 14.83 accesos por cada 100 habitantes).

Por otro lado, para BAF los accesos se basan típicamente en las tecnologías[‖] (Moya (2014)): 1) DSL: tecnología de transmisión por cable trenzado de cobre, su disponibilidad y la velocidad dependen de la distancia; 2) Cable coaxial: se forma por dos hilos de cobre cuya estructura permite permite, en generala más capacidad para transmitir información que el par trenzado de cobre; 3) Fibra óptica: formados por un medio de vidrio o polímero que permite el paso de haces de luz, pueden transmitir más de 10 Gbit/s hasta a 10 kilómetros de distancia, 4) Otras: incluye uso de ondas electromagnéticas como microondas, señales satélites y demás; usualmente tienen menor desempeño comparado a 1), 2) o 3), pero son viables en regiones de difícil acceso.

Con todo lo anterior, el objetivo del presente documento será plantear un modelo con métodos de aprendizaje de máquina que permitan identificar variables útiles para explicar el nivel de penetración de BAF (por ejemplo, datos geográficos y demográficos) así como de los factores que inciden en los despliegues de tecnologías capaces de dar servicios de Internet de alta velocidad.

En este sentido, con motivo de estudiar a los municipios que cuentan penetración de BAF basada en tecnologías capaces de dar servicios de velocidad alta, se enfocarán los indicadores de penetración presentados previamente sobre accesos correspondientes a tecnologías de cable coaxial o fibra óptica (es decir, se calcularan con respecto la cantidad de accesos resultado de sumar de los que correspondan a cable coaxial y aquellas de fibra óptica en cada municipio).

1. Revisión y análisis de fuentes de datos asociados a banda ancha fija

A continuación se resumen las fuentes de información consultadas con relación a servicios de BAF, junto con las consideraciones particulares derivadas de su exploración[**].

A. Revisión de fuentes de información. En términos generales, la revisión abarcó datos públicos de fuentes gubernamentales y organismos internacionales relativos a siguientes los ejes:

A.1. Identificación de municipios. Dado que la disponibilidad de información social y demográfica en fuentes públicas con desglose municipal se encuentra limitada a ejercicios estadísticos que abarcan hasta el año 2015 (INEGI (2015), CONAPO (2015), ONU (2015)), la identificación de los municipios se hizo de manera congruente el marco metodológico de la Encuesta Intercensal 2015 en donde se contabilizaron un total de 2,457 municipios.

A.2. Datos de accesos de banda ancha. El Banco de Información de Telecomunicaciones (IFT (2018b)) posee datos históricos (de 2013 a mediados 2019) sobre los accesos de banda ancha móvil y fija de México; sin embargo únicamente en el segundo caso se ofrece el detalle a nivel municipio. Dicha fuente provee datos de 29 empresas a las que pertenecen los accesos de BAF junto tecnología correspondiente (DSL, cable coaxial, fibra óptica, satelital y otras), sin proveer el desglose entre accesos residenciales o no residenciales. Se consideró relevante extraer los datos accesos en cada municipio, agregando los datos de todos los operadores por tecnología de acceso para el periodo con información más reciente (junio de 2019).

A.3. Datos socio-económicos. A través de la Encuesta Intercensal 2015 (INEGI (2015)), el INEGI reúne información de componentes que describen la evolución de la población, junto con sus viviendas y de sus condiciones socio-económicas.Tras estudiar esta fuente, los datos que se han considerado de interés para el estudio a nivel municipal la penetración de BAF son: 1) número de hogares, 2) número de habitantes, 3) porcentaje de viviendas que cuentan con disponibilidad de servicios de telecomunicaciones (es decir, a través de telefonía fija, telefonía celular, televisión de paga e Internet)[††].

A.4. Datos asociados a marginación. La Comisión Nacional de Población (CONAPO) diseño una herramienta denominada "índice de marginación" para realizar mediciones de las carencias que padece la población (CONAPO (2015)). Esta recoge una serie de dimensiones socio-económicas para su construcción: educación, vivienda, y distribución de la población e ingresos. Para explorar la relación de tales ejes con la penetración de servicios de BAF, se consideró relevante involucrar a las siguientes mediciones empleadas por CONAPO en 2015 para la construcción de dicho índice a) porcentaje de ocupantes en viviendas sin energía eléctrica, b) porcentaje de población en localidades con menos de 5,000 habitantes, y c) porcentaje de población ocupada con ingresos de hasta 2 salarios.

La elección de tales variables se basa en que: a) la energía eléctrica es una condición necesaria para el funcionamiento de los servicios de telecomunicaciones, b) los operadores deben de hacer mayores esfuerzos para atender zonas que son bajamente pobladas donde puede no ser rentable llevar servicios, y c) el ingreso de las personas constituye una condición necesaria para que puedan solventar el pago de un servicio de Internet.

A.5. Datos asociados a desarrollo humano. El Programa de las Naciones Unidas para el Desarrollo (PNUD) es una organización orientada a generar soluciones a los países que buscan alcanzar sus metas de desarrollo y lograr los objetivos compartidos con la comunidad internacional. Como parte de sus actividades en México ONU (2015), periódicamente evalúan el nivel de desarrollo de los municipios a través de la construcción del "Índice de Desarrollo Humano (IDH)", el cual considera los ejes de salud, educación e ingreso. Al respecto, puesto que el ingreso de la población es un componente relevante en el acceso a servicios de telecomunicaciones, se considerarán los datos del PNUD empleados para la estimación del IDH correspondientes al año 2015 relativos el ingreso bruto per cápita en poder de paridad de compra (PPC), expresado en dólares estadounidenses, de cada municipio[‡‡].

[‖] También existen configuraciones híbridas

[**] El procesamiento de la información se llevó a cabo a través de scripts en Bash, R y Python, véase apéndice

[††] Ante la ausencia de información pública sobre la presencia de infraestructura (ver Escobar (2017)) que permita proveer servicios de Internet de alta velocidad, por ejemplo la localización de centrales de redes de nueva generación o la presencia física de la red backhaul/core de los principales operadores de México o CFE, se estima que la información indirecta de la presencia de servicios constituye un proxy para la presencia de infraestructura necesaria, aunque no suficiente, para poder brindar tales servicios a usuarios finales. A su vez, dicha elección se apoya en que, como se ha mencionado previamente, los despliegues de redes fijas se realizan alrededor de donde se ubican los clientes potenciales, y a su vez en el hecho de que la presencia de servicios de telecomunicaciones (incluso distintos a BAF) constituye una señal de que existen condiciones positivas para que los operadores desarrollen una cadena de elementos de infraestructura y operación que le permitan atender en una zona específica ofreciendo servicios a la población

[‡‡] Estos datos no se encuentra disponibles para 11 municipios. A saber 5 del estado de Chihuahua (Buenaventura, Carichí, Santa Isabel, Temósachic, y Urique), 4 en Oaxaca (Matías Romero Avendaño, Santa María Chimalapa, Santa María Petapa y San Francisco Chindúa), uno en Puebla (San Nicolás de los Ranchos) y otro en Sonora (General Plutarco Elías Calles)

B. Consideraciones sobre la información reunida. En términos de la información reunida, la exposición previa (así como la metodología propuesta por IFT (2018c)), el análisis de cobertura para servicios de BAF partirá de las siguientes premisas:

- **Periodo de la información de accesos:** junio de 2019,
- **Nivel de desagregación geográfico:** municipal considerando los municipios de la Encuesta Intercensal 2015 (INEGI (2015)),
- **Indicador de cobertura:** se considerará la penetración por cada 100 habitantes (ver ec. 1),
- **Tipo tecnologías de acceso para establecer cobertura:** cable coaxial y fibra óptica, bajo la premisa de que pueden ser usadas para proveer servicios con velocidades mayores que otras opciones,
- **Identificación de municipios con cobertura de BAF:** aquellos que son reportados, de manera explícita en el BIT, como con presencia de accesos basadas en cable coaxial o fibra óptica[§§].
- **Identificación de municipios sin cobertura de BAF:** son aquellos para los que no hay reportes de presencia de accesos basadas en cable coaxial o fibra óptica.
- **Periodo de datos socio-económicos, de hogares y población:** conforme a la última disponibles, es decir, 2015.

Se debe reconocer que existe un desfase entre los datos socio-económicos recopilados (a 2015) y los datos de accesos de BAF (a junio de 2019); sin embargo, se considera que al menos, a alto nivel, a través de tales fuentes se pueden delinear patrones de hacia donde se han orientado los despliegues de los operadores en México que sustentan el estatus de cobertura de los servicios de banda ancha fija.

En este sentido, el estudio en cuestión se inclina a evaluar si la información socio-económica, de hogares y población puede ser de utilidad para inferir el nivel de cobertura los municipios del país, desde la perspectiva de modelos de aprendizaje de máquina, sin perjuicio de que, cuando se actualicen la información relevante, el análisis pueda complementarse y actualizarse con la mejor información disponible.

C. Análisis de información.

C.1. Datos de accesos de servicios de banda ancha fija. Se observó que a junio de 2019, existían 18.85 millones de accesos de BAF distribuidos sobre 1,604 municipios el país (de un total de 2,457), para los que en un porcentaje de 2.76% de los accesos no había datos de su ubicación. En complemento, el 37.1% de tales accesos son DSL, 38.3% cable coaxial, 21.8% fibra óptica, 0.125% satelitales, mientras que para el cerca de 2.6% se desconoce la tecnología.

En términos de los principales grupos de telecomunicaciones de México a los que pertenecen los accesos de BAF, se tiene la siguiente distribución: América Móvil[¶¶] 51.6%, Grupo Televisa[***] 23.3%, Megacable-MCM[†††] 16.1% y Totalplay 7.35%

Otro punto a destacar es que únicamente los empresas que pertenecen a dichos grupos han desplegado accesos basados en cable coaxial o fibra óptica en el país.[‡‡‡].

No se omite destacar la existencia de municipios con una cantidad de accesos inusualmente baja (e.g. hay 31 municipios con un sólo acceso, 7 que poseen únicamente dos accesos).

Table 1. Distribución de accesos de BAF por tecnología (Junio 2019)

Grupo	Coaxial	DSL	Fibra	Satelital	No especificado
América Móvil		71.9%	28.1%		
Grupo Televisa	95.6%			0.1%	4.3%
Megacable-MCM	99.9%				0.1%
TotalPlay			100%		

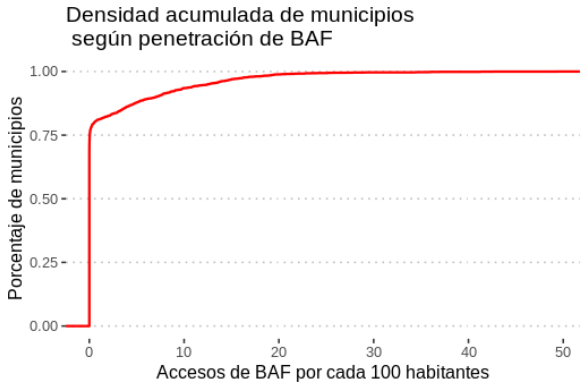

Fig. 1. Distribución acumulada de municipios según su penetración de BAF

C.2. Heterogeneidad de la penetración en los municipios. Para ilustrar el panorama de la diferencias en la cobertura en los diferentes municipios del país, se construyó la distribución acumulada empírica de los municipios según su penetración de accesos de BAF basados en cable coaxial y fibra óptica. Se aprecia que dicho parámetro oscila entre 0 y cerca de 50 accesos por 100 habitantes a nivel municipal, además casi el 75% de los municipios del país no cuenta con penetración de BAF. En contraste, únicamente 10% de los municipios supera los 10 accesos de BAF de este tipo por cada 10 habitantes.

En complemento, también se elaboraron mapas de calor[§§§] que responde a la cantidad de accesos de BAF (basados en fibra óptica y cable coaxial) por cada 100 hogares eligiéndose tres zonas diferentes del país: i) península de Baja California (figura 2) y iii) Ciudad de México (figura 4).

En términos generales, para las zonas de la península se aprecia mayor penetración en municipios fronterizos o zonas turísticas. Para la zona de la Ciudad de México, se aprecia una mayor penetración en la delegaciones Benito Juárez, Cuahutemoc y Miguel Hidalgo, mientras que Milpa Alta y Tláhuac son las zonas con menores niveles observados.

C.3. Estratificación de municipios de acuerdo a su nivel de cobertura BAF. Dado que, como se hizo notar en la sección previa, la

[§§] Como se verá posteriormente, para el 2.76% de los accesos se desconoce su localización; no se hacen hipótesis de pertenencia a un municipio pues se estima que la cantidad de accesos puede despreciarse

[¶¶] Telmex y Telnor

[***] Cablecom, Cablemas, Cablevision, Cablevision Red, Television Internacional y Sky

[†††] Megacable y MCM

[‡‡‡] Aunque no se tiene datos explícitos de fibra óptica para Grupo Televisa, la oferta comercial de "Izzi" si los considera : *"...hoy izzi ofrece servicio en más de 60 ciudades en 29 estados de la República Mexicana, mediante una red de más de 30,000 kilómetros de fibra óptica y 77,000 kilómetros de cable coaxial"*, consultado el 18 de Noviembre de 2019 en ver https://www.izzi.mx/nosotros#infraestructura

[§§§] Librearía *mxmaps*, disponible a través de https://www.diegovalle.net/mxmaps/. Por otra parte, una versión interactiva de este mapa se encuentra dispoinble para su consulta en el repositorio de Github que acompaña a este documento https://github.com/czammar/BandaAnchaFija/blob/master/Mapas/Penetracion_BAF_xcada100Habitantes_062019.html

cobertura de los municipios es heterogénea, para estudiarla se categorizará a cada municipio del país acuerdo a su nivel de penetración de banda ancha fija, con referencia al valor promedio de los países miembros de la OCDE (30.92 accesos por cada 100 habitantes), como se aprecia en la tabla 2.

Table 2. Niveles de penetración en un municipio

Nivel de penetración	Rango de penetración
Muy Alta	$Penetracion > Media\ OCDE$
Alta	$20 < Penetracion \leq Media\ OCDE$
Media	$10 < Penetracion \leq 20$
Baja	$0 < Penetracion \leq 10$
Nula	$Penetracion = 0$

C.4. Distribución de población y hogares. Como se ha mencionado previamente la distribución de población y hogares son factores que los operadores de telecomunicaciones toman en cuenta para realizar sus despliegues, y por lo tanto es un factor a considerar para explicar el nivel de penetración de BAF en municipios. Sin embargo, que un porcentaje de la población mayor al 16% habita 1,698 municipios donde no existen accesos de las tecnologías cable coaxial ni fibra óptica; en contraste los municipios con alta nivel de penetración agrupan únicamente a 2.4% de la población del país.

Table 3. Distribución de población y hogares según nivel de penetración de BAF en municipios

Penetracion	# Municipios	% Municipios	% Población	% Hogares
Nula	1,686	68.1	16.4	15.5
Baja	629	25.4	38.5	37.4
Media	134	5.4	35.4	36.3
Alta	20	0.8	7.42	7.96
Muy alta	8	0.3	2.4	2.85

C.5. Ingreso y penetración. Del análisis (ver figuras 5 y 6) se desprende una relación escalonada, en términos generales, entre el ingreso que de los municipios de México[¶¶¶], y el nivel de penetración de BAF basado en fibra óptica y cable coaxial, lo cual sugiere que los despliegues de ambas tecnologías se han focalizado en municipios donde hay mayores ingresos por habitante.

C.6. Densidad de habitantes y penetración. De acuerdo a la figura 7, aunque existe variabilidad considerable en la densidad de habitantes por municipio la información refleja un escalonamiento de conforme al nivel de penetración de BAF en los municipios.

2. Metodología para inferencia de niveles de cobertura de banda ancha fija en municipios

En linea con la exposición anterior, se busca explorar la posibilidad de que la información del entorno socio-económico, demográfico y tecnológico de los municipios sirva como una vía que permita inferir el nivel de cobertura actual de los servicios de banda ancha fija que los operadores hayan desplegado a la fecha.

Para ello, se propondrán modelos de aprendizaje de máquina que permitan clasificar a los municipios de acuerdo a las siguientes situaciones:

P1: Existe o no penetración de BAF basada en accesos de fibra óptica o cable coaxial.
P2: Nivel de penetración de BAF basada en accesos de fibra óptica o cable coaxial, de acuerdo a las especificaciones de la tabla 2.

En ambos casos, la metodología a seguir[17] es la siguiente:

A. Definir los modelos a considerar,
B. De entre todas la información reunida, establecer cuales son las variables relevantes a considerar,
C. Procesamiento de los datos de manera compatible con sus requerimientos de estos,
D. Para la ejecución de cada modelo, realizar un pipeline que permita calibrar los posibles hiper-parámetros de los modelos, considerando la evaluación de los mismos con un métrica establecida a través de validación cruzada,
E. Considerar el mejor modelo resultado del pipeline anterior de acuerdo a la métrica predefinida, y
F. Analizar los resultados interpretando el funcionamiento de los modelos de aprendizaje de máquina.

A continuación se exponen las consideraciones realizadas para llevar a cabo las fases anterior, en el entendido que la ejecución de estas se realizó a través de diferentes programas implementados en Python.

A. Modelos de aprendizaje de máquina a considerar. Dado que se trata de problemas clasificación , se estimó pertinente considerar un modelo de referencia (*baseline*) a partir de una regresión logística.

Ahora bien, dado en análisis exploratorio realizado se aprecia que la penetración de banda ancha no guarda necesariamente relaciones lineales con las diferentes variables se estimó que los árboles podrían ser de utilidad en el contexto de **P1** y **P2**, por lo cual se estimó pertinente usar los modelos que se describen a continuación:

A.1. Modelo de regresión. Los modelos regresión de regresión se emplean a problemas de clasificación de $K \geq 1$ clases; es decir se consideran conjuntos de puntos y etiquetas de la forma $\mathcal{D} = \{(x^{(i)}, y^{(i)}) \in \mathbb{R}^n \times \{0, 1, 2, \ldots, K-1\} : i \in [m]\}$ donde se asume la existencia de una relación funcional en términos probabilísticos $\mathbf{y} = \mathbf{W}^T\mathbf{x} + \epsilon$ donde $\mathbf{W}$ es un vector de pesos y ϵ representa un error residual.

En el caso en que la respuesta sea binaria ($y \in \{0, 1\}$) suele aplicarse el modelo de regresión logística el cual asume que y sigue una distribución Bernoulli, $p(\mathbf{y}|\mathbf{W}, \mathbf{x}) = Ber(y|\mu(\mathbf{x}))$ donde la media satisface $\mu(\mathbf{x}) = E[y|\mathbf{x}] = p(y = 1|x) = sigm(W^T x)$[18]. Dicho modelo predice la etiqueta $\hat{y}(x)$de un punto específico x a través de un umbral $p \in (0, 1)$; es decir decide que $\hat{y}(x) = 1 \leftrightarrow p(y = 1|x) > p$.

Por otra parte, el modelo de regresión multinomial resulta de una generalización de lo anterior, para tratar puntos con

[¶¶¶] Se reitera que en la fuente considerada no se cuentan con datos de 11 municipios del país.

[17] Aunque los pasos se presentan de manera lineal, la implementación requirió iteraciones entre ellos a efector de mejorar el funcionamiento global de las predicciones

[18] Función sigmoide dada por $sigm(\eta) = \frac{1}{1+e^{-\eta}}$

múltiples etiquetas ($K > 1$). En este caso, la probabilidad de que cada punto pertenezca una clase k se modela con la función $softmax$, esto es $P[y = k] = \frac{\exp(\mathbf{W}_k^T x)}{\sum_{i=1}^{K-1} \mathbf{W}_i^T x}$, donde W_i refiere a la i-ésima columna del vector de pesos. En ente sentido, esta familia de modelos predice la etiqueta $\hat{y}(x)$de un punto específico x a través asignándolo a la clase que maximice la probabilidad de pertenencia a ésta.

B. Bosques aleatorios. En el contexto de problemas de clasificación los árboles son modelos de aprendizaje de máquina que buscan realizar predicciones explorando a los individuos a través de cortes en sus características (variables de entrada) con el objeto de ir dividiendo la muestra, de manera que mediante cortes sucesivos se puedan refinar hasta establecer regiones que caractericen a las predicciones de manera acertada, lo cual 1) permite aproximar relaciones no lineales entre sus variables, y 2) otorga variabilidad en el proceso de su construcción.

Los bosques aleatorios son modelos de ensamble, pues permiten incorporar la información de predicciones de muchos árboles; para disminuir la varianza se recurre a un proceso que induce árboles con menor correlación entre sí. Para ello se considera una colección de muestras bootstrap[19] $\mathcal{D}_1^*, \mathcal{D}_2^*, \ldots, \mathcal{D}_B^*$ proveniente de los puntos de entrenamiento y sus etiquetas $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^n$, prefijando un número m de sus características. Para cada unas de tales muestras bootstrap se construye un árbol bajo el esquema de que: a) en cada nodo candidato a particionar, se escogen al azar m variables; 2) de entre ellas se elige la mejor variable y punto de corte (e.g. con un criterio de medida de impureza, coeficiente de Gini, entre otros), 3) repetir 1) y 2) hasta obtener un árbol suficientemente grande.

Ello permite consolidar un sistema de votación entre las predicciones de todos los árboles; de modo que la clase a la que pertenece un punto específico se estima como aquella a la que predijeron en mayor medida los árboles construidos en el proceso previamente descrito; es decir mediante la expresión $T^*(x) = argmax_k\{\#\{b|T_b^*(x) = k; b = 1, \ldots, B\}\}$.

C. Gradient tree boosting. El algoritmo de potenciación de gradiente de árboles (*gradient tree boosting*) es una técnica de aprendizaje de máquina que parte de la premisa de que al combinar de manera aditiva una cantidad de modelos base (árboles) se puede obtener un modelo con mejoras en sus capacidades de aprendizaje (Chen and Guestrin (2016)). Ello equivale a que dado un conjunto de datos,$\mathcal{D} = \{(\mathbf{x_i}, y_i) : \mathbf{x_i} \in \mathbb{R}^m, y_i \in \mathbb{R}\}$ con $|\mathcal{D}|$, la predicción resulta de un ensamble $\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in \mathcal{F}$, donde K es el número de árboles, f_k se refiere a una familia de modelos base de tipo árbol, $\mathcal{F} = \{f(\mathbf{x}) = w_{q(\mathbf{x})} : T \in \mathbb{N}, q : \mathbb{R}^m \to \{1, 2, \cdots, T\}, w \in \mathbb{R}^T\}$, siendo T el número de hojas de un árbol f_k y q una función que representa su estructura[20] y w una colección de pesos sobre sus hojas[21].

Desde una perspectiva teórica, el aprendizaje del modelo se logra a través de la optimización de una función objetivo $\mathcal{L}(\phi)$ que incorpora a una función de pérdida $l(y_i, \hat{y}_i)$ convexa y un término Ω que penaliza la complejidad del modelo y permite su regularización para evitar sobre-ajuste. Sin embargo, dado que puede ser difícil de minimizarla con métodos tradicionales, a través del algoritmo en comento se plantea aproximar dicha función en el proceso de entrenamiento de una manera aditiva, iniciando desde una predicción constante y añadiendo una función en cada paso; es decir para $\hat{y}_i^{(t)}$ la predicción del i-ésimo individuo en la t-ésima etapa de entrenamiento realizar la minimización de:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \sum_{k=1}^K \Omega(f_k) \qquad [2]$$

Pero considerando únicamente los términos de primer y segundo orden de la función de pérdida, $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$. Dicha estrategia permite encontrar una expresión simplificada de la aproximación de (2):

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^n [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \approx \mathcal{L}^{(t)} \qquad [3]$$

Lo cual equivale a que si $I_j := \{i|q(x_i) = j\}$, entonces:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{j=1}^T [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \qquad [4]$$

A partir de ella, para una estructura fija de un árbol $q(\mathbf{x})$, se puede obtener una expresión cerrada del peso w_j^* en la j-ésima hoja de dicho árbol, donde si $G_j := \sum_{i \in I_j} g_i$ y $H_j = \sum_{i \in I_j} h_i$ entonces:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \qquad [5]$$

Así como el cálculo del correspondiente valor óptimo:

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^T \frac{G_j^2}{H_j + \lambda} + \gamma T \qquad [6]$$

A partir de esta forma de medir qué tan bueno es un árbol, se puede plantear un algoritmo para crecer un árbol de la siguiente forma: 1) comenzar con un árbol de profundidad cero, 2) para cada nodo de hoja del árbol, intentar agregar una división (split), 3) evaluar la reducción de la perdida después de agregar la división para decidir si es aceptable con la ecuación:

$$\mathcal{L}_{split} = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma$$

XGBoost[22] es una de las librerías más populares que implementa el algoritmo recién descrito, encontrándose disponible para Python y R.

[19] Se refiere a una colección de muestras del conjunto de prueba, elegidas con reemplazo y con la misma cardinalidad que este conjunto en estudio

[20] Estableciendo una regla de asociación entre un punto del espacio y el índice de la hoja del árbol a la que corresponde

[21] En un ejemplo específico, se emplean reglas de decisión sobre la estructura de los árboles a elegir, para clasificarlos respecto a la elección de sus hojas y crear una predicción sumando los pesos de las mismas

[22] https://xgboost.readthedocs.io/en/latest/

D. Variables a considerarse en los modelos. En el caso del **P1** para clasificar a los municipios por la presencia o no de servicios de BAF basados en fibra óptica, se estimó pertinente considerar a las siguientes variables por municipio: 1) número de hogares, 2) número de habitantes, 3) superficie (en km), 4) densidad de $hogares/km^2$, 5) densidad de $habitantes/km^2$, 6) porcentaje de habitantes de municipio sin educación primaria, 7) años promedio de escolaridad por habitante, 8) porcentaje de hogares sin acceso a electricidad, 9) porcentaje de hogares sin acceso a energía eléctrica, 10) porcentaje de población que vive en localidades de menos de 5,000 habitantes, 11) porcentaje de la población que gana menos de 2 salarios mínimos, 12) ingreso promedio anual per cápita (poder de paridad de compra, expresado en dólares), 13) porcentaje de hogares con servicios de televisión de paga, 14) porcentaje de hogares con servicios de telefonía celular, y 15) porcentaje de hogares con disponibilidad de teléfono fijo.

Dado que en **P2** se busca clasificar con respecto al nivel específico de penetración de BAF en los municipios se estimó necesario proveer información adicional sobre el nivel de presencia de Internet y el nivel de competencia entre operadores para el despliegue de accesos basados en las tecnologías multicitadas; ello a través variables descritas para **P1** añadiendo 16) el porcentaje de hogares con disponibilidad de Internet, y 17) una variable que indica si hay dos o más operadores de telecomunicaciones que proveen servicios de BAF basados en cable coaxial o fibra óptica.

E. Procesamiento de los datos. Para los modelos de regresión las variables numéricas fueron estandarizadas aplicando una función z-score para evitar problemas numéricos asociados a su escala. En el caso de los modelos de árboles aleatorios y gradient tree boosting, las covariables no fueron transformadas en el entendido de que el funcionamiento de dichas herramientas es capaz de realizar la clasificación pretendida sin interferencias de escala.

Para **P1** la presencia o no de servicios de BAF basados en cable coaxial o fibra óptica se codificó como una variable binaria, que constituye la categoría a ser predecida. En el caso de **P2** los niveles de penetración se codificaron con una variable categórica donde 0 refiere a penetración "nula", 1 a penetración "baja" y así sucesivamente hasta que 5 alude a penetración "muy alta".

Cabe destacar que los datos de ingreso promedio anual per cápita no se encontraron disponibles para 4 municipios del país; tales fueron excluidos de la modelación.

Por otra parte, se realizó un partición de los datos de manera aleatoria para consolidar un conjunto de entrenamiento y prueba, guardando una relación, de 75% y 25%, de manera respectiva.

F. Ajuste de los modelos y calibración de sus hiper-parámetros. A manera de establecer una métrica que permita comparar entre los modelos a considerar, se estableció al $f1$-score.

Los modelos de regresión fueron ajustados con los datos descritos en la sección previa, usando el método de optimización *lbfgs* para **P1** y para **P2** se consideró a la función softmax para transforma la salida hacia las clases del problema de clasificación.

Para los modelos de árboles aleatorios se ajustaron los pesos de los árboles de manera inversamente proporcionales a las frecuencias de clase en los datos de los municipios para considerar el desbalance en los niveles de penetración de BAF. Además, tales modelos se calibraron los siguientes hiper-parámetros 1) la cantidad de árboles aleatorios a considerar, 2) la profundidad máxima de los árboles a considerar, 3) el mínimo número de muestras necesarias para dividir a un nodo de un árbol.

Por lo que hace a los modelos de gradient tree boosting, se ajustaron los siguientes hiper-parámetros: i) la tasa de aprendizaje, ii) la reducción de pérdida mínima requerida para realizar una partición adicional en un nodo hoja del árbol, iii) la profundidad máxima de los árboles a considerar, iv) la profundidad máxima de los árboles, v) la proporción del conjunto de entrenamiento que se empleará para muestra y construir los árboles, vi)la proporción de columnas del conjunto de entrenamiento que se usara para construir los árboles, así como vii) la cantidad de árboles aleatorios a considerar.

Para **P1** y P2, para la evaluación de la pertinencia de los hiper-parámetros se empleó validación cruzada de 6 hojas.

El detalle del procedimiento usado se puede consultar en los apéndices que acompañan al presente documento.

G. Análisis de resultados. En el caso de los modelos de regresión logística, se realizó el ajuste sobre el conjunto de entrenamiento, para posteriormente realizar la predicciones correspondiente sobre el conjunto de prueba.

Por su parte, los modelos de árboles, se ajustaron con los parámetros que obtuvieron mejores resultados sobre el conjunto de entrenamiento a través de validación cruzada y en función de la métrica descrita en la sección previa.

Realizando el ajuste de los modelos con los valores correspondientes a dichos hiper-parámetros sobre el conjunto de entrenamiento, se procedió a comparar los resultados obtenidos por todos los modelos sobre el conjunto de prueba.

Es así que los resultados de la metodología recién descrita para **P1** y **P2** se presentan a continuación:

G.1. P1. Recordemos que este problema aborda la clasificación de un municipio de acuerdo a si hay presencia de penetración de accesos de banda ancha fija basado en las tecnologías de cable coaxial y fibra óptica.

Table 4. Promedio ponderado de resultados obtenidos para los distintos modelos - P1

#	Modelo	Precisión	Recall	$F1$-score	Soporte
1	Logístico	0.87	0.87	0.86	612
2	Bosque Aleatorio	0.87	0.86	0.87*	612
3	Gradient Tree Boosting	0.87	0.86	0.86	612

A través de la tabla 4 se aprecia que el modelo de bosques aleatorio fue el que obtuvo mejor $F1$-score, a nivel promedio ponderado, sobre el conjunto de prueba, y que el resto de modelos obtuvo valores similares de desempeño.

Sin embargo, al analizar las matrices de confusión (ver figuras 8 , 9 y 10) se puede como el modelo logístico acertó en la predicción de más valores que los basados en árboles, mientras que los otros tuvieron menos falsos positivos que éste. Se destaca que las matrices de los modelos basados en árboles son prácticamente iguales, salvo en que la matriz del modelo de gradient tree boosting tuvo un falso positivo más que la proveniente del modelo de bosque aleatorio.

En complemento, a través de los diagramas de feature importance asociados al modelo de gradient tree boosting, (ver figuras 11, 12, 13 presentes en el apéndice) se puede apreciar que el número de hogares, el porcentaje de población del municipio que vive en localidades de menos de 5,000 hogares, la cantidad de habitantes, el ingreso anual per capita y la densidad de hogares son las variables que más aportaron en la reducción de pérdida a la hora de efectuar splits de los árboles del método, es decir, en términos relativos fueron las que más aportaron a la predicción de la existencia de cobertura de BAF.

G.2. P2. Mediante la tabla 5 se desprende que el modelo de bosques aleatorio fue el que obtuvo mejor $F1$-score, respecto al promedio ponderado, sobre el conjunto de prueba, y que el resto de modelos obtuvo valores similares de desempeño.

Table 5. Promedio ponderado de resultados obtenidos para los distintos modelos - P2

#	Modelo	Precisión	Recall	$F1$-score	Soporte
1	Logístico (1 vs all)	0.90	0.88	0.89	612
2	Logístico (all vs all)	0.90	0.88	0.89	612
3	Bosque aleatorio	0.90	0.89	0.89*	612
4	Gradient Tree Boosting	0.89	0.88	0.88	612

Se aprecia que respecto a los resultados obtenidos en el caso anterior, existe un aumento en la métrica de los valores predecidos; presumiblemente derivado de que dimos más información al modelos respecto a la presencia de servicios de Internet en los municipios y sobre la competencia a nivel local entre operadores.

Al revisar las matrices de confusión de esos modelos (véase las figuras 14, 15, 16 y 17 en el apéndice) se aprecia, en términos generales: a) en general, el modelo de gradient tree boosting fue el único que se equivocó prediciendo la penetración de un municipio a un nivel inmediato superior o inferior, b) ninguno de los modelos pudo predecir acertadamente el nivel de penetración del municipio con penetración muy alta.

Por otro lado, de acuerdo las figuras de feature importance asociados al modelo de gradient tree boosting, (ver figuras 18, **??**, 20 presentes en el apéndice) se desprender que la cantidad de operadores presentes en el municipio que tienen instalados servicios de BAF, el porcentaje de disponibilidad de Internet, que el número de hogares y la población son las variables que más aportaron en la reducción de pérdida a la hora de efectuar splits de los árboles del método, es decir, en términos relativos fueron las que más aportaron a la predicción del nivel de cobertura de BAF a nivel municipio.

H. Conclusiones. En este documento, exploramos el uso de modelos de aprendizaje de máquina para inferir el nivel de penetración de banda ancha de cierto tipo de tecnologías en los municipios del país, referido como la cantidad de acceso por cada 100 habitantes, a partir de datos socio-económicos provenientes fuentes públicas, basándonos en modelos de aprendizaje de tipo logístico y ensambles de árboles.

Aunque existen limitaciones en la precisión alcanzada por estos modelos, sobretodo al no involucrar detalle de la presencia de infraestructura a nivel geográfico, lo mostrado aquí sugiere que este tipo de modelos puede ser una alternativa para inferir el nivel de cobertura en una zona ante la ausencia de datos de despliegue de operadores, que pueda ser de utilidad a tomadores de decisiones, por ejemplo en la implementación de políticas públicas para aumentar el nivel de penetración de banda ancha, en el entendido de que al entrenarse con datos de México, estos modelos podrían ser aplicables a regiones que posean un nivel de madurez semejante para servicios fijos de telecomunicaciones.

References

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

CONAPO (2010-2015). Índice de marginación por entidad federativa y municipio 2010 y 2015. *Página electrónica de la Comisión Nacional de población.*

Escobar, R. (2017). Disponibilidad municipal de la Infraestructura de Telecomunicaciones. *Pagina electronica del Centro de Estudios del Instituto Federal de Telecomunicaciones.*

IFT (2017). Las telecomunicaciones a 4 años de la Reforma Constitucional en México. *Pagina electronica del Instituto Federal de Telecomunicaciones.*

IFT (2018a). Accesos a Banda Ancha Fija crecen 37% a partir de la Reforma en Telecomunicaciones (Comunicado 019/2018) 4 de marzo. *Pagina electronica del Instituto Federal de Telecomunicaciones.*

IFT (2018b). Banco de Información de Telecomunicaciones. *https://bit.ift.org.mx/BitWebApp/.*

IFT (2018c). Documento metodológico para el diagnóstico de cobertura de banda ancha a nivel municipal por Entidad Federativa. *Página electrónica del Instituto Federal de Telecomunicaciones.*

IFT (2018d). Manual de definiciones de los indicadores estadísticos de telecomunicaciones. *Pagina electronica del Instituto Federal de Telecomunicaciones.*

IFT (2019). México entre los países con mayor crecimiento de Banda Ancha Fija: OCDE (Comunicado 32/2019) 9 de julio. *Pagina electronica del Instituto Federal de Telecomunicaciones.*

INEGI (2015). Encuesta Intercensal 2015. *https://www.inegi.org.mx/programas/intercensal/2015/.*

INEGI (2018). Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares (ENDUTIH) 2018. *Pagina electronica del Instituto Nacional de Estadística y Geografía.*

Katz, R. (2012). Impact of broadband on the economy. *International Union of Telecommunications.*

Katz, R. (2018). La digitalización: Una clave para el futuro crecimiento de la productividad en América Latina. *Centro de Estudios de Telecomunicaciones de America Latina.*

Moya, J. (2014). *Telecomunicaciones. Tecnologías, Redes y Servicios. 2ª edición actualizada.* Profesional. Grupo Editorial RA-MA.

ONU (2010-2015). Informe de Desarrollo Humano Municipal 2010-2015. Transformando México desde lo local. *Página electrónica de la Comisión Nacional de población.*

Pradhan, R. P., Arvin, M. B., Norman, N. R., and Bele, S. K. (2014). Economic growth and the development of telecommunications infrastructure in the g-20 countries: A panel-var approach. *Telecommunications Policy*, 38(7):634 – 649.

Appendices

A. Repositorio Github

El desarrollo de este trabajo se consolidó a través de un repositorio en Github de acceso público: https://github.com/czammar/BandaAnchaFija; este contiene la totalidad de script desarrollados en Bash, R, Python, los mapas interactivos, las versiones en .tex de los documentos así como los datos empleados.

B. Figuras de análisis exploratorio

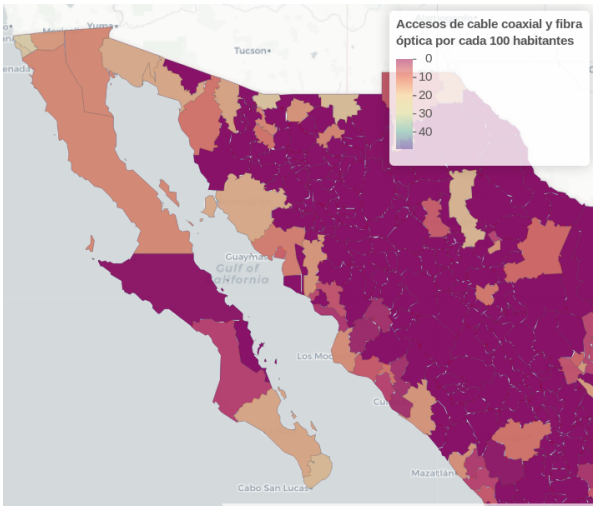

Fig. 2. Penetración de cable coaxial y fibra en la península de Baja California

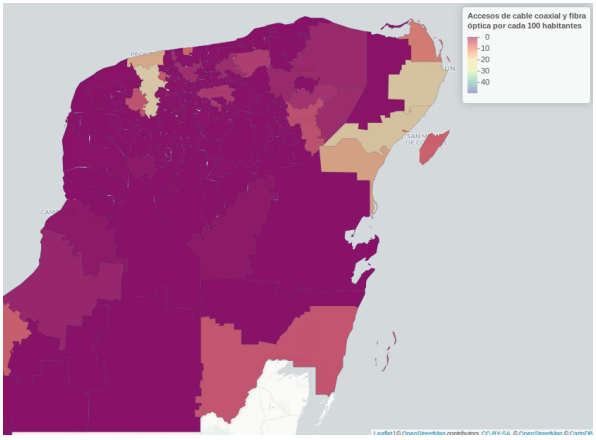

Fig. 3. Penetración de cable coaxial y fibra en la península de Yucatán

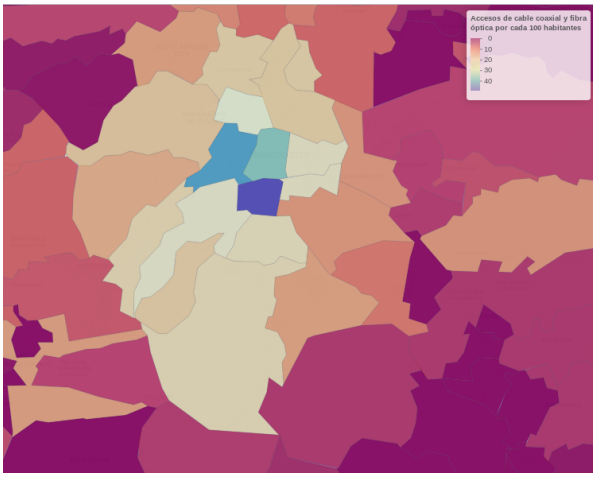

Fig. 4. Penetración de cable coaxial y fibra alrededor de CDMX

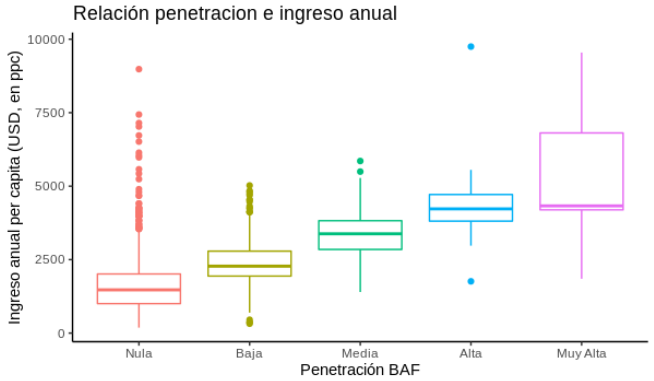

Fig. 5. Penetración BAF Junio 2018 (IFT (2018b)) vs Ingreso Anual per capita en municipios (ONU (2015))

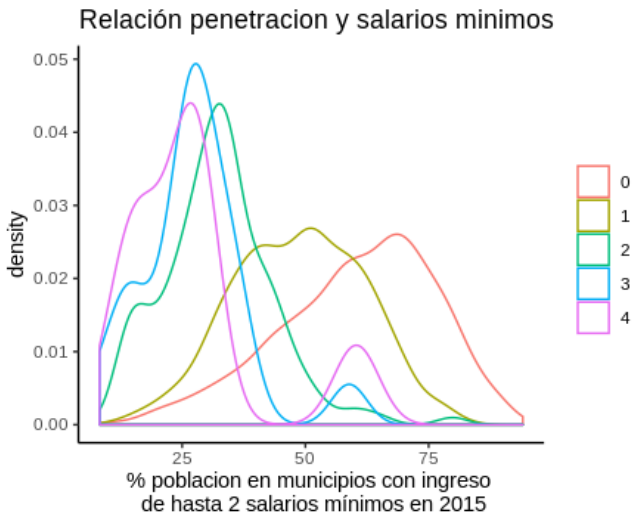

Fig. 6. Gráfico de densidad del porcentaje de población en municipios que ganaba hasta 2 salarios mínimos(CONAPO (2015)), desagregado por nivel de penetración de BAF. **Nota:** 0 = Pen. nula, 1= Pen. Baja, ..., 5= Pen. Muy alta

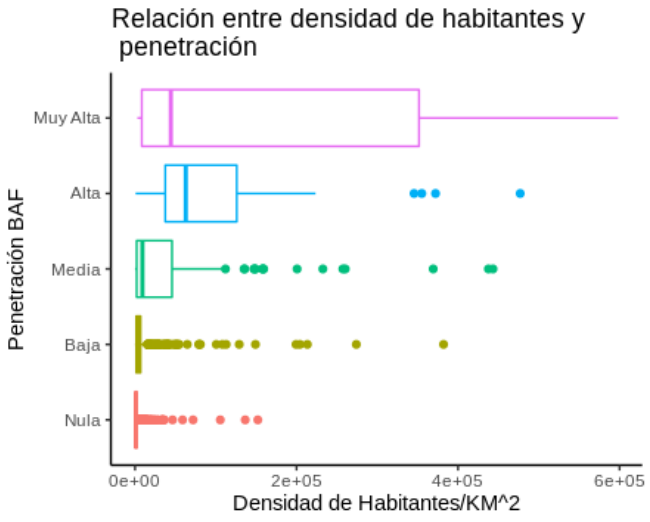

Fig. 7. Densidad de habitantes en municipios vs Penetración BAF Junio 2018 (IFT (2018b))

C. Resultados de modelos para P1

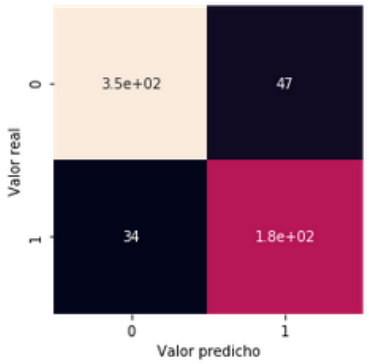

Fig. 8. Matriz de confusión del modelo logístico- P1

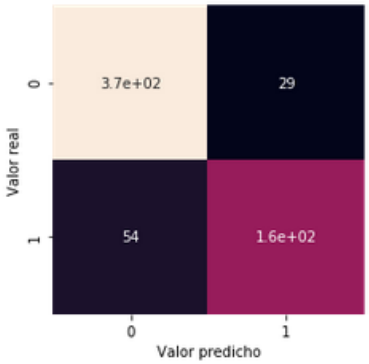

Fig. 9. Matriz de confusión del modelo de bosque aleatorio- P1

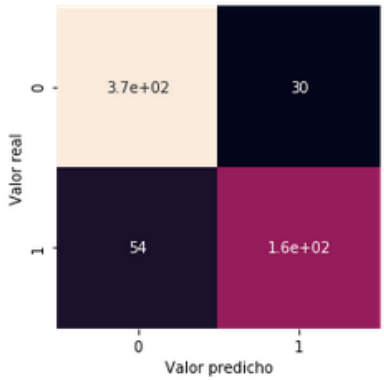

Fig. 10. Matriz de confusión del modelo gradient tree boost- P1

D. Resultados de modelos para P2

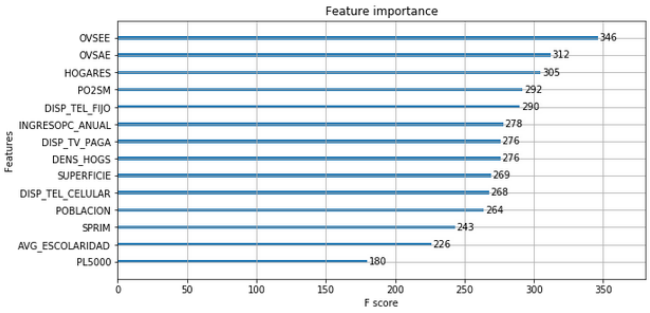

Fig. 11. Feature importace por "weight" del modelo gradiente tree boosting P1

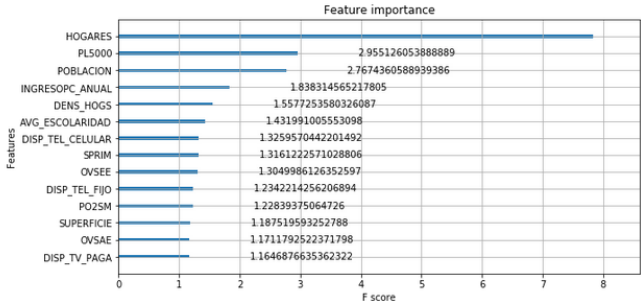

Fig. 12. Feature importace por "gain" del modelo gradiente tree boosting P1

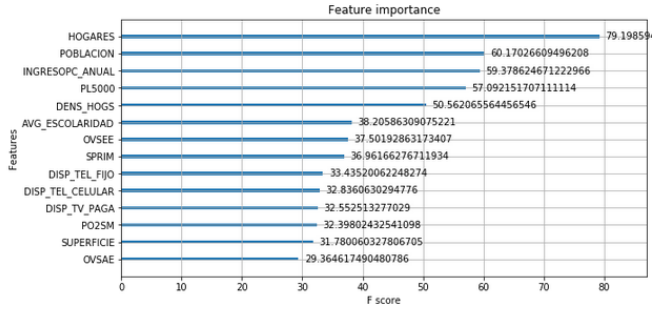

Fig. 13. Feature importace por "cover" del modelo gradiente tree boosting P1

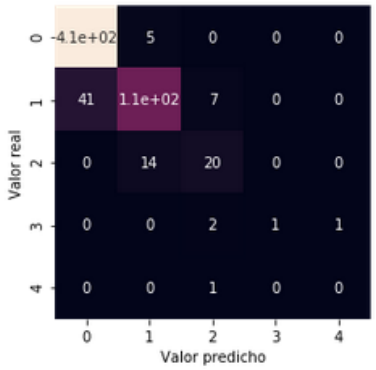

Fig. 14. Matriz de confusión del modelo logístico 1 vs all - P2

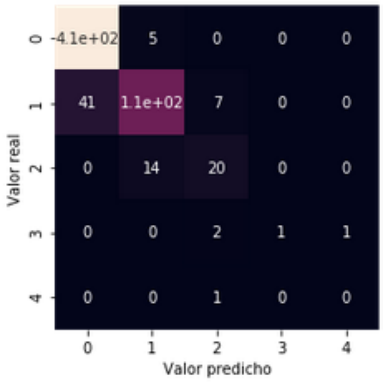

Fig. 15. Matriz de confusión del modelo logístico all vs all - P2

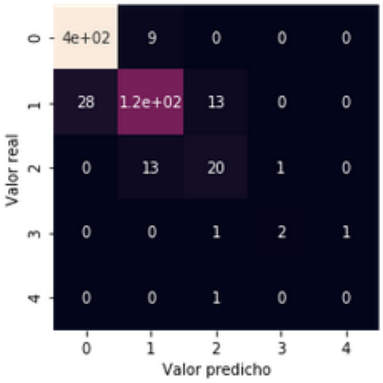

Fig. 16. Matriz de confusión del modelo de bosque aleatorio - P2

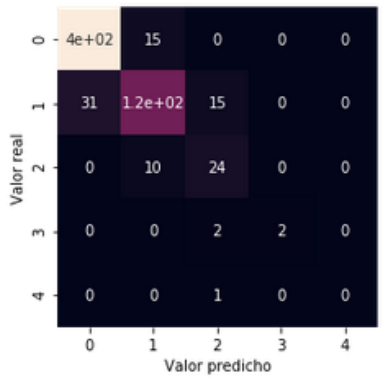

Fig. 17. Matriz de confusión del modelo gradient tree boosting - P2

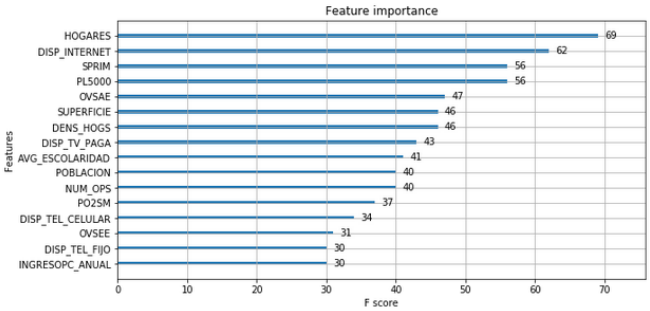

Fig. 18. Feature importace por "weight" del modelo gradiente tree boosting P2

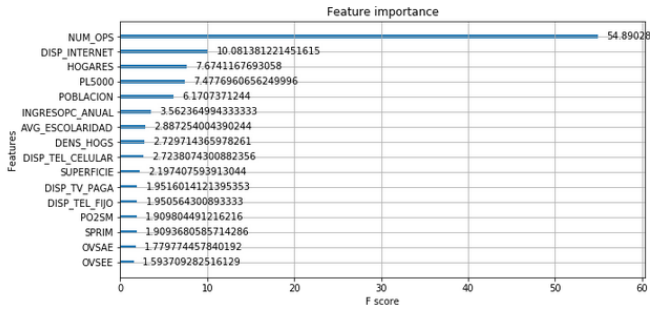

Fig. 19. Feature importace por "gain" del modelo gradiente tree boosting P2

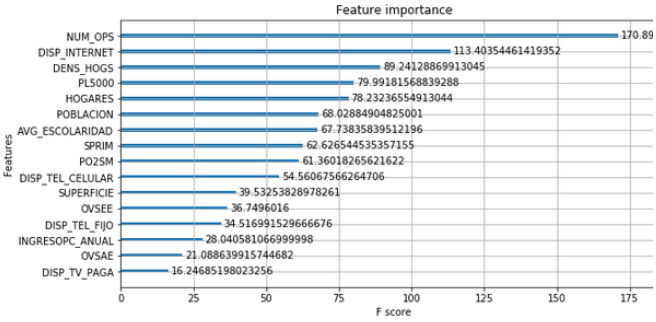

Fig. 20. Feature importace por "cover" del modelo gradient tree boosting P2