Minería de datos para perfilamiento de las brechas digital y educativa de ciudades en censos de población y vivienda

Sergio Coria, Ivania Orozco y Juan Luna

S. Coria, I. Orozco y J. Luna
Universidad de la Sierra Sur (UNSIS), Miahuatlán de Porfirio Díaz, Oaxaca., México
coria@unsis.edu.mx

M. Ramos.,V.Aguilera.,(eds.). Ciencias de la Ingeniería y Tecnología, Handbook -©ECORFAN- Valle de Santiago, Guanajuato, 2013.

Abstract

This work proposes and evaluates a data mining methodology to discover profiles of cities on the basis of educational characteristics of inhabitants and on discrete classes describing the presence of information and communication technologies (ICT) in households. City profiling involves discovering the variables and their corresponding values that allow distinguish classes of cities within a country regarding a determined target variable. Pattern discovery on the interaction among educational attainment of inhabitants, presence of ICT in households and other demographical and economical variables is relevant to researchers, public policy makers and private company managers. Using city (municipality) as analysis and modeling unit is novel in this research field because most of previous work has used the country level. In addition, the data mining approach is also novel in this area because prevailing approaches are based on creating composite quantitative indexes or on performing multivariate analyses.

3 Introducción

Existen diversas teorías sobre la noción de brecha digital. Esta investigación está basada en la definición establecida por (OECD, 2007): las diferencias entre individuos, hogares, negocios y áreas geográficas en distintos niveles socio-económicos respecto a sus oportunidades para acceder a tecnologías de información y comunicación (TIC), así como a su utilización de Internet para una amplia variedad de actividades. Por su parte, la brecha educativa involucra, principalmente, las diferencias entre el número de años de educación formal de los individuos de una región, país o conjunto de países. Generalmente, el fenómeno de la brecha educativa se investiga usando la medida denominada logro educativo (educational attainment) como medio de comparación entre diferentes contextos geográficos o socio-económicos. El logro educativo refiere al grado más alto completado dentro del nivel más avanzado atendido en el sistema educativo del país [u otra unidad territorial] donde se recibió la educación (OECD, 2007). Sobre esta base, un amplio número de países miden el logro educativo en términos del número de años acumulados de educación formal correspondiente al grado educativo más alto estudiado.

La investigación sobre la relación entre brecha digital y brecha educativa es relevante porque su entendimiento en contextos nacionales e internacionales puede conducir a explicaciones causales útiles para crear políticas públicas o para crear estrategias de empresas privadas que impacten en el mejoramiento de las condiciones socio-económicas de la población. Además de la disponibilidad de TIC y del logro educativo, otras variables demográficas y económicas que están involucradas en este fenómeno son, por ejemplo: poder adquisitivo de individuos o de familias, precios de los servicios de telefonía y de Internet, aspectos de regulación gubernamental sobre las TIC, etc. (Chinn & Fairlie, 2007).

Desde la década pasada, los censos de población y vivienda de una serie de países ofrecen online una amplia variedad de datos demográficos y socio-económicos que están disponibles en diferentes niveles geográficos de agregación y que incluyen información sobre bienes y servicios de TIC en hogares.

Uno de estos niveles de agregación es el de ciudad, que se ha vuelto más importante para propósitos de análisis y modelación porque los investigadores, los funcionarios gubernamentales y los administradores de empresas privadas requieren una mayor comprensión de los fenómenos de las brechas digital y educativa a este nivel.

El problema de investigación que abordamos en este trabajo consiste en determinar cómo las técnicas de minería de datos, particularmente los árboles de clasificación (AC, para abreviar) y el análisis de correlación pueden usarse para descubrir patrones no-triviales en el fenómeno que involucra la relación entre brecha digital, brecha educativa y otras variables demográficas y socio-económicas que están disponibles en censos de población y vivienda. El interés es en el nivel de ciudad (específicamente, el de *municipio*) debido a las razones expuestas anteriormente.

Respecto a las técnicas de AC (Han, Kamber & Pei, 2011), se eligió el algoritmo J4.8 (Witten & Frank, 2011), a su vez inspirado en el algoritmo C4.5 (Quinlan, 1993), porque además de las fortalezas de los algoritmos de AC en general, este está implementado en una herramienta de software libre: WEKA (Waikato Environment for Knowledge Analysis). El análisis de correlación, particularmente la regresión lineal simple (RLS), se aplica como técnica de análisis y modelación porque permite medir fácilmente y en modo general y compacto la interacción entre pares de variables numéricas.

Este artículo está organizado en las secciones que se enumeran a continuación: la sección 3.1 comenta los antecedentes y algunos trabajos relacionados. La sección 3.2 describe los datos fuente que son usados para crear y evaluar nuestra metodología. La sección 3.3 presenta la metdología de minería de datos que proponemos. La sección 3.4 presenta los resultados obtenidos. La sección 3.5 discute los resultados y, finalmente, la sección 3.6 presenta algunas conclusiones y sugiere trabajo de investigación futuro en este campo.

3.1 Antecedentes y trabajos relacionados

Las investigaciones sobre la correlación entre la disponibilidad y uso de TIC (tanto en hogares como en escuelas) por una parte, y el logro educativo por otra parte, ofrecen resultados contradictorios (OECD, 2010). Sin embargo, los resultados de la prueba PISA del año 2003 sugieren que los resultados más bajos son obtenidos por estudiantes con acceso limitado a las TIC, con menor experiencia en uso de TIC y menos confianza en su habilidad para usar computadoras. También, el uso más frecuente de TIC no siempre está correlacionado con mayores puntajes en pruebas PISA.

Por otra parte, respecto a la influencia del logro educativo sobre el grado de brecha digital de contextos territoriales determinados, otros trabajos, por ejemplo (Robinson, DiMaggio & Hargittai, 2003), establecen que el logro educativo podría estar asociado a la brecha digital más de lo que el poder adquisitivo de la población está asociado con esta última.

En un estudio sobre las formas de reducir la brecha digital para la población de bajos ingresos en Australia, (Yelland & Neal, 2013) descubrieron que el acceso a computadoras e Internet por jóvenes de edad escolar de familias desfavorecidas facilitaba la realización de tareas escolares, les permitía comunicarse con amigos y darles acceso a actividades de esparcimiento. Estos beneficios eran alcanzables debido a un entrenamiento para instalar y usar la computadora y a apoyo de línea de ayuda.

De acuerdo a la OECD (2012), los hogares de sus países miembros tienen altos porcentajes de acceso a Internet. En promedio, 67% de los hogares reportaron una suscripción de banda ancha en 2011. Sin embargo, México estaba debajo del porcentaje promedio, con aproximadamente 20% de los hogares con acceso a banda ancha. La OECD reporta que 47% de los usuarios de Internet en sus países miembros usan Internet para aprender. En México, solamente 10% lo usa para este propósito. La brecha digital es todavía un reto a resolver en México, tanto en hogares como en escuelas. En otros países, el acceso a TIC en escuelas está próximo a ser completo y la nueva preocupación es reducir la brecha en términos de competencias y habilidades para beneficiarse del uso de computadoras (OECD, 2012b).

Esta investigación está inspirada en las líneas generales descritas por (Coria et al., 2013), (Coria et al., 2013b) y (Coria et al., 2013c). Una metodología novedosa para analizar y modelar la brecha digital de ciudades (sin profundizar en el fenómeno de la brecha educativa) es propuesta por (Coria et al., 2013); ellos usan el algoritmo J4.8 para producir un conjunto de AC que describen perfiles de brecha digital de ciudades (municipios) usando datos del Censo Mexicano 2010. *Delta score* (Coria et al., 2013c) es propuesta como una medida multidimensional discreta de la brecha digital de ciudades. Esta medida puede interpretarse como un *ranking* de ciudades basado en la presencia de bienes y servicios de TIC, particularmente Internet, PC, teléfono fijo y teléfono celular. Su variante, *Delta+* (Delta Plus), incluye una medición del logro educativo a nivel de ciudad.

Delta y Delta+ pueden usarse también en inferencia estadística para propósitos de análisis y modelación de brecha digital de ciudades. En (Coria et al., 2013b) se hace una comparación entre los algoritmos J4.8 y PART (Frank & Witten, 1998) como medios para analizar y modelar la brecha digital de ciudades (la brecha educativa no es abordada); concluye que ambos algoritmos son útiles para este propósito, aunque PART puede producir modelos más compactos que presentan *accuracies* similares a las logradas por modelos J4.8.

3.2 Datos fuente

Los datos para crear y evaluar nuestra metodología provienen del Censo de Población y Vivienda de México del año 2010 (INEGI, 2011), que describe 2,456 municipios. Su correspondiente diccionario de datos (INEGI, 2011b) es altamente útil para explotar esta información. Sus aproximadamente 200 atributos están organizados en 14 categorías, como se enumera a continuación: identificación geográfica (9 attributes), población (47), fecundidad (1), migración (12), población indígena (13), discapacidad (9), educación (42), características económicas (12), servicios de salud (6), situación conyugal (3), religión (4), hogares censales (6), características de la vivienda (35) y tamaño de localidad (1).

Los items de identificación geográfica son: coordenadas geográficas (longitud, latitud, altitud), así como nombres e identificadores de entidad federativa, municipio y localidad (unidad territorial más pequeña que el municipio). La sección del censo que contiene el mayor número de atributos es la categoría de población; sus atributos consideran sexos y varios segmentos de edades de los habitantes.

La categoría de educación es la segunda con más atributos. Estos describen a la población considerando sexos y segmentos de edades que: asisten, o no, a la escuela; saben, o no, leer y escribir; tienen, o no, educación primaria incompleta; tienen, o no, secundaria incompleta y tienen, o no, educación post-básica. También, esta categoría incluye variables muy importantes para esta investigación: promedios de años de educación de población masculina, femenina y totalizada. Una de las variables más imporantes en esta categoría es GRAPROES (promedio de años de educación de la población, tomada aquí como una medición del logro educativo).

Los atributos sobre características de la vivienda son importantes para esta investigación porque incluyen cuatro variables descriptivas de bienes y servicios de TIC en hogares: número de hogares por municipio que tienen Internet (*VPH_INTER*), computadora personal (*VPH_PC*), teléfono fijo (*VPH_TELEF*) y teléfono celular (*VPH_CEL*). Además, otras variables interesantes en esta categoría abordan la presencia de: electricidad, receptor de radio, televisor, automóvil, refrigerador doméstico y lavadora de ropa.

Un aspecto importante de los datos del censo en relación a los propósitos de esta investigación es que no incluyen estrictamente información específica sobre el ingreso económico de los habitantes (p. ej. ingreso promedio por municipio). Aunque esta variable es importante para analizar los fenómenos de la brecha digital y del logro educativo, esta investigación se limita a explorar únicamente las variables disponibles en el censo. Esta variable podría incluirse en análisis y modelos a realizarse en trabajo futuro.

3.3 Metodología de minería de datos

La metodología de minería de datos que proponemos es altamente similar a la de (Coria et al., 2013) y (Coria et al., 2013b), pero está complementada con el uso de análisis de correlación, el cual no es utilizado por ninguna de estas dos.

Por lo tanto, nuestra metodología usa el algoritmo J4.8 para producir un conjunto de AC, y regresión lineal simple (RLS) para producir análisis de correlación.

Desde una perspectiva teórica, se eligió el algoritmo J4.8 porque produce automáticamente modelos clasificadores que describen las interacciones entre grandes conjuntos de datos, lo cual permite descubrir y representar patrones implícitos en el fenómeno de la interacción entre brechas digital y educativa de ciudades.

También, porque en comparación con otros algoritmos clasificadores (como el perceptrón multicapa), los modelos J4.8 son más expresivos y fáciles de entender para usuarios que no son expertos en aprendizaje automático.

Por otra parte, se aplica RLS porque permite evaluar qué tanto una variable específica podría estar asociada a otra en un fenómeno determinado. Se descartaron otras técnicas más sofisticadas para análisis de correlación porque el propósito principal en esta investigación es descubrir las correlaciones más sustanciales. La metodología está constituida por los siguientes pasos generales: 1) selección de datos, b) preprocesamiento, c) discretización, d) organización de datasets experimentales, e) generación y evaluación del conjunto de AC, y f) análisis de RLS. Estos pasos son explicados a continuación.

3.4 Selección de Datos

Al igual que en (Coria et al., 2013), se selecciona la gran mayoría de los atributos de la base de datos del censo. Los datos que no son seleccionados son: seis identificadores geográficos que no pueden contribuir al descubrimiento de patrones, incluyendo dos de localidad, dos de municipio y dos de entidad federativa; también, la variable de tamaño de localidad es descartada manualmente porque representa al número total de habitantes del municipio usando una escala numérica ordinal y preferimos usar el número total de habitantes disponible en la base de datos fuente.

El número total de atributos que se incorporan en cada dataset experimental depende de cuál sea la variable objetivo (*target*) para la creación de sus correspondientes modelos de AC. La razón es que una serie de targets específicos no necesitan (o no deben usar) a un subconjunto determinado de atributos para evitar así la generación de patrones triviales en los modelos de AC y de RLS. Por ejemplo, si el target para crear un AC es la variable discretizada que representará al promedio municipal de años de educación, entonces los atributos de porcentajes de población que tiene diversos niveles de educación son manualmente descartados del dataset experimental.

3.5 Reprocesamiento

Se realizan cálculos sencillos sobre las variables seleccionadas de los datos fuente para producir los porcentajes de hogares y de habitantes por municipio considerando la mayoría de los atributos del censo.

El porcentaje de hogares que tienen Internet en un municipio determinado es calculado dividiendo el número de hogares que tienen Internet entre el número total de hogares del municipio. Las cuatro variables de ICT se usan para producir estas variables porcentuales: *VPH_INTER_%*, *VPH_PC_%*, *VPH_TELEF_%* and *VPH_CEL_%*. Se realizan cálculos similares sobre todas las otras variables seleccionadas.

Las variables del censo que no necesitan ningún cálculo son, por ejemplo, las coordenadas geográficas, el total de habitantes del municipio, el promedio de hijos nacidos vivos, entre otras. El promedio municipal de años de educación es procesado realizando un simple redondeo (ver subsección *3.3 Discretización*).

3.6 Discretización

Se realiza una discretization sobre los atributos porcentuales *VPH_INTER_%*, *VPH_PC_%*, *VPH_TELEF_%* y *VPH_CEL_%* para producir cuatro variables nominales que se usan como target para los modelos de AC. Se usan varios tamaños de intervalo para hacer la discretización de cada atributo porcentual. A su vez, la discretización de *GRAPROES* se realiza como un simple redondeo hacia arriba (*half-up rounding*) para transformar ese número de tipo real en un entero de dos dígitos que se maneja como valor nominal.

Los valores nominales correspondientes a los atributos porcentuales se producen como sigue: con *VPH_INTER_%* se genera *inet_6* e *inet_25*; con *VPH_PC_%* se produce *pc_6* y *pc_25*; con *VPH_TELEF_%* se genera *telef_7* y *telef_25*; y con *VPH_CEL_%* se genera *cel_7* y *cel_25*. Los números 6, 7 y 25 en los nombres de los atributos nominales refieren al tamaño de intervalo usado para discretización en cada caso. Los valores para cada atributo nominal se definen usando un conjunto de etiquetas: c_1, c_2, c_3,… c_n, que representan a la *clase 1* (la clase con porcentajes más altos en cualquier variable), *clase 2*, *clase 3…* y *clase n* (la clase con porcentajes más bajos en cualquier variable). Por ejemplo, *inet_6* es el atributo nominal que describe la presencia de Internet en hogares de municipios considerando un tamaño de intervalo de 6 puntos porcentales; los municipios con porcentajes más altos de Internet corresponden a la *clase 1* (c_1) y aquellos con los porcentajes más bajos pertenecen a la *clase n* (c_n), donde *n* es igual al número total de clases de esa variable. La Tabla 3 muestra más detalles acerca del manejo de intervalos para los atributos discretizados en los diversos datasets experimentales. Con *GRAPROES* se produce *graproes_1*, *graproes_4*, *graproes_5* y *graproes_6*. Los números 1, 4, 5 y 6 también refieren a tamaños de intervalo, aunque en este caso no se trata de intervalos de porcentajes, sino de números de años de educación (ver Tabla 3). Por ejemplo, en *graproes_1* el tamaño de intervalo es de 1 año, y los valores nominales c_1 a c_n que se manejan para este atributo representan: c_1, a la clase de municipios con el mayor número de años de educación, y c_n a la clase de municipios con el menor número. Los tamaños de intervalo para discretizar los porcentajes de cada una de las cuatro variables de TIC se definieron en dos modalidades distintas: 1) usando la fórmula de (Sturges, 1926), y 2) estableciendo un tamaño convencional de 25 puntos porcentuales para manejar 4 clases nominales (c_1, c_2, c_3, c_4) en cada variable.

La fórmula de Sturges es: $C = rango / (1 + 3.322 \text{ Log } N)$, donde C es el tamaño óptimo de intervalo, *rango* es la diferencia entre los valores máximo y mínimo del atributo y N es el número de instancias en el dataset (2,456 municipios).

Tabla 3 Datasets Experimentales y Resultados de los Árboles de Clasificación Generados

Tema	Atributo target del dataset	No. de clases en dataset	Clases mayoritarias en dataset	ID del árbol	Comentarios sobre predictores seleccionados manualmente	Accuracy (%) del árbol	Kappa del árbol	No. de hojas del árbol	Atributo raíz del árbol
Internet	inet_6	12	c12 (63.6%) + c11 (18.4%) = 82.0%	i6pc	Con VPH_PC_%.	82.4	0.6797	120	VPH_PC_%
				i6nopc	Sin VPH_PC_%.	75.3	0.5484	179	P18YM_PB_F_%
	inet_25[a]	4	c4 (95.0%)	i25pc	Con VPH_PC_%.	98.6	0.8786	9	VPH_PC_%
				i25nopc	Sin VPH_PC_%.	97.7	0.7599	22	GRAPROES_F
PC	pc_6	12	c12 (34.5%) + c11 (26.1%) + c10 (16.9%) + c9 (9.6%) = 87.2%	p6i	Con VPH_INTER_%.	73.8	0.6594	183	VPH_INTER_%
				p6noi	Sin VPH_INTER_%.	63.7	0.5283	249	GRAPROES_F
	pc_25[a]	4	c4 (88.4%)	p25i	Con VPH_INTER_%.	96.5	0.8329	25	VPH_INTER_%
				p25noi	Sin VPH_INTER_%.	95.2	0.7619	32	P18YM_PB_M_%
Teléfono fijo	telef_7	13	c12 (18.2%) + c11 (16.9%) + c10 (15.1%) + c13 (13.6%)+ c9 (12.1%) + c8 (10.0%) = 86.0%	t7i	Con VPH_INTER_%.	32.5	0.2194	506	VPH_AUTOM_%
				t7noi	Sin VPH_INTER_%.	31.5	0.2069	503	VPH_AUTOM_%

	telef_25[a]	4	c4 (57.9%) + c3 (36.4%) = 94.3%	t25i	Con VPH_INTER_%.	78.9	0.5871		VPH_INTER_%
				t25noi	Sin VPH_INTER_%.	77.1	0.5655	180	VPH_REFRI_%
Teléfono celular	cel_7	13	c13 (16.7%) + c6 (11.3%) + c7 (9.6%) + c5 (8.9%) + c8 (8.9%) + c4 (7.0%) + c9 (7.0%) + c12 (6.9%) + c3 (6.3%) = 82.7%	c7	*****	29.4	0.2221	523	VPH_PC_%
	cel_25[a]	4	c4 (32.1%) + c2 (30.3%) + c3 (30.0%)	c25	*****	68.6	0.5573	217	VPH_PC_%
Promedio de años de educación	grapr oes_1	13	c9 (27.6%) + c10 (22.5%) + c8 (18.6%) + c7 (11.2%) = 80.0%	e1	Sin otros atributos de educación.	50.0	0.3831	341	VPH_PC_%
				d1	Solo delta_score.	44.8	0.3011	28	delta_score
	grapr oes_4	4	c3 (80.0%)	e4	Sin otros atributos de educación.	87.1	0.6108	85	VPH_PC_%
				d4	Solo delta_score.	83.7	0.3316	28	delta_score
	grapr oes_5	3	c3 (61.8%) + c2 (38.2%) = 99.9%	e5	Sin otros atributos de educación.	88.2	0.7496	83	VPH_PC_%

				d5	Solo delta_score.	81.9	0.6176	28	delta_score
	grapr oes_6	2	c2 (80.4%)	e6	Sin otros atributos de educación.	93.5	0.7925	46	VPH_PC_%
				d6	Solo delta_score.	90.8	0.6682	28	delta_score

El tamaño de intervalo no está calculado con la fómula de Sturges, que se basa en el rango de la variable, sino definido de modo convencional con base en una escala de 0% a 100%

Organización de Datasets Experimentales:Una vez que los porcentajes de las variables seleccionadas han sido calculados y los atributos *target* han sido discretizados, se crean 12 datasets (ver Tabla 1).

Cada dataset corresponde a un atributo target deseado para crear modelos de AC y de RLS. En cada dataset se incluyen casi todos los atributos porcentuales generados a partir de las cantidades de hogares y de habitantes correspondientes a cada categoría del censo. También, se incluyen otros atributos que no requieren ningún tratamiento aritmético y que son incorporados directamente de la fuente de datos original, como los mencionados en la sección preprocesamiento.

Generación y Evaluación de Árboles de Clasificación:Los modelos de AC son creados antes que los de RLS porque con los primeros se identifica cuáles son los pares de variables que podrían tener mayor potencial para estar correlacionadas en el fenómeno de nuestro interés. La Tabla 3.1 muestra los modelos generados con el software WEKA (Witten & Frank, 2011) para los diversos datasets.

Tabla 3.1 Análisis de correlación de variables clave del censo usando RLS

Análisis No.	Variables analizadas[b]	Ecuación lineal	R^2
1	VPH_INTER_% (y), VPH_PC_% (x)	y = 0.751x - 2.496	0.925 1
2	VPH_PC_% (y), P18YM_PB_M_% (x)	y = 2.496x - 3.030	0.799 0
3	VPH_PC_% (y), GRAPROES (x)	y = 6.209x - 29.243	0.786 1
4	VPH_INTER_% (y), P18YM_PB_F_% (x)	y = 1.782x - 4.689	0.739 0
5	VPH_INTER_% (y), GRAPROES (x)	y = 4.454x - 23.060	0.663 0
6	VPH_INTER_% (y), GRAPROES_F (x)	y = 4.221x - 20.910	0.645 0
7	VPH_CEL_% (y), GRAPROES (x)	y = 13.009x - 47.690	0.632 1
8	VPH_INTER_% (y), VPH_TELEF_% (x)	y = 0.409x - 3.160	0.573 0
9	VPH_TELEF_% (y), VPH_REFRI_% (x)	y = 0.462x - 5.699	0.520 0
10	VPH_INTER_% (y), VPH_CEL_% (x)	y = 0.240x - 2.782	0.519 0
11	VPH_TELEF_% (y), VPH_AUTOM_% (x)	y = 0.556x + 6.554	0.472 0
12	VPH_TELEF_% (y), GRAPROES (x)	y = 6.616x - 20.263	0.428 9

VPH_INTER_% es porcentaje de hogares con Internet; *VPH_PC_%* es porcentaje de hogares con PC; *P18YM_PB_M_%* es porcentaje de habitantes que tienen 18 o más años que son varones y que tienen educación post-básica; *GRAPROES* es el promedio de años de educación; *P18YM_PB_F_%* es el porcentaje de habitantes que tienen 18 o más años que son mujeres y que tienen educación post-básica; *GRAPROES_F* es promedio de años de educación de población femenina; *VPH_CEL_%* es porcentaje de hogares con teléfono celular; *VPH_TELEF_%* es porcentaje de hogares con teléfono fijo; *VPH_REFRI_%* es porcentaje de hogares con refrigerador; *VPH_AUTOM_%* es porcentaje de hogares con automóvil

Los criterios de aceptación (Coria et al., 2013) de los AC son: 1) accuracy mayor o igual que 75%, 2) Kappa (Cohen, 1960) mayor o igual que 0.67 y 3) número de clases mayoritarias mayor o igual que 2. Posteriormente, con base en (Coria et al., 2013) y (Coria et al., 2013b), de cada árbol aceptado se debe: 1) extraer sus reglas clasificadoras, 2) calcular sus respectivos valores de soporte (*support*, *coverage*) y de confianza (*confidence*), 3) ordenar las reglas con base en su valor de soporte para identificar las más significativas y, 4) agrupar las reglas correspondientes a cada clase para identificar los perfiles de los municipios.

Análisis de RLS:Una vez generados los AC e identificados los que cumplan los criterios de aceptación, se realiza un análisis de RLS para cada uno de aquellos usando, por ejemplo, Microsoft Excel, Calc de OpenOffice, R, etc.

Para cada AC elegido, su correspondiente análisis de RLS se realiza designando como variable independiente (x) a la variable numérica correspondiente al atributo que aparece como raíz del AC, y como variable dependiente (y) al atributo porcentual correspondiente al atributo nominal que se haya definido como *target*. Después, se calculan la ecuación lineal para y y la medida de correlación R^2, considerando los valores generalmente aceptadas en la literatura para su interpretación: -1.0, alta correlación negativa; 0.0, no existe correlación; +1.0, alta correlación positiva.

Por ejemplo, en la Tabla 3, en el árbol *i6pc*, el target es *inet_6* y la raíz es *VPH_PC_%*, así que su RLS se realiza designando a *VPH_PC_%* como x y a *VPH_INTER_%* como y. Los análisis se generan así porque el atributo que aparece como raíz en un AC es el que tiene el mayor poder discriminador entre todos los atributos de su dataset para clasificar instancias y, por lo tanto, tiene también el mayor potencial para estar correlacionado con el target. De este modo, se puede medir, por ejemplo, la correlación entre presencia de PC y presencia de Internet en hogares de los municipios.

3.7 Resultados

La Tabla 3 presenta las principales características de los datasets experimentales junto con los resultados de los 22 modelos de AC de estos (todos los modelos de AC completos están disponibles con los autores). La Tabla 3.1 muestra los resultados de RLS. En la Tabla 1, para cada aspecto del fenómeno de interés (Internet, PC, teléfono fijo, telefono celular y promedio de años de educación), se define una serie de targets en varios datasets considerando diferentes tamaños de intervalo que, a su vez, determinan al número de clases en el correspondiente dataset. Las clases mayoritarias son aquellas que, con base en análisis de Pareto, constituyen aproximadamente el 80% de las instancias del dataset.

Con cada dataset se producen varios modelos de AC, en los cuales se incorporan o se excluyen manualmente algunos atributos dependiendo del target. Por otra parte, en la Tabla 3 se presentan los resultados de RLS que tienen R^2 mayor o igual que +0.4; están ordenados con base en este valor en forma descendente, de modo que se identifican los casos que tienen las correlaciones positivas más significativas.

Resultados de Árboles de Clasificación (AC): Con base en los criterios de aceptación enumerados en la sección Generación y Evaluación de Árboles de Clasificación, la Tabla 1 muestra que solamente dos modelos de AC son aceptables: 1) *i6pc*: su target es la presencia de Internet con tamaño de intervalo 6, usando el atributo *VPH_PC_%* junto con otros predictores, y 2) *e5*: su target es el grado promedio de estudios con tamaño de intervalo 5, sin utilizar ninguno de los otros atributos predictores relacionados con educación. De los árboles aceptados, se muestran como ejemplo las reglas que tienen mayor soporte:

En *i6pc*: Si *VPH_PC_%*<=11.9 AND *VPH_PC_%*<=8.5, entonces *inet_6*: c_{12} (ocurre en 1,154 de 2,456 municipios, sin excepción). Significa: si los hogares con PC son menores o iguales al 11.9% y además son menores al 8.5%, entonces el municipio pertenece a la clase 12 de *inet_6*, teniendo entre 0% y 6% de presencia de Internet en hogares.

En *e5*: Si *VPH_PC_%*<=10.2 AND *VPH_PC_%*<=7.0 AND *PNACENT_F_%*>44.8, entonces *graproes_5*: c_3 (ocurre en 902 de 2,456 municipios, con 7 excepciones). Significa: si los hogares con PC son menores o iguales al 10.2% y además son menores al 7.0% y la población femenina nacida en la misma entidad federativa del municipio es mayor al 44.8%, entonces el municipio pertenece a la clase 3 de *graproes_5*, teniendo entre 2 y 6 años de educación.

Para aprovechar plenamente los dos modelos aceptados se requeriría analizar y organizar todas sus reglas, generando los perfiles de las clases de los municipios como se indica en la sección Generación y Evaluación de Árboles de Clasificación.

Resultados de Regresión Lineal Simple (RLS).La Tabla 3.1 muestra 12 análisis de RLS, incluyendo sus correspondienes pares de variables, ecuaciones lineales y valores de R^2. Cada *x* corresponde a la raíz de un AC, y su *y* corresponde al target de ese árbol. La R^2 más alta (0.9251) corresponde al par *VPH_INTER_%* y *VPH_PC_%*, lo cual sugiere que la presencia de Internet en hogares está asociada a la presencia de PC. Este patrón podría parecer trivial; sin embargo, no lo es porque otros dispositivos de TIC para acceder a Internet que son distintos a la PC tienen una presencia creciente en México.

El segundo R^2 (0.7990) corresponde al par: PC y población que tiene 18 años o más, que tiene educación post-básica y es de sexo masculino; esto sugiere que la presencia de PC en hogares está asociada con la educación media superior o superior de la población masculina. La tercera R^2 es 0.7861 y corresponde al par de: PC y promedio municipal de años de educación, lo cual sugiere que la presencia de PC en hogares está asociada con el logro educativo de los municipios.

3.8 Discusión

Los resultados generados por la metodología propuesta son consistentes con diversas teorías expuestas inicialmente y el enfoque es novedoso.

La interacción entre brecha digital, brecha educativa y características socio-económicas y demográficas de la población ha sido poco abordada en trabajos previos haciendo uso de técnicas de aprendizaje automático tales como los AC. Así como el uso de AC en (Coria et al., 2013) y (Coria et al., 2013b) facilita el descubrimiento de perfiles de brecha digital de ciudades, también facilita el descubrimiento de perfiles de brecha educativa, como puede observarse en los resultados del modelo *e5*.

El uso del nivel de ciudad (específicamente, de municipio) como unidad de análisis y modelación es original para abordar los fenómenos de interés de esta investigación. Además, es relevante en el contexto particular de México porque el municipio es una unidad territorial y política que tiene autonomía institucional con base en la Constitución Federal. Esto le da facultades para planeación y toma de decisiones que podrían aprovecharse para establecer políticas públicas de alcance local para enfrentar las problemáticas de las brechas digital y educativa.

En los análisis realizados con RLS, las fuertes correlaciones entre los atributos de Internet o de PC, con los atributos educativos, son consistentes con investigaciones internacionales sobre el tema, p. ej. (Robinson, DiMaggio & Hargittai, 2003). Una de las correlaciones descubiertas empíricamente en nuestra investigación es: presencia de Internet, y población que tiene 18 años o más con educación post-básica y es de sexo femenino (R^2=0.7390); es decir, la presencia de Internet en hogares podría estar asociada a la presencia de mujeres con educación media superior o superior. Algunas aplicaciones potenciales de los modelos de AC y de los análisis RLS que se generaron en esta investigación serían ser, por ejemplo: generar los perfiles completos de brecha digital y educativa de los municipios mexicanos (o de otros países usando sus respectivos datos) y definir prioridades y líneas de acción para contribuir a la solución de estas problemáticas. Además, con los modelos de AC generados se podrían desarrollar sistemas de soporte a las decisiones (DSS, *decision support systems*).

3.9 Conclusiones y trabajo futuro

Este artículo ha presentado una metodología de minería de datos para abordar la investigación sobre las relaciones entre brecha digital, brecha educativa y aspectos demográficos y socio-económicos de ciudades (municipios), haciendo uso de censos nacionales. Los resultados empíricos muestran que los árboles de clasificación generados con el algoritmo J4.8 y la regresión lineal simple son técnicas útiles para hallar patrones no triviales que describen interacciones entre el promedio de años de educación y la presencia de: Internet, PC, teléfono fijo, y teléfono celular en hogares de los municipios.

Las principales contribuciones científicas de esta investigación son: 1) una metodología que puede aplicarse sobre datos de censos nacionales de población y vivienda para analizar y modelar el fenómeno de la interacción entre brechas digital y educativa y otros aspectos socio-económicos y demográficos de las ciudades; 2) para el contexto específico de México, descubrimiento de patrones representados como árboles de clasificación que muestran los perfiles de municipios respecto a las brechas digital y educativa, y 3) identificación de pares de variables que presentan una alta correlación positiva dentro del fenómeno de la relación entre brechas digital y educativa en México.

En trabajo futuro se podría analizar y modelar datos de censos de otros países para descubrir sus respectivos patrones, pudiendo contrastarse estos con los de México. También, se podría probar PART y otros algoritmos clasificadores.

Otras variables que son relevantes para los fenómenos de las brechas digital y educativa, tales como el ingreso promedio por municipio, podrían incorporarse a los datasets experimentales generados para realizar nuevos modelos y análisis.

3.10 Referencias

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, vol. 20, pp. 37–46, 1960.

Chinn, D.M. & Fairlie, W. R. The determinants of the global digital divide: A crosscountry analysis of computer Internet penetration. Oxford Economic Papers, vol. 59, No. 1, pp.16–44, 2007.

Coria, S.R., Mondragón-Becerra, R., Pérez-Meza, M., Ramírez-Vásquez, S. K., Martínez-Peláez, R., Barragán-López, D. & Ávila-Barrón, O. CT4RDD: Classification trees for research on digital divide. Expert Systems with Applications, vol. 40, pp. 5779–5786, 2013.

Coria, S.R.; Pérez-Meza, M., Mondragón-Becerra, R., Barragán-López, D. Metodología de Minería de Datos para Perfilamiento Cuantitativo de la Brecha Digital de Ciudades.

Congreso Mexicano de Inteligencia Artificial (COMIA 2013). Francisco I. Madero, Hgo., Mexico, May 28-31, 2013. In: M. González-Mendoza y F. Castro-Espinoza (eds.), Research in Computing Science. Vol. 62. Avances en Inteligencia Artificial. México, pp. 3-13, 2013b.

Coria, S.R.; Ramirez-Vásquez, S.K., Luna-Trejo, J., Mondragón-Becerra, R., Pérez-Meza, M., Avila-Barron, O. Delta score: a novel simplified measurement for digital divide of cities. In: Procs. of the 14th Annual International Conference on Digital Government Research. Quebec, Canada, 2013c, pp. 102-110.

Frank, E. & Witten, I.H. Generating accurate rule sets without global optimization. In: 15th IMLS International Conference on Machine Learning, pp. 144–151. Morgan Kaufmann Publishers Inc., San Francisco (1998).

Han, J., Kamber, M. & Pei, J. Data mining: Concepts and techniques (3rd ed.). Waltham: The Morgan Kaufmann Series in Data Management Systems, 2011.

Instituto Nacional de Estadística y Geografía [INEGI]. (2011). Base de datos por localidad del censo nacional de población y vivienda 2010, http://www.inegi.org.mx

Instituto Nacional de Estadística y Geografía [INEGI] (2011b). Conformación de la base de datos por localidad del censo nacional de población y vivienda 2010, http://www.inegi.org.mx

Organisation for Economic Co-operation and Development [OECD] (2007). Glossary of statistical terms. http://stats.oecd.org/glossary/index.htm.

Organisation for Economic Co-operation and Development [OECD] (2010). The policy debate about technology in education. In: Are the New Millennium Learners Making the Grade?: Technology Use and Educational Performance in PISA 2006, OECD Publishing. http://dx.doi.org/10.1787/9789264076044-4-en Organisation for Economic Co-operation and Development [OECD] (2012). Internet adoption and use: Households and individuals. In: OECD Internet Economy Outlook 2012, OECD Publishing. http://dx.doi.org/10.1787/9789264086463-6-en

Organisation for Economic Co-operation and Development [OECD] (2012b). Equity and Equality of Opportunity. In: Education Today 2013: The OECD Perspective, OECD Publishing. http://dx.doi.org/10.1787/edu_today-2013-11-en
Quinlan, R. C4.5: Programs for machine learning. San Mateo: Morgan Kaufmann Publishers, 1993.

Robinson, J.P., DiMaggio, P. & Hargittai, E. New social survey perspectives on the digital divide. IT & Society, Vol. 1, No. 5, pp. 1-22, 2003.

Sturges, H.A. The choice of a class interval. Journal of the American Statistical Association, vol. 21, No. 153, pp. 65–66, 1926.

Yelland, N. & Neal, G. Aligning digital and social inclusion: A study of disadvantaged students and computer access. Education and Information Technologies (2013) 18:133–149. DOI 10.1007/s10639-012-9223-y.

Witten, I.H. & Frank, E. Data mining: Practical machine learning tools and techniques, 3rd ed. Burlington: Morgan Kaufmann Series in Data Management Systems, 2011.