ANÁLISIS DE VICTIMIZACIÓN Y VIOLENCIA EN MÉXICO

Daniel E. Vázquez*, Arturo de León Chapa,**, Agustín Santiago *, Janitza Arias *** y Mercedes Gaitán Angulo
*Facultad de Matemáticas de la UAGro.
**Unidad Académica de ciencia y tecnologías de la información de la UAGro.
***Centro de Investigaciones de la Escuela de Negocios (CIEN), Fundación Universitaria Konrad Lorenz, Colombia.

ABSTRACT
In recent years, in Mexico and in the world, surveys on victimization and perception of violence (ENVIPE, 2016) have become very important. This information is collected through a national sample collected by the Instituto Nacional de Geografía Estadística e Informática (INEGI, 2017), with the purpose of providing governments, and the authorities responsible for functions related to security and justice, information relevant on different aspects of this phenomenon, with the objective that decision making is made based on objective information. Unfortunately, most of these studies remain at the descriptive level and leave aside the multiple relationships that can occur between variables or sets of variables, whose understanding would give greater clarity about the phenomenon, by identifying factors, similarities or dissimilarities, dependency relationships or association, adjust models and make forecasts for intervention purposes. In order to identify some of these multiple relationships, a set of supervised methods of data mining, such as classification trees, regression trees and rule-based methods, are applied to the ENVIPE 2016 database (INEGI, 2016); in addition to some unsupervised methods such as Clustering. The procedures are executed in the software called WEKA, the models are adjusted using a training set, statistically significant attributes are identified and the models are adjusted. The selection of the models is done based on the criteria of goodness defined for them and later predictions are made in the cases in which it applies. The variables associated with victimization and perceptions of violence are identified, according to the association criterion, and predictions are made, in the probabilistic sense, about these phenomena.

KEYWORDS: Victimization, Data Mining, Decision Tree, Clustering.

MSC: 62-07 (62P25)

RESUMEN
En los últimos años, en México y en el mundo, ha cobrado gran relevancia la realización de encuestas sobre victimización y percepción de la violencia (ENVIPE, 2016): En México, Esta información se recopila a través de una muestra nacional recogida por del Instituto Nacional de Geografía, Estadística e informática (INEGI, 2017), con la finalidad de proporcionar a los gobiernos, y a las autoridades responsables de las funciones vinculadas con la seguridad y la justicia, información relevante sobre distintos aspectos de este fenómeno, y con el objetivo que la toma de decisiones se haga sobre la base de información objetiva. Lamentablemente la mayoría de estos estudios se queda en el nivel descriptivo y dejan de lado las múltiples relaciones que pueden presentarse entre variables o conjuntos de variables, cuya comprensión daría mayor claridad sobre el fenómeno, al identificar factores, similitudes o disimilitudes, relaciones de dependencia o asociación, ajustar modelos y hacer pronósticos con fines de intervención. Con la finalidad de identificar algunas de estas múltiples relaciones se aplican a la base de datos ENVIPE 2016 (INEGI, 2016), un conjunto de métodos supervisados de minería de datos, tales como Árboles de clasificación, Árboles de regresión y métodos basados en reglas; además de algunos métodos no supervisados como Clustering. Se ejecutan los procedimientos en el software denominado WEKA, se realiza el ajuste de modelos utilizando un conjunto de entrenamiento, se identifican atributos estadísticamente significativos y se ajustan los modelos. La selección de los modelos se realiza a partir de los criterios de bondad definidos para ellos y posteriormente se realizan predicciones en los casos en los que aplica. Se identifican las variables asociadas a la victimización y percepción de la violencia, según el criterio de asociación, y se hacen predicciones, en el sentido probabilístico, sobre estos fenómenos.

PALABRAS CLAVE: Victimización, Data Mining, Decision Tree, Clustering.

1. INTRODUCCIÓN

Uno de los problemas más graves que enfrenta la población mexicana es la seguridad. Esta se ha visto vulnerada en las últimas décadas debido al elevado nivel de violencia y delincuencia, convirtiéndose en el principal reto social para el gobierno y para la población que la padece. Según datos de la Tasa de incidencia delictiva por entidad federativa de ocurrencia por cada cien mil habitantes (SESNS, 2017), el número promedio de delitos en el país en 2010 fue de 30535, en el siguiente año se mantuvo aproximadamente en el

mismo valor (hubo un ligero decremento de 4.4%), pero en 2012 se presentó in incremento en la incidencia delictiva de aproximadamente el 20%, en promedio 35139 casos. Para el 2013 la situación empeoró, toda vez que, respecto del año anterior se presentó un aumento en 18% y tuvo un aumento marginal el siguiente año. En el año 2015 se registró una ligera disminución con respecto del año anterior (-15%), situándose en cifras similares al 2012. A partir de 2015 se han registrado aumentos significativos en la incidencia delictiva superiores al 5%. Los estados que mayor incidencia delictiva promedio han registrado, en este periodo, son el Estado de México, Ciudad de México, Baja California, Jalisco, Guerrero, Quintana Roo, Sonora, Morelos y Aguascalientes, con 60897, 49715, 42611, 41851, 39084, 37746, 37223, 36756 y 36406 casos, respectivamente. Con estas cifras, es natural que el gobierno implemente distintos programas para hacer frente a la delincuencia, como, por ejemplo, "Guerra contra el narcotráfico" del Presidente Felipe Calderón en 2006. A partir del año 2012, el presidente Enrique Peña Nieto, anunció durante la sesión del Consejo Nacional de Seguridad Pública (Olson, 2012), seis líneas de acción en materia de seguridad, entre las que destacan la creación de la Gendarmería Nacional, que contaría con 10 mil elementos, la revisión de la figura del arraigo y el fortalecimiento de la Procuraduría General de la República. Dijo que con el objetivo de apoyar labores de seguridad en municipios con "mayor debilidad", el territorio nacional se dividiría en cinco regiones operativas consensuadas con los gobiernos estatales. El mandatario estableció como una de las prioridades de su plan el fortalecimiento de los derechos humanos, la creación de un fondo de prevención del delito, así como una evaluación permanente, con indicadores claros y transparentes de los cuerpos de seguridad. Las líneas de acción contemplaron:

1) Planeación, para reducir la violencia y recuperar la paz de las familias mexicanas; así como disminuir los indicadores, de secuestro y homicidio.

Prevención. Seguido a la planeación, esta acción requiere la participación ciudadana. El mandatario explicó que hay que atender las causas del fenómeno delictivo y no sólo de los efectos. Así también anunció un programa de rescate de los espacios públicos, en el que solicitaría un fondo de 115,600 millones de pesos y un fondo para víctimas, de donde se destinará para mejorar el entorno urbano, entre otros planes.
La tercera línea de acción, protección de los derechos humanos, con el Programa Nacional de Derechos Humanos y el fortalecimiento de las dependencias federales.
En un cuarto punto llama a depurar y reestructurar el Instituto Nacional de Migrantes y a la instrumentación de una policía pública para atender casos de personas no localizadas.
En su quinto punto anunció la creación de la Gendarmería Nacional, que apoyará labores de seguridad en municipios con 'mayor debilidad', así como la creación del mando único de policías estatales.
En su último punto se destaca que la evaluación del cuerpo policíaco sería fundamental para devolver la paz a los mexicanos.
De estos puntos se derivaron un conjunto de acciones que incluyen la participación del ejército, la marina, la policía federal, las policías estatales y las policías municipales, sin embargo, según los resultados derivados de la Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública (ENVIPE, 2016), la victimización y percepción sobre seguridad, está lejos de disminuir, por lo que en muchos casos la población se ha visto en la necesidad de auto defenderse, originando la aparición de las llamadas "policías ciudadanas", ante la percepción manifiesta de que son ineficaces las acciones del estado para hacer frente a este mal nacional.
La ENVIPE, cuya realización data del 2008, pero que utiliza el mismo cuestionario estandarizado a partir del 2010 (los resultados se presentaron en 2011), presenta resultados descriptivos sobre victimización del hogar y victimización personal, se estima el número de víctimas en el año, el número de delitos, la cifra negra, se mide la percepción de seguridad del lugar donde las personas viven, el grado de confianza en las instituciones, cambios en actividades y hábitos, costos de la delincuencia en personas, repercusiones del delito sobre las víctimas, identifica y mide actitudes y experiencias de las víctimas con las instituciones de seguridad pública y de procuración de justicia. A partir de esta información se pretende que el estado implemente políticas públicas en materia de seguridad y victimización.
Consideramos que, para que la tarea de diseño de políticas públicas en materia de seguridad y victimización sea efectiva, es necesario contestar los "por qué" de las distintas situaciones que se presentan en este problema tan complejo. Por esta razón, hemos iniciado un análisis de la base de datos de la ENVIPE, con el objetivo de identificar grupos de variables que se agrupan, a partir de similitudes, e intentar explicar a partir de las hipótesis que se deriven de tal asociación, las múltiples relaciones que se pueden presentar a partir de la concurrencia de las variables en dichos grupos.

Una de las razones para analizar el crimen es para maximizar el uso limitado de los recursos para la aplicación de la ley (Swadi, 2011), por tal razón, si identificamos las variables que pueden caracterizar un tipo de delito, estaremos en condiciones de proponer a los tomadores de decisiones recomendaciones óptimas sobre el tema. El análisis se realizará a partir de métodos de aprendizaje no supervisado de minería de datos, tales como análisis cluster, mediante el algoritmos K-means y el algoritmo EM (Agarwal J., 2013), para explicar los distintos fenómenos asociados a la victimización se han utilizado métodos de aprendizaje supervisado, basados principalmente en árboles de decisión. En el presente trabajo se ejecutan los procedimientos en el software denominado WEKA, se realiza el ajuste de modelos utilizando un conjunto de entrenamiento, se identifican atributos estadísticamente significativos y se ajustan los modelos. La selección de los modelos se realiza a partir de los criterios de bondad definidos para ellos y posteriormente se realizan predicciones en los casos en los que aplica. Se identifican las variables asociadas a la victimización y percepción de la violencia, según el criterio de asociación, y se hacen predicciones, en el sentido probabilístico, sobre estos fenómenos. A partir de lo anterior, se identifican grupos de variables asociadas a la victimización y percepción de violencia, con la finalidad de incidir en que las políticas públicas en esa materia se orienten a la eliminación, mitigación o control de las causas de ese fenómeno social.
Específicamente se plantea: Identificar mediante procedimientos de minería de datos las variables asociadas a la victimización y percepción de la violencia; demás de proponer modelos de clasificación que permitan estimar la probabilidad de que ocurra victimización o violencia, dadas las variables que se suponen asociadas al fenómeno y finalmente seleccionar el mejor modelo de clasificación a partir de la eficiencia en la clasificación.

2. METODOLOGÍA.

2.1. Descripción de la base de datos.

La base de datos de la encuesta ENVIPE 2016, consta de 6 tablas principales relacionadas, por lo que se realizó la construcción de una consulta, para obtener la primera versión de la "vista minable". La estructura de la base de datos responde a la estructura del cuestionario utilizado, de tal manera que se han generado las tablas siguientes:

- Tabla de la vivienda seleccionada (**TVivienda**): Esta tabla contiene las características de las viviendas que fueron captadas por la carátula del cuestionario, así como el recuento de los residentes de la vivienda y el número de hogares, es decir, la sección I del cuestionario.
- Tabla del hogar (**THogar**): Esta tabla contiene el resultado de la entrevista, el renglón del informante elegido, el factor hogar, entre otras características del hogar.
- Tabla de residentes del hogar (**TSDem**): Esta tabla contiene las características sociodemográficas de los integrantes del hogar, esto corresponde a la sección III del cuestionario.
- Tabla de percepción sobre seguridad y desempeño institucional (**TPer_Vic1**): Esta tabla recoge información sobre la percepción sobre seguridad pública y el desempeño institucional. Es decir, las secciones IV y V del cuestionario personal.
- Tabla de percepción sobre seguridad, desempeño institucional y victimización (**TPer_Vic2**): En esta tabla se encuentra la información de la victimización en el hogar y la victimización personal. Esto es, las secciones VI y VII del cuestionario personal.
- Tabla del módulo sobre victimización (**TMod_Vic**): Tabla que incluye información sobre los delitos ocurridos de los que fue víctima la persona elegida y su hogar durante el año de referencia.

El procedimiento para generar la consulta fue el siguiente: Primero se revisó el tamaño de cada una de las tablas (número de registros), se eliminan columnas redundantes, innecesarias y las que tienen solo valores nulos, por lo que la consulta final con todas las tablas fue

```
USE envipe2016
SELECT *
FROM
 ((((dbo.TSDem
 LEFT JOIN dbo.TPer_Vic1 ON dbo.TSDem.ID_PER = dbo.TPer_Vic1.ID_PER)
 LEFT JOIN dbo.TPer_Vic2 ON dbo.TSDem.ID_PER = dbo.TPer_Vic2.ID_PER)
```

```
LEFT JOIN dbo.THogar ON dbo.TSDem.ID_HOG = dbo.THogar.ID_HOG)
LEFT JOIN dbo.TMod_Vic ON dbo.TSDem.ID_PER = dbo.TMod_Vic.ID_PER)
LEFT JOIN dbo.TVivienda ON dbo.TSDem.ID_VIV = dbo.TVivienda.ID_VIV;
```

Para los métodos de aprendizaje supervisado se utilizó la base de datos completa, generada a partir de la consulta anterior y, para el caso de los métodos de aprendizaje no supervisado se ha utilizado la base de datos de la ENVIPE 2016, específicamente el módulo sobre victimización, de la cual se tomó la tabla Tper_Vic1 y las variables aream (área metropolitana), ap4_6_1, ap4_6_2 y ap4_63. En el documento "Estructura de la base de datos de la encuesta envipe 2016", en su página 21, sección 4.6

Tabla 1: En lo que resta de 2016, por los lugares donde transita o por el tipo de actividades que realiza,

¿Cree que a usted le pueda ocurrir...

Un robo o asalto en la calle o en el transporte público.	AP4_6_1	Numérico	1 2 3 9	Sí No No aplica No sabe / No responde
lesiones por una agresión física?	AP4_6_2	Numérico	1 2 3 9	Sí No No aplica No sabe / No responde
una extorsión o secuestro para exigirle dinero o bienes?	AP4_6_3	Numérico	1 2 3 9	Sí No No aplica No sabe / No responde

Continuando con la tabla Tper_vic1, se consideró la sección 4.2 del cuestionario en donde preguntan: De los temas que le voy a mostrar, ¿cuáles son los tres que le preocupan más? En esta sección se proporcionan 13 variables y quienes responden solo pueden elegir 3 de estas.
Esto generó una serie de patrones de los cuales se buscaron cuáles fueron las principales preocupaciones por área metropolitana. Las variables y sus posibles valores se muestran a continuación:

Tabla 2: De los temas que le voy a mostrar, ¿cuáles son los tres que le preocupan más?

VARIABLE	NOMBRE	VALORES	ETIQUETAS
Pobreza	AP4_2_01	0 1	No se declaró como opción afirmativa Sí
Desempleo	AP4_2_02	0 1	No se declaró como opción afirmativa Sí
Narcotráfico	AP4_2_03	0 1	No se declaró como opción afirmativa Sí
Aumento de precios	AP4_2_04	0 1	No se declaró como opción afirmativa Sí
Inseguridad	AP4_2_05	0 1	No se declaró como opción afirmativa Sí
Desastres naturales	AP4_2_06	0 1	No se declaró como opción afirmativa Sí
Escases de agua	AP4_2_07	0 1	No se declaró como opción afirmativa Sí
Corrupción	AP4_2_08	0 1	No se declaró como opción afirmativa Sí
Educación	AP4_2_09	0 1	No se declaró como opción afirmativa Sí
Salud	AP4_2_10	0 1	No se declaró como opción afirmativa Sí
Falta de castigo a delincuentes	AP4_2_11	0 1	No se declaró como opción afirmativa Sí
Otro	AP4_2_12	0 1	No se declaró como opción afirmativa Sí
No sabe/no responde	AP4_2_99	0 1	No se declaró como opción afirmativa Sí

3. PROCEDIMIENTO DE ANÁLISIS DE ENTREVISTAS CON VICTIMIZACIÓN Y SIN VICTIMIZACIÓN.

El objetivo de este primer ejercicio consistió en identificar el patrón de respuesta que tiene la población de cada una de las áreas metropolitanas del país. Aunque se debe considerar que el área metropolitana cubre también los municipios alrededor de esta. Se crearon clústeres con las diferentes combinaciones de respuesta disponibles en las variables de la base de datos y posteriormente se determina a que patrón de respuesta corresponde cada área metropolitana. Para conseguir esto se tomó la tabla Tper_Vic1 y se escogieron solo aquellos registros que cumplan con las siguientes características:
La variable resul_h tenga un valor de A, y que no tenga en blanco alguna columna con las que se va a trabajar. El valor A de la variable resul_h, nos dice que contiene: "**entrevista completa con victimización**", es decir, de aquellos que concluyeron la entrevista con algún grado de victimización y que son el conjunto de individuos que se desean estudiar por el momento. La consulta para obtener estos registros fue la siguiente:

```
select aream,ap4_6_1,ap4_6_2,ap4_6_3,ap4_6_1 from tper_vic1 where resul_h='A' and aream! =”” and ap4_6_1!=”” and ap4_6_2!=”” and ap4_6_3!=””
```

De esta consulta se obtuvieron, solo 11,121 registros que cumplieron con el criterio. Para su análisis se requería que una de las variables categóricas, es por eso que la variable aream se sustituyó su valor numérico por el nombre del área que le corresponde en el catálogo de áreas metropolitanas. Este procedimiento se realizó con apoyo de Libre Office 5.0.6.2. El análisis se realizó para los algoritmos EM y K- means sobre el archivo resultante.
De forma análoga se realizó el análisis de entrevistas sin victimización.

4. RESULTADOS Y DISCUSIÓN

4.1. Análisis de victimización con el algoritmo k-means.

El método k-medias es muy eficiente en términos de ejecución y funciona muy bien con grandes volúmenes de datos (Maimon & Rokach, 2010). (Dean, 2014), afirma que es el algoritmo más utilizado por su simplicidad y velocidad.
Mediante este procedimiento se tomará un máximo de 33 clústeres, ya que es máximo de áreas metropolitanas disponibles y bajo el supuesto que cada una de estas caiga dentro de un cluster diferente.
Se arribó al resultado en dos iteraciones, utilizando la distancia euclidiana, con una suma de cuadrados dentro de cluster de 0.128, generando 33 clústeres. Los valores perdidos son reemplazados por la media o la moda y se construye el modelo a partir de un conjunto de entrenamiento obteniendo 8 clústeres como se indica en la siguiente tabla.

Tabla 3: Clúster con proporción de instancias distintas de cero.

Cluster	Instancias	Porcentaje
0	5167	46%
1	1790	16%
2	1214	11%
3	767	7%
4	1416	13%
5	246	2%
6	122	1%
7	154	1%

El resultado indica que hay un máximo de 8 clústeres probables, sin embargo, después de varias pruebas, se consideraron un máximo de 8 clústeres y un mínimo de 4, para realizar este primer estudio. Al parecer, con 5 clústeres se agrupan los datos de una manera que quedan mejor distribuidos. Ejecutamos el procedimiento utilizando 5 clústeres, consiguiendo convergencia en 7 iteraciones, una suma de cuadrados dentro de clústeres de 233.84 y centroides dados en la siguiente tabla.

Tabla 4: Centroides para los clústeres finales sobre victimización en el hogar

Attribute	Full Data (11121.0)	0 (5292.0)	1 (2001.0)	2 (1378.0)	3 (1031.0)	4 (1419.0)
ap4_6_1	1.221	1.0242	1.077	2.4006	1.2386	1

ap4_6_2	1.401	1	1	2.3657	2.1232	2
ap4_6_3	1.5168	1	2.1844	2.4187	1	2.0021

Tabla 5: Clústeres finales

Cluster	Instancias	Porcentaje
0	5292	48%
1	2001	18%
2	1378	12%
3	1031	9%
4	1419	13%

Este resultado nos indica se han detectado básicamente 4 configuraciones de respuesta que predominan en los cuestionarios y que son:

Tabla 6: Interpretación de las configuraciones dentro de cluster.

Cluster	Patrón	Interpretación
0	1, 1, 1	El peor de todos los casos, toda vez que el entrevistado considera que, durante un trayecto dentro de su comunidad o por la actividad que realiza puede ser víctima de un robo o asalto en la calle o el transporte público, puede ser víctima de alguna lesión o de una extorsión o secuestro.
1	1, 1, 2	Aquí los entrevistados consideran que durante un trayecto dentro de su comunidad, o por la actividad que realiza, puede ser víctima de un robo o asalto en la calle o el transporte público, puede ser víctima de alguna lesión pero no de un secuestro o extorsión.
2	2, 2, 2	Este es el mejor escenario, puesto que el entrevistado considera que no pueden ser víctimas de alguno de los delitos mencionados.
3	1, 2, 1	En esta combinación de respuestas el entrevistado considera que durante un trayecto dentro de su comunidad o por la actividad que realiza, puede ser víctima de un robo o asalto en la calle o el transporte público o de una extorsión o secuestro.
4	1, 2, 2	Aquí, consideran que durante un trayecto dentro de su comunidad o por la actividad que realiza, sólo pueden ser víctima de un robo o asalto en la calle o el transporte público.

Fuente: Elaboración propia

Si clasificamos estos clústeres de acuerdo al nivel de percepción de inseguridad (1, 1, 1), que corresponde al cluster 0, clasificaría como *Alto*, los cluster 1 (1, 1, 2) y 3 (1, 2, 1), clasifican como *Medio*, el cluster 4 (1, 2, 2) clasifica como *Bajo* y el cluster 2 (2, 2, 2) como *Muy bajo*.

Los resultados anteriores serían de utilidad para las autoridades y a partir de ella podrían diseñar políticas públicas para mejorar la percepción o para prevenir el delito, si existe posibilidad identificar zonas en las cuales la incidencia del delito es mayor. Para conseguir esto se investigó si con la información que se tiene, es posible relacionar algunas de las zonas metropolitanas con alguno de los clústeres. Se ejecuta en Weka la opción "classes for cluster evaluation" y se selecciona la variable aream, obteniendo el resultado:

Tabla 7: Distribución de las instancias de los cluster en área metropolitana.

Clústeres					Aream	Clústeres				
0	1	2	3	4		0	1	2	3	4
224	108	78	42	43	Aguascalientes	4.23%	**5.40%**	**5.66%**	4.07%	3.03%
178	70	57	43	31	Frontera Tijuana	3.36%	3.50%	4.14%	4.17%	2.18%
111	59	77	30	31	La paz	2.10%	2.95%	**5.59%**	2.91%	2.18%
79	30	35	26	21	Campeche	1.49%	1.50%	2.54%	2.52%	1.48%
81	38	25	18	17	Saltillo	1.53%	1.90%	1.81%	1.75%	1.20%
183	76	97	34	51	Colima	3.46%	3.80%	**7.04%**	3.30%	**3.59%**
65	25	27	10	12	Tuxtla Gutiérrez	1.23%	1.25%	1.96%	0.97%	0.85%
104	21	41	20	29	Chihuahua	1.97%	1.05%	2.98%	1.94%	2.04%
803	357	137	118	395	Ciudad de México	**15.17%**	**17.84%**	**9.94%**	**11.45%**	**27.84%**
119	53	43	31	33	Durango	2.25%	2.65%	3.12%	3.01%	2.33%
116	69	28	23	32	León	2.19%	3.45%	2.03%	2.23%	2.26%
94	26	15	23	46	Acapulco	1.78%	1.30%	1.09%	2.23%	3.24%
60	23	28	20	14	Pachuca	1.13%	1.15%	2.03%	1.94%	0.99%
299	144	64	59	105	Guadalajara	**5.65%**	**7.20%**	4.64%	**5.72%**	**7.40%**
40	14	3	15	12	Toluca	0.76%	0.70%	0.22%	1.45%	0.85%
138	52	20	15	16	Morelia	2.61%	2.60%	1.45%	1.45%	1.13%
229	42	37	60	48	Cuernavaca	4.33%	2.10%	2.69%	**5.82%**	3.38%
68	26	41	26	27	Tepic	1.28%	1.30%	2.98%	2.52%	1.90%
343	105	67	29	53	Monterrey	**6.48%**	**5.25%**	4.86%	2.81%	**3.74%**

76	34	11	28	17	Oaxaca	1.44%	1.70%	0.80%	2.72%	1.20%
257	109	49	46	69	Puebla	**4.86%**	**5.45%**	3.56%	4.46%	4.86%
286	83	63	53	58	Querétaro	**5.40%**	4.15%	4.57%	5.14%	**4.09%**
215	75	51	33	50	Cancún	4.06%	3.75%	3.70%	3.20%	3.52%
132	54	31	23	33	San Luis potosí	2.49%	2.70%	2.25%	2.23%	2.33%
147	25	20	15	17	Culiacán	2.78%	1.25%	1.45%	1.45%	1.20%
104	45	23	23	29	Hermosillo	1.97%	2.25%	1.67%	2.23%	2.04%
121	30	14	9	10	Villahermosa	2.29%	1.50%	1.02%	0.87%	0.70%
97	21	27	12	14	Tampico	1.83%	1.05%	1.96%	1.16%	0.99%
207	61	49	57	48	Tlaxcala	3.91%	3.05%	3.56%	**5.53%**	3.38%
47	14	13	12	15	Veracruz	0.89%	0.70%	0.94%	1.16%	1.06%
207	98	96	71	33	Mérida	3.91%	4.90%	**6.97%**	**6.89%**	2.33%
62	14	11	7	10	Zacatecas	1.17%	0.70%	0.80%	0.68%	0.70%
5292	2001	1378	1031	1419	Totales					

Fuente: Elaboración propia

Recordemos que el cluster 0 representa el 48% de las instancias y corresponde al peor escenario en cuanto a seguridad. Si consideramos la frecuencia relativa de las instancias dentro de cluster, éstas se pueden ordenar, por lo que las cinco áreas metropolitanas en las que se percibe mayor inseguridad son, la Ciudad de México, Monterrey, Guadalajara, Querétaro y Puebla; el cluster 1 y 3, que en conjunto representan el 27% de las instancias, identifican como áreas metropolitanas con percepción media de inseguridad a la Ciudad de México, Guadalajara, Puebla, Aguas Calientes y Monterrey, sumándose Cuernavaca y Tlaxcala; el cluster 4, que corresponde a percepción baja de inseguridad, incluye a las áreas metropolitanas Ciudad de México, Guadalajara, Puebla, Querétaro y Monterrey; el cluster con nivel de inseguridad muy baja incluye la Ciudad de México, Mérida, Cuernavaca, Guadalajara y Tlaxcala. En general podemos afirmar que las áreas metropolitanas que se consideran más seguras son Toluca y Zacatecas, puesto que en percepción de inseguridad obtienen la puntuación más baja. Observe que la Ciudad de México tiene presencia en todos los clústeres, la razón de ello es que, en una ciudad de ese tamaño, según la opinión de la gente, se reproducen todos los escenarios.

4.2. Análisis de victimización con el algoritmo EM.

El objetivo básico del algoritmo EM consiste en determinar las medias y desviaciones estándar de cada cluster, para sí poder maximizar su verosimilitud (Lewicki & Hill, 2005). Ejecutamos el algoritmo en Weka y se han obtenido los siguientes resultados, excluyendo, como en el caso anterior, la variable aream, seleccionando, en cada caso, las tres principales preocupaciones de los entrevistados, considerando un conjunto de entrenamiento.

Tabla 8: Clústeres iniciales generados mediante algoritmo EM para los problemas que preocupan más.

Atribute	Statistic	Cluster						
		0	1	2	3	4	5	6
ap4_2_01	Mean	0.1004	0.1487	0.1897	0.2725	0.4387	0.4307	0.1524
	std. dev.	0.3005	0.3558	0.3921	0.4453	0.4962	0.4952	0.3594
ap4_2_02	Mean	0.1896	0.3142	0.3491	0.4324	0.6956	0.6338	0.1958
	std. dev.	0.392	0.4642	0.4767	0.4954	0.4602	0.4818	0.3968
ap4_2_03	Mean	0.2177	0.2894	0.001	0.1338	0.0421	0.3344	0.1445
	std. dev.	0.4127	0.4535	0.0317	0.3404	0.2009	0.4718	0.3516
ap4_2_04	Mean	0.2238	0.2342	0.1933	0.2259	0.329	0.5228	0.1822
	std. dev.	0.4168	0.4235	0.3949	0.4182	0.4699	0.4995	0.386
ap4_2_05	Mean	0.6747	0.9173	0.6272	0.5647	0	0.8319	0.4772
	std. dev.	0.4685	0.2754	0.4835	0.4958	0.0031	0.374	0.4995
ap4_2_06	Mean	0	0	0	0	0	0	0.8659
	std. dev.	0.1828	0.1828	0	0.1828	0.1828	0.1828	0.3408
ap4_2_07	Mean	0.0915	0	0.1108	0.0797	0.008	0.2209	0.1349
	std. dev.	0.2883	0	0.3139	0.2709	0.0893	0.4149	0.3416
ap4_2_08	Mean	0.3731	0.9872	0.0893	0.4225	0.4579	0	0.2109
	std. dev.	0.4836	0.1123	0.2852	0.494	0.4982	0.4812	0.4079
ap4_2_09	Mean	0.0949	0	0.4188	0.8594	0.0041	0	0.1327
	std. dev.	0.2931	0	0.4934	0.3476	0.0638	0.4244	0.3393

ap4_2_10	Mean	0.1392	0.1038	1	0	0.8286	0	0.2243
	std. dev.	0.3462	0.3049	0.431	0.431	0.3769	0.431	0.4171
ap4_2_11	Mean	0.8886	0	0.0139	0	0.0556	0	0.1488
	std. dev.	0.3146	0.439	0.1172	0.439	0.2292	0.439	0.3559
ap4_2_12	Mean	0	0	0	0	0	0	0.1215
	std. dev.	0.0695	0.0695	0.0695	0.0695	0.0695	0.0695	0.3267
ap4_2_99	Mean	0	0	0	0	0.0457	0	0
	std. dev.	0.0328	0.0328	0.0328	0.0328	0.2088	0.0328	0.0328

Tabla 9: Frecuencia relativa de instancias dentro de clústeres.

Cluster	Instancias	Porcentaje
0	1722	15%
1	3244	29%
2	4571	41%
3	347	3%
4	94	1%
5	771	7%
6	372	3%

Como se observa, de estos siete clústeres, existen 3 en los cuales se concentra el 85% de instancias. Cada uno de los cluster se pueden interpretar de acuerdo al peso relativo (Mean) de los atributos dentro de cluster. Así, considerando los principales problemas tenemos:
Cluster 0: Inseguridad (0.67), Falta de castigo a delincuentes (0.88):
Cluster 1: Inseguridad (0.9173), Corrupción (0.9872):
Cluster 2: Inseguridad (0.6272), Salud (1):
Cluster 3: Inseguridad (0.5647), Educación (0.8594),
Cluster 4: Desempleo (0.6956), Salud (0.8286):
Cluster 5: Desempleo (0.6338), Inseguridad (0.8319)
Cluster 6: Desastres naturales (0.8659):
Como hemos observado en la estadística anterior el 41% de las instancias se ajustan de alguna forma al cluster 2, por lo que la mayor preocupación es la salud y en segundo término la inseguridad. El cluster 1, que concentra el 29%, tiene como principales preocupaciones la corrupción y la inseguridad, el cluster 0, que representa el 15%, tiene como principal preocupación la falta de castigo a delincuentes, es decir, la impunidad y, en segundo lugar, se percibe como principal problema la inseguridad. Como se observa, la inseguridad es una preocupación presente en cinco clústeres, aunque no es la principal preocupación.
Ahora, fragmentando la información por área metropolitana se obtienen los siguientes resultados, destacando que en la Ciudad de México están presentes, en distinto grado, todas las preocupaciones, por lo que, además de esta área metropolitana podemos identificar las representativas de cada cluster.
Cluster 0 <-- Ciudad de México, Mérida
Cluster 1 <-- Ciudad de México, Monterrey
Cluster 2 <-- Ciudad de México, Guadalajara
Cluster 3 <-- Ciudad de México, Guadalajara, Puebla
Cluster 4 <-- Ciudad de México, Aguascalientes, Mérida
Cluster 5 <-- Ciudad de México, Guadalajara
Cluster 6 <-- Ciudad de México, La paz, Frontera Tijuana
Se confirma la clasificación que nos entrega el algoritmo EM, aunque no es contundente, el cluster 2 es el que predomina en todas las áreas metropolitanas con las excepciones de Monterrey, Frontera Tijuana, Veracruz y Cancún en donde predomina de alguna forma el cluster 1.

4.3. Arboles de clasificación.

4.3.1. Algoritmo J48

Se ejecuta el procedimiento en Weka considerando 131520 instancias, que corresponde a una muestra de 40% de las instancias totales al realizar la consulta en las tablas de ENVIPE, 265 atributos, lo cual permitió abreviar tiempo de ejecución del algoritmo, que converge al alcanzar un tamaño de árbol de 44.
Los resultados son los siguientes:

Tabla 10: Stratified cross-validation J48

Correctly Classified Instances	131056	99.6472 %
Incorrectly Classified Instances	464	0.3528 %
Kappa statistic		0.9917
Mean absolute error		0.0054
Root mean squared error		0.0505
Relative absolute error		1.2764 %
Root relative squared error		10.9435 %
Total Number of Instances	131520	

Observe que las instancias clasificadas correctamente representan el 99% de la población, lo cual se confirma a partir del índice Kappa de acuerdo en la clasificación, cuyo valor es también de 0.9917, muy cercano a 1. La precisión en la clasificación se resume en la siguiente tabla.

Tabla 11: Detailed Accuracy by Class J48

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC	Area Class
	0.995	0.001	1.000	0.995	0.997	0.992	1.000	1.000	B
Weighted Avg.	0.999	0.005	0.990	0.999	0.994	0.992	1.000	0.999	A
	0.996	0.002	0.997	0.996	0.996	0.992	1.000	1.000	

Recuerde que A, representa la encuesta completa con victimización y B, encuesta completa sin victimización.

Tabla 12: Confusion Matrix J48

a	b	classified as
90594	423	a = B
41	40462	b = A

Según esta salida, las variables asociadas al resultado de victimización (RESULT_V), mediante este árbol binario fueron:

Tabla 13: Variables incluidas en el árbol.

RESULT_V (≤1)		RESULT_V (>1)	
AP2_2	¿Todas las personas que viven en esta vivienda comparten un mismo gasto para comer?	AP6_7	Durante 2015, ¿algún integrante de este hogar incluido usted, sufrieron alguna de las situaciones del grupo B?
AP6_7	Durante 2015, ¿algún integrante de este hogar incluido usted, sufrieron alguna de las situaciones del grupo B?	TOT_PER	Total de personas en el hogar.
AP6_4_03	En términos de delincuencia, dígame si se siente seguro o inseguro en la calle	NIV	¿Hasta qué año o grado aprobó en la escuela?
AP6_4_04	En términos de delincuencia, dígame si se siente seguro o inseguro en la escuela	DOMINIO	Urbano, completamente urbano y Rural
AP6_4_02	En términos de delincuencia, dígame si se siente seguro o inseguro en su trabajo	AP2_1	¿Cuántas personas viven normalmente en esta vivienda, contando a los niños chiquitos y a los ancianos? (Incluya a los trabajadores domésticos que vivan aquí)
AP6_4_01	En términos de delincuencia, dígame si se siente seguro o inseguro En su casa		
Edad			

La pregunta AP2_2 trata de establecer si en la vivienda donde se realiza la entrevista, existe un núcleo familiar o más y, es a partir de ella que se indaga si la familia o alguien de la familia o ambos, han sido víctima de la delincuencia señalando algunas de las situaciones del grupo B; que tienen que ver con Robo o asalto, fraude bancario, fraude al consumidor, extorsión, amenazas verbales, lesión física provocada por alguien con actitud abusiva, secuestro, agresión mediante hostigamiento sexual, manoseo, exhibicionismo o intento de violación, violación sexual y otros delitos distintos a los anteriores. Las respuestas a esta pregunta muestran evidencia que, en términos de delincuencia, la gente se siente insegura en la calle, en la escuela, en su trabajo y en menor medida en su casa, aunque habría que destacar que las respuestas dependen de la edad de la persona entrevistada, en este caso las personas mayores de 17 años.

En la rama derecha del árbol, la victimización se encuentra asociada a las situaciones descritas en el grupo B y depende del total de personas en el hogar, el nivel de estudios, si vive en el medio urbano, completamente urbano o rural.

De acuerdo con estos resultados es posible predecir la probabilidad de que la persona entrevistada sea víctima de algún delito, en nuestro caso el 31.37% de la población y el 68.63% clasifica en el grupo de los que no han sido víctima del delito.

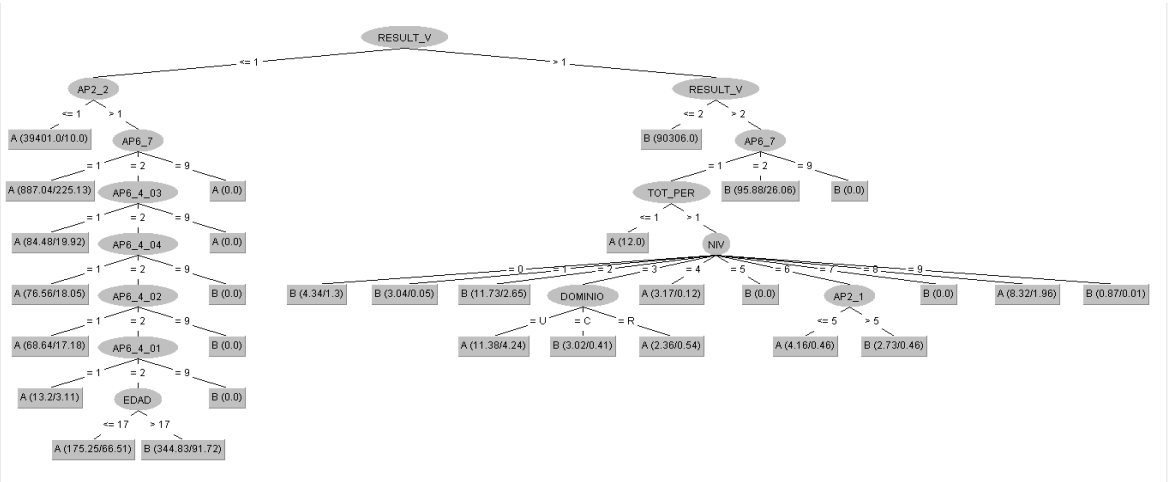

Ilustración 1: Árbol de clasificación generado con J48

4.3.2. Algoritmo REPTree

Se ejecuta el procedimiento con 131520 instancias y 265 atributos y una validación cruzada con 10-fold y se obtiene un tamaño de árbol de 163, un árbol grande.

Tabla 14: Stratified cross-validation REPTree

Correctly Classified Instances	130941	99.5598 %
Incorrectly Classified Instances	579	0.4402 %
Kappa statistic		0.9897
Mean absolute error		0.0059
Root mean squared error		0.0544
Relative absolute error		1.3799 %
Root relative squared error		11.7901 %
Total Number of Instances	131520	

Las instancias correctamente clasificadas son del 99% y la índice kappa es también elevado, en nuestro caso es de 0.99, lo que significa consenso en la clasificación. Las siguientes tablas indican la bondad de la clasificación, si observamos la tasa de clasificaciones correctas encontramos que para A y para B, es de aproximadamente el 99%

Tabla 15: Detailed Accuracy by Class REPTree

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC	Area Class
	0.995	0.003	0.998	0.995	0.997	0.990	1.000	1.000	B
Weighted Avg.	0.997	0.005	0.989	0.997	0.993	0.990	1.000	0.999	A
	0.996	0.004	0.996	0.996	0.996	0.990	1.000	0.999	

Tabla 16: Confusion Matrix REPTree

a	b	classified as
90577	440	a = B
139	40364	b = A

5. ALGORITMOS BASADOS EN REGLAS

5.3.1. Algoritmo OneR-B6

Este algoritmo es una manera muy simple de generar reglas de clasificación desde un conjunto de datos, y por lo general es bastante bueno para caracterizar estructuras de datos y frecuentemente alcanza una exactitud muy alta (Whitten & Eibe, 2017).

Se ejecuta el procedimiento en Weka sobre 131520 instancias y 265 atributos, con una validación cruzada de 10-fold, consiguiendo convergencia en

Tabla 17: Stratified cross-validation para OneR-B6

Correctly Classified Instances	130832	99.4769 %
Incorrectly Classified Instances	688	0.5231 %
Kappa statistic		0.9878
Mean absolute error		0.0052
Root mean squared error		0.0723
Relative absolute error		1.2273 %
Root relative squared error		15.667 %
Total Number of Instances	131520	

Las instancias correctamente clasificadas representan el 99% del total, además que la índice kappa de acuerdo en la clasificación es bastante buena (0.99, aproximadamente)

Tabla 18: Detailed Accuracy by Class OneR-B6

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC	Area Class
	0.993	0.001	0.999	0.993	0.996	0.988	0.996	0.997	B
Weighted Avg.	0.999	0.007	0.985	0.999	0.992	0.988	0.996	0.984	A
	0.995	0.003	0.995	0.995	0.995	0.988	0.996	0.993	

Tabla 19: Confusion Matrix OneR-B6

a	b	classified as
90387	630	a = B
58	40445	b = A

5.3.2. Algoritmo JRip -F 3 -N 2.0 -O 2 -S 1

Se ejecuta el algoritmo con 131520 instancias y 265 atributos, consiguiendo convergencia con 26 reglas y validación cruzada de 10-fold.

Tabla 20: Stratified cross-validation JRip -F 3 -N 2.0 -O 2 -S 1

Correctly Classified Instances	130991	99.5978 %
Incorrectly Classified Instances	529	0.4022 %
Kappa statistic		0.9906
Mean absolute error		0.0059
Root mean squared error		0.0565
Relative absolute error		1.3835 %
Root relative squared error		12.2343 %
Total Number of Instances	131520	

Tabla 21: Detailed Accuracy by Class JRip -F 3 -N 2.0 -O 2 -S 1

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC	Area Class
	0.997	0.006	0.998	0.997	0.997	0.991	0.997	0.998	B
Weighted Avg.	0.994	0.003	0.993	0.994	0.993	0.991	0.997	0.997	A
	0.996	0.005	0.996	0.996	0.996	0.991	0.997	0.998	

Tabla 22: Confusion Matrix JRip -F 3 -N 2.0 -O 2 -S 1

a	b	classified as
90715	302	a = B
227	40276	b = A

Con la finalidad de seleccionar el mejor modelo, se comparan las raíces cuadradas de los errores cuadráticos medios de cada uno. Estos se pueden observar en la siguiente tabla, en la cual se incluye la ganancia en precisión del modelo de menor error cuadrático medio. Como se observa en la tabla 23, el mejor modelo es el que corresponde al algoritmo J48, seguido de los algoritmos REPTree, JRip -F 3 -N 2.0 -O 2 -S 1 y OneR-B6, para los cuales J48 tiene una ganancia en eficiencia de 7.7%, 11.9% y 43%, respectivamente.
Aunque JRip -F 3 -N 2.0 -O 2 -S 1 y OneR-B6 son algoritmos de clasificación, lo hacen basados en reglas, por lo que, la comparación de la eficiencia debería realizarse entre ellos. Comparando estos algoritmos, la ganancia en precisión de JRip -F 3 -N 2.0 -O 2 -S 1 sobre OneR-B6 es de 28%, aproximadamente.

Tabla 23: Ganancia en precisión de los modelos.

Algoritmo	ECM	Ganancia
J48	0.0505	
REPTree	0.0544	7.72277228
JRip -F 3 -N 2.0 -O 2 -S 1	0.0565	11.8811881
OneR-B6	0.0723	43.1683168

6. CONCLUSIONES

Los resultados obtenidos permiten concluir que las variables asociadas a la victimización y violencia son los relacionados con victimización en el hogar, en la calle, en el trabajo y en la escuela, además de las principales preocupaciones de la gente. Una de las variables recurrente en los grupos es precisamente la inseguridad, falta de castigo a delincuentes, corrupción, salud, educación, desempleo y desastres naturales.
Con los árboles de clasificación, encontramos que las variables asociadas a la victimización y violencia son aquellas que tienen que ver con el núcleo familiar y las tareas que tienen que desempeñar en la calle, en la escuela, el trabajo, donde temen ser víctimas del delito. Otras variables relacionadas al fenómeno son la edad, el total de personas en el hogar, el nivel de escolaridad y si la zona donde vive clasifica como urbana, complemento urbana o rural.
Respecto de los modelos propuesto para estimar la probabilidad de victimización, dadas el conjunto de variables mencionadas anteriormente, los cuatro permiten clasificar correctamente aproximadamente el 99% de las instancias, por lo que, los cuatro son buenos, pero cuando comparamos sus errores cuadráticos medios el mejor algoritmo es el J48, toda vez que al menos tiene una ganancia en precisión del 8%, respecto de su más cercano competidor.
Por lo anterior, el mejor algoritmo en la clasificación es el J48, el cual permite estimar la probabilidad de victimización para la población de aproximadamente 0.31, lo cual indica que 31 personas de cada cien, han sido víctimas de la violencia en México.

RECEIVED: JULY, 2018.
REVISED; NOVEMBER, 2018.

REFERENCES

[1] AGARWAL J., N. R. (2013): Crime Analysis using K-Means Clustering. International. **Journal of Computer Applications**, 10, 1-4.
[2] DEAN, J. (2014): **Big Data, Data Mining and Machine Learning**. John Wiley, Chichester.
[3] INEGI. (2016): **Encuesta Nacional de Victimización y Percepción sobre Seguridad Pública.** Obtenido de http://www.beta.inegi.org.mx/proyectos/enchogares/regulares/envipe/2016/
[4] INEGI. (2017): **Instituto Nacional de Estadística y Geografía.** Obtenido de http://www.beta.inegi.org.mx
[5] LEWICKI, P., and HILL, T. (2005): **Statistics: Methods and Aplications**. StatSoft, USA.
[6] MAIMON, O. and ROKACH, L. (2010): **Data Mining and Knowledge Discovery Handbook**. Springer, N. York.
[7] OLSON, G. (2012): **Excelsior Hagamos equipo contra crimen: EPN; define estrategia de seguridad en seis ejes**. Obtenido de https://www.excelsior.com.mx/2012/12/18/nacional/875492
[8] ONU. (2009): **Manual de encuestas de victimización**. Ginebra, Suiza: ONUDD-CEE.
[9] SESNS. (2017): **Secretriado Ejecutivo del Sistema Nacional de Seguridad.** Obtenido de http://secretariadoejecutivo.gob.mx/docs/pdfs/tasas%20por%20cada%20100%20mil%20habitantes/Tasas072017.pdf
[10] SWADI, K. A. (2011): A Proposed Frameword For Analyzing Crime Data Set Using Decision Tree And Simple K-Means Mining Algorithms. **Journal Kufa for Mathematics and Compute***r*, 15, 8-24.
[11] WHITTEN, I. H. and EIBE, F. (2017): **Data Mining: Practical Machine Learning Tools and Techniques**. Morgan Kaufmann Publisher, Calcutta.