ISSN 1870-4069

Clasificación del género de peatones y conductores basada en su comportamiento en la vía pública

Francisco Fernando Torres Rosas, Maricela Quintana López,
Héctor Rafael Orozco Aguirre, Saul Lazcano Salas

Universidad Autónoma del Estado de México,
Centro Universitario UAEM Valle de México,
México

fer04.torres92@gmail.com,
{mquintanal,hrorozcoa,slazcanos}@uaemex.mx

Resumen. Los accidentes vehiculares representan un problema de salud pública a nivel mundial, los traumatismos por accidentes de tránsito son la primera causa de muerte en niños y adultos jóvenes. En México, a través de la estadística de accidentes elaborada por el Instituto Nacional de Estadística y Geografía se logra tener un panorama de los accidentes viales en el ámbito nacional, así como las consecuencias humanas y materiales que conlleva. Sin embargo, no se tiene a detalle cuáles fueron los factores o comportamientos de los peatones o conductores, que contribuyeron a que el accidente ocurriera. En este artículo, se presenta la clasificación del género basado en el comportamiento que los actores viales, peatón y conductor, exhiben dentro de la vía pública. Para ello, se usa el algoritmo J48 que proporciona árboles de decisión que permiten examinar el conocimiento adquirido. El mejor clasificador de género, para los peatones obtuvo un 71.61% de acierto usando validación cruzada de 5 pliegues. De los 13 atributos, solo se utilizan 6 para la clasificación en el árbol de decisión, dando lugar a 7 reglas, 3 para la clase masculina y 4 para la femenina. En el caso de los conductores, el mejor clasificador obtuvo un 73.39% de acierto, pero lo logra a costa de la clase femenino. En ambos actores viales, los atributos que proporcionan más información para la separación de las clases son cruzar o conducir bajo los efectos del alcohol y conducir cansado.

Palabras clave: Clasificación, árboles de decisión, minería de datos.

Gender Classification of Pedestrians and Drivers based on their Behavior on Public Road

Abstract. Car accidents represent a public health problem worldwide, road traffic injuries are the leading cause of death in children and young adults. In Mexico, through the statistics of accidents prepared by the National Institute of Statistics and Geography, it is possible to have an overview of road accidents at the national level, as well as the human and material consequences that they entail. However, the factors or pedestrians and drivers' behavior that contributed to the accident occurred are not detailed. In this article, the gender classification

based on the road actors' behavior, pedestrians and drivers, exhibit on public roads is presented. For this, the J48 algorithm is used, which provides decision trees that allow examining the acquired knowledge. The best gender classifier, for pedestrians obtained a 71.61% success rate using 5-fold cross validation. Of the 13 attributes, only 6 are used for classification in the decision tree, giving rise to 7 rules, 3 for the male class and 4 for the female one. In the case of drivers, the best classifier obtained a 73.39% success rate, but it does so at the expense of the female class. In both road actors, the attributes that provides more information for separating classes are crossing or driving under the influence of alcohol and driving tired.

Keywords: Classification, decision trees, data mining.

1. Introducción

Los accidentes viales representan un problema de salud pública a nivel mundial, por ello la Organización de las Naciones Unidas (ONU) aprobó el "Decenio de Acción para la Seguridad Vial 2011-2020" y lo asignó a la Organización Mundial de la Salud (OMS) para su implementación [1]. Dentro de sus objetivos se encuentra el de compilar el estimado de la cantidad de víctimas mortales de tránsito en todo el mundo, para posteriormente, implementar estrategias y acciones sobre seguridad vial, en el ámbito regional, nacional y mundial, que disminuyan dicha estimación en un 50%; es decir, que se busca reducir la cantidad de víctimas mortales y traumatismos para el año 2020, la cual se aproxima anualmente a 1.3 millones, y de 20 a 50 millones respectivamente.

De acuerdo con el Informe de Estado Global sobre Seguridad Vial [2], los traumatismos por accidentes de tránsito son la octava causa de muerte para todos los grupos de edades y la primera en niños y adultos jóvenes de 5 a 29 años. Estos decesos tienen repercusiones sociales y económicas, ya que el costo aproximado que tienen los accidentes viales para un país está entre el 1% y el 3% de su producto nacional bruto [1].

México tomó acción sumándose al grupo de países que desean lograr un cambio significativo, por ello los gobiernos federales, estatales y regionales están implementando cada vez más medidas para prevenir estos percances, mejorando la infraestructura vial, limitando la velocidad en lugares como escuelas, integrando la información existente sobre accidentes viales y regularizando los estándares de seguridad mínimos con los que un vehículo debe contar [3].

Por otro lado, en México, a través de la estadística de Accidentes de Tránsito Terrestre en Zonas Urbanas y Suburbanas (ATUS) elaborada por el Instituto Nacional de Estadística y Geografía (INEGI) [4], se logra tener un panorama acerca de los accidentes viales en el ámbito nacional, así como las consecuencias humanas y materiales que conlleva [5]. Sin embargo, no se tiene detalle de cuáles fueron los factores o comportamientos de los peatones o conductores, que contribuyeron a que el accidente ocurriera.

Se llama actor vial tanto al peatón que es la persona que va a pie por la vía pública, como al conductor, que en este caso es aquel que conduce un vehículo particular automotor dentro de la vía pública. El tener el comportamiento de un actor vial, permite modelarlo usando un sistema multiagente [6] para poder realizar posteriormente una

simulación en la que, cuando ocurra un accidente, se puedan examinar los comportamientos de los involucrados. De esta manera, será posible tener información puntual que permita crear estrategias o campañas enfocadas a reducir los comportamientos viales que producen un accidente y, en consecuencia, si las personas acatan las recomendaciones, reducir el número de accidentes.

En este artículo, se presenta la forma en que se recolectó la información para permitir obtener los comportamientos de peatones y conductores y la generación de un clasificador para distinguir el comportamiento por género. El resto del artículo se organiza como sigue: en la sección 2 se muestran los trabajos relacionados, mientras que la metodología utilizada en esta investigación se presenta en la sección 3, y en la 4 el desarrollo del clasificador. En la sección 5, se presentan los experimentos realizados y los resultados obtenidos. Finalmente, en la sección 6 se presentan las conclusiones y trabajo futuro.

2. Trabajos relacionados

Comprender las circunstancias en las que los actores viales, conductores y peatones, tienen más probabilidades de morir o sufrir lesiones graves en un accidente automovilístico, o bien, estar inmersos en una situación de riesgo, es de particular preocupación para que los gobiernos a distinto nivel implementen programas encaminados a la mejora de la seguridad vial. Existen trabajos que han utilizado algoritmos de clasificación, como es el caso de Romero [7], donde usan C4.5 para organizar y clasificar la información de accidentes de tránsito, con base en las causas que los generaban como el clima, la infraestructura e infracciones de tránsito entre otras.

En [8] hacen uso de árboles de decisión y regresión CART, para determinar qué zonas son de mayor accidentalidad y qué variables se presentan en esos accidentes. Zhang [9] utilizó los algoritmos ID3 y C4.5 para el análisis de colisiones de tráfico en la zona de Saskatchewan, Canadá, donde identifican los factores que contribuyen a las colisiones y su gravedad. Encontraron que la bebida y no respetar las reglas de tránsito son los factores que más contribuyen para los accidentes. Por último, en cuanto al género, para hombres el factor es la bebida, los errores al conducir y las reglas de tránsito y para mujeres fueron reglas de tránsito, ambiente y bebida. En Jain [10], usan los algoritmos J48, Naive Bayes y Bayes Net para analizar accidentes de tránsito y determinar qué áreas son más propensas a accidentes de tránsito en la India. Uno de los factores que más influyen es la velocidad con la se transita.

En [11] se analizan las causas de los accidentes mediante los algoritmos Naive Bayes, Random Forest, MLP y AdaBoost, añadiendo algunas características que pueden ser medidas en tiempo real, como el clima y condiciones de la vialidad, de este modo, generan una predicción sobre la posible fatalidad de los accidentes. Dogru & Subasi [12], realizan una propuesta basada en el intercambio de información entre vehículos (velocidades, ubicación, trayectorias, tipo de vialidad, etc.) en un entorno de IoT y usando Random Forest, realizan una predicción de accidentes en vías de alta velocidad.

En [13], se utilizó una red bayesiana, árboles de decisión y redes neuronales artificiales, para detectar factores con la mayor influencia en accidentes

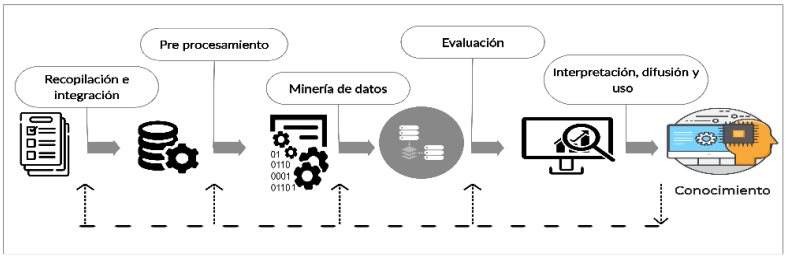

Fig. 1. Etapas del proceso de extracción del conocimiento

automovilísticos. Para ello, se llevó a cabo un experimento con datos de accidentes de tráfico en Reino Unido. De este experimento, se dedujo que los tres factores más frecuentes en un accidente vial son las condiciones de luz, la maniobra del vehículo y el tipo de carretera. Por su parte, en [14] para comprender los factores que conducen a la gravedad de los accidentes automovilísticos en la ciudad de Riad, Arabia Saudí, se utilizan tres técnicas de clasificación: CHAID, J48 y Naive Bayes. De este trabajo, se resaltan el peligro de distracción mientras se conduce y que los accidentes con autos más viejos tienen más probabilidades de provocar lesiones o muertes.

Si bien se sabe que algunos de los factores de riesgo, como el consumo de alcohol y drogas, afectan la gravedad, un modelo exacto de sus influencias sigue siendo un tema de investigación abierto. Sin embargo, no hay una precisión lo suficientemente confiable como para conocer todas las causas involucradas en un accidente vial, lo que sugiere que estas son complejas y se requiere más investigación. Por ende, hay varios factores de riesgo presentes esperando ser descubiertos o analizados.

3. Metodología

Para poder llevar a cabo esta investigación, se utilizó la metodología para el descubrimiento del conocimiento y minería de datos [15] propuesta en 1996 por Fayyad y que otros autores como Hernández han complementado o ampliado para adaptarla a sus necesidades particulares, aunque la esencia es la misma [16]. Una diferencia principal, es que no en todos los problemas se parte de una base de datos, por lo que se agrega la etapa de adquisición o recolección de datos, como es el caso de esta investigación. Considerando lo anterior, en la Fig. 1, se muestran las etapas de la metodología para la extracción de conocimiento, mismas que se explican a continuación.

3.1. Etapas del proceso de extracción del conocimiento

El proceso de extracción del conocimiento es tanto iterativo como interactivo; iterativo porque en cualquier etapa es posible regresar a una etapa anterior y repetir el proceso, mientras que la interacción se refiere a que un experto o usuario de interés ayude a la validación del conocimiento extraído [16]. Las etapas son las siguientes:

- **Recopilación e integración:** los datos necesarios para un análisis pueden estar almacenados en diferentes bases de datos por lo que la diversidad o dispersión

en una organización, puede ser un gran problema, además existen casos en los que los datos no han sido recolectados o son insuficientes para analizarlos y es necesario contar con un mecanismo para adquirirlos.

- **Preprocesamiento:** para poder utilizar los datos es necesario realizar una selección basada en la calidad de los datos que se tienen con el fin de utilizar los más relevantes para el problema. De igual forma, una vez que se han seleccionado los datos, es necesario eliminar el ruido que estos puedan tener o considerar situaciones como el manejo de valores faltantes o perdidos ya que puede conducir a resultados imprecisos. Otra tarea que también se debe realizar es la transformación de atributos, que consiste en realizar operaciones o funciones a los atributos originales.
- **Minería de datos:** el objetivo de esta fase es producir conocimiento, esto se hace construyendo un modelo para observar patrones y relaciones entre los datos que pueden usarse para hacer predicciones que ayuden a entender situaciones. Aquí se determina qué tipo de tarea se va a realizar, la representación final del conocimiento adquirido y en consecuencia el algoritmo a utilizar.
- **Evaluación:** el resultado de la etapa de minería de datos idealmente debe cumplir con al menos tres características: ser comprensibles, interesantes y precisos [16]. Las primeras dos características son evaluadas por el experto y son parte de la interpretación. Para determinar la precisión existen varias formas de evaluación, entre estas la validación simple y la validación cruzada.
- **Interpretación y uso:** etapa donde un analista o experto hace una recomendación con base en el modelo y sus resultados, o bien, aplica tal modelo a otro conjunto de datos.

4. Desarrollo del clasificador

A continuación, se presenta el desarrollo del clasificador de género basado en el comportamiento vial. Se muestran cada una de las etapas de la metodología aplicada al problema de clasificar a hombres y mujeres basándose en su comportamiento vial.

4.1. Recopilación e integración

Se examinó la información de la página del INEGI [4] respecto a los accidentes de tránsito en México, la cual consta de 46 atributos referentes al accidente, pero poco se habla del comportamiento de los involucrados. Sólo se maneja si el conductor se encontraba en estado de ebriedad y si usaba el cinturón de seguridad [17]. Por esta razón, para la adquisición de los datos, en esta investigación, se entrevistó a agentes de tránsito, y con base en sus respuestas y consideraciones, se diseñó una encuesta para los actores viales, que se aplicó a la población estudiantil del Centro Universitario UAEM Valle de México de la Universidad Autónoma del Estado de México (UAEM).

Las entrevistas se realizaron a 41 agentes de tránsito tanto estatal como municipal con el fin de que dijeran qué aspectos de comportamiento, tanto del peatón como del conductor, están presentes cuando un accidente ocurre. En la Tabla 1, se presentan algunos de los factores y el porcentaje en que estos agentes consideran que influyen en

Tabla 1. Factores que influyen en un accidente vial según los agentes de tránsito.

Factor	Porcentaje
Exceso de velocidad	46.34 %
Alcohol, medicamento y/o estupefacientes por parte del conductor	63.41 %
Alcohol, medicamento y/o estupefacientes por parte del peatón	29.26 %
Conducción distraída	65.85 %
Conducción con fatiga o sueño	65.85 %
Tratar de rebasar por la derecha por parte de los conductores	48.78 %
Cruzar por el medio de la calle o avenida por parte del peatón	58.53 %
Intención de ganarle al semáforo o vehículo por parte del peatón	63.41 %
Cruce distraído por parte del peatón	53.65 %

un accidente vial. Aproximadamente el 63% de los agentes, consideran que la intoxicación por parte del conductor, así como la conducción distraída o con sueño o fatiga influyen de manera importante en que ocurra un accidente.

Por otro lado, el que un peatón intente ganarle al semáforo, cruzar la calle o avenida por en medio de esta o cruce distraído es considerado como un factor de riesgo aproximadamente por el 60% de los agentes. En cuanto al alcohol, los agentes están de acuerdo que es un factor más relevante para conductores que para peatones en un accidente vial.

Con base en la información proporcionada por los agentes de tránsito, se diseñó una encuesta, que de acuerdo con Akerkar [18] es una técnica de adquisición del conocimiento relativamente fácil de realizar y recomendada para automatizar el procesamiento eficiente de los conocimientos recopilados y analizar los resultados.

La encuesta consta de 32 preguntas que se responden en una escala de Likert, que va del 1 al 5, donde 1 es nunca, y 5 es siempre, y que es apropiada para medir actitudes o comportamientos [19].

El público para recolectar los datos fueron estudiantes de entre 18 y 30 años. Considerando la matrícula en el ciclo escolar 2019B, que fue de 3811 alumnos, y considerando un intervalo de confianza del 95%, la cantidad de encuestados debió ser al menos de 350, esto de acuerdo con la ecuación de muestreo aleatorio simple (ver ecuación 1):

$$Tamaño\ de\ la\ muestra = \frac{\frac{Z^2 * p(1-p)}{e^2}}{1 + \frac{z^2 * p(1-p)}{e^2 N}}, \tag{1}$$

donde N= tamaño de la muestra, e = margen de error (porcentaje expresado con decimales) y z= puntuación z, número de desviaciones estándar por encima o por debajo de la media de población.

La encuesta se aplicó a un total de 587 estudiantes, 384 peatones (238 hombres y 146 mujeres) y 203 conductores (149 hombres y 54 mujeres). En la Tabla 2, se presentan las preguntas, así como la media y desviación estándar de las respuestas dadas.

Tabla 2. Resultados de la encuesta de comportamiento vial de peatones y conductores.

Comportamiento de peatones		**Comportamiento de conductores**	
Mirar a ambos lados $\bar{x} = 4.8, \sigma = 0.54$	Uso de teléfono $\bar{x} = 1.6, \sigma = 0.87$	Respeto por la luz roja $\bar{x} = 4.7, \sigma = 0.62$	Conducción con sueño o fatiga $\bar{x} = 4.6, \sigma = 1.14$
Calle de un solo sentido mira a ambos lados $\bar{x} = 4.3, \sigma = 0.99$	Uso de audífonos $\bar{x} = 2.4, \sigma = 1.26$	Respeto por el límite de velocidad $\bar{x} = 4.2, \sigma = 0.92$	Conducción bajo los efectos del alcohol $\bar{x} = 2.3, \sigma = 0.99$
Respeto semáforo peatonal $\bar{x} = 4.5, \sigma = 0.75$	Cruzar bajo efectos del alcohol $\bar{x} = 1.9, \sigma = 1.06$	Respeto por el paso peatonal $\bar{x} = 4.7, \sigma = 0.69$	Conducción bajo efectos de una droga $\bar{x} = 1.7, \sigma = 0.88$
Semáforo vehicular en verde $\bar{x} = 2.3, \sigma = 1.12$	Cruzar corriendo $\bar{x} = 3.1, \sigma = 0.91$	Uso del teléfono $\bar{x} = 1.8, \sigma = 0.95$	Conducción distraída $\bar{x} = 1.4, \sigma = 1.05$
Uso paso cebra $\bar{x} = 3.9, \sigma = 1.03$	Caminar por la banqueta $\bar{x} = 4.2, \sigma = 0.83$	Uso de los manos libres $\bar{x} = 2.6, \sigma = 1.47$	Uso del cinturón de seguridad $\bar{x} = 2.0, \sigma = 0.78$
Cruzar por esquinas $\bar{x} = 3.8, \sigma = 0.88$	Caminar distraídos $\bar{x} = 2.7, \sigma = 1.10$	Respeto por la señalización $\bar{x} = 4.5, \sigma = 0.77$	-
Uso puente peatonal $\bar{x} = 4.4, \sigma = 0.84$	-	-	-

4.2. Preprocesamiento de datos

La encuesta aplicada constó de un total de 32 campos o atributos, ya que además de las preguntas mostradas en la Tabla 2, también se le preguntó al actor vial, si se ha visto involucrado en un accidente vial y se le pidió que describiera algunos aspectos de este. Para esta etapa se seleccionaron únicamente 24 preguntas, 13 de peatones y 11 de conductores, y se separaron en dos archivos. Fueron pocas las encuestas que presentaron valores faltantes, sólo 3 y lo que se realizó para corregir, fue utilizar el valor entero más cercano a la media aritmética de la pregunta en cuestión.

4.3. Minería de datos

La minería de datos consiste en la extracción de información sensible, previamente desconocida, que reside de manera implícita en los datos. La minería de datos prepara y explora los datos para obtener información oculta en ellos [8]. Para esta etapa se eligió el algoritmo J48 que es la implementación del algoritmo C4.5, creado por Ross Quinlan como una extensión de su antecesor el ID3, y uno de los más utilizados en la minería de datos [20]. Este algoritmo utiliza el enfoque de divide y vencerás para generar el clasificador, calcula la ganancia de cada atributo y elige el de mayor ganancia para que sea la raíz del árbol, los datos son divididos con base en el dominio del atributo, de manera iterativa, se repite el proceso con cada partición [16].

El factor de confianza, en la construcción del árbol de decisión, es el parámetro que influye en el tamaño y capacidad de predicción del árbol construido, ya que está asociado a las operaciones de poda del árbol. Para cada operación, se compara la cantidad de error que sufriría el árbol de decisión antes y después. A menor probabilidad, se exigirá que la diferencia en los errores de predicción sea más

significativa para no podar. El valor por defecto es 25%. Según disminuya este valor, se aprueba una poda mayor, el tamaño del árbol disminuye y en consecuencia el error puede crecer [8], [21]. Los experimentos realizados y resultados en esta etapa se presentan en la sección 5 debido a su extensión e importancia.

4.4. Evaluación

El método que se usó para la evaluación fue la validación cruzada con n pliegues, para este caso n se manejó con los valores 10 y 5 pliegues para todos los experimentos realizados. La validación cruzada consiste en dividir aleatoriamente los datos en n grupos. Un grupo es para el conjunto de prueba y con el otro conjunto n-1 restante se usa para construir un modelo y se utiliza para pronosticar el resultado de los datos del grupo reservado. Por ejemplo, usar 5 pliegues equivale a utilizar el 80% de los datos para construir el clasificador y 20% para probarlo. El proceso se repite n veces, utilizando un conjunto de datos diferente para cada prueba. El error se calcula como la media aritmética de los errores en cada repetición [16, 22].

4.5. Interpretación y uso

Para esta etapa se contempla utilizar el modelo que entrega el algoritmo J48 para elegir el comportamiento que presentarán los agentes dentro de una simulación con el fin de poder registrar las conductas que estos exhibirían al ocurrir un accidente vial, y posteriormente, analizar los datos para determinar las conductas que más influyen en los diferentes niveles de riesgo de sufrir un accidente vial.

5. Experimentos y resultados

Para esta etapa se realizaron experimentos con el algoritmo J48 en el software WEKA en su versión 3.8.3 [23], con dos objetivos principalmente: el primero determinar si se podía generar un clasificador con un resultado aceptable que determine el género del actor vial basado en su comportamiento y el segundo determinar de qué forma son diferentes, en su comportamiento vial, los masculinos de los femeninos, y esto se logra a través de examinar los árboles construidos. Se presentan los resultados y el análisis de los experimentos realizados con los datos de peatones y conductores.

5.1. Peatones

En el caso de los peatones, se tiene un total de 384 encuestas, 238 hombres y 146 mujeres. Debido a que las clases no están balanceadas, un resultado adecuado con el cual poder discernir si el clasificador es aceptable, es que su porcentaje de acierto sea de un 62% o más, ya que ese es el resultado obtenido por omisión cuando se considera la clase mayoritaria y todas las instancias de género femenino tendrían una clasificación incorrecta dando un error del 38%. Se realizaron varias ejecuciones del algoritmo J48, en el que se modifican tanto el factor de confidencia como el número de pliegues utilizados para la validación cruzada. Los resultados se presentan en la Tabla 3. Se

Tabla 3. Experimentos con el algoritmo J48 para peatones.

N.º	Pliegues	Confidencia	Acierto	Error
1	10	0.25	66.92% (257)	33.07% (127)
2	10	0.20	68.22% (262)	31.77% (122)
3	10	0.15	68.75% (264)	31.25% (120)
4	10	0.10	68.75% (264)	31.25% (120)
5	10	0.05	69.79% (268)	30.20% (116)
6	5	0.25	68.48% (263)	31.51% (121)
7	5	0.20	70.83% (272)	29.16% (112)
8	5	0.15	71.35% (274)	28.64% (110)
9	5	0.10	71.61% (275)	28.38% (109)
10	5	0.05	70.31% (270)	29.68% (114)

Tabla 4. Experimento 9 con un 71.61% de acierto.

Masculino	Femenino	% de acierto por clase
202	36	84.87%
73	73	50%

puede observar que los mejores resultados se obtienen al usar la validación cruzada de 5 pliegues, y que el mejor, basado en porcentaje de acierto, es el experimento 9 seguido del experimento 8. Para poder determinar si el experimento sirve para el segundo objetivo es necesario ver más a detalle los resultados de dichos experimentos. Para ello, se muestra en la tabla 4 la matriz de confusión, y en la Fig. 2 el árbol de decisión, el cual tiene 13 nodos, de los cuales 7 son hojas.

La matriz de confusión, mostrada en la Tabla 4, indica que, de las 238 instancias masculinas, 202 se clasifican correctamente y 36 de forma incorrecta dando un porcentaje de acierto para la clase masculina del 84.87%. Por otro lado, de las 146 mujeres, solo se logra clasificar de manera correcta a 73 dando un acierto por clase del 50%.

Analizando el árbol de la Fig. 2, se muestra que el nodo raíz "caminar bajo los efectos del alcohol", divide a las 384 instancias en 2 conjuntos, 199 que nunca cruzan bajo los efectos del alcohol (rama izquierda) y 185 que cruzan bajo los efectos del alcohol, de estos últimos, 155 cruzan casi nunca o a veces bajo los efectos del alcohol, y los 30 restantes casi siempre o siempre, siendo 27 de ellos masculinos.

En las hojas del árbol, se observa la clase y la cantidad de instancias que llegaron a la misma, seguido del número de instancias mal clasificadas. Así, en la rama derecha la conclusión es que el género es masculino, cayendo en esta rama 185 instancias, de entre las cuales hay 41 sujetos femeninos mal clasificados.

De los 13 atributos que se utilizan en la clasificación, sólo 6 aparecen en el árbol final, el cuál al ser recorrido proporciona las siguientes reglas:

Reglas de comportamiento para peatones masculinos

1. Si a veces cruza bajo los efectos del alcohol entonces es masculino

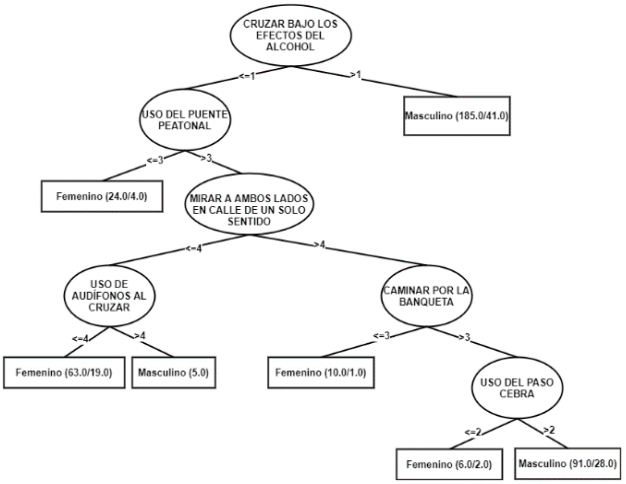

Fig. 2. Árbol de decisión del comportamiento vial de peatones.

2. Si nunca cruza bajo los efectos del alcohol, casi siempre o siempre usa un puente peatonal, a veces o casi siempre al cruzar mira a ambos lados de la calle, aun cuando es de un solo sentido, y siempre usa audífonos entonces es masculino.
3. Si nunca cruza bajo los efectos del alcohol, casi siempre o siempre usa un puente peatonal, siempre al cruzar mira a ambos lados de la calle, aun cuando es de un solo sentido, casi siempre o siempre usa la banqueta y a veces, casi siempre o siempre usa el paso cebra entonces es masculino.

Reglas de comportamiento para peatones femeninos

1. Si nunca cruza bajo los efectos del alcohol y a veces, o casi nunca o nunca usa un puente peatonal entonces es femenino.
2. Si nunca cruza bajo los efectos del alcohol, casi siempre o siempre usa un puente peatonal, a veces o casi siempre al cruzar mira a ambos lados de la calle, aun cuando es de un solo sentido, y a veces usa audífonos entonces es femenino.
3. Si nunca cruza bajo los efectos del alcohol, casi siempre o siempre usa un puente peatonal, siempre al cruzar mira a ambos lados de la calle, aun cuando es de un solo sentido, y a veces usa la banqueta entonces es femenino.
4. Si nunca cruza bajo los efectos del alcohol, casi siempre o siempre usa un puente peatonal, siempre al cruzar mira a ambos lados de la calle, aun cuando es de un solo sentido, siempre o casi siempre usa la banqueta y nunca o casi nunca usa el paso cebra entonces es femenino.

5.2. Conductores

En el caso de los conductores, se encuestó a un total de 203 personas, 149 hombres y 54 mujeres. Las clases no están balanceadas, por lo que, en este caso un resultado adecuado con el cual poder discernir si el clasificador es aceptable, es que su porcentaje de acierto sea igual o esté por arriba del 73.39%, ya que ese es el resultado obtenido

Tabla 5. Experimentos con el algoritmo J48 para conductores.

N.°	Pliegues	Confidencia	Acierto	Error
1	10	0.25	70.44% (143)	29.55% (60)
2	10	0.20	71.42% (145)	28.57% (58)
3	10	0.15	72.90% (148)	27.09% (55)
4	10	0.10	73.39% (149)	26.60% (54)
5	10	0.05	73.39% (149)	26.60% (54)
6	5	0.25	70.44% (143)	29.55% (60)
7	5	0.20	72.41% (147)	27.58% (56)
8	5	0.15	73.39% (149)	26.60% (54)
9	5	0.10	73.39% (149)	26.60% (54)
10	5	0.05	73.39% (149)	26.60% (54)

Tabla 6. Matrices de confusión en los experimentos para conductores

N.°	Matrices de confusión		% acierto por clase	% acierto
	Masculino	Femenino		
1	131	18	87.9	70.44
	42	12	22.2	
2	142	7	95.3	71.42
	51	3	5.5	
3	148	1	99.3	72.9
	54	0	0	
6	136	13	91.2	70.44
	47	7	12.9	
7	147	2	98.6	72.41
	54	0	0	

por omisión cuando se considera la clase mayoritaria, es decir, en el que se clasifica para masculino todas las instancias y las instancias de género femenino tendrían una clasificación incorrecta dando un error del 26.61%. En la Tabla 5, se presentan los resultados de las ejecuciones del algoritmo J48. Se observa que los experimentos con mejor porcentaje de acierto son el 4, 5, 8, 9 y 10, ya que tuvieron el número mayor de instancias correctamente clasificadas, pero se observa que absorben todas las instancias de género femenino como error, por lo que, si bien cumplen con el primer objetivo, no cumplen con el segundo objetivo de diferenciar entre los dos géneros.

A fin, de poder rescatar las diferencias de comportamiento por género entre los conductores, se procede a examinar las matrices de confusión de los experimentos 1, 2, 3 6, y 7, mismas que se muestran en la Tabla 6. Las columnas 2 y 3 presentan la matriz de confusión de cada experimento, el primer renglón corresponde a la clase masculino mientras que a la clase femenino corresponde el segundo renglón. Por ejemplo, en el experimento 1, de los 149 conductores masculinos, sólo 131 se clasificaron correctamente y 18 se clasificaron erróneamente como femeninos. Por otro lado, de las 54 conductoras femeninas, solo 12 se clasifican correctamente y 42 de forma incorrecta.

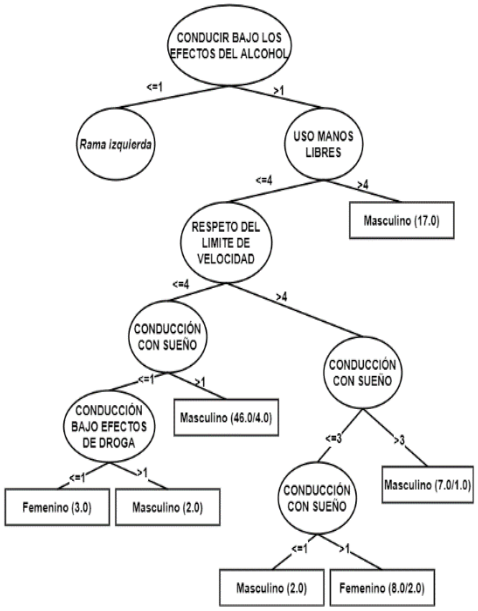

Fig. 3. Árbol de decisión del comportamiento vial de conductores.

El porcentaje de acierto de la clase masculino es de 88% mientras que de la clase femenina es de 29%. Se observa en los experimentos del 1 al 3, que fueron generados con validación cruzada de 10 pliegues, que conforme va aumentando el nivel de acierto del clasificador, se va disminuyendo el porcentaje de acierto de la clase femenino.

El árbol del experimento 1, provee más información para diferenciar las clases, tiene 20 hojas, 7 de la clase femenino y 13 de la clase masculino. En la Fig. 3, se observa que la raíz del árbol nuevamente está relacionada con el consumo de alcohol. Dividiéndose las 203 instancias en 118 que nunca conducen bajo los efectos del alcohol (ver Fig. 4) y 85, la rama derecha, que si lo hacen. También de las 85 instancias que si conduce bajo los efectos del alcohol sólo 9 son de género femenino, y de las 76 de género masculino, 42 manejan cansados. En la rama izquierda dada en la Fig. 4, se encuentran los 118 que nunca conducen bajo los efectos del alcohol. Se puede observar que hay mayor profundidad en el árbol por tratar de definir las hojas que pertenecen al género femenino. En los dos primeros niveles del subárbol hay dos hojas con 59 instancias, de las 118 que inmediatamente clasifican a masculino, teniendo 11 instancias mal clasificadas.

6. Conclusiones y trabajo futuro

El objetivo de este trabajo fue clasificar a los actores viales por su género, basándose en su comportamiento vial. Los actores que se incluyeron fueron el peatón y el conductor de un vehículo particular. Con los experimentos conducidos utilizando el algoritmo C4.5 (J48) en el software WEKA, se encuentra que el mejor clasificador de

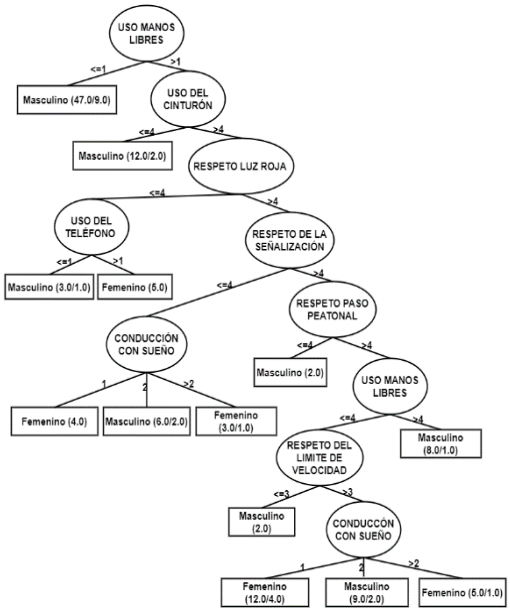

Fig. 4. Rama izquierda del árbol de decisión para conductores.

género, para los peatones obtuvo un 71.61% de acierto usando validación cruzada de 5 pliegues. De los 13 atributos, sólo se usan 6 para el árbol de decisión, dando lugar a, 3 reglas para la clase masculino y 4 para la femenina. En el caso de los conductores, el mejor clasificador que tiene el 73.39% de acierto, lo logra a costa de la clase femenino. Esto es porque la clase masculina es mayoritaria, sólo el 26.61% son mujeres. Al parecer, los datos que se tienen son insuficientes para diferenciar a los conductores en hombres y mujeres basados en su comportamiento.

Vale la pena mencionar que, en ambos actores viales, el atributo que proporciona más información para la separación de las clases es la de cruza o conduce bajo los efectos del alcohol, de los 185 peatones que lo hacen, 144 son hombres, mientras que, en los conductores de 85 instancias, 76 son hombres. Otro atributo que aparece varias veces en el árbol de los conductores es el de la conducción con sueño. Este atributo junto con el de conducción bajo los efectos del alcohol, fueron evaluados por los agentes de tránsito como los que más influyen en un accidente vial.

En este artículo, se considera que los resultados son alentadores en el caso de los peatones, y que para los conductores es necesario un mayor número de encuestados y repetir el experimento para discernir si la razón por la que no fue posible obtener un buen clasificador es porque ambos géneros, masculino y femenino, se comportan igual, o bien, esto es provocado por el desbalanceo de las clases obtenidas en los experimentos.

Como trabajo futuro se considera realizar un análisis comparativo empleando otros métodos basados en árboles, como Random Forest, así como algoritmos basados en reglas como PART, para mejorar la precisión de la clasificación.

Además, se pueden generar reglas de asociación para descubrir otras regularidades en los conjuntos por género.

Referencias

1. Sminkey, L.: Plan mundial para el decenio de acción para la seguridad Vial 2011-2020. World Health Organization (2011)
2. World Organization Health: Global status report on road safety 2018: Summary (2018)
3. OPS México: Organización panamericana de la salud (2018)
4. INEGI: Instituto Nacional de Estadística y Geografía (2016)
5. Secretaría de Salud: Informe sobre la situación de la seguridad vial (2017)
6. Osorio, L.A.: Modelo de simulación vial basado en agentes de software (2013)
7. Romero, A.G.: Minería de datos en el análisis de accidentes de tránsito en el Ecuador (2019)
8. Calderón, D.H, Sora, D.F.: Análisis de accidentalidad vehicular usando técnicas de minería de datos (2019)
9. Zhang, X.F., Fan, L.: A decision tree approach for traffic accident analysis of Saskatchewan highways. In: Canadian Conference on Electrical and Computer Engineering (2013)
10. Jain, A., Ahuja, G., Mehrotra, A.: Data mining approach to analyse the road accident in India. IEEE, pp. 175–179 (2016)
11. Xia, X.L., Nan, B., Xu, C.: Real-time traffic accident severity prediction using data mining technologies. In: Proceedings of the International Conference on Network and Information Systems for Computers, ICNISC, pp. 242–245, Institute of Electrical and Electronics Engineers Inc. (2017)
12. Dogru, N., Subasi, A.: Traffic accident detection using random forest classifier. In: 15th Learning and Technology Conference, pp. 40–45, Institute of Electrical and Electronics Engineers Inc. (2018)
13. Castro, Y., Kim, Y.J.: Data mining on road safety: factor assessment on vehicle accidents using classification models. International Journal of Crashworthiness, 21, pp. 104–111 (2016)
14. Al-Turaiki, I., Aloumi, M., Aloumi, N., Alghamdi, K.: Modeling traffic accidents in Saudi Arabia using classification techniques. In: 4th Saudi International Conference on Information Technology (Big Data Analysis), KACSTIT, Institute of Electrical and Electronics Engineers Inc. (2016)
15. Fayyad, U., Piatetsky-Shapiro, G., Smith, P.: From data mining to knowledge discovery and data mining. AI Magazine, 17(1), pp. 37–54 (1996)
16. Hernández, J., Ramírez, M.J., Ferri, C.: Introducción a la minería de datos (2004)
17. INEGI: https://www.inegi.org.mx/programas/accidentes/default.html#Documentacion (2018)
18. Akerkar, R., Srinivas, P.: Knowledge-based system (2010)
19. Vallejo, P.: Medición de actitudes en psicología y educación: construcción de escalas y problemas metodológicos (2006)
20. Quinlan, J., Kumar, V., Wu, X.: The top 10 algorithms in data mining. Knowledge and Information Systems 14(1), pp. 1–37 (2008)
21. Drazin, S..: Decision tree analysis using Weka. Machine Learning-Project II (2012)
22. Hoyos, G., Sleyther, Y.: Evaluación del riesgo crediticio en entidades bancarias en el área de microfinanzas utilizando árboles de decisión. pp. 1–67 (2017)
23. Weka: https://www.cs.waikato.ac.nz/ml/weka/ (2018)