Predicción del divorcio por medio de técnicas inteligentes

Omar Danilo Castrillón

Universidad Nacional de Colombia, Facultad de Ingeniería y Arquitectura, Departamento de Ingeniería Industrial, Campus La Nubia Bloque Q piso 2, Manizales - Colombia. (Correo-e: odcastrillong@unal.edu.co)

Recibido Feb. 25, 2021; Aceptado May. 3, 2021; Versión final May. 24, 2021, Publicado Oct. 2021

Resumen

El objetivo de este trabajo es estimar las relaciones entre las variables más importantes que pueden llevar a un divorcio. Se realiza utilizando la plataforma de aprendizaje automático y minería de datos Weka, la cual emplea el algoritmo de clasificación (J48), que es de comportamiento similar a un algoritmo bayesiano. Inicialmente se modela una base de datos existente, con 173 registros y 55 variables, de las cuales 54 son independientes y una variable dependiente (divorcio). Se seleccionan las variables independientes que producen una influencia moderada sobre la variable dependiente. Adicionalmente, se realiza una segunda reducción de variables mediante un proceso de selección estadística. Como resultado se identifican las cuatro variables más influyentes con un éxito del 100 % cuando se emplea un porcentaje de partición superior al 90 %. Se concluye que por medio del árbol de decisión se pueden observar de forma clara y concisa las variables más importantes que pueden causar un proceso de divorcio.

Palabras clave: minería de datos; predicción; bayesiano; divorcio; Weka

Predicting divorce through intelligence techniques

Abstract

The objective of this study is to estimate the relationships among variables that can lead to a divorce. The automatic learning and data-mining platform Weka is employed. Weka uses the classification algorithm J48, which behaves similarly to a Bayesian algorithm. First, an existing database is modeled with 173 records and 55 variables, of which 54 are independent and one is dependent (divorce). The independent variables that produce a moderate influence on the dependent variable are selected. Then, a second variable reduction is performed through a statistical selection process. The results show that the four most influential variables are identified with a 100% success rate when a partition percentage greater than 90% is used. It is concluded that through a decision-making tree, the most important variables that can lead to a divorce process can be determined clearly and accurately.

Keywords: data mining; prediction; Bayesian; divorce; Weka

INTRODUCCIÓN

El divorcio, un proceso con una larga evolución en Colombia (Ruiz, 2020), puede entenderse como la separación de dos personas ó la ruptura de su vínculo matrimonial. Pero, también es un indicador de transformación social y familiar, el cual tiene diferentes determinantes e impactos según el sexo de la persona (Tamez-Valdez y Ribeiro-Ferreira, 2016; Bucheli y Vigna, 2005). Por un lado, estos autores pueden ver el divorcio de una forma positiva. Igualmente, otros autores han encontrado una relación directa entre el divorcio y el aumento de la felicidad tanto en hombres como mujeres, especialmente en las etapas posteriores al mismo (Cavapozzi et al., 2020). No obstante, y en contraste con la posición anterior, un proceso de divorcio puede estar fundamentado en múltiples causas, las cuales pueden ser percibidas como un factor negativo que atenta contra la institucionalidad de la familia, el matrimonio y la misma sociedad, principalmente en aquellas sociedades con enfoques más tradicionalistas, en los cuales se concibe el matrimonio para toda la vida y ven un incremento notorio en sus tasas de divorcio (Ribeiro et al., 2012).

La Inteligencia Artificial ha incurrido en casi todos los campos de la vida diaria, y el campo del derecho no ha sido la excepción (Samaca, 2016; Martínez, 2012; Cáceres, 2006), hasta el punto que existen programas inteligentes (sistemas expertos de sentencias) los cuales pueden ayudar en los procesos de divorcio, sin que ello implique que el juez sea reemplazado. Así mismo, también es factible encontrar sistemas inteligentes que puedan ayudar a emitir diversas propuestas de divorcio, las cuales pueden o no ser acogidas por los cónyuges: En este mismo sentido, es factible encontrar sistemas para la predicción de divorcios basados en técnicas inteligentes y procesos de correlación estadística (Yöntem, et al., 2019). El anterior sistema por medio de los predictores de divorcio establecidos en la terapia de pareja de Gottman, permite determinar las características más influyentes en un proceso de divorcio, logrando establecer 6 características principales con un porcentaje de acierto del 98.82%, por medio de una red neuronal. De forma similar se logran muy buenos resultados al emplear otras técnicas inteligentes.

En este trabajo se estiman por medio de técnicas de inteligencia artificial, las relaciones más influyentes entre las variables, determinantes de un proceso de divorcio, constituyéndose este aspecto en el objetivo fundamental del mismo. Específicamente se emplea una técnica de minería de datos, basada en el algoritmo J48, de comportamiento similar a un algoritmo bayesiano. Esta investigación ayudará a mitigar el impacto de un divorcio en la sociedad, especialmente en aquellas familias o sociedades que ven el divorcio como algo negativo. Estas personas, podrán establecer con objetividad los aspectos en los cuales se deben centrar si desean prevenir un divorcio.

En este proceso se parte de una base de datos previamente diseñada en Internet, la cual puede ser encontrada en el siguiente vínculo (<https://archive.ics.uci.edu/ml/datasets/Divorce+Predictors+data+set#>). En esta base de datos, se analizan 54 factores que influyen en un proceso de divorcio, los cuales mediante una técnica de minería de datos Implementada por la plataforma WEKA (Witten, 2017) y un proceso de selección estadística determinan los factores más influyentes en este proceso. Como resultado se identifican con una efectividad del 100%, las 4 características, más determinantes, cuando se emplea un porcentaje de partición (entrenamiento) superior al 90%, en el algoritmo de clasificación.

Desde el punto de vista de la computación aplicada al campo del derecho, las técnicas de inteligencia artificial y las técnicas matemáticas - computacionales, también han sido empleadas. Es así como se pueden encontrar algoritmos genéticos que pueden ser empleados en la resolución de conflictos al crear soluciones factibles para las partes involucradas en el mismo (Carneiro et al., 2013). En este mismo orden de ideas, otros autores (Dijk et al., 2020) también han investigado, en este campo del derecho, logrando por medio de un análisis multivariado – multinivel, basado en ecuaciones estructurales, calcular la correlación que existe entre las variables, conflicto interparental, crianza y ajuste psicosocial de los niños. Como resultado de este proceso, estos autores, han encontrado una correlación significativa entre el conflicto interparental, la crianza y el ajuste psicosocial de los niños. Por su parte otros artículos investigativos (Balestrino et al., 2013) han analizado matemáticamente el impacto y la probabilidad que el divorcio ha tenido en el matrimonio, lográndose establecer porque el divorcio no ha limitado la disminución de matrimonios.

En complemento con el párrafo anterior, pero desde un punto de vista social, otros autores (Jerskey et al., 2010; Schaan et al., 2019) muestran como los procesos de divorcio pueden estar influenciados por factores genéticos, ambientales y hasta desórdenes mentales. En este mismo sentido, diversos autores han desarrollado otros estudios similares (Buyukkececi y Leopold, 2021; Sorek, 2020; Lan, 2021), donde muestran los efectos de la interacción social en el matrimonio, divorcio, calidad de vida de los hijos y perfiles parentales. Otros autores (Spikic et al., 2021; Angelini et al., 2019; Klurfeld, 2020) analizan si un divorcio puede producir cambios en la personalidad de un individuo, factores que aumenten el riesgo de divorcio, y sus efectos entre otros aspectos. La anterior revisión literaria, elaborada en este documento, muestra que la mayoría de las investigaciones en este campo han sido realizadas desde un punto de vista social, encontrándose muy pocos trabajos desde un punto de vista matemático computacional, siendo aún más

escasos los trabajos relacionados con sistemas de predicción de divorcios basados en técnicas inteligentes, computacionales y matemáticas. Específicamente bajo la plataforma de aprendizaje automático y minería de datos denominado Weka solo se encontró un trabajo realizado (Yöntem et al., 2019). Los anteriores aspectos justifican aún más el desarrollo de este proyecto de investigación, dado que ha sido un campo poco explorado.

De otro lado, si se tiene presente la gran cantidad de atributos que pueden resultar en esta investigación (Tabla 1), es necesario desarrollar un proceso de selección de atributos el cual tiene como objetivo reducir la dimensionalidad del conjunto de datos original. Estas técnicas de reducción manejan un método de selección y una función de evaluación (Rodríguez et al., 2014), aspecto de gran relevancia en este caso. Actualmente, existen diversas formas de realizar esta selección las cuales están basadas en: Correlaciones, aprendizajes basados en casos, búsquedas exhaustivas, reglas, Maquinas de soporte vectorial, vecino más cercano, ranking, entre otros (Toledo, 2016). En este mismo sentido los métodos de evaluación están asociados a diferentes técnicas, algunas de ellas basadas en minería de datos como: árboles de decisión, redes neuronales, redes bayesianas, bosques aleatorios, arboles de funciones, algoritmos genéticos, algoritmos C4.5 entre otros. Los anteriores algoritmos pueden ser complementados por diferentes funciones de evaluación como porcentaje de aciertos y errores, falsos positivos, falsos negativos, matrices de confusión, índice Kappa, error medio absoluto, error cuadrático medio, entre muchas otras funciones.

Para su presentación este artículo ha sido estructurado de la siguiente forma: inicialmente, se expone la metodología empleada para abordar el problema objeto de estudio, en esta metodología se explica detalladamente la base de datos utilizada y cada una de sus características o variables influyentes en el proceso de divorcio. Seguidamente, en la sección de resultados se establecen las variables principales que afectan un proceso de divorcio y su porcentaje de efectividad. Posteriormente, se encuentra, en el documento, una sección denominada discusión de resultados, en la cual se realiza una comparación de los resultados obtenidos, frente a otros trabajos similares. Finalmente, se incluye una sección de conclusiones, las cuales son una consecuencia lógica de los resultados. Además, se presentan los correspondientes agradecimientos y referencias empleadas. Esta metodología se estructura con base en trabajos previos del mismo autor (Rodas y Castrillón, 2019; Quesada et al., 2020)

MATERIALES Y METODOS

El desarrollo de esta metodología se hace con base en la Plataforma de Minería de datos denominada Weka. Aunque existen muchos algoritmos de clasificación en esta plataforma, específicamente, se trabaja con el algoritmo J48 para predecir el comportamiento de la variable dependiente denominada Divorcio. Esta predicción, se realiza con base en las siguientes variables independientes, encontradas en el vínculo relacionado previamente en la introducción (Tabla1). Las variables relacionadas en la Tabla 1, se encuentra en una escala de 0 a 4, donde 0 representa el peor estado y 4 el mejor estado posible.

Tabla1: Variables influyentes en el divorcio. (Dua y Graff, 2019; Yöntem et al., 2019)

Variable (V)	Descripción	Máximo	Mínimo	Promedio	Desviación
1	Si alguien se disculpa, cuando la discusión empeora, la discusión termina	0	4	1.77	1.62
2	Se pueden ignorar las diferencias aun en situaciones difíciles	0	4	1.65	1.46
3	Se pueden retomar las discusiones desde el principio y corregirlas	0	4	1.76	1.41
4	Cuando discuto con mi cónyuge, funcionara contactarlo	0	4	1.48	1.50
5	El tiempo que paso con mi cónyuge es especial para nosotros	0	4	1.54	1.63
6	No pasamos tiempo en casa como socios	0	4	0.74	0.90
7	Somos dos extraños que comparten el mismo ambiente en casa más que en familia	0	4	0.49	0.89
8	Disfruto los festivos con mi cónyuge	0	4	1.45	1.54
9	Disfruto viajar con mi cónyuge	0	4	1.45	1.55
10	La mayoría de nuestros objetivos son comunes	0	4	1.57	1.42
11	Pienso que un día en el futuro, cuando mire hacia atrás, veré que cada uno ha estado en armonía.	0	4	1.68	1.64
12	Tenemos valores similares referentes a la libertad	0	4	1.65	1.46
13	Tenemos formas similares de entretenernos	0	4	1.83	1.47
14	La mayoría de los objetivos para las personas son los mismos	0	4	1.57	1.50
15	Nuestros sueños con mi cónyuge son similares y armoniosos	0	4	1.57	1.50
16	Somos compatibles sobre lo que debería ser el amor	0	4	1.47	1.50
17	Compartimos los mismos puntos de vista sobre cómo ser felices en nuestra vida conyugal	0	4	1.65	1.61
18	Tenemos ideas similares sobre como debería ser el matrimonio	0	4	1.51	1.56
19	Tenemos ideas similares sobre como deben ser los roles en el matrimonio	0	4	1.64	1.64
20	Tenemos valores similares de confianza	0	4	1.45	1.55
21	Se exactamente que le gusta a mi cónyuge	0	4	1.38	1.45
22	Se como mi conyuge quiere que lo cuiden cuando está enfermo	0	4	1.24	1.44

Tabla 1: continuación.

Variable (V)	Descripción	Máximo	Mínimo	Promedio	Desviación
23	Conozco la comida favorita de mi cónyuge	0	4	1.41	1.61
24	Puedo decir que tipo de estrés enfrenta mi cónyuge	0	4	1.51	1.50
25	Conozco el mundo interior de mi cónyuge	0	4	1.62	1.53
26	Conozco las ansiedades básicas de mi cónyuge	0	4	1.48	1.50
27	Conozco las fuentes actuales de estrés de mi cónyuge	0	4	1.4	1.45
28	Conozco las esperanzas y deseos de mi cónyuge	0	4	1.30	1.46
29	Conozco muy bien a mi cónyuge	0	4	1.49	1.59
30	Conozco los amigos de mi cónyuge y sus relaciones sociales	0	4	1.49	1.50
31	Me siento agresivo cuando discuto con mi cónyuge	0	4	2.12	1.64
32	Cuando discuto usualmente uso expresiones como: usted siempre, usted nunca.	0	4	2.05	1.62
33	Uso frases negativas acerca de la personalidad de mi cónyuge en una discusión	0	4	1.80	1.78
34	Use expresiones ofensivas en una discusión	0	4	1.9	1.63
35	Insulto a mi cónyuge en una discusión	0	4	1.60	1.84
36	Puedo ser humillado cuando discutimos	0	4	1.60	1.79
37	Las discusiones con mi cónyuge no son calmadas	0	4	2.08	1.71
38	Odio la forma en que mi cónyuge abre un tema	0	4	1.85	1.73
39	Las discusiones a menudo ocurren de repente	0	4	2.08	1.71
40	Comenzamos una discusión antes de saber que está pasando	0	4	1.87	1.79
41	Cuando hablo con mi conyugue sobre algo, de repente rompo mi calma	0	4	1.99	1.72
42	Cuando discuto con mi conyugue solo salgo y no digo una sola palabra	0	4	2.15	1.57
43	Me quedo en silencio para calmar el ambiente	0	4	2.70	1.34
44	Pienso que es bueno que me vaya de casa un rato	0	4	1.94	1.68
45	Prefiero quedarme en silencio que discutir con mi conyugue	0	4	2.45	1.49
46	Incluso si tengo razón permanezco en silencio para no lastimar a mi conyugue	0	4	2.55	1.37
47	Cuando hablo con mi cónyuge, me quedo en silencio porque tengo miedo de no poder controlar mi ira	0	4	2.27	1.58
48	Me siento bien en nuestras discusiones	0	4	2.74	1.13
49	No tengo nada que ver con lo que me acusan	0	4	2.38	1.54
50	No soy culpable de lo que me acusan	0	4	2.42	1.40
51	No soy el que se equivoca con los problemas en casa	0	4	2.47	1.26
52	No dudaría en decirle a mi conyugue sobre su insuficiencia	0	4	2.51	1.47
53	Cuando discuto le recuerdo a mi cónyuge su insuficiencia	0	4	2.24	1.50
54	No tengo miedo de decirle a mi conyugue sobre su incompetencia	0	4	2.01	1.66

La anterior predicción permite diseñar planes de prevención tendientes a minimizar este problema, como se comenta en la introducción. La metodología propuesta en esta investigación, es estructurada en siete pasos: (1) Construcción de la base de datos; (2) validación del cuestionario (3) Correlación y selección de variables; (4) importancia de las variables seleccionadas; (5) Diseño del archivo arff; (6) Éxito y generación del árbol de clasificación; (7) Identificación de las principales causas de divorcio. El desarrollo de esta investigación se hace tomando como referencia la metodóloga desarrollada por el mismo autor en trabajos previos (Rodas y Castrillón, 2019; Quesada et al., 2020)

Paso 1. Depuración de la base de datos.

Para la depuración de la base de datos, se tomaron 170 registros. Estos registros se captaron para cada una de las 54 variables independientes, existentes en la base de datos, según se muestra en la Tabla 1. Adicionalmente, en esta base de datos, se incluyó una variable dependiente denominada divorcio. Las variables independientes son medidas en una escala de 0 a 4, donde cero es el valor mínimo y 4 el máximo.

Paso 2. Validación del cuestionario

Con el fin de validar el cuestionario y establecer la consistencia interna de mismo, se calcula el índice alfa de Cronbach. Este cálculo se realiza por medio del complemento XRealStats de Excel. El resultado de este índice será un valor entre 0 y 1. Cuando el índice es superior a 0.7, se puede decir que el cuestionario tiene consistencia interna, siendo más consistente entre más cercano sea su valor a 1.

Paso 3. Correlación y Selección de Variables.

Tomando como referencia la base de datos construida en el Paso 1, se analiza la correlación existe entre cada una de las variables independiente respecto a la variable dependiente, con el fin de verificar que la correlación entre estas variables no sea insignificante o demasiado alta, en este sentido, se seleccionan aquellas variables con una correlación mayor que 0.3 y menor que 0.8. Las variables que no cumplan con esta premisa son descartadas, dado que su influencia sobre la variable dependiente es escasa o presentan

una influencia demasiado grande. Una influencia muy alta condicionaría el comportamiento de la variable dependiente a estas variables independientes.

Paso 4. Importancia de las variables seleccionadas.

Con base en la misma plataforma Weka, específicamente en el algoritmo J48, se establece la efectividad de las variables seleccionadas en el paso 2 de la metodología, previa elaboración del archivo .arff el cual es interpretado por medio esta esta plataforma. Posteriormente, por medio de la plataforma Weka (algoritmo J48), se mira la influencia de cada una estas variables seleccionadas respecto a su porcentaje de clasificación. Para determinar esta influencia, se suprime una variable a la vez y se determina el nuevo porcentaje de éxito obtenido con las variables restantes. Un menor porcentaje de éxito indica una mayor relevancia de la variable suprimida.

Paso 5. Diseño del archivo .Arff

Con base en la información obtenida en el paso anterior, se seleccionan las variables más influyentes, con el fin de construir el respectivo archivo Arff, para ser interpretado por medio del clasificador bayesiano J48, desde la plataforma de aprendizaje automático y minería de datos denominada Weka. El diseño de este archivo se estructura en dos partes encabezado y datos.

Paso 6. Éxito y generación del árbol de clasificación

El archivo elaborado en el paso 4, se interpreta por medio de la plataforma Weka, empleando el algoritmo de clasificación bayesiano J48, mediante las técnicas validación cruzada (90-10) y “*Porcentaje Split*” (Porcentaje de partición). En esta última técnica el sistema es entrenado con porcentajes del 90%, 80% y 60% de los datos y validado con los porcentajes complementarios, respectivamente. Estas técnicas permiten generar el respectivo árbol de decisión, sobre el cual es posible analizar las principales causas de un divorcio. Finalmente, en este punto se expresa que, aunque el algoritmo J48 no es un algoritmo bayesiano, presenta un comportamiento muy similar a estos algoritmos, siendo importante resaltar que los algoritmos de clasificación bayesiana permiten obtener muy buenos resultados con pocos datos (Valencia et al., 2015), lo cual hace que la muestra analizada sea suficiente.

Paso 7. Identificación de las principales causas de divorcio.

Con base en el árbol de clasificación obtenido en el paso anterior mediante el algoritmo bayesiano J48, se determinan las principales causas, representadas en las variables independientes (nodos del árbol) que inciden en un resultado de divorcio o variable dependiente (hojas del árbol).

RESULTADOS

Como consecuencia del desarrollo de cada uno de los pasos descritos en la metodología anterior, se obtuvieron los siguientes resultados:

Paso 1. Depuración de la base de datos.

En el desarrollo de este paso, se tomaron todos los registros de la base de datos relacionada en el paso 1 de la metodología. Las Tablas 2, 3 y 4, ilustran un pequeño fragmento de esta base de datos. Por razones de espacio no se ilustran todos los registros. La Tablas 2, 3 y 4 son elaboradas con la información existente (Dua, y Graff, 2019; Yöntem, et al., 2019). Los nombres de las variables ilustradas en las Tablas 2, 3 y 4 son descritos en la Tabla 1.

Tabla 2: Primer conjunto de variables

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
2	2	4	1	0	0	0	0	0	0	1	0	1	1	0	1	0	0	0	1
4	4	4	4	4	0	0	4	4	4	4	3	4	0	4	4	4	4	3	2
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2	2	2	2	1	3	2	1	1	2	3	4	2	3	3	3	3	3	3	2
3	2	3	2	3	3	3	3	3	3	4	3	3	4	3	3	3	3	3	4
2	2	1	1	1	1	0	0	0	0	0	1	0	1	1	1	1	1	2	1

Tabla 3: Segundo conjunto de variables

V21	V22	V23	V24	V25	V26	V27	V28	V29	V30	V31	V32	V33	V34	V35	V36	V37	V38	V39	V40
0	0	0	0	0	0	0	0	0	1	1	2	1	2	0	1	2	1	3	3
1	1	0	2	2	1	2	0	1	1	0	4	2	3	0	2	3	4	2	4
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
1	0	1	2	2	2	2	2	3	2	3	3	1	1	1	1	2	1	3	3
1	1	1	1	2	1	1	1	1	3	2	3	2	2	1	1	3	3	4	4
1	0	0	0	0	2	1	2	1	1	1	1	1	1	0	0	0	0	2	1

Tabla 4: Tercer conjunto de variables

V41	V42	V43	V44	V45	V46	V47	V48	V49	V50	V51	V52	V53	V54	D(1=S,0=N)
2	1	1	2	3	2	1	3	3	3	2	3	2	1	1
2	2	3	4	2	2	2	3	4	4	4	4	2	2	1
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	3	2	3	2	3	2	3	1	1	1	2	2	2	1
0	2	3	0	2	2	1	2	3	2	2	2	1	0	1

Paso 2. Validación del cuestionario.

Como resultado de este paso, se establece que el índice de Alfa de Cronbach es igual a 0.9932. Como se expresó en el paso 2 de la metodología, este índice es calculado por medio del complemento XRealstats de Excel. Este índice establece una muy buena consistencia interna del cuestionario.

Paso 3. Correlación y Selección de Variables.

Por medio del programa Excel, se estableció la correlación entre cada una de las variables independientes respecto a la variable dependiente como se ilustra en las Tablas 5, 6 y 7. Las Tablas 5, 6 y 7 muestran 13 variables (variables señaladas con un *) con una correlación moderada mayor que 0.3 y menor que 0.8, estas variables son: V6, V7, V42, V43, V45, V46, V47, V48, V49, V50, V51, V52, V53.

Tabla 5: Correlación del primer conjunto de variables independientes respecto a la variable dependiente.

	V1	V2	V3	V4	V5	*V6	*V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20
Divorcio	0,9	0,8	0,8	0,8	0,9	0,4	0,5	0,9	0,9	0,8	0,9	0,9	0,8	0,9	0,9	0,9	0,9	0,9	0,9	0,9

Tabla 6: Correlación del segundo conjunto de variables independientes respecto a la variable dependiente.

	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30	V31	V32	V33	V34	V35	V36	V37
Divorcio	0,9	0,8	0,8	0,8	0,9	0,9	0,9	0,8	0,9	0,9	0,8	0,8	0,9	0,8	0,9	0,9	0,9

Tabla 7: Correlación del tercer conjunto de variables independientes respecto a la variable dependiente.

	V38	V39	V40	V41	*V42	*V43	V44	*V45	*V46	*V47	*V48	*V49	*V50	*V51	*V52	*V53	V54
Divorcio	0,9	0,9	0,9	0,9	0,7	0,6	0,8	0,5	0,4	0,7	0,6	0,7	0,8	0,7	0,7	0,7	0,8

Cuando la selección de variables se realiza por medio de otros métodos (Tabla 8) los resultados obtenidos en el proceso de clasificación no son tan buenos como los resultados que se obtiene cuando se usa un filtrado por ranking empleando el método de correlaciones. Método con el cual se llega a obtener hasta el 100% de éxito con solo 6 variables. Los resultados de la Tabla 8 pueden ser obtenidos empleando el menú de selección de atributos de Weka, sin embargo, en ningún caso se logra superar el 99%, razón por la cual se selecciona un filtrado por ranking empleando el método de correlaciones, como se expresó antes.

Tabla 8: Resultados obtenidos por medio de otras formas de clasificación (Weka)

Seleccionador -Evaluador	Búsqueda	Éxito	Observación
CfsSubsetEval	Bestfirst	79.4%	Se busca un conjunto de características las cuales funcionen bien juntas. (Todo el conjunto para entrenamiento)
ChiSquaredAttributeEval	Ranker	98.82%	Evalúa los atributos según la distribución Chi Cuadrado. Se obtiene la siguiente clasificación de atributos: V18, V20, V40, V17, V19, V11, V9, V16, V26, V15, V29, V30, V39, V12, V25, V14, V36, V5, V41, V27, V28, V8, V4, V21, V38, V37, V1, V44, V13, V22, V35, V10, V24, V33, V2, V31, V23, V34, V32, V54, V3, V50, V42, V51, V49, V53, V7, V48, V47, V52, V43, V45, V6, V46. Se seleccionan las 44 primeras variables - 90% de la información. Se entrena con todo el conjunto
ConsistencySubsetEval	Genetic Search	98.23%	Se valoran los atributos según el nivel de consistencia y se obtiene la siguiente selección: V3, V4, V11, V13, V15, V18, V42. Se entrena con todo el conjunto.
CostSensitiveAttributeEval	Ranker	98.82%	Valoración basada en un meta conjunto evaluador con un subconjunto de base sensitivo al costo.

Paso 4. Importancia de las variables seleccionadas.

En una primera predicción del comportamiento de la variable dependiente (Divorcio) con base en las variables independientes seleccionadas en el paso anterior, empleando el algoritmo J48 de la plataforma Weka, se obtiene un resultado de éxito del 91.76% por medio de un proceso de validación cruzada (90-10) (Tabla 9).

Tabla 9: Porcentaje de Clasificación. Validación cruzada (90 -10)

Nombre	Valor	Percentage
Instancias clasificadas correctamente	156	91.7647%
Instancias clasificadas incorrectamente	14	8.2353
Estadística Kappa	0.8352	
Error medio absolute	0.1141	
Error cuadrático Medio	0.2653	
Error absoluto relative	22.8195 %	
Raíz cuadrática del error relativo	53.0519%	
Número total de instancias	170	

Posteriormente como se ilustró en el paso 3 de la metodología, por medio de la plataforma Weka (algoritmo J48), se encuentra la influencia de cada una de las 13 variables seleccionadas. En este proceso se suprime una variable a la vez y se determina el nuevo porcentaje de éxito obtenido con las 12 variables restantes. Un menor porcentaje de éxito indica una mayor relevancia de la variable suprimida (Tabla 10). Según la información suministrada por la Tabla 10, las variables que presentan una mayor influencia sobre la variable dependiente son: V6, V7, V43, V45, V46, V47. (Variables señaladas con un asterisco en la Tabla 10)

Tabla 10: Influencia de cada variable.

V6*	V7*	V42	V43*	V45*	V46*	V47*	V48	V49	V50	V51	V52	V53	% Acierto
	x	x	x	x	x	x	x	x	x	x	x	x	*89.41
x		X	x	x	x	x	x	x	x	x	x	x	*87.64
x	x		x	x	x	x	x	x	x	x	x	x	91.76
x	x	x		x	x	x	x	x	x	x	x	x	*89.41
x	x	x	x		x	x	x	x	x	x	x	x	*89.41
x	x	x	x	x		x	x	x	x	x	x	x	*88.11
x	x	x	x	x	x		x	x	x	x	x	x	*90.58
x	x	x	x	x	x	x		x	x	x	x	x	91.76
x	x	x	x	x	x	x	x		x	x	x	x	91.76
x	x	x	x	x	x	x	x	x		x	x	x	92.35
x	x	x	x	x	x	x	x	x	x		x	x	91.76
x	x	x	x	x	x	x	x	x	x	x		x	91.76
x	x	x	x	x	x	x	x	x	x	x	x		91.76

Paso 5. Diseño del archivo Arff

Con el fin de analizar el comportamiento de las 6 variables seleccionadas en el paso anterior, desde la plataforma de aprendizaje automático y minería de datos denominada Weka, se toman estas 6 variables seleccionadas y se estructura el respectivo archivo arff para ser interpretado por esta plataforma. La estructura de este archivo, se muestra en dos Tablas encabezado (Tabla 11) y datos (Tabla 12) como se ilustra a continuación. Por razones de espacio solo se presenta parte del archivo ilustrado en la Tabla 11.

Tabla 11: Encabezado archivo .arff

@relation	relation
@attribute	V6 {0,1,2,3,4}
@attribute	V7 {0,1,2,3,4}
@attribute	V43 {0,1,2,3,4}
@attribute	V45 {0,1,2,3,4}
@attribute	V46 {0,1,2,3,4}
@attribute	V47 {0,1,2,3,4}
@attribute	Divorcio {SI,NO}

Tabla 12: Datos archivo .arff

V6	V7	V43	V45	V46	V47	D
0	0	3	2	2	2	SI
3	2	2	2	3	2	SI
.	.	.	.	.	.	.
0	0	4	4	4	4	SI

Paso 6. Éxito y generación del árbol de clasificación.

Con base en el archivo Arff generado en el paso anterior, se construye el respectivo árbol de clasificación para ser interpretado por medio del algoritmo J48, de comportamiento similar a los Bayesianos. Aunque esta técnica no es un algoritmo bayesiano, presenta un comportamiento similar a los mismos con resultados superiores o iguales a los demás métodos de clasificación, como se ilustra en la Tabla 13. El árbol construido

es mostrado en Figura 1. La clasificación lograda por este algoritmo, permite un nivel de aciertos del 90%, mediante un proceso de validación cruzada (90-10). La Tabla 14. Ilustra estos resultados. Por otro lado, cuando se entrena el sistema empleando diferentes porcentajes de partición, se puede lograr un nivel de éxito del 100%. Este aspecto es ilustrado en la Tabla 15.

Tabla 13: Resultados de clasificación por medio de otras técnicas

Técnica	Resultado	Observación
AdTree	86%	Árbol de decisión alterno
BFTree	86%	Árbol de decisión con los mejores primeros
DecisionStump	80%	Árbol de decisión de un nivel
FT	86%	Clasificador basado en árboles de funciones
ID3	86%	Árbol de decisión sin podar basado en el algoritmo ID3
J48	100%	Árbol de decisión sin podar basado en el algoritmo C4.5
J48Graft	93.33%	Árbol de decisión podado basado en el algoritmo C4.5
Logistic	80%	Regresión Logística con un modelo estimador
RBFNetwork	86.66%	Redes Neuronales
BayesNet	86.66%	Redes Bayesianas
NaiveBayes	86.66%	Naive de Bayes
RandomForest	86.66%	Random Forest

Tabla 14: Porcentaje de Clasificación. Variables V6, V7, V43, V45, V46, V47

Nombre	Valor	Percentage
Instancias clasificadas correctamente	153	90 %
Instancias clasificadas incorrectamente	17	10 %
Estadística Kappa	0.7999	
Error medio absolute	0.1581	
Error cuadrático Medio	0.2924	
Error absoluto relativo	31.6165	
Raíz cuadrática del error relativo	58.461%	
Numero total de instancias	170	

Tabla 15: Éxito según el porcentaje de partición.

% de partición (Entrenamiento)	% de Éxito
Mayor al 90%	100%
80%	91.17%
60%	86.76%

Paso 7. Identificación de las principales causas de divorcio

Según lo ilustrado en la Figura 1, se identifican 4 causas principales para un divorcio, las cuales son desglosadas en la Tabla 16.

Tabla 16: Descripción principales causas de divorcio.

Código	Descripción
V7	Somos dos extraños que comparten el mismo ambiente en casa más que en familia
V47	Cuando hablo con mi cónyuge, me quedo en silencio porque tengo miedo de no poder controlar mi ira
V46	Incluso si tengo razón permanezco en silencio para no lastimar a mi conyugue
V43	Me quedo en silencio para calmar el ambiente

Fig. 1: Árbol de clasificación

DISCUSION

El árbol de clasificación ilustrado en la Figura 1, determina 4 variables influyentes en un divorcio, todas ellas relacionadas con los procesos de comunicación en pareja. En orden de importancia estas variables son: V7, V47, V46, V43. La primera variable V7, muestra que si en una relación, las dos personas se sienten como extrañas, así sea en algún grado ($V7 \geq 1$), este hecho puede conducir a un divorcio. Solo en el caso en que las personas no se sientan como extraños ($V7 = 0$), empiezan a tener importancia las demás variables. En este sentido, la variable V47 muestra que, si alguno de los miembros de la pareja nunca se queda en silencio, perdiendo el control de su temperamento o lastimando a su pareja ($V47 \geq 4$, $V46 \geq 4$), se puede producir un divorcio. Igualmente, cuando la persona prudentemente, permanece algunas veces en silencio para calmar el ambiente ($V43 = 3$), el divorcio puede ser evitado.

Por otro lado, en la literatura descrita en este artículo solo se encontró un trabajo en este mismo sentido (Yöntem et al., 2019), como se ilustra en este documento. Este autor logra porcentajes superiores al 98% por medio de redes neuronales, seleccionado 6 de las 54 características como predominantes. En esta investigación se obtiene un porcentaje de clasificación del 90% cuando se emplea una validación cruzada (90-10) o un porcentaje de acierto del 100% cuando se emplea un porcentaje de partición superior al 90%, para entrenar la base de datos. Aunque en principio los resultados pueden diferir entre los dos trabajos esto, se debe a la diferencia en la selección de las variables independientes empleadas en ambas investigaciones. En este trabajo se seleccionaron las variables con una correlación moderada, frente a la variable dependiente. Posteriormente empleando la misma plataforma Weka se estableció la importancia de cada una de las variables seleccionadas y se realizó un segundo proceso de selección. En el trabajo referenciado, se seleccionan las variables independientes con la correlación más significativa, frente a la variable dependiente.

Si bien los dos métodos de selección de variables son válidos, en este trabajo se decidió descartar aquellas variables independientes con poca influencia o una gran influencia la cual pudiera condicionar el resultado de la variable dependiente. Finalmente, es importante resaltar que los resultados son válidos en el caso de este estudio analizado, por lo cual es necesario extender la aplicación de la metodología a una muestra más amplia en muchos otros contextos, especialmente en Colombia con el fin de encontrar factores propios de la cultura que podrían incidir en este problema, aspecto que podría llevar al surgimiento de otras causas diferentes de un divorcio, las cuales complementen este proceso investigativo. Desafortunadamente por la emergencia sanitaria que vive el mundo como consecuencia del COVID -19 esto no fue posible, sin embargo, se deja la metodología planteada para que la misma pueda ser aplicada posteriormente en cualquier contexto.

CONCLUSIONES

Por medio del árbol de decisión se puede observar de forma clara y concisa las variables más importantes que pueden causar un proceso de divorcio, según la metodología propuesta: V7, V47, V46, V43. Variables que son descritas en la Tabla 16. El árbol de clasificación obtenido, en esta investigación, permite obtener una mirada mucho más clara de la información objeto de análisis. Este árbol, también permite predecir, futuros divorcios con un porcentaje de acierto que puede llegar hasta el 100%, como se ilustró en este trabajo al aplicar la metodología descrita en la sección denominada “*Materiales y Métodos*”. Una correcta predicción de un divorcio dependerá de los eventos que se presenten, sobre las 4 variables seleccionadas y su posible corrección de forma oportuna. Trabajos similares, podrán ser aplicados en la predicción de diferentes problemas sociales, en diversos países y comunidades. Finalmente, se resalta que replicar este trabajo es fácil, para esto solo se debe seguir cada uno de los pasos estructurados en la metodología, algo más factible de lograr si se tiene presente que esta investigación está desarrollada bajo plataforma de software libre denominada Weka.

AGRADECIMIENTOS

Se agradece la colaboración a la Universidad Nacional de Colombia Sede Manizales. Igualmente se agradece a: Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

REFERENCIAS

- Angelini, V., Bertoni, M. y otros 2 autores, The Ant or the Grasshopper? The Long-Term Consequences of Unilateral Divorce Laws on Savings of European Households, <https://doi.org/10.1016/j.eurocorev.2019.07.002>, European Economic Review, 119, 97–113 (2019)
- Balestrino, A., Ciardib, C. y Mamminia, C., On the Causes and Consequences of Divorce, <http://dx.doi.org/10.1016/j.socrec.2013.02.006>, The Journal of Socio-Economics, 45, 1– 9 (2013)
- Bucheli, M. y Vigna, A., Un Estudio de los Determinantes del Divorcio en Uruguay, Desarrollo y Sociedad, 56, 1-21 (2005)

- Buyukkececi, Z. y Leopold, T., Sibling influence on Family Formation: A Study of Social Interaction Effects on Fertility, Marriage, and Divorce, <https://doi.org/10.1016/j.alcr.2020.100359>, *Advances in Life Course Research*, 47(1), 100359 (2021)
- Cáceres, E., Inteligencia Artificial, Derecho y E-Justice (el Proyecto iij-Conacyt)*, *Boletín Mexicano de Derecho Comparado*, nueva serie, año XXXIX, 116, 593-611 (2005)
- Carneiro, D., Novais, P. y Neves, J., Using Genetic Algorithms to Create Solutions for Conflict Resolution, <http://dx.doi.org/10.1016/j.neucom.2012.03.024>, *Neurocomputing*, 109, 16–26 (2013)
- Cavapozzi, D., Fiore, S. y Pasini, G., Divorce and Well-Being. Disentangling the Role of Stress and Socio Economic Status, <https://doi.org/10.1016/j.jeoa.2019.100212>, *The Journal of the Economics of Ageing*, 16, 100212 (2020)
- Dijk, R., Valk, I. y otros 2 autores, A Meta-Analysis on Interparental Conflict, Parenting, and Child Adjustment in Divorced Families: Examining Mediation Using Meta-Analytic Structural Equation Models, <https://doi.org/10.1016/j.cpr.2020.101861>, *Clinical Psychology Review*, 79, 101861 (2020)
- Dua, D. y Graff, C., UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science, (2019)
- Jerskey, B., Panizzon, M. y otros 7 autores, Marriage and Divorce: A Genetic Perspective, <http://doi:10.1016/j.paid.2010.05.007>, *Personality and Individual Differences*, 49, 473–478 (2010)
- Klurfeld, Z.B., Buqo, T. y otros 2 autores, Comparing the Nature of Grief and Growth in Bereaved Divorced and Unemployed Individuals, <https://doi.org/10.1016/j.jad.2020.06.008>, *Journal of Affective Disorders*, 274, 1126–1133 (2020)
- Lan, X., Disengaged and Highly Harsh? Perceived Parenting Profiles, Narcissism, and Loneliness Among Adolescents from Divorced Families, <https://doi.org/10.1016/j.paid.2020.110466>, *Personality and Individual Differences*, 171, 110466 (2021)
- Martínez, G., La Inteligencia Artificial y su Aplicación al Campo del Derecho, *Alegatos*, 82, 827-846 (2012)
- Quezada, M., Tobón, A. y Castrillón, O., Minería de Datos: Una Aplicación para Determinar cuáles Factores Socio-Económicos Influyen en el Embarazo Adolescente, <http://dx.doi.org/10.4067/S0718-07642020000600053>, *Información Tecnológica*, 30(6), 53-60 (2020)
- Ribeiro, M., Landero, R. y Thierry, B., *El Divorcio Procesos Causas y Consecuencias*, 1^{era} Ed., Clave Editorial, México, México (2012)
- Rodas, L. y Castrillón, O., Predicción de Fallos Mecánicos en Equipos de Envolvura, <http://dx.doi.org/10.4067/S0718-07642019000600111>, *Información Tecnológica*, 30(6), 111-122 (2019)
- Rodríguez, Y., Fernández, Y. y otros 2 autores, Selección de Atributos Relevantes Aplicando Algoritmos que Combinan Conjuntos Aproximados y Optimización en Colonias de Hormigas, *Rev cuba cienc informat*, 8(1), 140-155 (2014)
- Ruiz, P., La Construcción del Divorcio en Colombia desde las Normas Jurídicas a Partir del Siglo XIX. Diferencias de Género e Influencia Política y Religiosa, *Revista de Derecho Privado*, 39, 109-139 (2020)
- Samaca, A., Inteligencia Artificial Aplicada al Derecho, Tesis de grado, <http://hdl.handle.net/11634/9376>, Universidad Santo tomas, Facultad de Derecho, Bogotá, Colombia (2016)
- Schaan, V., Schulz, A. y otros 2 autores., Parental Divorce is Associated with an Increased Risk to Develop Mental Disorders in Women, <https://doi.org/10.1016/j.jad.2019.06.071>, *Journal of Affective Disorders*, 257, 91–99 (2019)
- Sorek, Y., Grandparental and Overall Social Support as Resilience Factors in Coping with Parental Conflict Among Children of Divorce, <https://doi.org/10.1016/j.childyouth.2020.105443>, *Children and Youth Services Review*, 118, 105443 (2020)
- Spikic, S., Mortelmans, D. y Pasteels, I., Does Divorce Change your Personality? Examining the Effect of Divorce Occurrence on the Big Five Personality Traits Using Panel Surveys From Three Countries, <https://doi.org/10.1016/j.paid.2020.110428>, *Personality and Individual Differences*, 171, 110428 (2021)
- Tamez-Valdez, B. y Ribeiro-Ferreira, M., El Divorcio, Indicador de Transformación Social y Familiar con Impacto Diferencial entre los Sexos: Estudio Realizado en Nuevo León, DOI: <http://dx.doi.org/10.22185/24487147.2016.90.040>, *papeles de población*, 90, 229-263 (2016)
- Toledo, A., Métodos de Selección de Atributos para Clasificación Supervisada Basados en Teoría de Información, Instituto Superior Politécnico “José Antonio Echeverría” Facultad de Ingeniería Informática, La Habana, Cuba (2016)
- Valencia, M., Correa, J. y Díaz, F., Métodos Estadísticos Clásicos y Bayesianos para el Pronóstico de Demanda. Un Análisis Comparativo, <https://doi.org/10.15446/rev.fac.cienc.v4n1.49775>, *Revista Facultad de Ciencias Universidad Nacional de Colombia*, 4(1), 52 -67 (2015)
- Witten, I., Frank, E. y otros dos autores, *Data Mining Practical Machine Learning Tools and Techniques*, Morgan and Kaufman publication (Elsevier), ISBN-13 978-0128042915, Cambridge, USA (2017)
- Yöntem, M., Adem, K. y otros dos autores, Divorce Prediction Using Correlation Based Feature Selection and Artificial Neural Networks, *Nevşehir Hacı Bektaş Veli University SBE Dergisi*, 9(1), 259-273 (2019)