Aplicación de Machine Learning para análisis de los fenómenos de violencia intrafamiliar en el departamento del Atlántico

Machine Learning application for analysis of domestic violence phenomena in the department of Atlantico

K. Chamorro*, N. Laza*, H. Noriega*, R. Rojano*, J. Vega* & D. Heredia**
{kenner.chamorro, nicolas.laza , harold.noriega, Royber.rojano, juan.vega} @unisimon.edu.co –
{dianahv} @unisimonbolivar.edu.co
**Estudiante de Ingeniería de sistemas **Profesor investigador del grupo ingebiocaribe*
Universidad Simón Bolívar, Barranquilla-Colombia.

Resumen La finalidad de este artículo es presentar los resultados de una búsqueda global y caracterizar los hechos de violencia intrafamiliar presentados en el departamento del Atlántico; al recolectar información de estudios hechos a nivel mundial, en los continentes europeos, el continente asiático, y el continente americano, al igual, cuáles son los países más afectados por la violencia intrafamiliar, se ubica a Colombia en el puesto 147 de 162 países evaluados, siendo el número 1 el puesto correspondiente al país más pacífico. Partiendo de esto, hicimos un enfoque en el país, más específicamente en el departamento del atlántico. A su vez a los datos recolectados se les aplicaron técnicas de Machine Learning.

Palabras clave: *Machine Learning, violencia intrafamiliar, estudios de hechos.*

Abstract The purpose of this article is to present the results of a global search and to characterize the acts of domestic violence presented in the department of Atlántico; When collecting information from studies carried out worldwide, in the European continents, the Asian continent, and the American continent, as well as which countries are most affected by domestic violence, Colombia is ranked 147 out of 162 countries evaluated , number 1 being the position corresponding to the most peaceful country. Starting from this, we focused on the country, more specifically on the department of the Atlantic. In turn, Machine Learning techniques were applied to the collected data.

Keywords: *Machine Learning, domestic violence, factual studies.*

Para referenciar este artículo (IEEE):

K. Chamorro *, N. Laza *, H. Noriega*, R. Rojano *, J. Vega* & D. Heredia, "Aplicación de Machine Learning para análisis de los fenómenos de violencia intrafamiliar en el departamento del Atlántico", *Investigación y Desarrollo en TIC,* vol. 12, no. 1, pp. 1-12 2021.

I. INTRODUCCIÓN

La violencia intrafamiliar es un fenómeno social que ha sido legitimado por décadas. A pesar de su extensión y gravedad, no se le ha dado la importancia que merece, en gran parte porque hasta hace muy poco el espacio de lo familiar había sido reconocido como un espacio que pertenece a la intimidad, y los comportamientos violentos se llegan a legitimar como herramientas útiles para educar, mantener el control, o como mecanismo válido para resolver sus conflictos. La violencia dentro de la familia, entonces, ha sido considerada como «funcional» porque reforzaba los roles de poder y relaciones jerárquicas y de fuerza que se dan en la misma, así como en algunos casos se legitiman patrones de crianza violentos, como parte del proceso educativo de niños y niñas [1].

Colombia ha experimentado una creciente violencia, militarización y violación de los derechos humanos. La construcción cultural de las identidades ha estado afectada por una amplia violencia en el campo de lo doméstico, la cual se superpone con otras expresiones políticas y sociales de violencia [2]

El machine Learning es una rama en evolución de los algoritmos computacionales que están diseñados para emular la inteligencia humana aprendiendo del entorno circundante. Se les considera el caballo de batalla en la nueva era de los llamados big data. Las técnicas basadas en el aprendizaje automático se han aplicado con éxito en diversos campos que van desde el reconocimiento de patrones, la visión por computadora, la ingeniería de naves espaciales, las finanzas, el entretenimiento, la educación, seguridad ciudadana y la biología computacional hasta aplicaciones biomédicas y médicas, entre muchas otras. [3] [4] [5]

Además, la clasificación de datos es una de estas tareas, la cual consiste en organizar, en diferentes clases, datos escritos en lenguaje natural. Entre las diferentes maneras de abordar este problema, aquí nos limitaremos a tratar con técnicas de aprendizaje automático, es decir, procesos que construyen automáticamente clasificadores. Estos clasificadores aprenden las características de cada clase gracias a un entrenamiento previo sobre un conjunto de datos preclasificados. También daremos cabida en nuestro estudio a las diferentes técnicas de preprocesamiento de los datos, las cuales deben ser realizadas con cuidado con el fin de resolver eficazmente cualquier problema de clasificación propuesto.

II. ESTADO DEL ARTE

La violencia intrafamiliar (en adelante vif), constituye una problemática compleja y vigente en la sociedad colombiana. Según el último informe del Instituto de Medicina Legal y Ciencias Forenses, en el año 2013 se presentaron 68.230 casos de vif, de los cuales el 77,58% fue contra mujeres y 22,42% contra hombres. El primer puesto dentro la vif lo ocupa la violencia de pareja seguida de la violencia entre otros familiares; en tercer lugar, la violencia contra niños, niñas y adolescentes; y en último lugar la violencia hacia personas mayores. Las mujeres son las víctimas más frecuentes, salvo en la violencia contra personas mayores donde los hombres ocupan el primer lugar. Algunos de los estudios analizados, demuestran la utilidad de Machine Learning en el análisis de la violencia presente en Colombia especialmente en el ámbito de los homicidios, encontrando así información relevante que pueda generar nuevos conocimientos que permitan crear nuevos objetivos y estrategias para la disminución de estos [6].

Uno de los métodos utilizados para el tratamiento de los datos fue Simple k-means y el método de clasificación de Random Forest, esto debido a que son los métodos que más se acoplan a nuestra investigación y para poner en práctica estos métodos, utilizamos WEKA una herramienta utilizada para el aprendizaje automático y minería de datos.

III. MACHINE LEARNING

Es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos. La máquina que realmente aprende es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros. Automáticamente, también en este contexto, implica que estos sistemas se mejoran de forma autónoma con el tiempo, sin intervención humana.

Dentro del Machine Learning encontramos múltiples herramientas, entre estas tenemos Weka, Pytorch, Rapidminer, Knime, Acord.net, scikit-Learn. Dentro de estas herramientas podemos encontrar las distintas técnicas de clasificación y regresión de Machine Learning. Estas son:

A. Simplekmeans

Uno de los métodos más utilizados para particionar un conjunto de datos en grupos de patrones. Sin embargo, la mayoría de los métodos de k-means, requiere cálculos de centroides para lograr la convergencia [7] [8].

B. Random Forest

son una combinación de predictores de árboles de manera que cada árbol depende de los valores de un vector aleatorio muestreado de forma independiente y con la misma distribución para todos los árboles del bosque. El error de generalización para bosques converge a.s. hasta un límite a medida que aumenta el número de árboles en el bosque [9]. [10]

C. Clustering

se enmarca en el aprendizaje no supervisado; es decir, que para esta técnica solo disponemos de un conjunto de datos de entrada, sobre los que debemos obtener información sobre la estructura del dominio de salida, que es una información de la cual no se dispone [11] [12].

D. Árbol de decisión

Un árbol de decisión es una forma gráfica y analítica de representar todos los eventos (sucesos) que pueden surgir a partir de una decisión asumida en cierto momento. Nos ayudan a tomar la decisión más "acertada", desde un punto de vista probabilístico, ante un abanico de posibles decisiones. Estos árboles permiten examinar los resultados y determinar visualmente cómo fluye el modelo. Los resultados visuales ayudan a buscar subgrupos específicos y relaciones que tal vez no encontraríamos con estadísticos más tradicionales. Los árboles de decisión son una técnica estadística para la segmentación, la estratificación, la predicción, la reducción de datos y el filtrado de variables, la identificación de interacciones, la fusión de categorías y la discretización de variables continuas [13] [14].

E. WEKA

Weka es un acrónimo de Waikato Environment for Knowledge Analysis, es un entorno para experimentación de análisis de datos que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario [11].

IV. METODOLOGÍA DE LA INVESTIGACIÓN

Fase 1: identificar las técnicas de Machine Learning usadas en otros estudios para caracterizar fenómenos de violencia y elegir algunas de ellas.

Dentro de esta fase encontramos las técnicas utilizadas en otros estudios para el análisis de datos, de estos estudios obtuvimos que las técnicas más factibles eran Random Forest, Simplekmeans y clustering.

Fase 2: Obtener y caracterizar un dataset de violencia en Colombia y/o en el departamento del Atlántico, a partir del cual se realizará el estudio.

En esta fase obtuvimos el dataset, este fue extraído de la página de datos abiertos de Colombia [15, 16, 17, 18]. Este dataset contaba con 458,914 registros; posteriormente se le hizo un filtrado al departamento del Atlántico dejando como resultado 16,825 registros. A este dataset se le aplicaron algunas técnicas de preprocesamiento de datos, como lo son:

1) Se agregaron 4 columnas una indicando el día de la semana y la fecha se separó en las 3 columnas restantes en día, mes, año respectivamente.

2) Se le colocó un filtro a cada columna para que sea más fácil la búsqueda.

3) Se cambiaron las vocales con tilde, por vocales sin tilde para que no arrojara error al momento de ingresar el archivo a Weka.

4) Se cambiaron las letras “ñ” por “n” para que no arrojara error al momento de ingresar el archivo a Weka.

También se realizaron algunas observaciones.

1) Dentro de la columna “armas medios” existen 2 valores llamados “no reporta” y “no reportado”, se asume que hacen referencia a lo mismo. Se encuentran 76.340 registros con estos valores.

2) En la columna “día semana”, que hace referencia al día de la semana en que ocurrió el hecho, se aprecian valores de 1 a 7, que hacen referencia de domingo a sábado respectivamente.

Fase 3: Aplicar las técnicas elegidas al dataset preparado y Analizar los resultados obtenidos e interpretarlos. Después de hacer el filtrado y eliminar columnas poco relevantes, estas fueron las columnas que quedaron junto con la distribución de los clústeres (Luego de haber ejecutado clustering con 4 clústeres, se decidió agregarlo como atributo clasificador):

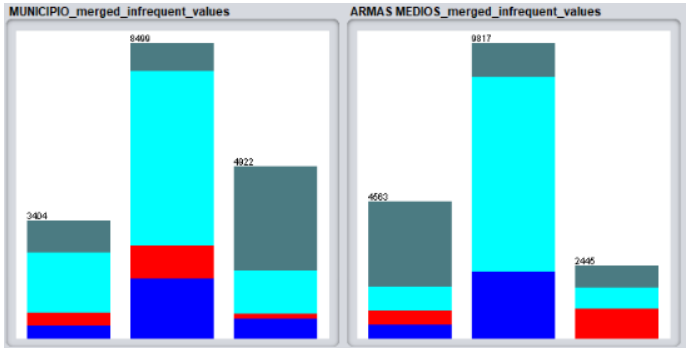

FIGURA 1. APLICACIÓN DE CLÚSTER A MUNICIPIOS Y ARMAS MEDIOS

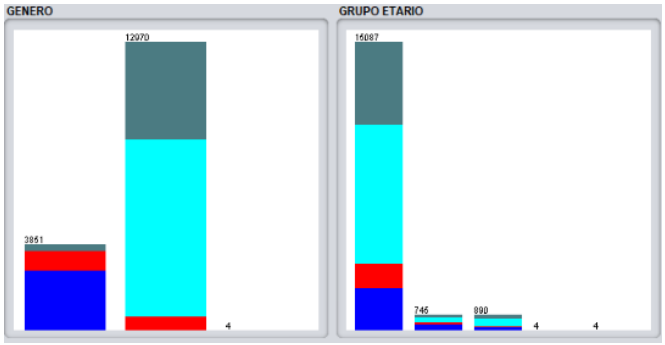

FIGURA 2. APLICACIÓN DE CLÚSTER A GÉNERO Y GRUPO ETARIO

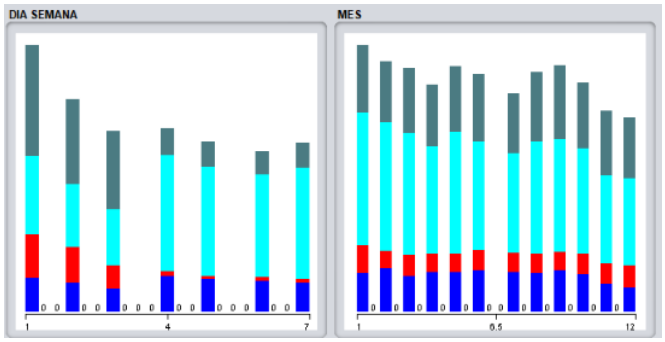

FIGURA 3. APLICACIÓN DE CLÚSTER A DÍA DE SEMANA Y MES

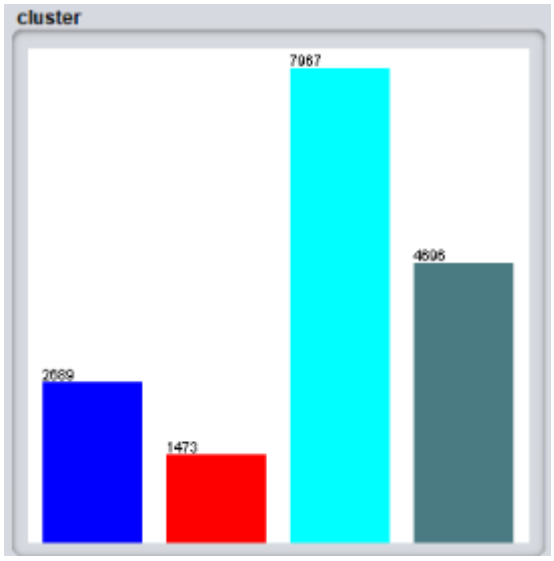

FIGURA 4. DISTRIBUCIÓN DE LOS CLÚSTERES

Dentro de esta fase, se le aplicó la técnica de clustering, árbol de decisión J48 y Random Forest con Simple k-means y obtuvimos los siguientes resultados

V. RESULTADOS

Primero le aplicamos las tecnicas anteriormente mencionadas a lso datos a nivel nacional y estos fueron los resultados.

Resultados a nivel nacional con SimpleKmeans:

Final cluster centroids:

Attribute	Full Data (321239.0)	Cluster# 0 (134488.0)	1 (186751.0)
DEPARTAMENTO	CUNDINAMARCA	CUNDINAMARCA	ANTIOQUIA
MUNICIPIO	BOGOTA D.C. (CT)	BOGOTA D.C. (CT)	MEDELLIN (CT)
ARMAS MEDIOS	CONTUNDENTES	SIN EMPLEO DE ARMAS	CONTUNDENTES
FECHA HECHO	1/01/2020	1/03/2020	1/01/2020
GENERO	FEMENINO	FEMENINO	FEMENINO
GRUPO ETARIO	ADULTOS	ADULTOS	ADULTOS
CANTIDAD	1.6803	1.9866	1.4597
DIA SEMANA	3.7816	3.8125	3.7593
DIA	15.3086	15.3302	15.293
MES	6.2347	6.4235	6.0987
ANO	2016.6681	2017.9051	2015.7773

FIGURA 5. RESULTADOS CON 2 CLÚSTERES A NIVEL NACIONAL

Se puede observar dentro de esta prueba que colocó los centroides en Antioquia y en Cundinamarca, se asume que fueron escogidos ya que cuentan con la mayor cantidad de casos.

	Cluster#				
Full Data	0	1	2	3	4
(321239.0)	(77451.0)	(61541.0)	(64615.0)	(65757.0)	(51875.0)
CUNDINAMARCA	CUNDINAMARCA	CUNDINAMARCA	BOYACA	VALLE	BOLIVAR
OGOTA D.C. (CT)	BOGOTA D.C. (CT)	BOGOTA D.C. (CT)	TUNJA (CT)	CALI (CT)	CARTAGENA (CT)
CONTUNDENTES	CONTUNDENTES	CONTUNDENTES	CONTUNDENTES	SIN EMPLEO DE ARMAS	CONTUNDENTES
1/01/2020	29/11/2020	31/01/2020	1/10/2020	1/01/2021	1/01/2020
FEMENINO	FEMENINO	MASCULINO	FEMENINO	FEMENINO	FEMENINO
ADULTOS	ADULTOS	ADULTOS	ADULTOS	ADULTOS	ADULTOS
1.6803	2.3146	1.457	1.3536	1.6487	1.445
3.7816	3.1851	3.6082	5.5723	4.0198	2.3454
15.3086	15.5949	15.3039	15.3402	15.3036	14.8535
6.2347	7.6543	6.2953	6.1517	6.1288	4.281
2016.6681	2017.8792	2016.6514	2015.7775	2017.0713	2015.4782

FIGURA 6. RESULTADOS CON 5 CLÚSTERES A NIVEL NACIONAL

Al haber colocado una mayor cantidad de clústeres, distribuye los centroides de una mejor manera, ubicándolos en los departamentos de Cundinamarca, Boyacá, Valle y Bolívar.

Posterior a estas pruebas a nivel nacional, separamos al departamento del Atlántico para hacerle las pruebas correspondientes y junto a esto, se omitieron las columnas de fecha, año, departamento y día, los resultados fueron los siguientes.

	Cluster#		
Full Data	0	1	2
(11777.0)	(3127.0)	(2441.0)	(6209.0)
BARRANQUILLA (CT) CONTUNDENTES SIN EMPLEO FEMENINO ADULTOS	SOLEDAD DE ARMAS FEMENINO ADULTOS	BARRANQUILLA (CT) CONTUNDENTES MASCULINO ADULTOS	BARRANQUILLA (CT) CONTUNDENTES FEMENINO ADULTOS
1.5438	1.5574	1.1512	1.6913
3.6913	2.7077	3.7358	4.1691
6.2352	6.3438	6.2966	6.1564

FIGURA 7 RESULTADO CON 3 CLÚSTERES ATLÁNTICO

Se puede observar que los centroides, los coloca en Barranquilla y Soledad junto con esto, se puede apreciar que el género más afectado es el femenino

Resultados con 4 clústeres

Final cluster centroids:

Attribute	Cluster# 0 (2733.0)	1 (1360.0)	2 (7888.0)	3 (4844.0)
MUNICIPIO_merged_infrequent_values	BARRANQUILLA (CT)	BARRANQUILLA (CT)	BARRANQUILLA (CT)	SOLEDAD
ARMAS MEDIOS_merged_infrequent_values	CONTUNDENTES	SIN EMPLEO DE ARMAS	CONTUNDENTES	-147914876
GENERO	MASCULINO	MASCULINO	FEMENINO	FEMENINO
GRUPO ETARIO	ADULTOS	ADULTOS	ADULTOS	ADULTOS
DIA SEMANA	3.9718	2.3103	4.3489	2.8697
MES	6.262	5.8676	5.9433	6.9238

FIGURA 8 RESULTADO CON 4 CLÚSTERES

En este resultado se encuentran los centroides de una manera mejor distribuida, y se observa que los dos primeros clústeres son de género masculino y los dos últimos de género femenino.

Seguidamente, se le aplica Árbol de Decisión dando como resultado la siguiente matriz de confusión

```
=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1149    0    0    0 |    a = cluster1
    0  812    0    0 |    b = cluster2
    0    0 2511    1 |    c = cluster3
    0    0    0 1247 |    d = cluster4
```

FIGURA 9 MATRIZ DE CONFUSIÓN DE ÁRBOL DE DECISIÓN MÉTODO MJ48

En el resultado anterior se puede observar una tupla que quedó mal clasificada, que se encontraba en el clúster 3 y la clasificó en el clúster 4

Por último, se le aplicó Random Forest y a diferencia de Árbol de Decisión aquí se encuentran 2 tuplas mal clasificadas dando como resultado la siguiente matriz de confusión:

```
=== Confusion Matrix ===

    a    b    c    d   <-- classified as
  923    0    0    0 |    a = cluster1
    0  493    0    1 |    b = cluster2
    0    0 2714    1 |    c = cluster3
    0    0    0 1588 |    d = cluster4
```

FIGURA 10 MATRIZ DE CONFUSIÓN RANDOM FOREST

Se puede observar que se encuentran 2 tuplas que eran de los clústeres b y c respectivamente y se clasificaron ambas dentro del clúster d. Esto quiere decir que el Random Forest es un poco más sensible en ciertos casos dándonos un 99.96% de las tuplas clasificadas correctamente.

VI. CONCLUSIONES

Después de haber hecho las pruebas vistas anteriormente se pueden llegar a las siguientes conclusiones.

Dado que en el archivo original la gran mayoria de casos reportados en el departamento del Atlántico, se dan en Barranquilla y Soledad, siendo los otros municipios con una cantidad considerablemente menor, todos los cluster se van a centar en los dos municipios anteriormente mencionados, por lo tanto, los hechos presentados en los demás municipios, no serán contundentemente clasificados, por el contrario quedarán vagamente clasificados, lo cual representa una desventaja bastante grande.

Resulta en evidencia que la mayor parte de los casos de violencia intrafamiliar se presentan contra mujeres con armas contundentes, los datosde fecha, no se pueden considerar demasiado relevantes ya que se encuentran distribuidos de una manera más uniforme.

Las técnicas de machine learning nos pueden brindar una visión general de los fenómenos de violencia en cuanto a quién va dirigida la mayor cantidad de hechos, qué tipo de arma es mayormente utilizada y en qué sitios se presentan con mayor frecuencia, sin embargo no será tan contundente al momento de hacer predicción ya que los modelos construidos se centraron en los sitios donde el archivo historico presentaba mayor cantidad de casos.

VII. REFERENCIAS

[1] J. Pineda y L. Otero, «Género, violencia intrafamiliar e intervención pública en Colombia,» *Revista de estudios sociales,* nº 17, pp. 19-31, 2004.

[2] C. Caicedo, «LUCHA CONTRA LA VIOLENCIA INTRAFAMILIAR : PERSPECTIVAS DESDE LA EXPERIENCIA COLOMBIANA,» Diciembre 2005. [En línea]. Available: http://www.americalatinagenera.org/documentos/roster/ros_32_Lucha%20contra%20la%20violencia%20intrafamiliar.pdf. [Último acceso: Mayo 2021].

[3] E. N. I. y M. M.J., «What Is Machine Learning?,» de *Machine Learning in Radiation Oncology*, Springer International Publishing Switzerland, 2015, pp. 3-11.

[4] D. Heredia, Y. Amaya y E. Barrientos, «Student dropout predictive model using data mining techniques,» *IEEE Latin America Transactions,* vol. 13, nº 9, pp. 3127-3134, 2015.

[5] D. Heredia, J. Castillo, P. Sanmartin y V. Quintero, «Aplicación de técnicas de minería de datos sobre datos georreferenciados para obtener un modelo predictivo: Caso de Estudio Hurtos en la Ciudad de Barranquilla,» de *Datos, Información, Tendencias, tres miradas sobre un contexto cambiante*, Barranquilla, Ediciones Universidad Simón Bolívar, 2019, pp. 77-91.

[6] E. Fernandez y Y. Gomez, «Metodología para el análisis de la violencia en el departamento de Bolívar mediante técnicas de machine learning,» 2018. [En línea]. Available: http://repositorio.utb.edu.co/handle/20.500.12585/1118. [Último acceso: Febrero 2021].

[7] M. Hung, J. Wu, J. Chang y Y. D, «An Efficient k-Means Clustering Algorithm Using Simple Partitioning,» *JOURNAL OF INFORMATION SCIENCE AND ENGINEERING,* nº 21, pp. 1157-1177, 2005.

[8] E. Kulkarni y R. Kulkarni, «WEKA Powerful Tool in Data Mining,» de *National Seminar on Recent Trends in Data Mining*, 2016.

[9] L. Breiman, «Random Forests,» de *Machine Learning*, Berkley, Springer, 2001, p. 5–32.

[10] Z. Masetic y A. Subasi, «Congestive heart failure detection using random forest classifier,» *Computer Methods and Programs in Biomedicine,* vol. 130, pp. 54-64, 2016

[11] C. Corso, «Aplicación de algoritmos de clasificación supervisada usando Weka.,» 2009. [En línea]. Available: https://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/CNIT_2009_Aplicacion_Algoritmos_Weka.pdf. [Último acceso: Mayo 2021].

[12] S. Kapil y M. Chawla, «Performance evaluation of K-means clustering algorithm with various distance metrics,» de *IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, 2016.

[13] R. M. V. R. Berlanga V., 1 Enero 2013. [En línea]. Available: http://hdl.handle.net/2445/43762. [Último acceso: Mayo 2021].

[14] P. Kapoor y R. Rani, «Efficient Decision Tree Algorithm Using J48 and Reduced Error Pruning,» *International Journal of Engineering Research and General Science,* vol. 3, nº 3, pp. 1613-1621, 2015.

[15] Policía Nacional de Colombia, «Datos abiertos,» Febrero 2021. [En línea]. Available: https://www.datos.gov.co/Seguridad-y-Defensa/Reporte-Delito-Violencia-Intrafamiliar-Polic-a-Nac/vuyt-mqpw. [Último acceso: Febrero 2021].

[16] C.A. Gutiérrez, R. Almeida., y W. Romero, “Diseño de un modelo de migración acloud computing para entidades públicas de salud”, Investigación e Innovación en Ingenierías, vol. 6, n°. 1, pp. 10 - 26., 2018. DOI: https://doi.org/10.17081/invinno.6.1.2772

[17] Y. L. Coronel Montaguht, J. P. Dolcey, y Morales González K., «Economi?a naranja: ¿cua?nto se puede exprimir?», *ADGNOSIS*, vol. 6, n.º 6, pp. 195–199, dic. 2017.

[18] H. G. Hernandez Palma, J. Solórzano Movilla, y J. Jinete Torres, "La Teoría de restricciones para los procesos de gestión y control en las IPS del Caribe Colombiano", Investigación e Innovación en Ingenierías, vol. 8, n.º 1, pp. 54–68, 2020. DOI: https://doi.org/10.17081/invinno.8.1.3624

[19] E. J. De la Hoz Domínguez, T. J. . Fontalvo Herrera, y A. A. Mendoza Mendoza, "Aprendizaje automático y PYMES: Oportunidades para el mejoramiento del proceso de toma de decisiones", Investigación e Innovación en Ingenierías, vol. 8, n.º 1, pp. 21–36, 2020. DOI: https://doi.org/10.17081/invinno.8.1.3506

[20] W. A. Ceballos Betancur, «Tendencias de la responsabilidad social universitaria (RSU) de las instituciones de educacio?n superior (IES) en la ciudad de Medelli?n - Colombia», ADGNOSIS, vol. 6, n.º 6, pp. 85–100, dic. 2017.