EXTRACCIÓN DE CONOCIMIENTO A PARTIR DEL ANÁLISIS DE LOS DATOS EN EL PERÍODO 2013-2017 DEL MINISTERIO DE SALUD PÚBLICA EN ECUADOR

Oscar J. Alejo Machado[1*], Tatiana Tapia Bastidas[**], Maikel Yelandi Leyva Vázquez[***]
[*]Instituto Superior Tecnológico Bolivariano de Tecnología, Guayaquil, Guayas, Ecuador.
[**]Instituto Superior Tecnológico Bolivariano de Tecnología, Guayaquil, Guayas, Ecuador.
[***]Universidad Politécnica Salesiana, Guayaquil, Guayas, Ecuador.

ABSTRACT
The databases of the Ministry of Public Health of Ecuador in the 2013-2017 period contain valuable information that can be used to determine the strengths, weaknesses, potential problems, among others, that affect the public health of the country. This knowledge can serve to draw better public health policies. This paper aims to propose a methodology that allows us to extract knowledge from these databases and at the same time to obtain association rules based on the combination of algorithms such as FP-growth and k-means. In summary, the methodology consists of the following steps: first, the dataset is stored in 5 files in the SPSS (Statistical Package for the Social Sciences) format, and then the disease-related attributes are grouped and encoded, according to the code ICD-10, for this purpose it is proposed to apply the WEKA software. Finally, the FP-Growth algorithm is used to extract association rules from frequent items with the support of RAPIDMINER, which has the advantage of allowing us the use of WEKA algorithms. The methodology is illustrated with an example that shows how to use it and its usefulness to extract association rules in real-life situations from medical databases. With these representations of the information, morbidity and incidence behavior analysis of the registered groups and diseases can be made.

KEYWORDS: Data mining, Artificial Intelligence in medicine, unsupervised learning, associating rule, clustering.

MSC:68T05. 68T10, 68T30, 68T37, 97M60

RESUMEN
Las bases de datos del Ministerio de Salud Pública de Ecuador en el período 2013-2017 contienen una valiosa información que puede ser utilizada para determinar los puntos fuertes, lo débiles, los problemas potenciales, entre otros, que afectan la salud pública del país. Este conocimiento puede servir para trazar mejores políticas de salud pública. Este artículo tiene como objetivo proponer una metodología que permita extraer conocimientos de estas bases de datos y a la vez obtener reglas de asociación basadas en la combinación de algoritmos tales como *FP-growth* y *k-means*. De forma resumida la metodología consiste en lo siguiente: primero, los datos se almacenan en 5 grandes ficheros en formato SPSS (*Statistical Package for the Social Sciences*), luego, se agrupan y codifican los atributos relacionados con las enfermedades, según el código CIE-10, para ello se propone aplicar el software WEKA. Finalmente se utiliza el algoritmo *FP-Growth* para extraer reglas de asociación a partir de *itemsets* frecuentes con ayuda de RAPIDMINER, que tiene la ventaja de permitir utilizar los algoritmos de WEKA. La metodología se ilustra con un ejemplo que muestra la manera de utilizar la metodología y su utilidad para extraer reglas de asociación en situaciones de la vida real en bases de datos médicas. Con estas representaciones de la información se podrán hacer análisis de comportamiento de morbilidad e incidencia de los grupos y enfermedades registradas.

PALABRAS CLAVES: Minería de datos, Inteligencia Artificial en la medicina, aprendizaje no supervisado, regla de asociación, agrupamiento.

1. INTRODUCCIÓN

La minería de datos dentro de las ciencias de la computación, tiene como objetivo descubrir patrones en grandes volúmenes de datos, véase [7]. O sea, cuando se cuenta con una base de datos de gran volumen, para que esta sea útil es necesario descubrir las regularidades y los modelos que caracterizan estos datos, y esta es la razón de aplicar técnicas de la minería de datos. Algunas técnicas de la minería de datos provienen de la Inteligencia Artificial, [18]. En especial, esta investigación propone una metodología para extraer conocimientos almacenados en las bases de datos del Ministerio de Salud Pública de Ecuador durante el período de tiempo 2013-2017.
La Inteligencia Artificial por su parte es una disciplina multidisciplinaria, que depende de la Filosofía, las Ciencias de la Computación, la Lógica, entre otras, y tiene como objetivo la creación de software o entes artificiales que emulen el comportamiento de la inteligencia humana, [18].

[1]Email:oalejo@bolivariano.edu.ec

El uso de la inteligencia artificial no es nuevo dentro de la medicina y ha sido aprovechada por esta de múltiples maneras. Algunos ejemplos de esto son el uso de Sistemas Expertos, sobre todo basados en la Lógica Difusa que se aplican en el diagnóstico y en los sistemas biológicos. Esto se debe a que tales modelos permiten la modelación de la imprecisión y la incertidumbre que forman parte de estas ramas de la medicina.
Otra aproximación es el uso de la computación evolutiva. Algunos algoritmos como el algoritmo genético, imitan el comportamiento evolutivo de los seres vivos. Específicamente, este ha sido aplicado para predecir la perspectiva en pacientes con estadios críticos de una enfermedad, la segmentación de imágenes de tejidos para determinar las áreas de tumor, para medir la eficacia de estrategias de tratamiento, entre otras, [5][20].
Es necesario destacar el célebre sistema experto MYCIN desarrollado por Edward Shortliffe a principios de la década de 1970, que fuera programado en el lenguaje Lisp y tal que realizaba diagnósticos de enfermedades de la sangre, [19].
Los centros de salud de todo el mundo almacenan grandes bases de datos que contienen los datos recogidos sobre cada paciente frecuentemente durante largos períodos de tiempo. Esta información se almacena en forma de datos numéricos, textos e imágenes. La importancia de aplicar las técnicas de minería de datos en estas bases de datos ayuda a la extracción de conocimiento imprescindible que es de una utilidad fundamental, teniendo en cuenta que se aplica en un campo tan sensible como es la salud humana. Algunas aproximaciones a la minería de datos aplicada a la medicina se pueden encontrar en [9][11][21].
El objetivo de este artículo es el diseño de una metodología que permita extraer conocimiento de las bases de datos del Ministerio de Salud Pública de Ecuador, lo que permitirá determinar las tendencias que sufre la salud pública en el país, y permitirá trazar políticas que optimicen los recursos humanos y financieros para mejorar la calidad de los servicios de salud. Además de obtener reglas de asociación a partir de las bases de datos guardadas en el período 2013-2017. Para ello se utilizan software que incluyen herramientas o paquetes que permiten el cálculo con minería de datos, como Weka ([3][4]), Rapidminer ([10]) entre otros.
El presente artículo continúa con una sección de materiales y métodos donde se hace un resumen de la teoría sobre los algoritmos que se tratan en el artículo. La sección 3 describe la metodología que se propone y la sección 4 se dedica a dar las conclusiones.

2. MATERIALES Y MÉTODOS

Esta sección contiene las bases teóricas para comprender los resultados que se proponen en este artículo. Sea $I = \{a_1, a_2,\ldots,a_m\}$ un conjunto de ítems, y una *base de datos transaccional* $BD = <T_1, T_2,\ldots, T_n>$, donde T_i ($i \in \{1, 2, \ldots, n\}$) es una transacción que contiene un conjunto de ítems en I. El *soporte* (o frecuencia de ocurrencia) de un *patrón* A, donde A es un conjunto de ítems, es el número de transacciones de BD que contienen A. Un patrón A es *frecuente* si el soporte de A no es menor que un *umbral de soporte mínimo* predefinido, denotado por ξ, véase [2][14].
Una estructura de datos compacta se puede diseñar basada en las observaciones siguientes:

1. Se realiza un escaneo de las transacciones de BD para identificar el conjunto de ítems frecuentes (El *conteo de frecuencia* se obtiene como un subproducto).
2. Si el conjunto de ítems frecuentes de cada transacción se puede almacenar en alguna estructura compacta, se puede evitar el escaneo repetido de la BD original.
3. Si múltiples transacciones comparten un conjunto de ítems frecuentes, es posible fusionar los conjuntos compartidos mediante el número de ocurrencias registradas como *conteo*. Es fácil controlar si dos conjuntos son idénticos cuando los ítems frecuentes en todas las transacciones se listan siguiendo un orden fijo.
4. Si dos transacciones comparten el mismo prefijo, según cierto orden de ítems frecuentes, las partes compartidas se pueden fusionar usando una estructura de prefijo tan larga como es registrado el *conteo*.

De esta manera se obtienen estructuras de la forma $<(a_1:n_1), (a_2:n_2), \ldots, <(a_m:n_m)>$ donde $(a_i:n_i)$ denota el ítem con su conteo de frecuencia.
Definición 1. Un *árbol de patrón-frecuente* (o árbol PF) es un árbol con la estructura definida a continuación:

1- Contiene una raíz etiquetada como *nula*, un conjunto de *subárboles ítem-prefijos* como hijos de la raíz y una *tabla de encabezamiento de ítems frecuentes*.
2- Cada nodo en el subárbol ítem-prefijo contiene tres campos: un *ítem-nombre*, *conteo* y *nodo-conexión*, donde el *ítem-nombre* registra cual ítem se representa por este nodo, el *conteo* registra el número de transacciones representadas por la porción de pasos que alcanzan este nodo y el *nodo-*

conexión se conecta al próximo nodo que contiene el mismo ítem-nombre, o a *nulo* si este no existe.

3- Cada entrada en la tabla de encabezamiento de ítems frecuentes consiste en dos campos: de *ítem-nombre* y *cabeza de nodo-conexión* que es un puntero que apunta al primer nodo en el árbol PF que lleva ese ítem-nombre.

Dos algoritmos que se utilizan para crear reglas de asociación en BD a partir de árboles PF son el *Algoritmo a priori* ([1][12]) y el *FP-growth*.

El *algoritmo a priori* comienza buscando los conjuntos frecuentes unitarios, a cada uno de ellos se les añade un ítem adicional y se seleccionan entre los secundarios los frecuentes, a continuación se les añade a estos los ítems formando los terciarios entre los cuales se seleccionan los frecuentes, se repite cada iteración hasta no obtener ningún conjunto frecuente. Este algoritmo asume un orden lexicográfico entre los ítems.

FP-growth pre-procesa la BD como sigue: un escaneo inicial determina la frecuencia de los ítems,con soporte de conjuntos con un solo ítem. Todos los ítems infrecuentes se eliminan de la transacción. Aparte se tiene que los ítems en cada transacción se organizan, tal que estén en orden descendente con respecto a su frecuencia en la base de datos.

El algoritmo k-medias o en inglés *k-means* parte de un conjunto de observaciones $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \Re^d$. El objetivo es obtener una partición de las observaciones en k conjuntos disjuntos $C_1, C_2, \ldots, C_M$, por similaridad, véase [8][13][15]. Matemáticamente consiste en resolver el siguiente problema de optimización:

$$\min E(m_1, m_2, \ldots, m_M) = \sum_{i=1}^{N}\sum_{k=1}^{M} I_i \|x_i - m_k\|^2 \text{ donde } I_i = \begin{cases} 1, & \text{si } x_i \in C_k \\ 0, & \text{en otro caso} \end{cases}$$

, mientras que m_1, m_2,…, m_M son los centroides y $\|\cdot\|$ es una norma para medir distancia entre vectores. Para más detalle de estos algoritmos y estás teorías en general véase [2][13][15].

3. RESULTADOS

Esta sección se dedica al diseño de la metodología que se propone para la aplicación de Minería de Datos a las BDs perteneciente al Ministerio de Salud Pública de Ecuador.

3.1. Procesamiento de los datos con ayuda de RAPIDMINER

Los Datos fueron entregados en 5 grandes ficheros en formato (SPSS, *Statistical Package for the Social Sciences*), véase [10]. Una vez cargados en este programa estadístico informático, se pudo determinar cantidad de variables y registros por año. Se realizó, por cada año de manera específica, un análisis de los valores asignados a cada etiqueta, su codificación, tipología, significado y relación general con el conjunto de datos, véase Tabla 1.

Año	Variables	Registros	Reg. Diarios Aprox.
2013	72	3.430.611	9.398,93
2014	53	3.390.454	9.288,92
2015	55	3.722.221	10.197,87
2016	92	3.428.706	9.393,72
2017	78	3.638.168	9.967,58
		17.610.160	**9.649,40**

Tabla 1. Resumen de los datos almacenados en los registros del Ministerio de Salud Pública del Ecuador.

Para el procesamiento de los ficheros y la obtención de reglas de asociación con RAPIDMINER, los atributos relacionados con las enfermedades fueron agrupados y codificados según el CIE-10, acrónimo de la Clasificación Internacional de Enfermedades, 10.ª edición, véase [16]. A continuación se muestran algunos ejemplos de este código:

- (A00–B99) Ciertas enfermedades infecciosas y parasitarias
- (C00–D48) Tumores [neoplasias]
- (D50–D89) Enfermedades de la sangre y de los órganos hematopoyéticos, y ciertos trastornos que afectan el mecanismo de la inmunidad

Los conjuntos de datos originales para Weka (formato ARFF) fueron transformados y ajustados para su trabajo con RAPIDMINER, a partir de la siguiente estructura:

Para el procesamiento de cada uno de los conjuntos de datos se diseñó un proceso como se muestra en la

Figura 1, donde se utilizó el algoritmo *FP-Growth* para extraer reglas de asociación a partir de *itemsets* frecuentes.

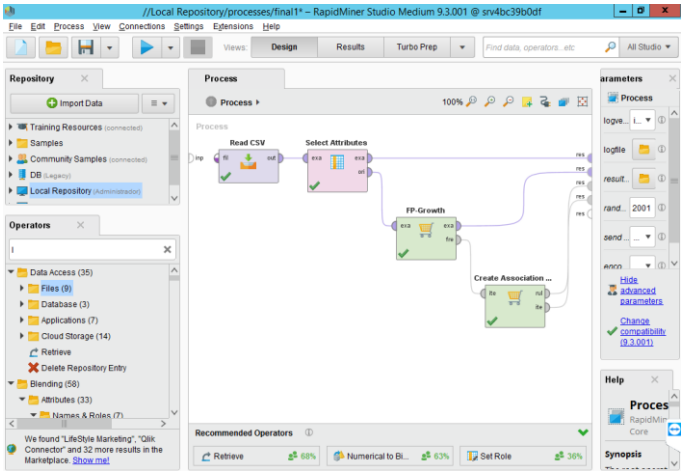

Figura 1. Extracción de reglas de asociación mediante *FP-Growth* con ayuda de RAPIDMINER.

En términos generales, el algoritmo emplea una estructura de árbol (*Frequent Pattern Tree*), véase [6], donde almacena toda la información de las transacciones. Esta estructura permite comprimir la información de una base de datos de transacciones hasta 200 veces, haciendo posible que pueda ser cargada en memoria RAM. Una vez que la base de datos ha sido comprimida en una estructura *FP-Tree*, se divide en varias bases de datos condicionales, cada una asociada con un patrón frecuente. Finalmente, cada partición se analiza de forma separada y se concatenan los resultados obtenidos. En la mayoría de casos, *FP-Growth* es más rápido que el algoritmo clásico Apriori.

La configuración de los parámetros de entrada del algoritmo se realizó de la forma que muestra la Figura 2.

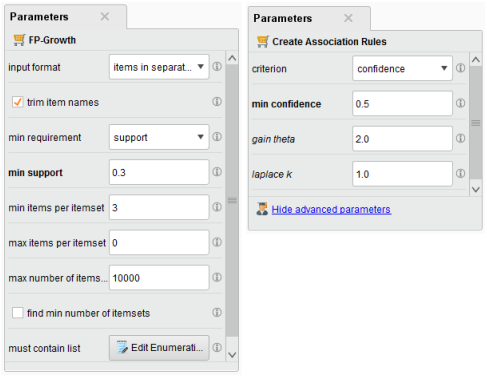

Figura 2. Configuración de los parámetros de entrada para el algoritmo *FP-Growth*.

Finalmente, se obtuvieron reglas de asociación en formato tabular, gráfico y en descripción lógica, como se muestra a continuación, comenzando por la Figura 3.

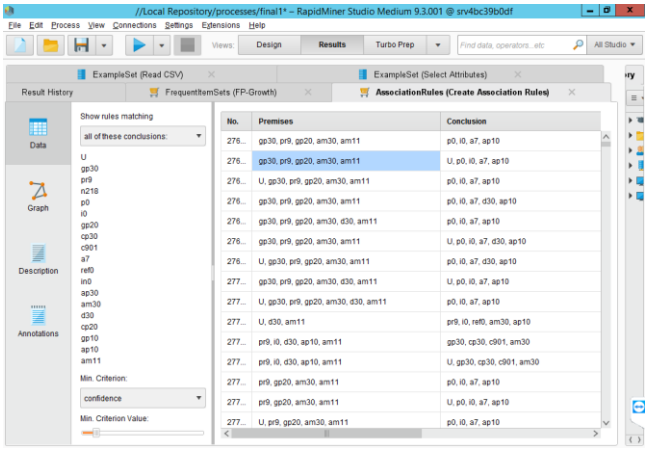

Figura 3. Reglas de asociación obtenidas en formato de descripción lógica.

Considerando que una regla define una relación entre dos conjuntos de elementos (*itemsets*) X e Y que no tienen elementos en común, X->Y significa que si se tiene X en una transacción, entonces se puede tener

Y en la misma transacción.
El formato de presentación de las reglas encontradas es:

Nro.* | *Premises* | *Conclusion* | *Support* | *Confidence* | *LaPlace* | *Gain* | *P-S* | *Lift* | *Conviction

Donde:

Nro.: orden numérico de la regla de asociación generada.
Premises: indica conjuntos de elementos considerados como premisa de la regla.
Conclusion: indica conjunto(s) de elemento(s) considerados como la conclusión de una regla.
Support: Probabilidad de encontrar elementos o conjuntos de elementos X e Y en una transacción. Se estima por el número de veces que ambos elementos o conjuntos de elementos se encuentran en todas las transacciones disponibles. Este valor se encuentra entre 0 y 1.
Confidence: Probabilidad de encontrar un elemento o conjunto de elementos Y en una transacción, sabiendo que el elemento o conjunto de elementos X está en la transacción. Se estima por la frecuencia correspondiente observada (número de veces que X e Y se encuentran en todas las transacciones, dividido por el número de veces que se encuentra X). Este valor se encuentra entre 0 y 1.
LaPlace: Cuando se selecciona esta opción, el Laplaciano se calcula utilizando el laplaciano de parámetro k.
Gain: Cuando se selecciona esta opción, se calcula la ganancia utilizando la ganancia de parámetro theta.
P-S: Cuando se selecciona esta opción los criterios ps se utilizan para la selección de reglas.
Lift: La importancia de una regla, que es simétrica (importancia (X-> Y) = importancia (Y-> X)), es el soporte (*support*) del conjunto de elementos que agrupa X e Y, dividido por el soporte de X y el soporte de Y. Este valor puede ser cualquier número real positivo. Una *lift* mayor que 1 indica un efecto positivo de X en Y (o Y en X) y por lo tanto la significación de la regla. Un valor de 1 significa que no hay efecto, y es como si los elementos o conjuntos de elementos fueran independientes. Una *lift* menor que 1, significa que hay un efecto negativo de X en Y o viceversa, como si fueran excluyentes entre sí.
Conviction: *Conviction* es dependiente de la dirección de la regla, o sea conv(X implica Y) no es lo mismo que conv(Y implica X). *Conviction* está más o menos inspirada en la definición lógica de implicación y se propone medir el grado de implicación de una regla. *Conviction* se define como conv(X implica Y) =(1 - supp(Y))/(1 - conf(X implica Y))

Ejemplo 1. Ejemplo de ***Premisa*** y ***Conclusión***:
Premises Conclusion
p0, gp30, i0, am30, ap20, e6, c901, 90112 => n218, gp20, pr9, d30, cie3NA, U
A partir del procesamiento realizado con cada conjunto de datos se obtuvieron los siguientes resultados:
Total, de reglas encontradas en el año 2013: 678184.
Por el volumen de información, y para una mejor organización, tales reglas se almacenaron en formato Excel, en 7 hojas de cálculo o particiones (*FOLD*), con el siguiente tamaño:

	Hoja 1 FOLD 1	Hoja 2 FOLD 2	Hoja 3 FOLD 3	Hoja 4 FOLD 4	Hoja 5 FOLD 5	Hoja 6 FOLD 6	Hoja 7 FOLD 7
# Reglas	1000000	1000000	1000000	1000000	1000000	1000000	781849

Total de reglas encontradas en el año 2014: 5965591.

	Hoja 1 FOLD 1	Hoja 2 FOLD 2	Hoja 3 FOLD 3	Hoja 4 FOLD 4	Hoja 5 FOLD 5	Hoja 6 FOLD 6
# Reglas	1000000	1000000	1000000	1000000	1000000	965592

Total, de reglas encontradas en el año 2015: 2203039.

	Hoja 1 FOLD 1	Hoja 2 FOLD 2	Hoja 3 FOLD 3
# Reglas	1000000	1000000	203040

Total, de reglas encontradas en el año 2016: 2012061.

	Hoja 1 FOLD 1	Hoja 2 FOLD 2	Hoja 3 FOLD 3
# Reglas	1000000	1000000	12062

Total de reglas encontradas en el año 2017: 2349092.

	Hoja 1 FOLD 1	Hoja 2 FOLD 2	Hoja 3 FOLD 3
# Reglas	1000000	1000000	349093

3.2. Procesamiento de los datos con ayuda de WEKA

Para un procesamiento efectivo de DATA se utilizó Weka (*Waikato Environment for Knowledge Analysis*) como plataforma de software para el aprendizaje automático y la minería de datos. Weka es software libre distribuido bajo la licencia GNU-GPL, escrito en Java y desarrollado en la Universidad de Waikato en Nueva Zelanda, véase Figura 4.

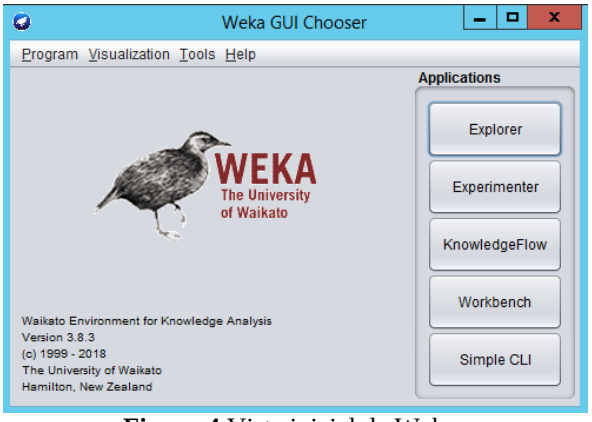

Figura 4. Vista inicial de Weka.

Se trabajó en el módulo "*Associate*" dentro de *Explorer*, cargando y haciendo uso del método *HotSpot*, el cual aprende un conjunto de reglas (que se muestran en una estructura similar a un árbol) que maximizan / minimizan una variable / valor objetivo de interés.

La configuración de parámetros se realizó de la manera mostrada en la Figura 5.

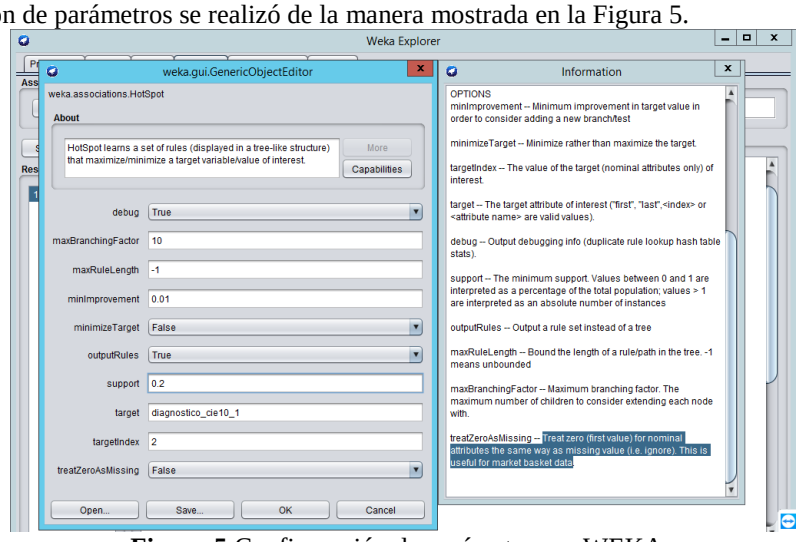

Figura 5. Configuración de parámetros en WEKA.

3.3. Agrupamiento

El agrupamiento o *clustering* juega un papel muy importante en aplicaciones de minería de datos, tales como exploración de datos científicos, recuperación de la información y minería de texto, aplicaciones sobre bases de datos espaciales (tales como GIS o datos procedentes de astronomía), aplicaciones Web, marketing, diagnóstico médico, análisis de ADN en biología computacional y muchas otras.

En el contexto de este proyecto, se ha utilizado el *Clustering*, como técnica de minería de datos, para dividir los vectores de datos (instancias generadas por el Software del *registro diario automatizado de atenciones y consultas* o RDACAA, [17]) en grupos de objetos similares, con lo cual se logra una representación más simple e interpretable del volumen de información bajo estudio.

Por la estructura y semántica de los datos, no existe en esta investigación una clase o atributo objetivo en específico, lo cual justifica el uso de estos tipos de procedimientos de aprendizaje automático no

supervisado para extraer conocimiento de los datos.
Para el procesamiento de los ficheros y la obtención de agrupamientos (*clustering*) de casos representativos se utilizaron los archivos que aparecen en la Tabla 2.

Año	# Atributos Originales	# Atributos Seleccionados	# Instancias Originales	# Instancias Seleccionadas	Estructura del Fichero (.arff)
2013	72	30	3.430.611	1.866.773	Anexo 1.
2014	53	30	3.390.454	3.390.453	Anexo 2.
2015	55	35	3.722.221	3.722.221	Anexo 3.
2016	92	37	3.428.706	3.421.433	Anexo 4.
2017	78	44	3.638.168	3.636.305	Anexo 5.

Tabla 2. Resumen del agrupamiento de datos obtenidos.

En la transformación y estructuración de los datos, los atributos relacionados con las enfermedades fueron agrupados y codificados según el CIE-10.
Se trabajó en el módulo "*Cluster*", cargando y haciendo uso del método Simple *K-Means*.
Este algoritmo define el número de *clusters* que se desean obtener, así se convierte en un algoritmo voraz para particionar. Los pasos básicos para aplicar el algoritmo son muy simples. Primeramente se determina la cantidad de *clusters* en los que se quiere agrupar la información. Luego se asume de forma aleatoria los centros por cada *cluster*. Una vez encontrados los primeros centroides el algoritmo hará los tres pasos siguientes:

1. Determina las coordenadas del centroide.
2. Determina la distancia de cada objeto a los centroides.
3. Agrupa los objetos basados en la menor distancia.

En el procesamiento del archivo "seleccion2013.arff" se obtuvieron *clusters* con diferentes tamaños: K={2,3,4,5,6,7,8,9,10,20,50,100}. En la Figura 6 se presentan los centroides obtenidos para K=100. A esta memoria documental se adjuntan los ficheros de 2013 resultantes de cada procesamiento, donde se pueden visualizar y analizar los indicadores en relación a comportamiento de atributo versus *cluster*.

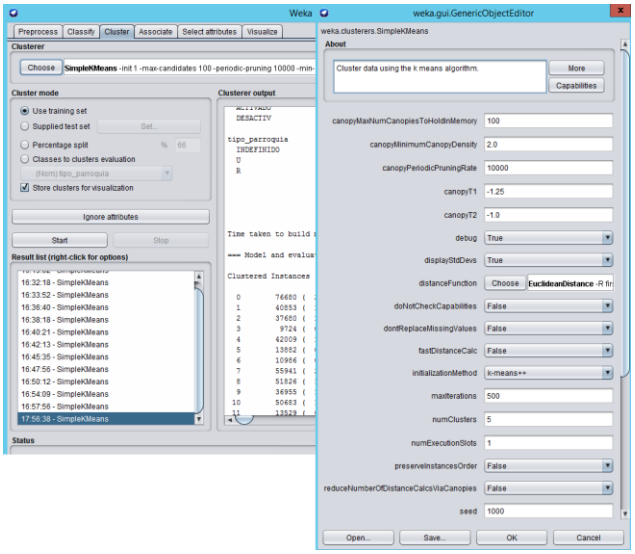

Figura 6. Imagen de la obtención de agrupamientos de diferentes tamaños en WEKA.

4. CONCLUSIONES

Este artículo se dedicó al diseño de una metodología que permite la obtención de reglas de asociación para extraer conocimiento útil de las bases de datos del Ministerio de Salud Pública de Ecuador en el período 2013-2017. Para ello se incluyeron algoritmos muy populares como *FP-growth* y *k-means* que se combinan mediante software de probada eficacia como RAPIDMINER y Weka. Se obtuvo un número considerable de reglas de asociación mediante esta metodología. Esta es una contribución para el trazado de políticas de salud públicas más eficaces y eficientes, tanto desde el punto de vista financiero, como humano.

RECEIVED: NOVEMBER, 2019.
REVISED: FEBRUARY, 2020.

REFERENCIAS

[1] BODON, F. (2005): A Trie-based APRIORI Implementation for Mining Frequent Item sequences. in: B. Goethals, S. Nijssen y M.J. Zaki, (Eds.), **First International Workshop on Open Source Data Mining Frequent Pattern Mining Implementations,** 56-65.

[2] BORGELT, C. (2005): An Implementation of the FP-growth Algorithm. in: B. Goethals, S. Nijssen, and M.J. Zaki, (Eds.), **First International Workshop on Open Source Data Mining Frequent Pattern Mining Implementations,** 1-5.

[3] BOUCKAERT, R.R., FRANK, E., HALL, M., KIRBY, R., REUTERMANN, P., SEEWALD, A. y SCUSE, D. (2014): **WEKA Manual for Version** 3-6-11, 2014.10

[4] HALL, M., FRANK, E. y WITTEN, I.H. (2011): Practical Data Mining. Tutorial 1: Introduction to the WEKA Explorer, Disponible en: http://www.cs.waikato.ac.nz/~ml/weka/book.html, Consultado: 13-12-2018.

[5] HAMET, P. y TREMBLAY, J.(2017): Artificial intelligence in medicine. **Metabolism**, 69, S36-S40.

[6] HAN, J., PEI, J., YIN, Y. y MAO, R. (2004): Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. **Data Mining and Knowledge Discovery**, 8, 53–87.

[7] HAND, D., MANNILA, H. y SMYTH , P. (2001):**Principles of Data Mining**, The MIT Press, Cambridge.

[8] JAIN, A.K. y DUBES, R.S. (1988):**Algorithms for Clustering Data**, Prentice Hall, Englewood, New Jersey.

[9] JOTHI, N. , RASHID, N.A.A. y HUSAIN, W.(2015): Data Mining in Healthcare – A Review. **Procedia Computer Science**, 72, 306 – 313.

[10] JUNGERMANN, F. (2009): Information Extraction with RapidMiner, GSCL Symposium'Sprachtechnologie und Humanities, 50-61.

[11] KOH, H.C. y TAN, G.(2011): Data Mining Applications in Healthcare. **Journal of Healthcare Information Management** , 19, 64-72.

[12] LA ROSA SEGURA, J.M. (2012): **M2-Apriori. Software para la generalización de reglas de asociación**, Facultad de Ciencias Técnicas Departamento de Informática, Universidad de Las Tunas, Las Tunas, Cuba.

[13] LANG, J.-S.R., SUN, C.-T. y MIZUTANI, E. (1997):**Neuro-Fuzzy and Soft Computing: A computational Approach to Learning and Machine Intelligence**, Prentice Hall, Upper Saddle River, New Jersey.

[14] LI, H., WANG, Y., ZHANG, D., ZHANG, M. y CHANG, E. (2008): PFP: Parallel FP-Growth for Query Recommendation, **ACM Conference on Recommender Systems**, 107-114.

[15] LIKAS, A., VLASSIS, N.y VERBEEK, J.J. (2003): The global k-means clustering algorithm. **Pattern Recognition**,36, 451-461.

[16] OMS (2010): CIE-10 en español, descarga o consulta, Disponible en: http://www.cie10.org/. Consultado el 12-12-2018.

[17] PÁEZ MEDINA, G.A. (2017): **Estrategias para el mejoramiento de la calidad del Registro Diario Automatizado de Atenciones y Consultas Ambulatorias (RDACAA) en la unidad operativa de salud Augusto Egas, Provincia de Santo Domingo de los Tsáchilas**, Facultad de Ciencias Médicas, Universidad Regional Autónoma de Los Andes, Ambato, Ecuador.

[18] PANNU, A.(2015): Artificial Intelligence and its Application in Different Areas. **International Journal of Engineering and Innovative Technology**, 4, 79-84.

[19] SHORTLIFFE, E.H. (1976): **Computer-Based Medical Consultations: MYCIN**, Elsevier, New York.

[20] TOPOL, E.J. (2019): High-performance medicine: the convergence of human and artificial intelligence. **Nature medicine**, 25, 44-56.

[21] YOO, I., ALAFAIREET, P., MARINOV, M., PENA-HERNANDEZ, K., GOPIDI, R., CHANG, J.F.y HUA, L. (2012): Data mining in healthcare and biomedicine: a survey of the literature. **Journal of Medical Systems**, 36, 2431-2448.