Clustering basado en el algoritmo K-means para la identificación de grupos de pacientes quirúrgicos

Clustering technique based on k- means algorithm for the identification of clusters of surgical patients

Javier Hernández Cáceres
Docente Facultad de Ingeniería Industrial
Universidad Santo Tomás, seccional Bucaramanga
Bucaramanga, Colombia
Javier.hernandez@ustabuca.edu.co

Resumen **- La caracterización de pacientes permite mejorar la calidad de atención hospitalaria y representa un cumplimiento normativo para la estrategia de gobierno electrónico del Estado colombiano. La presente investigación busca apoyar la Gestión Estratégica del Proceso de Cirugía en un hospital público de alta complejidad, a partir del agrupamiento de pacientes quirúrgicos cuyos diagnósticos estuvieron asociados a tumores malignos. *Metodología*: se definió conjunto de datos con atributos sociodemográficos y clínicos y se utilizó WEKA como herramienta de minería de datos para aplicar la técnica Clustering basada en el algoritmo K-means y distancia Euclídea. *Resultados*: descripción exploratoria de atributos y clústeres que identifican grupos de pacientes que fueron sometidos a intervenciones quirúrgicas durante los períodos 2007 al 2015. Los clústeres obtenidos permitieron evidenciar la presencia de diagnósticos asociados a cáncer, agrupando la población por: edad, grupo etario, régimen de salud, género, zona de residencia, estrato, etnia, estado civil, grupo CIE-10, grupo quirúrgico y tipo de atención. *Conclusiones*: La evaluación exhibe altos niveles de aceptación por parte de las partes interesadas. El resultado de agrupación de este estudio no es una solución universal a todos los datos del paciente, para generalizar el resultado, un mayor conjunto de datos debería ser adoptado en la investigación futura. Los principales tipos de tumores malignos en hombres eran tumores del estómago, colon y la laringe. Entre las mujeres, las principales fueron los tumores de colon, estómago y cerebro. Aunque el grupo más afectado fue el de más de 60 años también se observó la presencia significativa de tumores a partir de 40 años.**

Palabras claves: Data Mining, Clustering, CRISP-DM, K-means, Minería de datos en Salud.

Abstract **- The characterization of patients allows improving the quality of hospital care and represents a regulatory compliance for Colombia's e-government strategy. This research aims at supporting the Strategic Management of Surgical Procedures in a high complexity public hospital by using a clustering technique of surgical patients whose diagnoses were associated with malignant tumors. *Methodology*: data with sociodemographic and clinical attributes was defined. Weka was used as a data mining tool to implement the clustering technique based on K-means algorithm and Euclidean distance. *Results*: An exploratory description of attributes and clusters to identify groups of patients who underwent surgical procedures from 2007 to 2015 was conducted. Clusters obtained allowed demonstrating the presence of groups of diagnoses associated with cancer respect to age range, health system, gender, place of residence, social stratum, ethnic group, marital status, ICD-10 group, surgical group and type of care. *Conclusions*: The evaluation exhibits high levels of acceptance by the interested parties. The grouping result of this study is not a universal solution to all patient data, to generalize the result, a larger dataset should be adopted in the future research. The main malignant tumors types occurring in men were tumors of the stomach, colon and larynx. Among women, the main tumors were of the colon, stomach and brain. Although the most affected group was more than 60 years also significant presence of tumors was observed from 40 years.**

Keyword: Data Mining, Clustering Technique, CRISP-DM, K-means, Data Mining in Health Care.

I. INTRODUCCIÓN

Gracias a los avances de las plataformas tecnológicas, hemos incrementado exponencialmente la generación y almacenamiento de datos, y por tal motivo, es necesario aplicar técnicas que mejoren la posibilidad de su aprovechamiento [1]

Este artículo es producto de la investigación "Solución de inteligencia de negocios basada en minería de datos, para apoyar la toma de decisiones en el proceso de cirugía del Hospital Departamental Universitario Santa Sofía de Caldas", para optar al título de Magister en Gestión y Desarrollo de Proyectos de Software de la Universidad Autónoma de Manizales (Manizales – Colombia). Línea Gestión de Proyectos de Software - inteligencia de negocios (BI). Autor Tesis: Wilson Alejandro Rojas Calvo. Director: Javier Hernández Cáceres. Fecha Inicio: 01/06/2015. Fecha Fin: 31/05/2016

Según [2], dentro de estas enormes masas de datos existe una gran cantidad de información "oculta" de interés estratégico, a la cual no se puede acceder por las técnicas clásicas de recuperación de la información. El descubrimiento de esta información es posible gracias a la Minería de Datos (DM) por su sigla en inglés de Data Mining. Como expresan [3] la minería de datos permite elevar los niveles de competencia, con base en los rápidos procesamientos y extracción de información relevante del mismo, descubriendo conocimiento y patrones en las bases de datos.

En Colombia, el Sistema único de Acreditación, mediante el Anexo técnico de la Resolución 123 de 2012 y específicamente en el grupo de estándares Gerencia de la Información, dispone que los hospitales deberán implementar procedimientos de Minería de datos para la toma de decisiones orientada a mejorar la calidad en la atención al usuario y su familia. Recomendación complementada por la estrategia e-government del Estado colombiano (Decreto 2573 de 2014), la cual a través de su lineamiento marco de referencia LI.UA.02 establece la necesidad de caracterizar e identificar grupos de usuarios para análisis y toma de decisiones.

El Hospital Departamental Universitario Santa Sofía de Caldas realizó 7.665 cirugías durante la vigencia 2013, 8545 para 2014 y 9.052 para 2015[1], lo cual le permite ubicarse como una Institución con gran experiencia en el tratamiento de diversas especialidades quirúrgicas

La minería de datos ha sido utilizada de manera satisfactoria en el sector de la salud, proporcionando beneficios a las Instituciones Prestadoras de Servicios (IPS) e impactando positivamente la mejora en la atención a los pacientes [4]

II. ANTECEDENTES

[5] Aplicaron técnicas de clustering para el agrupamiento de tipos de procedimientos quirúrgicos que facilitaran la gestión de agendas en quirófanos; por otro lado, [6] nos presentan el uso del software para minería de datos WEKA, la aplicación de técnicas de clustering y específicamente el algoritmo K-Means para la obtención de grupos de pacientes a los cuales se les suministraron ciertos medicamentos. Algoritmo que fue igualmente validado para el agrupamiento y descubrimiento de conocimiento biológicamente relevante [7]. [8] Aplica el algoritmo k-means para agrupar pacientes de características comunes y optimizar la gestión de recursos en una sala de emergencias.

III. MARCO TEÓRICO

A. Data Warehouse.

De acuerdo con [31], un Data Warehouse es una colección de datos, orientados a hechos relevantes del negocio, integrados, que incluyen el tiempo como característica importante de referencia y no volátiles para el proceso de toma de decisiones. [30] nos plantea la definición de Data Warehouse como una colección de datos en forma de una base de datos, que guarda y ordena información extraída directamente desde los sistemas operacionales y datos externos. Ambos autores nos presentan los Data Warehouses (DW) como una fuente de información confiable y consolidada que está orientada al negocio. Metodología.

B. Data Mining

Surge como una tecnología emergente que sirve de soporte para el descubrimiento de conocimiento, que se revela a partir de patrones observables en datos estructurados o asociaciones que usualmente eran desconocidas [7].

Existen diferentes técnicas para llevar a cabo minería de datos, su elección dependerá del objetivo del negocio. Las técnicas descriptivas nos permiten conocer y como su nombre lo indica describir los datos utilizando resúmenes estadísticos, medias, desviaciones estándar, visualización y gráficos, buscando relaciones y vínculos potenciales entre variables [9].

La minería de datos ofrece la oportunidad de descubrir patrones en los datos que pueden ayudar a predecir el comportamiento de los clientes, productos y procesos [9].

Se aprecia el uso de técnicas para la detección de enfermedades crónicas y de salud pública como el consumo de tabaco en adolescentes [10], eficiencia en determinación de factores de riesgo en diabetes tipo II [11] [12] y enfermedades con alto grado de mortalidad como cáncer de mama [13].

Los investigadores indagan por métodos que apoyen labores de asignación de camas en procesos de atención con calidad [14], eficiencia en cuidados de la salud en Unidades de Cuidados Intensivos [15], control y trazabilidad de los días de estancia hospitalaria, que tiene gran influencia en seguridad del paciente y la mitigación de riesgos de infecciones asociadas a la atención [16]. Casos de estudio de gran importancia, teniendo en cuenta que todos estos servicios terminan derivando atenciones quirúrgicas

Es importante notar que no existe un “mejor” modelo o algoritmo de minería de datos, depende del problema en estudio y de los datos disponibles para decir cuál entrega resultados más confiables [17].

C. Tipos de técnicas de Minería de Datos

No supervisadas o descriptivas: Estos modelos no cuentan con un resultado conocido y por ello se conocen como modelos de aprendizaje no supervisado y se va ajustando de acuerdo a las observaciones o datos entregados [17].

Supervisadas o predictivas: Los modelos predictivos requieren ser “entrenados”, utilizando un conjunto de datos de cuyo valor de variable objetivo es conocido [17]. Los resultados obtenidos se comparan con los valores conocidos en el entrenamiento.

[1] E.S.E Hospital Departamental universitario Santa Sofía de Caldas. Evaluación Anual del desempeño de acuerdo al Convenio de desempeño 0188 de 2004. Informes vigencias 2013-2015.

D. Clustering

El análisis de conglomerados o Clustering, es una técnica que permite analizar y examinar datos que no se encuentran etiquetados, formando conjuntos de grupos a partir de su similitud [18].

El principal objetivo del análisis clúster es dividir un conjunto de objetos en dos o más grupos, basándose en la similitud de un conjunto de variables que los caracterizan [19].

La similitud puede medirse a través de funciones de distancia, las cuales juegan un papel crucial, ya que individuos cercanos deberían ir para el mismo grupo [20] [21]. Se agrupan los objetos de acuerdo a todas las variables y por ello, una variable irrelevante puede generar ruido en los resultados obtenidos [19].

E. Medidas de Similitud

Las medidas de similitud establecen la forma en que se determina la proximidad que hay entre los datos. Miden la distancia entre dos objetos [17] [20] en su libro "Introducción a la Minería de Datos" nos explica como las medidas de distancia formalizan a través de métricas la similitud entre los objetos. Nos indica que las medidas de distancia tradicionales, se aplican sobre dos instancias o ejemplos numéricos *x* e *y* de dimensión *n* y describe la distancia euclidea como longitud de la recta que une dos puntos:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

F. Algoritmo K-Means

Es un algoritmo particional, es decir, divide los objetos en un número de clústeres pre especificado, sin atender a una estructura jerárquica [22], puede aplicarse para problemas de "agrupación por similitud" y puede ayudar al investigador a una comprensión cualitativa y cuantitativa de grandes cantidades de datos N-dimensionales [23].

Funciona de forma iterativa, dividiendo óptimamente el conjunto inicial de datos en un número (K) de clústeres, el cual se indica como parámetro. Está basado en la minimización de la distancia interna. [20] nos propone una especificación matemática para describir el procedimiento llevado a cabo por el algoritmo K-Means, de la siguiente manera:

a) Se calcula para cada ejemplo de x_k, el prototipo más próximo A_g y se incluye en la lista de dicho prototipo:

$$A_g = \underset{A_i}{\operatorname{argmin}} \{d(x_k, A_i)\} \quad \forall_i = 1..n \quad (2)$$

b) Luego de introducir todos los ejemplos, cada prototipo A_k, tendrá un conjunto de elementos

$$l(A_k) = \{x_{k1}, x_{k2} \dots x_{km}\} \quad (3)$$

c) El prototipo es desplazado hacia el centro de su conjunto de ejemplos:

$$A_k = \frac{\sum_{i=1}^{m} x_{k_i}}{m} \quad (4)$$

d) Se repite el procedimiento hasta que ya no se desplazan los prototipos. Los ejemplos de entrada k, se dividen en regiones y el prototipo de cada región estará en el centro de cada una para reducir las distancias cuadráticas euclídeas entre los patrones de entrada y su centro más cercano. Minimizando así el valor J.

$$J = \sum_{i=1}^{k}\sum_{n=1}^{m} M_{i,n} d_{EUCL}(x_n - A_i)^2 \quad (5)$$

e) Siendo (m) el conjunto de patrones, d_{EUCL} $d_{EUCL} d_{EUCL}$ la distancia euclidea, x_n el ejemplo de entrada, A_i el prototipo de la clase *i* y $M_{i,n}$ la función que indica la pertenencia del ejemplo *n* a la región *i*. Valiendo 1 si el prototipo A_i A_i es el más cercano al ejemplo x_n y 0 en caso contrario. $M_{i,n}$

$$M_{i,n} = \begin{cases} 1 \text{ si } d_{EUCL}(x_n - A_i) < d_{EUCL}(x_n - A_s) \ \forall_s \neq i, s = 1,2,\dots,k \\ 0 \text{ en caso contrario} \end{cases} \quad (6)$$

IV. METODOLOGÍA

A. Unidad de Análisis

La unidad de análisis correspondió a los registros clínicos de pacientes quirúrgicos entre los años 2007 y 2015, cuyos diagnósticos iniciales estaban relacionados con cáncer, de acuerdo a la Clasificación Internacional de enfermedades versión 10 (CIE-10). Dada la característica de confidencialidad de la historia clínica (Resolución 1995 de 1999) se obtuvo aprobación del Comité de Ética Institucional, puesto que ninguno de los datos personales de los pacientes se expuso durante la investigación. Se exploraron datos sociodemográficos generales.

B. Metodología del proyecto de minería

Las metodologías permiten llevar a cabo el proceso de minería de datos en forma sistemática y no trivial [24]. Existen diferentes metodologías aplicables, entre las cuales se revisaron: CRISP-DM (Cross Industry Standard Process for Data Mining), SEMMA (Sample – Explore – Modify – Model – Assess), KDD (Knowledge Discovery and Data Mining), Metodología Catalyst (P3TQ) y Metodología Berry y Linoff. [25] nos presenta una comparación entre las diversas metodologías.

CRISP-DM estructura el proceso en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación [26].

Tiene como objetivo realizar proyectos de minería de datos, menos costosos, fiables, repetibles, manejables, y ágiles [27]; CRISP-DM ha crecido como estándar y define un conjunto de pasos secuenciales que pretenden ser guía en la implementación de aplicaciones de minería de datos [3], de acuerdo con [28] CRISP-DM define los procesos y tareas que se deben realizar para desarrollar en forma exitosa un proyecto de explotación de información y [29] la define como un proceso para el desarrollo de proyectos de minería de datos iterativo, abierto, personalizable y de gran reconocimiento por la industria y la academia. Sin desvirtuar ni desconocer el potencial y suficiencia de las demás metodologías y sin que la comparación fuese el fin último de la investigación, el estudio de caso siguió recomendaciones de la metodología CRISP-DM.

C. *Modelo de datos*

El modelo de datos desde el cual se pudieron obtener los set de datos para minería, se basó en el Esquema de Constelación de Hechos propuesto por [30], integrando datos relacionados con pacientes, actos quirúrgicos, procedimientos quirúrgicos, diagnósticos, ubicaciones geográficas y aseguradoras.

D. *Documentación de las fases de la metodología*

Se diseñaron y aplicaron formatos para documentar cada una de las fases propuestas por CRISP-DM: la *Comprensión del negocio* permitió alinear la investigación con objetivos y requerimientos del hospital, la *Comprensión de los datos* facilitó la exploración de los orígenes de datos y su relevancia para la investigación, a través de la *Preparación de los datos* se estructuró el data warehouse luego de realizar tareas de selección, limpieza y formateo de datos y se generaron las vistas minables (ver "Fig. 1.")que permitieron la aplicación de la técnica de minería durante la fase de *Modelado*. La *Evaluación* permitió revisar los resultados de cada fase y facilitar la etapa de validación por parte de los interesados y en el *Despliegue* se socializaron los resultados obtenidos.

E. *Proceso de minería*

Durante la fase I "comprensión del Negocio", se identificó como objetivo de minería de datos: Aplicar técnica Clustering de minería de datos, para agrupar pacientes que han sido sometidos a procedimientos quirúrgicos durante los años 2007-2015 y cuyos diagnósticos estaban asociados a tumores malignos.

Se estableció un set de datos que incorporó 2268 instancias y 11 atributos como: edad, grupo CIE-10, Grupo_Cirugía, TipoAtencion, Etareo, ZonaResidencia, Regimen, Genero, EstadoCivil, Etnia y Estrato.

Para cada atributo se estableció su descripción, tipo de dato, valores permitidos, valor mínimo y máximo, así como su relevancia en la investigación.

Se realizó un análisis exploratorio de los datos para identificar sus características, datos válidos, frecuencia, porcentaje acumulado y por último se procedió a aplicar la técnica de minería Clustering, específicamente el algoritmo K-Means, utilizando el software Weka (Waikato Environment for Knowledge Analysis) versión 3.8.0. Ver "Fig. 2."

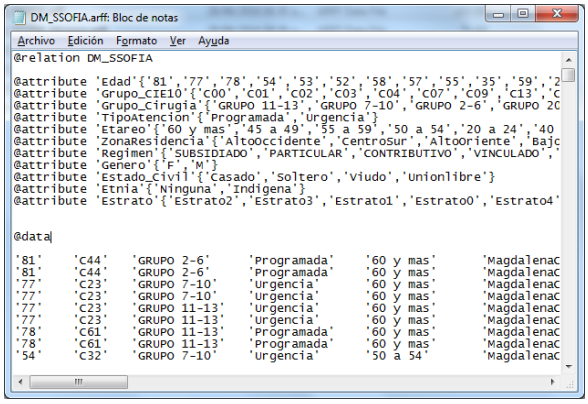

Fig. 1. Ejemplo Archivo arff - set de datos

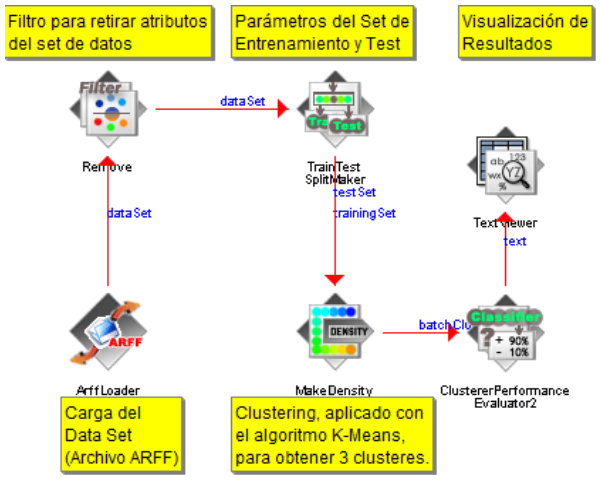

Fig. 2. Data Mining Processes. Flujo de trabajo en Weka

V. RESULTADOS

El análisis de la escala discreta en años permitió establecer que la edad mínima fue de 13 años y la máxima de 89, la edad más observada en estos procedimientos fue de 62 años con una frecuencia de 85, que equivale al 3,7% del total de la población; de acuerdo al Código Internacional de clasificación de enfermedades, los grupos más representativos fueron: C16 (Tumor maligno del estómago) con una frecuencia de 396 y equivalente al 17,5% del total de la información, seguido de C18 (Tumor maligno del colon) para un 9%, C76 (Tumor maligno de otros sitio y sitios mal definidos) presentó el 5,6%, C25 (Tumor maligno de páncreas) un 5,5% y C71 (Tumor maligno del encéfalo) con 5,2%. Para los grupos quirúrgicos se obtuvo un 46% para el grupo 7 al 10, seguido de 25,65% para

el grupo 11 al 13 y 20,1% para procedimientos quirúrgicos grupo 20 al 23.

Las cirugías programadas superaron con un 72,4% a las urgencias, que alcanzaron un 27.6%. El grupo etáreo con mayor frecuencia correspondió a "60 años y más" con 1239 instancias y un 54,6%.

La zona "Centro Sur" arrojó las mayores cifras de atención quirúrgica con un 71,6%, seguida del "Alto Occidente" y "Magdalena Caldense" con un 7,3% y 6,2% respectivamente.

Dada la naturaleza pública del Hospital, el 53,1% de la población correspondió al régimen "Subsidiado", mientras que el 30,2% al régimen "Contributivo". De forma similar, el estrato más atendido fue el estrato 1 con 40,2%, seguido estrato 0 con un 26,8%. Aunque en la zona del Alto Occidente del departamento de Caldas existen asentamientos indígenas, sólo el 2.2% de los observados pertenecen a dicha etnia, a diferencia de un 97,8% de atención que no registró etnia. Para terminar, el 50.7% de la población correspondió al género femenino y el 49,3% masculino y un 75,2% de la población perteneció al estado civil "Casado".

Inicialmente se obtuvieron dos clústeres, el primero con un 67% de la población, agrupados bajo el código CIE-10 C16, grupo de cirugía 7 al 10, tipo de atención programada, régimen subsidiado, hombres casados mayores de 60 años , sin etnia y de estrato 0. El segundo clúster con un 33% de la población, de más de 60 años, de sexo femenino, se agrupan alrededor de la enfermedad bajo el código C16, grupo de cirugía 7 al 10, tipo de atención programada, régimen contributivo, sin etnia y de estrato 1.

Al obtener los Clústeres anteriores, se pudo validar la información estadística sobre la patología de cáncer de estómago[2] (C16) la cual de acuerdo a la información epidemiológica, tiene presencia importante en la zona de influencia.

Se procedió entonces a excluir el grupo CIE-10 (C16), analizando 1872 registros y obteniendo dos nuevos clústeres, que ratifican la prevalencia de tumores malignos en personas mayores de 60 años, pero con intervenciones clasificadas en grupos quirúrgicos del 11 al 13. Estos nuevos clústeres hacen énfasis en tumores del colon (C18) y presentan un clúster con el 56% y el otro con el 44% de la población. Ver "Fig. 3."

		Cluster#	
Attribute	Full Data	0	1
	(1872.0)	(1045.0)	(827.0)
Edad	62	62	66
Grupo_CIE10	C18	C18	C18
Grupo_Cirugia	GRUPO 7-10	GRUPO 11-13	GRUPO 7-10
TipoAtencion	Programada	Programada	Programada
Etareo	60 y mas	60 y mas	60 y mas
ZonaResidencia	CentroSur	CentroSur	CentroSur
Regimen	SUBSIDIADO	CONTRIBUTIVO	SUBSIDIADO
Genero	F	F	M
Estado_Civil	Casado	Casado	Casado
Etnia	Ninguna	Ninguna	Ninguna
Estrato	Estrato1	Estrato1	Estrato0
Clustered Instances			
0	1045 (56%)		
1	827 (44%)		

Fig. 3. Clusteres sin incuir grupo CIE10 C16 (Cáncer de estómago)

Buscando agrupar personas de otros rangos de edad, se excluyó el grupo etario de más de 60 años, obteniendo 861 registros y otros dos clústeres, el primero con un 62% de la población, con grupo diagnóstico tumores malignos del encéfalo (C71), grupo de edad entre 45 y 49 años del régimen contributivo y cuyas cirugías pertenecieron al grupo 7 al 10, de sexo femenino, sin etnia y de estrato 1. Al segundo clúster se asignó el 38% de la población con presencia de tumores del colon (C18), grupo de cirugías entre el 7 y el 10, rango de edad entre 55 y 59 años, en su mayoría del sexo masculino y del estrato 0. Ver "Fig. 4."

Attribute	Full Data	0	1
	(861.0)	(533.0)	(328.0)
Edad	48	48	55
Grupo_CIE10	C71	C71	C18
Grupo_Cirugia	GRUPO 7-10	GRUPO 7-10	GRUPO 7-10
TipoAtencion	Programada	Programada	Programada
Etareo	50 a 54	45 a 49	55 a 59
ZonaResidencia	CentroSur	CentroSur	CentroSur
Regimen	SUBSIDIADO	CONTRIBUTIVO	SUBSIDIADO
Genero	F	F	M
Estado_Civil	Casado	Casado	Casado
Etnia	Ninguna	Ninguna	Ninguna
Estrato	Estrato1	Estrato1	Estrato0
Clustered Instances			
0	533 (62%)		
1	328 (38%)		

Fig. 4. Clusteres personas menores de 60 años.

La zona "Centro Sur" incluyó el 71,6% de las instancias, por lo tanto se separó del set de datos, para analizar otras sub-regiones. De las 641 instancias analizadas se obtuvieron tres clústeres. Ver "Fig. 5."

Final cluster centroids:

		Cluster#		
Attribute	Full Data	0	1	2
	(641.0)	(389.0)	(162.0)	(90.0)
Grupo_CIE10	C16	C16	C16	C32
Grupo_Cirugia	GRUPO 7-10	GRUPO 7-10	GRUPO 7-10	GRUPO 20-23
Etareo	60 y mas	60 y mas	60 y mas	50 a 54
ZonaResidencia	AltoOccidente	AltoOccidente	BajoOccidente	MagdalenaCaldense
Regimen	SUBSIDIADO	SUBSIDIADO	SUBSIDIADO	SUBSIDIADO
Genero	M	F	M	M
Estrato	Estrato0	Estrato0	Estrato2	Estrato1

[2] Sistema de información para la Calidad del Hospital.

Fig. 5. Clústeres sub-regiones diferentes a Centro Sur.

Aunque continua el grupo diagnóstico C16 para dos de los tres clústeres, aparece en la región del "Magdalena Caldense" el código C32 (Tumor maligno de la Laringe), y un grupo etario entre 50 y 54 años, de género masculino y estrato 1, a quienes se les ha venido realizando cirugías de alta complejidad (grupo 20 al 23).

Se puede apreciar que la zona del "Alto occidente" agrupa 389 instancias, para un 60,7% del total, seguida del "Bajo Occidente" con un 25,3% y "Magdalena Caldense" con 14%.

El clúster principal por otras subregiones está conformado por mujeres, mayores de 60 años, de estrato 0 (comunidades rurales), del régimen subsidiado. El segundo clúster, se caracteriza por grupo de hombres, pertenecientes al estrato2, de más de 60 años y régimen subsidiado.

En la "Fig. 6.", se puede visualizar la concentración de instancias obtenidas durante el análisis. Existe mayor concentración de grupos de diagnóstico C16-C20 para edades mayores a 60 años, sin embargo, estos diagnósticos se hacen presentes desde los 40 años.

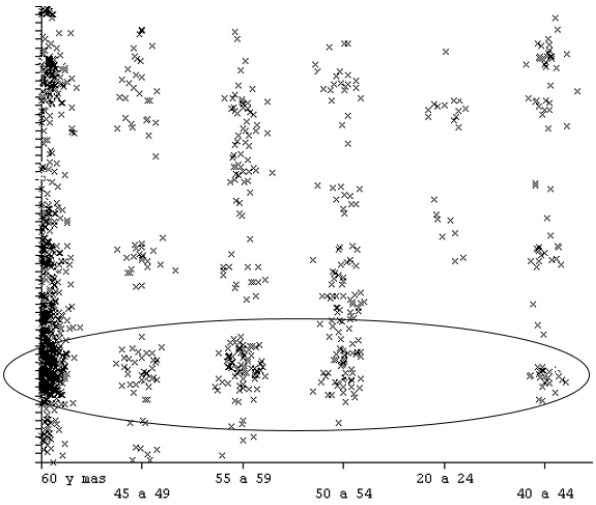

Fig. 6. Grupos CIE frecuentes en mayores de 40 años (C16-C20)

En la "Fig. 7." Se observa que, para los pacientes de otros municipios diferentes a la zona "Centro sur", el grupo de edades 45 a 49 años, presenta concentración de instancias en cuanto a diagnósticos de los grupos (C73-C76) Glándulas endocrinas, tiroides y médula espinal y (C32-C34) laringe, traquea y pulmón.

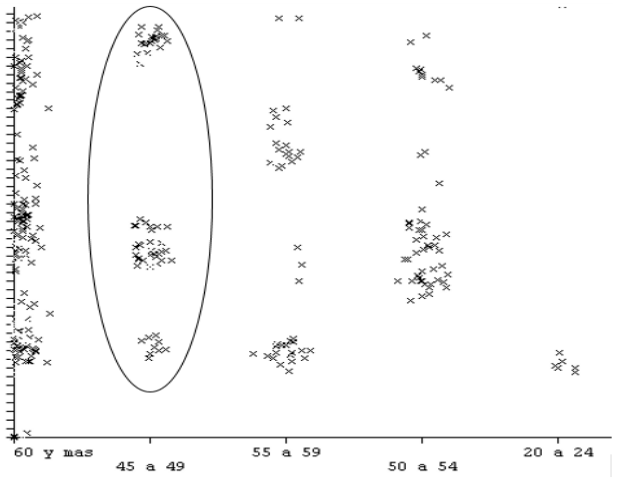

Fig. 7. Grupos CIE frecuentes en pacientes de 45 a 49 años. Subregiones diferentes a zona Centro Sur.

VI. CONCLUSIONES

La anterior información ratifica elementos conceptuales que se han venido tratando al interior de la Institución frente al incremento de atenciones relacionadas con cáncer.

El estudio permite complementar las cifras expresadas por [32] que demuestran para el departamento de Caldas, unas tasas ajustadas por edad, por 100.000 habitantes de 24.5 para hombres y 12,2 para mujeres en cuanto al cáncer de estómago, seguidas de cáncer de colon, recto y ano que muestra tasas de 15,9 y 16,2 respectivamente.

Tanto para hombres, como mujeres mayores de 60 años se observó presencia de tumores malignos del estómago y colon, siendo los primeros con mayor presencia en hombres y el segundo con mayor presencia en mujeres. En cuanto a mujeres menores de 60 años, se apreció un clúster para el grupo etario 45 a 49 años en el cual se registraron tumores relacionados con tumores malignos del encéfalo y los hombres de 55 a 59 Años continuaron presentes los tumores malignos del estómago.

A nivel de sub regiones, se mantiene la tendencia hacia el cáncer del estómago, sin embargo aparecen los tumores malignos de la laringe en hombres de 50 a 54 años del Magdalena Caldense, requiriendo cirugías de alta complejidad grupos 20 al 23.

Teniendo en cuenta la tasa de mortalidad observada para hombres en cuanto a cáncer de estómago, que corresponde a 18,6 y 9,2 para mujeres [32], se exalta la importancia de avanzar en investigaciones que apoyen la detección temprana de dichas patologías. Trabajos futuros podrían incorporar la aplicación de técnicas de minería en la detección temprana de dichas patologías.

La técnica de minería *Clustering*, bajo el algoritmo *K-Means* y medida de similitud *Euclidea*, permitió obtener grupos de pacientes cuyas intervenciones diagnósticos estaban asociados a tumores malignos.

A partir del agrupamiento y caracterización de pacientes, las Instituciones Prestadoras de Servicios de Salud pueden avanzar

en la definición de estrategias de promoción de la salud y prevención de la enfermedad, así como avanzar en la adherencia normativa que exige un mayor conocimiento sobres sus pacientes.

AGRADECIMIENTOS

A la Empresa Social del Estado (E.S.E) Hospital Departamental Universitario Santa Sofía de Caldas, por haber permitido realizar el presente trabajo de investigación y facilitar los recursos técnicos necesarios para alcanzar los objetivos.

REFERENCES

[1] HOWE, Bill. (2014) Introduction to Data Science. Consultado [11/08/2015]. Universidad de Washington. (https://www.coursera.org/course/datasci?from_restricted_preview=1&course_id=972772&r=https%3A%2F%2Fclass.coursera.org%2Fdatasci-002%2Fclass).

[2] Valcárcel Asencios. (2004). Data Mining y el descubrimiento del conocimiento Industrial Data, vol. 7, núm. 2, julio-diciembre, pp. 83-86, Universidad Nacional Mayor de San Marcos. Perú. http://www.redalyc.org/pdf/816/81670213.pdf (Consultado 16/07/2014)

[3] Albarrán Trujillo SE, Salgado Gallegos M. (2013) La Inteligencia Analítica y la Competitividad en las Empresas. RECAI Revista de Estudios en Contaduría, Administración e Informática [Internet]. 30–4 [cited 2015 Apr 22];0(0). Available from: http://www.revistarecai.mx/index.php/recai/article/view/18

[4] Tremblay, M. C., Hevner, A. R., & Berndt, D. J. (2012). Design of an information volatility measure for health care decision making. Decision Support Systems, 52(2), 331–341. doi:10.1016/j.dss.2011.08.009

[5] Van Oostrum, J. M., Parlevliet, T., Wagelmans, A. P. M., & Kazemier, G. (2011). A Method for Clustering Surgical Cases to Allow Master Surgical Scheduling. INFOR: Information Systems and Operational Research, 49(4), 254–260. http://doi.org/10.3138/infor.49.4.254

[6] Rivas IT, Rivera MR, Lizama ER. (2007) Una metodología para sectorizar pacientes en el consumo de medicamentos aplicando datamart y datamining en un hospital [Internet]. Industrial Data. [cited 2015 Jul 16]. Available from: http://www.redalyc.org/articulo.oa?id=81610114

[7] Taié Armando. (2008). Desarrollo de una metodología de extracción de conocimientos a partir de datos de micromatrices de DNA basada en ontologías genéticas. Universidad de Buenos Aires. [cited 15/07/2015]. Available from: http://inta.gob.ar/personas/taie.armando

[8] Xu, M., Wong, T. C., & Chin, K. S. (2014). A medical procedure-based patient grouping method for an emergency department. Applied Soft Computing, 14, Part A, 31–37. http://doi.org/10.1016/j.asoc.2013.09.022

[9] Two Crows Corporation. (2005). Introduction to Data Mining and Knowledge Discovery. Third Edition. ISBN: 1-892095-02-5. [cited 2015 Jul 22] Available from: http://www.twocrows.com/intro-dm.pdf

[10] Montaño Moreno, Juan J.; Gervilla García, Elena; Cajal Blasco, Berta; Palmer, Alfonso. (2014) Técnicas de clasificación de data mining: una aplicación al consumo de tabaco en adolescentes Anales de Psicología, vol. 30, núm. 2, mayo-agosto, pp. 633-641 Universidad de Murcia Murcia, España http://www.redalyc.org/pdf/167/16731188027.pdf

[11] Rezaei, Mansour. Eghbal, ZanKarimi. Amirthossein Hasheiman. (2013) Comparison of Artificial Neural Network, Logistic Regression and Discriminant Analysis Efficiency in Determining Risk Factors of Type 2 Diabetes. World Applied Sciences Journal 07/2013; 23(11):1522-1529. DOI: 10.5829/idosi.wasj.2013.23.11.1119

[12] Bellazzi R, Abu-Hanna A. (2009) Data Mining Technologies for Blood Glucose and Diabetes Management. J Diabetes Sci Technol [Internet]. [cited 2014 Aug 16];3(3):603–12. Available from:http://dst.sagepub.com/content/3/3/603

[13] Mclraren, Christine. Pin Chen , Wen. Nie, Ke. Su, Ming-Ying. (2009) Prediction of malignant breast lesions from MRI features: a comparison of artificial neural network and logistic regression techniques. Department of Epidemiology, University of California. Academic radiology (Impact Factor: 2.09); 16(7):842-51. Elsevier. ISSN: 1878-4046.

[14] Oliveria, Sérgio. Portela, Filipe. Manuel F. Santos. José Machado. António Abelha. (2013) Predictive Models for Hospital Bed Management. Using Data Mining Techniques. University of Minho, Guimarães, Portugal.

[15] Sarmiento, Xavier. Guardiola, Juan. Roca, J. Soler, M. Toboso, J.M. Klamburg, J. Artigas, Antonio. (2013) Servicio Medicina Intensiva, Hospital Universitario Germans Trias i Pujol, Badalona, Barcelona, España. Medicina Intensiva (Impact Factor: 1.32). 37(3):132–141. DOI: 10.1016/j.medin.2012.03.006

[16] Kudyba, S., & Gregorio, T. (2010). Identifying factors that impact patient length of stay metrics for healthcare providers with advanced analytics. Health Informatics Journal, 16(4), 235–245. doi:10.1177/1460458210380529

[17] Martínez Álvarez, Clemente Antonio. (2012). Aplicación de Técnicas de Minería de Datos para Mejorar el Proceso de Control de Gestión en Entel [Internet]. Universidad De Chile; Available from: http://repositorio.uchile.cl/bitstream/handle/2250/112065/cf-martinez_ca.pdf?sequence=1

[18] Xu, Rui. Donald, Wunsch. (2005). “Survey of Clustering Algorithms.” IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council 16(3):645–78.

[19] Aldás Manzano, Joaquin. (2002) El análisis clúster. Universitat de València Dpto. de Dirección de Empresas “Juan José Renau Piqueras”.

[20] Hernández Orallo, José. Ramírez Quintana, M José. Ferri Ramírez, César. (2005) Introducción a la Minería de Datos. Editorial Pearson, ISBN: 84 205 4091 9. p. 575.

[21] Rivas IT, Rivera MR, Lizama ER. (2007) Una metodología para sectorizar pacientes en el consumo de medicamentos aplicando datamart y datamining en un hospital [Internet]. Industrial Data. [cited 2015 Jul 16]. Available from: http://www.redalyc.org/articulo.oa?id=81610114

[22] Xu, Rui. Donald, Wunsch. (2009). Clustering. Wiley-IEEE Press. ISBN 9780470276808

[23] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Presented at the Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, The Regents of the University of California. Retrieved from http://projecteuclid.org/euclid.bsmsp/1200512992

[24] Moine, Juan Miguel. Haedo, Silvia. Gordillo, Silvia. (2010). Estudio comparativo de metodologías para minería de datos. UTN Rosario Facultad de Ciencias Exactas, Universidad Nacional de Buenos Aires Facultad de Informática, Universidad Nacional de La Plata. http://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf%3Fsequence%3D1 12/02/2015.

[25] Vanrell JÁ, Bertone RA, García Martínez R. (2010) Un modelo de procesos de explotación de la información. [cited 2015 Apr 22]. Available from: http://hdl.handle.net/10915/19462

[26] Chapman, Pete. Clinton, Julian. Kerber, Randy. Khabaza Thomas. Reinartz, Thomas. Shearer, Colin. Wirth Rüdiger. SPSS, NCR, DaimlerChrysler. (2000) Guía paso a paso de Minería de Datos. Copyright © 1999, 2000. http://the-modeling-agency.com/crisp-dm.pdf 12/02/2015

[27] Wirth, Rüdiger. Hipp, Jochen. (2000) CRISP-DM: Towards a Standard Process Model for Data Mining. DaimlerChrysler Research & Technology. University of Tübingen. Germany. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5133&rep=rep1&type=pdf 21/04/2015

[28] Tomasello M, Pytel P, Rodríguez D, Arboleya H, Pollo Cattaneo MF, Britos PV, et al. (2011) Estimación de proyectos de explotación de información. [cited 2015 Apr 22]. Available from:http://hdl.handle.net/10915/20041

[29] Cobos, Carlos. Zuñiga, Jhon. Guarin, Juan. León, Elizabeth. Mendoza, Marha. (2010) CMIN - herramienta case basada en CRISP-DM para el soporte de proyectos de minería de datos. Ingeniería e Investigación. http://www.scielo.org.co/pdf/iei/v30n3/v30n3a04.pdf 20/04/2015

[30] Kimball, Ralph. Ross, Margy. (2013) The Data Warehouse Toolkit. The definitive guide to Dimensional Modeling. Third Edition. Kimball Group. Editorial Wiley. ISBN: 978-1-118-53080-1.

[31] Inmon, W. H., (2005) Building the Data Warehouse., 4 edition, John Wiley Publishing.

[32] Pardo Ramos, C. Cendales Duarte, R. (2011) Incidencia, mortalidad y prevalencia de Cáncer en Colombia 2007-2011Ministerio de Salud y Protección Social. Instituto Nacional de Cancerología ESE. p. 52-53. Cited [01/03/2016] http://www.cancer.gov.co/files/libros/archivos/incidencia1.pdf