MINERÍA DE DATOS: APORTES Y TENDENCIAS EN EL SERVICIO DE SALUD DE CIUDADES INTELIGENTES

Efraín Alberto Oviedo Carrascal [1], Ana Isabel Oviedo Carrascal [2], Gloria Liliana Vélez Saldarriaga [3]

[1] Ingeniero Electrónico, Estudiante de Maestría en Tecnologías de la Información y la Comunicación, Grupo de Investigación GIDATI. Correo electrónico: eaoc46@gmail.com.
[2] Doctora en Ingeniería Electrónica – Énfasis en Descubrimiento de Conocimiento, Profesora titular en la Facultad de Ingeniería en Tecnologías de la Información y la Comunicación, Grupo de Investigación GIDATI. Correo electrónico: ana.oviedo@upb.edu.co.
[3] Magister en Gestión Tecnológica, Profesora titular en la Facultad de Ingeniería en Tecnologías de la Información y la Comunicación, Grupo de Investigación GIDATI. Correo electrónico: gloria.velez@upb.edu.co.
[1, 2, 3] Universidad Pontificia Bolivariana, circular 1 No. 70-01, Medellín, Colombia.

RESUMEN

Entre las numerosas aplicaciones de la minería de datos se destacan los aportes al servicio de salud en ciudades inteligentes. Dichas aplicaciones tienen por objetivo mejorar la calidad de vida de los ciudadanos, prevenir enfermedades, facilitar la toma de decisiones y analizar datos provenientes de las instituciones de salud. Con el objetivo de apoyar el desarrollo de ciudades inteligentes, en este trabajo se presenta una revisión de avances y tendencias de la minería de datos en el servicio de salud. Entre los principales avances en minería de datos se pueden encontrar diversas técnicas, metodologías y plataformas que han sido utilizadas en el sector salud. Entre las tendencias se pueden identificar algunos desafíos como: análisis de textos e imágenes, metodologías con etapas de preprocesamiento e indexamiento de datos no estructurados y herramientas con soporte a minería multimedia.

Palabras clave: Servicio de salud en ciudades inteligentes, minería de datos, minería de texto y minería de imágenes.

Recibido: 15 de Septiembre de 2014. Aceptado: 10 de junio de 2015.
Received: September 15th, 2014. *Accepted: June 10th, 2015.*

DATA MINING: CONTRIBUTIONS AND TRENDS IN THE HEALTH SERVICE OF SMART CITIES

ABSTRACT

Among the applications of data mining, the contributions to health services in smart cities are highlighed. These applications are intended to improve the life quality of citizens, prevent disease, facilitate decision making and analyze data from health institutions. In order to support the development of smart cities, this paper presents a review in developments and trends of data mining in health services. In the data mining developments we can found techniques, methodologies and platforms that have been used in the health sector. In the data mining trends we can found some challenges in the health services: text and images analysis, data mining methodologies with a stage of unstructured data preprocessing and indexing, and data mining tools to support multimedia.

Keywords: *Health service in Smart Cities, data mining, text mining and image mining.*

1. INTRODUCCIÓN

Las ciudades inteligentes tienen como objetivo mejorar la calidad de vida de los ciudadanos. Para ello se utilizan las tecnologías de la información y la comunicación (TICs) como herramientas para transformar y mejorar los procesos y actividades de la administración [1]. En los últimos años, este concepto ha tenido una gran acogida alrededor del mundo y se ha elevado considerablemente el número de ciudades que se han preocupado por realizar actividades de investigación y desarrollo al respecto. Como muestra de esta tendencia, en [2] se realiza un estudio de los planes de desarrollo de 415 ciudades inteligentes de diferentes países del mundo.

El *International Data Corporation* (IDC) propone algunas áreas en las cuales las ciudades inteligentes deben centrar sus esfuerzos [3]. Estas áreas son: gobierno, construcción, movilidad, energía, medio ambiente y servicios. El servicio de salud en las ciudades inteligentes se enmarca en las áreas de gobierno y servicio, buscando prevenir enfermedades y mejorar la salud de los ciudadanos. Buscando aportar al servicio de salud en una ciudad inteligente, en la actualidad se desarrollan numerosas aplicaciones que permiten analizar datos provenientes de las instituciones de salud para apoyar la toma de decisiones en la ciudad. Dichas aplicaciones son desarrolladas con técnicas de minería de datos, las cuales permiten descubrir conocimiento en grandes volúmenes de datos.

Desde hace más de 60 años se han publicado gran cantidad de artículos en conferencias y revistas sobre la minería de datos [4]. Sin embargo, es un área que aún no encuentra estabilidad, ya que se siguen publicando nuevos métodos de selección de atributos, nuevas técnicas de modelaje y nuevas medidas de evaluación de resultados. Esta inestabilidad se ve reflejada en las plataformas de minería de datos, ya que cada herramienta utiliza métodos y medidas de evaluación diferentes. Adicionalmente, la minería de datos es un campo con nuevos requerimientos en la actualidad ya que Big Data trae nuevos y exigentes requerimientos para el área [5].

En este trabajo se presenta una revisión de aplicaciones de minería de datos orientadas al servicio de salud en el marco de las ciudades inteligentes. La organización del artículo es la siguiente. En la sección 2 se abordan los temas relacionados con minería de datos como tareas, técnicas, metodologías y plataformas. En la sección 3 se presenta una revisión de las aplicaciones de minería para el sector salud. En la sección 4 se analizan las tendencias y los nuevos requerimientos que nacen en el sector salud para las aplicaciones de minería de datos. Finalmente en la sección 5 se presentan las conclusiones y trabajo futuro.

2. MINERÍA DE DATOS

La minería de datos se puede definir como el proceso a través del cual se descubre conocimiento no trivial en forma de patrones, asociaciones, cambios, anomalías y estructuras significantes de grandes cantidades de datos almacenados en bases de datos, bodegas de datos u otros repositorios de información. Para realizar este proceso se suele utilizar técnicas de la inteligencia artificial y la estadística [6].

2.1 Tipos de Análisis

En la minería de datos se pueden desarrollar dos tipos de análisis: predictivo y descriptivo. Dichos análisis permiten desarrollar diferentes tareas como la clasificación [7], la predicción [8], la segmentación [9] y la asociación [10].

2.1.1 Análisis Predictivo

Algunas de las aplicaciones comúnmente desarrolladas con análisis predictivo son: predecir riesgos, predecir activación de nuevos clientes, predicción de ventas, entre otras [4]. Este tipo de análisis se caracteriza porque requiere un conjunto de entrenamiento, el cual está formado por un histórico de datos. En el análisis predictivo se pueden desarrollar tareas de predicción discreta y predicción continua.

La predicción discreta también recibe el nombre de clasificación, donde el conjunto de entrenamiento está conformado por atributos y una variable discreta que representa la clase a la cual pertenece cada registro, como se muestra en la figura 1.

En la predicción continua, los registros en el conjunto de entrenamiento están conformados por atributos y una variable de predicción continua (numérica), como se muestra en la figura 2.

Id	Atributo 1	Atributo 2	...	Atributo n	Clase
1	10	alto		35	Cliente Oro
2	35	bajo		54	Cliente Plata
3	43	medio		28	Cliente Bronce
4	26	bajo		65	Cliente Bronce
5	87	alto		32	Cliente Oro
6	45	alto		29	Cliente Plata
7	76	bajo		55	Cliente Bronce
8	5	medio		46	Cliente Oro
9	12	medio		43	Cliente Bronce
10	54	bajo		27	Cliente Plata

Fig.1. Conjunto de entrenamiento utilizado en el análisis predictivo discreto (clasificación).

Id	Atributo 1	Atributo 2	...	Atributo n	Predicción
1	10	alto		35	54.678
2	35	bajo		54	100.500
3	43	medio		28	27.000
4	26	bajo		65	9.800
5	87	alto		32	23.600
6	45	alto		29	65.400
7	76	bajo		55	78.000
8	5	medio		46	43.500
9	12	medio		43	78.900
10	54	bajo		27	93.000

Fig.2. Conjunto de entrenamiento utilizado en el análisis predictivo continuo.

2.1.2 Análisis Descriptivo

Algunas de las aplicaciones más comunes del análisis descriptivo son: análisis del perfil de personas, detección de anomalías, detección de reglas que condicionen la venta de productos, entre otras [4]. En este tipo de análisis se pueden desarrollar tareas de agrupación (clustering) y de asociación. El conjunto de datos requerido está conformado por los atributos que se desean analizar para encontrar similitudes o asociaciones entre los datos. Un ejemplo de un conjunto de datos para análisis descriptivo se presenta en la figura 3.

Id	Atributo 1	Atributo 2	...	Atributo n
1	10	alto		35
2	35	bajo		54
3	43	medio		28
4	26	bajo		65
5	87	alto		32
6	45	alto		29
7	76	bajo		55
8	5	medio		46
9	12	medio		43
10	54	bajo		27

Fig.3. Conjunto de datos utilizado en el análisis descriptivo.

2.2 Técnicas de Minería de Datos

Existen diversas técnicas de minería de datos [11], la elección de una de ellas depende básicamente de dos condiciones: el tipo de atributos y el objetivo de la minería [4]. De forma general, las técnicas se pueden agrupar en técnicas supervisadas y no supervisadas. Aunque existen gran cantidad de técnicas, a continuación se presenta una breve descripción de las técnicas más utilizadas en las aplicaciones de minería de datos.

2.2.1 Técnicas Supervisadas

Las técnicas supervisadas son aplicadas en el análisis predictivo. Algunas técnicas supervisadas son: Redes Neuronales, Árboles de Decisión, Máquinas de Soporte Vectorial, Métodos de Regresión, Método Bayesiano y Métodos basados en Ejemplos.

Las Redes Neuronales imitan el funcionamiento del cerebro humano para realizar tareas de aprendizaje. Tienen una arquitectura organizada en capas de neuronas, las cuales tienen pesos asignados a sus interconexiones. El aprendizaje de la red consiste en ajustar los pesos mediante una regla que indica cómo modificar los pesos en función de los datos de entrenamiento [12].

Los Árboles de Decisión representan reglas en una estructura de árbol, en la cual los nodos internos son configurados con los atributos, las ramas representan los valores del atributo y las hojas del árbol identifican las clases. La clasificación se realiza descendiendo en el árbol hasta alcanzar una hoja, la cual indica la clase a la cual pertenece cada registro de la base de datos [13]. También

existen árboles de predicción, que permiten analizar la salida en variables continuas.

Las Máquinas de Soporte Vectorial mapean los datos de entrada a un espacio de características de más alta dimensión, donde se puede construir un hiperplano que separe los datos que pertenecen a la clase, de los que no pertenecen a ella. El mapeo de los datos se realiza por medio de una función kernel (por ejemplo: lineal, polinominal, función de base radial, sigmoidal, entre otras). Los datos más próximos al hiperplano de separación son conocidos como muestras críticas o vectores soporte del modelo [14].

La Regresión también es utilizada como técnica supervisada en la minería de datos. La regresión lineal permite predecir la salida continua de una variable dependiente. Por su parte, la regresión logística es utilizada para predecir la clase a la que pertenece cada registro de la base de datos según variables predictoras independientes entre sí [15].

Los métodos Bayesianos se basan en el teorema de Bayes para pasar de la probabilidad a priori de un suceso P(suceso) a la probabilidad a posteriori P(suceso/observaciones). El aprendizaje en el clasificador bayesiano consiste en estimar las diferentes probabilidades en términos de sus frecuencias sobre los registros de la base de datos. La probabilidad de que un registro pertenezca a una clase está dada por el teorema de probabilidad condicional de Bayes [16].

Los métodos basados en ejemplos también son llamados clasificadores perezosos ya que no realizan ninguna labor en la etapa de entrenamiento, sólo almacenan los datos. El algoritmo de los K – vecinos más cercanos, KNN (K-Nearest Neighbor), es el más utilizado. Cuando se tienen nuevos datos para clasificar, el algoritmo busca los k registros más cercanos según funciones de distancia. Finalmente, el algoritmo asigna la clase a la que pertenece la mayoría de los registros vecinos [15].

2.2.2 Técnicas No Supervisadas

Por otro lado, las técnicas no supervisadas son aplicadas en el análisis descriptivo. Algunas técnicas no supervisadas son: Método Particional, Método Jerárquico, Método Probabilístico, Redes Neuronales y Reglas de Asociación.

Los método particionales dividen el conjunto de datos en un número predefinido de clusters [17] [18]. K-means es el algoritmo más popular por su simplicidad y eficacia. El objetivo del algoritmo es encontrar k centroides, uno por cada cluster, de tal manera que los centroides sean lo más alejados posibles según funciones de distancia y los datos son asociados al centroide más cercano [19].

Los métodos jerárquicos permiten encontrar estructuras de clustering de forma recursiva, utilizando dendogramas o árboles binarios. En el dendograma la raíz representa la población, los nodos intermedios simbolizan la proximidad entre los datos y las hojas representan los datos de la población [17] [18].

Los métodos probabilísticos asumen que los datos son generados de acuerdo a distribuciones de probabilidad [17] [20]. Expectation – Maximization es el algoritmo probabilístico más comúnmente usado, el cual asigna una distribución de probabilidad a cada cluster y ajusta los parámetros con los datos.

Las redes neuronales también son utilizadas en la búsqueda de clusters. El algoritmo más utilizado es SOMs (Self Organizing Maps), el cual tiene una arquitectura de neuronas hexagonal o rectangular. Las neuronas están conectadas entre sí con una relación de vecindad y se usa la regla de aprendizaje de Kohonen para buscar la neurona más cercana a cada uno de los datos [21].

Finalmente, las Reglas de Asociación se utilizan para descubrir relaciones frecuentes entre los datos [22]. Apriori es el algoritmo más ampliamente utilizado para detectar asociaciones, el cual se basa en el conocimiento previo de los datos en cada iteración.

2.3 Metodologías de Minería de Datos

Existen diversas metodologías que proporcionan una serie de pasos a seguir con el fin de realizar una implementación adecuada de la minería de datos. Según sondeos publicados en KDnuggets[1], las metodologías más utilizadas son: CRISP-DM, SEMMA, KDD y Catalyst.

CRISP-DM (*Cross Industry Standard Process for Data Mining*) fue concebida desde un enfoque

[1] KDnuggets: Data Mining Community Top Resource <http://www.kdnuggets.com/>

práctico de acuerdo la experiencia de sus creadores: un consorcio de empresas europeas, incluyendo SPSS de IBM. Actualmente CRISP-DM es la guía de referencia más utilizada en el desarrollo de proyectos de minería de datos [23], [24], [25]. Está constituida por seis fases: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue.

La metodología SEMMA (*Sample, Explore, Modify, Model and Assess*) es la propuesta de *SAS Analytics Solutions* para desarrollar proyectos de minería de datos. La metodología establece cinco fases: muestreo, exploración, modificación, modelado y evaluación [26]. Se caracteriza por incluir una fase de muestreo estadístico que no se considera en otras metodologías.

KDD (*Knowledge Discovery in Database*) se conoce como el descubrimiento de conocimiento en bases de datos como un proceso no trivial donde se identifican patrones válidos, novedosos, potencialmente útiles y en última instancia entendibles en los datos [27]. Algunos autores consideran a la minería de datos como una etapa en el de KDD [4]. Sin embargo, según las encuestas de KDnuggets, se está utilizando KDD como metodología para hacer minería de datos.

Catalyst también es conocida como la metodología P3TQ (*Product, Place, Price, Time, Quantity*). Las relaciones ente estas variables buscan mantener el producto correcto, en el lugar adecuado, en el momento adecuado, en la cantidad correcta y con el precio correcto. Esta metodología plantea la formulación de dos modelos: el modelo de negocios y el modelo de minería de datos [28].

2.4 Plataformas de Minería de Datos

Las plataformas de minería de datos son herramientas que facilitan la aplicación de las técnicas de la minería de datos. Algunas plataformas son: WEKA, RapidMiner, R, SPSS *Modeler* y SAS *Enterprise Miner*.

WEKA (*Waikato Environment for Knowledge Analysis*) ha sido diseñada por un grupo de desarrolladores de la universidad de Waikato en Nueva Zelanda, y se distribuye bajo licencia GNU, es decir que es posible modificar el código fuente para adicionar nuevas funcionalidades [29]. La herramienta WEKA permite realizar tareas de clasificación, regresión, clustering, asociación y visualización. Una de las características más atractivas es su capacidad de extensibilidad, es decir, que añadir nuevas funcionalidades es una tarea sencilla [30].

RapidMiner es una herramienta de minería de datos desarrollada en el año 2001 por el departamento de inteligencia artificial de la Universidad de Dortmund. Entre sus principales ventajas se destaca que es multiplataforma, de código abierto y con licencia GPL. RapidMiner permite analizar y extraer datos a través de unos operadores, utilizando para ello un entorno gráfico [31].

R es un software desarrollado para realizar análisis de datos y presentar como resultado cálculos estadísticos y gráficas que permiten extraer información valiosa de los datos. Fue desarrollada por los científicos Robert Gentleman y Ross Ihaka del departamento de estadística de la Universidad de Auckland de Nueva Zelanda [32].

SPSS es un paquete estadístico que contiene una serie de herramientas que permiten realizar análisis de datos. Una de estas herramientas está diseñada para realizar tareas de la minería de datos, se trata de SPSS Modeler [33], esta herramienta permite desarrollar modelos predictivos orientados a mejorar la toma de decisiones.

SAS *Institute* comercializa diferentes paquetes y productos, entre ellos se encuentra SAS *Analytics*, el cual permite el modelado predictivo y descriptivo en minería de datos [34]. Esta herramienta se complemente con módulos de visualización, investigación de operaciones, estadística y procesos de calidad.

Se han realizado diversas comparaciones de las herramientas para hacer minería de datos [35] [36]. Las características comúnmente comparadas son: cantidad de descargas de internet, popularidad, área de trabajo, capacidad para integrarse con otro software, tipo de licencia y capacidad para manejar extensa cantidad de registros. En estas comparaciones se resaltan las herramientas WEKA, R y RapidMiner por ser las más descargadas desde Internet y con una alta popularidad entre los profesionales.

3. APORTES DE LA MINERÍA DE DATOS AL SECTOR SALUD

A continuación se presenta una revisión sobre estudios de minería de datos en el área de la salud organizados en dos tipos de aplicaciones: estudios de enfermedades, estudios sobre la prestación del servicio de salud.

3.1 Estudios de Enfermedades

Se han realizado numerosas aplicaciones de minería de datos al estudio de enfermedades, diagnósticos y tratamientos. Algunas de ellas son: cáncer de próstata, enfermedades cardiovasculares, hipertensión arterial, cáncer de mama, parkinson y enfermedades tumorales, cáncer de cuello uterino, diabetes, dengue, entre otras. Estas aplicaciones se describen a continuación.

En [37] se presenta un estudio predictivo para determinar la eficacia de la braquiterapia en el tratamiento de cáncer de próstata utilizando minería de datos. Para tratar enfermedades complejas como el cáncer la elección del tratamiento se debe tener en cuenta factores como los riesgos de la terapia, la edad de los pacientes y la calidad de vida luego de realizar el tratamiento. Esta situación evidencia la necesidad de herramientas para mejorar la toma de decisiones al momento de escoger el tratamiento adecuado para un paciente. En este estudio se utilizan los árboles de decisión como técnica de clasificación.

En [7] se presenta un estudio predictivo que compara las técnicas de clasificación de la minería de datos aplicada a las enfermedades cardiovasculares. Para realizar esta comparación se utilizó un conjunto de datos de pacientes con enfermedades cardiovasculares que cuenta con 14 atributos y 303 registros. Como resultado del estudio se obtuvieron mejores resultados al utilizar las máquinas de soporte vectorial como método clasificador. En [38] también se presenta un estudio predictivo para el diagnóstico de enfermedades cardiovasculares. En este estudio se utilizan las redes bayesianas y los arboles de decisión para realizar tareas de predicción relacionadas con esta enfermedad. Como resultado del estudio se presenta una alternativa para determinar si se debe realizar un procedimiento clínico a un paciente basándose en variables como la presión arterial.

En [39] se presenta un estudio cuyo objetivo es realizar la predicción de la hipertensión arterial, este estudio toma importancia al conocer que el 30% de las muertes a nivel mundial son producidas por enfermedades cardiovasculares. Para realizar este estudio se tomó una muestra de 138 personas entre los 20 y 34 años de edad, la técnica utilizada en este caso fueron las redes neuronales.

En [40] se realiza un estudio predictivo con una caracterización y análisis de las base de datos de cáncer de mama del programa de vigilancia, epidemiología y resultados finales (SEER) del Instituto Nacional de Cáncer de los Estados Unidos por medio de series de tiempo temporales. Uno de los objetivos del estudio consiste en establecer los factores que más influyen en la enfermedad. Como resultado del estudio se resalta que el cáncer de mama es una de las principales causas de muerte en mujeres menores de 30 años.

En [8] se presenta un prototipo para la predicción de Parkinson y enfermedades tumorales primarias, utilizando técnicas de minería de datos. En este estudio se platea una comparación de los resultados utilizando técnicas como redes neuronales, árboles de decisión y métodos bayesianos, obteniendo un mejor resultado al utilizar las redes neuronales.

En [41] se realiza un estudio descriptivo para descubrir patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino. En este estudio se utiliza la asociación para determinar los factores socioeconómicos y clínicos asociados a la supervivencia de mujeres con esta enfermedad. El objetivo de este estudio es apoyar la toma de decisiones relacionadas con políticas públicas y programas de protección para las mujeres que padecen esta enfermedad.

En [42] se presenta una aplicación de la minería de datos a pacientes pre-diabéticos. El objetivo del estudio es detectar patrones de comportamiento para apoyar la toma de decisiones en futuros casos de personas con esta enfermedad. Para ello, el estudio, extrajo información de los expedientes clínicos para descubrir y conocer las características como edad, sexo, niveles de colesterol, triglicéridos, glucosa e insulina. Para en análisis de datos se utilizó un clustering con el método K-means y una predicción con árboles de decisión.

En [10] se presenta un estudio descriptivo donde se pretende extraer las reglas de asociación para minimizar los efectos del dengue utilizando minería de texto. En este estudio se destaca la importancia de estudiar la enfermedad, ya que el 40% de la población mundial vive en zonas donde hay transmisión del dengue. El objetivo del estudio es sugerir de forma proactiva las próximas ubicaciones geográficas donde esta enfermedad pueda llegar a tener influencia, esto con el fin de que los gobiernos puedan tomar medidas al respecto.

3.2 Estudios de la Prestación del Servicio de Salud

Diferentes estudios de minería de datos han logrado mejorar el servicio de urgencias de los hospitales. En el trabajo presentado en [43], se utilizan las redes bayesianas en el triaje de pacientes de urgencias. Un análisis similar se presenta en [44], donde se utiliza el algoritmo de clúster k-means para encontrar patrones de atenciones en el servicio de urgencias y se utiliza árboles de decisión para analizar el triaje de los pacientes.

En [45], se realiza un modelo de minería de datos para estimar la demanda de la sala de urgencias de un hospital pediátrico público de la ciudad de Santiago de Chile. Para el estudio se utilizan las técnicas de regresión lineal, red neuronal y regresión de soporte vectorial.

En [46] se desarrolla un modelo de minería de datos para mejorar el servicio de emergencias del Hospital Guayaquil. En el estudio se utiliza una regresión para predecir la cantidad de emergencias esperadas, se utiliza clustering para analizar el personal médico y se utilizan algoritmos genéticos para la planeación de guardias médicas.

4. TENDENCIAS DE LA MINERÍA DE DATOS EN EL SECTOR SALUD

Con el objetivo de establecer nuevas líneas de trabajo de la minería de datos para apoyar el sector salud en el marco de ciudades inteligentes, en esta sección se discuten algunas tendencias encontradas en la revisión bibliográfica que se traducen en desafíos investigativos, los cuales son impulsados por el área de minería multimedia, explosión de *big data* en las ciudades e iniciativas nacionales e internacionales de liberar datos de la ciudad para beneficio de todos (*open data*) [2].

4.1 Análisis de Tipos de Datos No Estructurados

En el área de la salud un gran porcentaje de los datos se encuentran en imágenes o texto. Sin embargo, la minería de datos se aplica convencionalmente en datos estructurados, es decir información organizada en bases de datos. Los análisis de minería de datos NO estructurados son un requerimiento nuevo y exigente que permitirá procesar información multimedia, dando lugar a nuevas áreas de interés: minería de texto y minería de imágenes. Estas áreas pretenden desarrollar análisis predictivos y descriptivos a información multimedia de la salud.

4.1.1 Minería de Texto

La minería de texto [47] [48], hace referencia al proceso mediante el cual se puede extraer patrones o conocimiento no trivial, a partir de documentos de texto. Esta área ha encontrado diversas aplicaciones en el área de la salud. En [49] se reconoce que los instrumentos actuales para el tratamiento de la información médica que se tiene disponible en formato electrónico, no son los más adecuados para realizar estas tareas. Así mismo se reconoce en la minería de texto una herramienta valiosa para gestionar grandes volúmenes de información y generar nuevo conocimiento a partir de dicha información.

4.1.2 Minería de Imágenes

La minería de imágenes [50] [51] se interesa por extraer patrones característicos a partir de un gran número de imágenes. La minería de imágenes también proporciona soluciones a los problemas del sector salud. En [52] se presenta un sistema que permite detección inteligente de ojos somnolientos. Este sistema fue pensado como una herramienta para ayudar a los conductores de vehículos a mantener su atención en la vía y no quedarse dormidos al volante. Para ello se capturaron cerca de dos mil imágenes de los ojos y se probaron varias técnicas de la minería de datos con el fin de encontrar un sistema que pueda funcionar en tiempo real y dar solución al problema planteado.

[2] MinTIC Colombia: Datos Abiertos para el país <http://www.mintic.gov.co/portal/604/w3-article-5664.html>.

4.2 Metodologías con Fases de Preprocesamiento e Indexamiento de Datos No Estructurados

Aunque las metodologías en su mayoría incluyen fases de preparación de los datos, dicha preparación sólo incluye análisis estadísticos y transformaciones. Para analizar datos multimedia es necesario incluir etapas de preprocesamiento e indexamiento de los datos, donde se pueda representar la información multimedia en vectores de características que puedan ser procesados por las técnicas de minería de datos. Como un acercamiento a esta etapa de preprocesamiento, algunos autores han modificado la metodología CRISP-DM para realizar minería multimedia. En [53] se presenta una aplicación de algoritmos de clasificación de minería de textos, a pesar de tratarse de datos no estructurados, se utiliza la metodología CRISP-DM. Un caso similar se presenta en [36] donde se utiliza la minería de texto en el diseño de un modelo de clasificación de opiniones subjetivas utilizando la metodología CRISP-DM.

4.3 Herramientas con Soporte a Minería Multimedia

Dentro de las investigaciones revisadas sobre minería de datos aplicada a la salud, se ha notado cierta preferencia por la herramienta WEKA para aplicar las técnicas de la minería de datos. En [54] se resalta el prestigio y la popularidad de la herramienta WEKA, y se emplea esta herramienta con algunas técnicas de minería de datos aplicadas al diagnóstico de enfermedades y servicios de salud. Aunque WEKA recientemente incluye métodos para el procesamiento de textos en español, no presenta soporte para minería de imágenes. Otras herramientas como RapidMiner, SPSS Modeler y SAS Analytics también incluyen módulos para minería de textos, pero no para minería de imágenes. Este es un requerimiento nuevo y exigente para las herramientas.

5. CONCLUSIONES

Basados en que las ciudades inteligentes tienen como objetivo mejorar la calidad de vida de los ciudadanos, en este trabajo se presentó una revisión de aportes y tendencias en el análisis de datos y la toma de decisiones en los servicios de salud por medio de minería de datos.

La minería de datos es un área que presenta avances en diferentes líneas de trabajo como técnicas, metodologías y plataformas. En las aplicaciones relacionadas con servicios de salud se pueden observar las siguientes preferencias:

- En el análisis predictivo se utilizan con mayor frecuencia las redes neuronales y los árboles de decisión.
- En el análisis descriptivo se utiliza frecuentemente el método particional k-means.
- Como metodología de minería de datos para aplicaciones de salud se utiliza con mayor frecuencia CRISP-DM.
- Finalmente, en la etapa de desarrollo WEKA se desataca como la herramienta preferida en aplicaciones de salud.

La revisión en aplicaciones de minería de datos a los servicios de salud permitió establecer las siguientes tendencias y requerimientos para el área:

- Análisis de tipos de datos no estructurados como texto e imágenes, ya que gran porcentaje de los datos utilizados en los servicios de salud se encuentran en formato multimedia.
- Metodologías con etapas de preprocesamiento e indexamiento de datos no estructurados, ya que las metodologías actuales sólo incluyen análisis estadísticos y transformaciones de las bases de datos.
- Herramientas con soporte a minería multimedia, aunque algunas herramientas ya cuentan con soporte a procesamiento de texto, es necesario incluir el procesamiento de imágenes para así realizar modelamiento de servicios de salud con imágenes diagnósticas.

Cada una de las tendencias encontradas permite establecer una ruta para trabajos futuros en el área. Otros trabajos futuros son: soporte para *big data* y herramientas con conectividad a bases de datos abiertas de las ciudades.

6. RECONOCIMIENTOS

Los autores expresan su agradecimiento al grupo de investigación GIDATI de la Universidad Pontificia Bolivariana. Este documento es resultado del proyecto “Plataforma de Minería de Datos Estructurados y No Estructurados - Caso de Estudio Salud Pública”.

7. REFERENCIAS

[1] Rodríguez C, Gil S. Ciudades amigables con la edad, accesibles e inteligentes, CEAPAT-IMSERSO, 2014.

[2] Chen CC. The Trend towards “Smart Cities", International Journal of Automation and Smart Technology, 4(2), 63-66, 2014.

[3] Achaerandio R, Curto J, Bigliani R, Gallotti G. Análisis de las ciudades inteligentes en España 2012 - El viaje a la ciudad inteligente. En: IDC España - Analize the future, 2012.

[4] Riquelme J, Ruiz R, Gilbert K. Minería de datos: Conceptos y tendencias, Revista Iberoamericana de Inteligencia Artificial, 10(29), 11-18, 2006.

[5] Wu X, Zhu X, Wu GQ, Ding W. Data mining with big data, IEEE Transactions on Knowledge and Data Engineering, 26(1), 97-107, 2014.

[6] Mena J. Data mining your website, Digital Press, 1999.

[7] Kumari M, Sunila G. Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction, 2, 2011.

[8] Shaikh T. A Prototype of Parkinson’s and Primary Tumor Diseases Prediction Using Data Mining Techniques, International Journal of Engineering Science Invention, 3(4), 2014.

[9] Jain, A. K., Murty, M. N., y Flynn, P. J. Data clustering: a review.,ACM computing surveys (CSUR), 31(3), 264-323, 1999.

[10] Amin A, Takib R, Raza S, Javed S. Extract association rules to minimize the effects of dengue by using a text mining technique. 3(4), 2014.

[11] Han J, Kamber M. Data Mining Concepts and Techniques, Morgan Kaufmann; 2006.

[12] Wiener E, Jan P, Weigend A. A Neural Network Approach to Topic Spotting, En: 4th Annual Symposium on Document Analysis and Information Retrieval; Las Vegas. p. 317-332, 1995.

[13] Apté C, Weiss S. Data mining with decision trees and decision rules, Future Generation Computer Systems, 197-210, 1997.

[14] Joachims T, Hofmann T, Yue Y, Yu CN. Predicting structured objects with support vector machines, Communications of the ACM. 52(11), 97-104, 2009.

[15] Yang Y, Liu X. A re-examination of text categorization methods, En: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 42-49, 1999.

[16] Wettig H, Grünwald P, Roos T, Myllymäki P, Tirri H. On supervised learning of Bayesian network parameters, Helsinki Institute for Information Technology (HIIT), 2002.

[17] Xu R, Wunsch D. Survey of clustering algorithms,Neural Networks, IEEE Transactions on 16(3), 2005.

[18] Filippone M, Camastra F, Masulli F, Rovetta S. A survey of kernel and spectral methods for clustering, Pattern recognition.41(1), 176-190, 2008.

[19] Steinley D. K-means clustering : A half-century synthesis, British Journal of Mathematical and Statistical Psychology. 59(1), 1-34, 2006.

[20] François O, Ancelet S, Guillot G. Bayesian clustering using hidden Markov random fields in spatial population genetics, Genetics.174(2), 805-816, 2006.

[21] Meireles M, Almeida P, Godoy M. A comprehensive review for industrial applicability of artificial neural networks, IEEE Transactions on Industrial Electronics, 50(3), 2003.

[22] Slimani T, Amor L. Efficient Analysis of Pattern and Association Rule Mining Approaches, arXiv preprint arXiv:1402.2892, 6(3), 70-81, 2014.

[23] Moine J. Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo [Tesis de Maestría]. Argentina: Universidad Nacional de la Plata, 2013.

[24] Torres D. Diseño y aplicación de una metodología para análisis de noticias policiales utilizando minería de textos [Tesis de Pregado]. Chile: Universidad de Chike, 2013.

[25] Corrales D, Ledesma A, Peña A, Hoyos J, Figueroa A, Corrales J. A new dataset for coffee rust detection in Colombian crops base on classifiers. Revista S&T, 9-23, 2014.

[26] Azevedo A, Rojão L. KDD, SEMMA and CRISP-DM: a parallel overview. ; 2008.

[27] Usama F, Piatetsky-Shapiro G, Padhraic S, Uthurusamy R. Advances in knowledge discovery and data mining, The MIT Press, 1996.

[28] Pyle D. Business modeling and data mining: Morgan Kaufmann, Morgan Kaufmann, 2003.

[29] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann , Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explorations, 10-18, 2009.

[30] Azoumana K. Análisis de la deserción estudiantil en la Universidad Simón Bolívar, facultad Ingeniería de Sistemas, con técnicas de minería de datos, Pensamiento Americano, 41-51, 2013.

[31] Hofmann M, Ralf K. RapidMiner: Data Mining Use Cases and Business Analytics Applications, CRC Press, 2013.

[32] Zhao, Y. R and data mining: Examples and case studies, Academic Press, 2012.
[33] Devi B, Rao K, Setty S, Rao M. Disaster Prediction System Using IBM SPSS Data Mining Tool, International Journal of Engineering Trends and Technology (IJETT), 3352-3357, 2013.
[34] Fernandez, G. Data mining using SAS applications. CRC press, 2002.
[35] Mikut, R. and Reischl, M. Data mining tools, WIREs Data Mining Knowl Discov, 1: 431–443, 2011.
[36] Tapia M, Ruiz O, Chirinos C. Modelo de clasificación de opiniones subjetivas en redes sociales, Ingeniería: Ciencia, Tecnología e Innovación, 2014.
[37] Reparaz D, Merlino H, Rancán C, Rodríguez D, Britos P, García R. Determinación de la eficacia de la braquiterapia en tratamiento de cáncer basda en minería de datos. En: X Workshop de Investigadores en Ciencias de la Computación, 2008.
[38] Solarte R, Castro YV. Modelo híbrido para el diagnóstico de enfermedades cardiovasculares basado en inteligencia artificial, Tecnura, 35-52, 2012.
[39] Pérez A. Aplicación de la red de probabilidad neuronal y escala de framingham para predicción de la hipertension arterial. En: Memorias Convención Internacional de Salud Públic,; La Habana, 2012.
[40] Molero G, Céspedes Y, Campaña M. Caracterización y análisis de la base de datos de cáncer de mama SEER-DB. En: IX Congreso Internacional Informática en Salud, 2013.
[41] Timarán R, Yépez M. La minería de datos aplicada al descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino, Universidad y salud, 117-129, 2012.
[42] Hernández H. Aplicación de minería de datos a información de pacientes pre-diabéticos, Revista Iberoamericana de Producción Académica y Gestión Educativa, 2014.
[43] Abad Grau M, Lerache J, Cervino C. Aplicación de Redes Bayesianas en el modelado de un sistema experto de triaje en servicios de urgencias médicas En: IX Workshop de Investigadores en Ciencias de la Computación, 2007.
[44] Vergara Silva CL. Mejora en la gestión de recursos y calidad del servicio en el proceso de atención de urgencias en el Hospital Dr. Sótero del Río [Tesis de Maestría]. Chile: Universidad de Chile, 2012.
[45] Reveco C, Weber R. Gestión de Capacidad en el Servicio de Urgencia en un Hospital Público, Revista Ingeniería de Sistemas XXV, 2011.
[46] Echeverria Briones PF, Aviles Monroy JA, Navarro T, Toapaxi C. Sistema De Predicción Y Clasificación Para La Utilización De Recursos Humanos Para El Área De Emergencias De Un Hospital [Tesis de Pregrado]. Ecuador: Escuela Superior Politécnica del Litoral, 2009.
[47] Sebastiani F. Machine learning in automated text categorization, ACM Computing Surveys, 34(1), 1-47, 2002.
[48] Zhong N, Li Y, Wu ST. Effective Pattern Discovery for Text Mining, Transactions on knowledge and data engineering, 24(1), 30-44, 2010.
[49] Piedra D, Antoni F, Joaquim G. Minería de textos y medicina: utilidad en las enfermedades respiratorias, Archivos de Bronco neumología, 50(3), 113-119, 2014.
[50] Choubassi M, Nefian A, Kozintsev I, Bouguet J, Wu Y. Web image clustering, Acoustics, Speech and Signal Processing, 2007.
[51] Hsu W, Mong LL, Ji Z. Image Mining: Trends and Developments, Journal of Intelligent Information Systems, 19(1), 7-23, 2002.
[52] Emam A. Intelligent drowsy eye detection using image mining, Information Systems Frontiers, 1-14, 2014.
[53] Santana P, Costaguta R, Missio D. Aplicación de Algoritmos de Clasificación de Minería de Textos para el Reconocimiento de Habilidades de E-tutores Colaborativos, Revista Iberoamericana de Inteligencia Artificial, 57-67, 2014.
[54] Dávila F, Sánchez Y. Técnicas de minería de datos aplicadas al diagnóstico de enfermedades clínicas, Revista Cubana de Informática Médica, 2012.