APLICACIÓN DEL MODELO DE CLUSTERIZACIÓN BASADO EN EL ALGORITMO DE K-MEANS PARA LA SEGMENTACIÓN DE LA MORBILIDAD MATERNA EN EL HOSPITAL SAN BARTOLOMÉ DE LA CIUDAD DE LIMA-2012

Autores: Keyla De La Cruz Gutiérrez, Joel Cieza Rivasplata, Caleb Flores Sahuanga
Universidad Peruana Unión - LIMA
E-mails: dervyth19@gmail.com;jocir@gmail.com; everthflores@gmail.com

Resumen:

La presente investigación tiene como objetivo segmentar las causas de la morbilidad materna del hospital San Bartolomé, aplicando el modelo de clustering basado en el algoritmo de K-Means de SQL Server. La metodología usada fue Crisp-DM. Los resultados del algoritmo mostraron el modelo de minería de datos presentado por diez grupos segmentados de acuerdo a la mayor similaridad que presentan entre ellos y las relaciones que se dan entre cada grupo.

Palabras claves: Clustering, K-means, Morbilidad, CRISP-DM, Data Mining.

1. INTRODUCCIÓN

En los últimos años el desarrollo y los avances tecnológicos han marcado a la sociedad actual, volviéndola cada vez más compleja, por lo que son necesarias herramientas y algoritmos de inteligencia artificial para identificar patrones de los pacientes. Por otro lado, la gran cantidad de información que comienzan a manejar las empresas se incrementan día a día, lo que significa que para efectuar el análisis se requiere de potentes herramientas que extraigan la información necesaria [1].

Data Mining es una de las herramientas más utilizadas para determinar patrones de comportamiento. Dentro de la minería de datos, basadas en el aprendizaje automático, están los modelos de clusterización que permiten identificar grupos donde los atributos guardan similitudes entre sí y diferentes características con los otros.

La principal característica de esta técnica es la utilización de una medida de similaridad que, en general está basada en los atributos que describen a los objetos y se define usualmente por proximidad en el espacio multidimensional. Uno de los algoritmos más utilizados para hacer clustering es el K-Means, que se caracteriza por su sencillez y su objetivo fue reducir la cantidad de datos mediante la caracterización o agrupamiento de datos según las características similares.

Por esta razón, a la base de datos de morbilidad materna aplicamos el algoritmo de K-Means para determinar las características similares de las madres embarazadas que se atiende en el hospital San Bartolomé, de esa manera conocer las causas que ocasionan la morbilidad de los pacientes.

En el primer apartado se presenta un marco teórico que facilita una mejor compresión del artí-

culo. El segundo contiene la descripción general de la metodología a desarrollar en el trabajo, seguido del algoritmo a utilizar; posteriormente se presentan los resultados y el análisis, finalmente las conclusiones de la investigación.

2. MARCO TEÓRICO

A continuación se presentan algunas definiciones que facilitan una mejor comprensión de los apartados de este artículo.

- Agrupamiento o Clustering: Es un procedimiento de agrupación de una serie de ítem permitiendo utilizar múltiples atributos para identificar grupos similares de una manera no supervisada, sin la necesidad de etiquetar los grupos[2][3]. Por ello al análisis de clustering se le conoce como método de clasificación automática no supervisada que consiste "en encontrar la partición más adecuada del conjunto de entrada a partir de similitudes entre sus ejemplos" [4]. Es decir que, la idea de formar clústeres es agrupar elementos en conjuntos homogéneos en función de algunas semejanzas entre ellos y diferentes a los que pertenecen a otros grupos.

- Algoritmo de K-Means: Creado por MacQueen en 1967. Este algoritmo es popular por ser de fácil implementación, y porque su complejidad es del orden del número de objetos [4]. El mayor problema con este algoritmo es que es muy sensible a la partición inicial seleccionada, y puede converger fácilmente. Sin embargo, es el más conocido y utilizado ya que es de muy simple aplicación y eficaz. Sigue un procedimiento simple de clasificación de un conjunto de objetos en un determinado número de clústeres, determinado a prioridades [4]. K- means representa cada uno de los clústeres por la media de sus puntos, es decir, por su centroide, y así cada clúster es caracterizado por este, el cual se encuentra en el centro de los elementos que componen el clúster [3].

  K- means es traducido como K-medias y se realiza en 4 etapas [5] fundamentales:

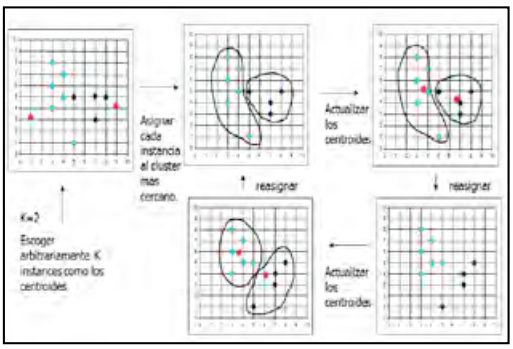

Figura 1. – Funcionamiento algoritmo de K-Means

Etapa 1: Consiste en elegir aleatoriamente K objetos que forman así los K clústeres iniciales. Para cada uno de los clúster k, el valor inicial del centro es igual xi, siendo estos valores xi únicos objetos de Dn= (x1, x2, x3…, xn) pertenecientes al clúster, para todo i real.

Etapa 2: En esta etapa se reasignan los objetos del clúster. Para cada objeto x, el prototipo que se le asigna es el más próximo al objeto que se encuentra, según la medida de distancia (regularmente la medida euclidiana). Esta medida se realiza en la etapa de modelización de la metodología de CRISP-DM y muy importante debido a que gracias a ella se verá la validez del modelo.

Etapa 3: Una vez que todos los objetos son colocados, se recalculan nuevamente los centros de los K clúster (los baricentros).

Etapa 4: Repetir las etapas 2 y 3 hasta que no se hagan más reasignaciones. Aunque el algoritmo termina siempre, no se garantiza el obtener la solución óptima. En efecto, el algoritmo es muy sensible a la elección aleatoria de los k centros iniciales. Esta es la razón por la que, se utiliza el algoritmo de K-means numerosas veces sobre el mismo conjunto de datos para intentar minimizar este efecto, sabiendo que a centros iniciales los más espaciados posibles dan mejores resultados.

- Morbilidad: Según la Real Academia Española (RAE) la morbilidad es definida como una proporción de personas que enferman en un sitio y tiempo determinado. La morbilidad es, entonces, un dato estadístico de altísima importancia para poder comprender la evolución y avance o retroceso de una en-

fermedad, así también como las razones de su surgimiento y las posibles soluciones.

3. MATERIALES Y MÉTODOS

3.1. Analysis Service (SASS) de SQL Server 2008

El algoritmo de k-medias clustering es el referente principal entre los diversos métodos para seleccionar grupos representativos entre los datos. El componente SASS de SQL Server 2008 incluye el algoritmo de K-medias clustering para segmentar en grupos. El último lanzamiento de Microsoft SQL Server ofrece una plataforma de datos completa, más segura, confiable, administrable y escalable para aplicaciones críticas. Permite que los desarrolladores creen aplicaciones nuevas, capaces de almacenar y consumir cualquier tipo de datos en cualquier dispositivo, y que todos los usuarios tomen decisiones informadas en base a conocimientos relevantes. Incluye varios algoritmos de aprendizaje automático para analizar grandes bases de datos y de esa manera generar conocimiento.

3.2. Metodología Crisp-DM

La metodología Crisp – DM, es un estándar para la realización de proyectos de minería de datos, en donde el ciclo de un proceso de minería de datos está estructurado en 6 fases. Algunas de estas fases son bidireccionales, lo que significa que algunas fases permitirán revisar parcial o totalmente las fases anteriores [7][8]

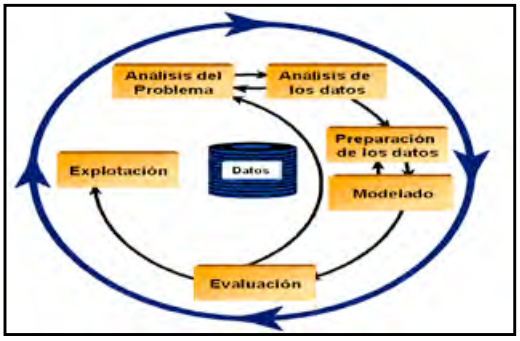

Figura 3. -Modelo de la metodología CRISP - DM

Fase de análisis del negocio o problema. Es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto [9].

Fase de análisis de los datos. Es necesario familiarizarse con los datos teniendo presente los objetivos del negocio, para ello debe cumplir las siguientes actividades. Recopilación inicial de datos, descripción de los datos, exploración de los datos, verificación de calidad de datos [7].

Fase de preparación de los datos. En esta etapa se desarrolló la selección, limpieza, e integración de los datos. Y una de las herramientas más eficaces para integrar los datos es el Analysis Services de SQL Server.

Fase de modelado. En esta fase se realizará la selección de la técnica de modelado, seguido del diseño de la evaluación, construcción del modelo y posteriormente la evaluación del modelo.

Fase de evaluación. Aquí se realizó la evaluación de los modelos de las fases anteriores para determinar si son útiles a las necesidades del negocio y debe considerarse lo siguiente: Evaluación de resultados y revisar el proceso.

Fase de explotación o implementación. Una vez que el modelo ha sido construido y validado, se transformará el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación del modelo y sus resultados, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso [4].

4. APLICACIÓN DE LA METODOLOGÍA Crisp-DM

Comprensión del negocio. Se hizo un estudio a los datos de las madres que sufren de morbilidad de la base de datos del hospital San Bartolomé. Luego, se buscó la familiarización con los términos que se utilizaron a lo largo del proyecto, estableciendo los objetivos del proyecto y la elaboración del plan del proyecto [10].

Comprensión de los datos. En esta fase se realizó el estudio de los datos obtenidos, y se evaluó la lógica de relación entre los datos; además se logró dar sentido a las variables que se utilizan en la data.

Preparación de los datos. En esta etapa se realizó la selección y limpieza de los datos más relevantes que se presentan en el cuadro 1, mostrando las variables más influyentes en la morbilidad de las madres gestantes, como también su descripción [11].

Limpieza de los datos. En esta fase se procedió a seleccionar los datos de las madres gestantes, luego se realizó la limpieza de los datos teniendo como actividades: eliminación de los datos nulos y conversión de los datos para un mejor significado y/o entendimiento.

Normalización de los datos. En esta parte se procedió a relacionar los datos que fueron seleccionados y limpiados, luego, se procedió a realizar una pequeña base de datos relacional para un mejor entendimiento de los datos.

Desnormalización de los datos. Esta fase consistió en volver a juntar todos los datos que se seleccionaron y limpiaron para prepararlos para la selección de indicadores.

Encontrar los indicadores. En esta fase, se realizó la selección de indicadores que nos servirán para la construcción del modelo de clustering. En nuestro caso se eligió como variable de entrada y predictora a la morbilidad. Según este atributo vamos a generar el modelo de clustering.
Modelado. Para esta etapa se utilizó el componente de SQL Server 2008 el Analysis Services. Esta herramienta ya tiene incluido el algoritmo de K-medias, la cual agrupa los datos obtenidos según las características más comunes entre ellas. En el presente trabajo de investigación se establecieron las variables de entrada y la de predicción presentado en el cuadro 1. En este caso la variable de predicción vendría a ser la morbilidad de las gestantes.

1. RESULTADOS Y ANÁLISIS

El modelo de clusterización basado en el algoritmo de K – medias muestra los grupos que arrojo el modelo en función a la población total y tomando como variable predictor a la morbilidad. Se puede apreciar que existe mucha igualdad en sus características en los clústeres 1, 2, 3, 4 y la relación más fuerte se da entre los clústeres 1-8, 3-7 y 9-5.

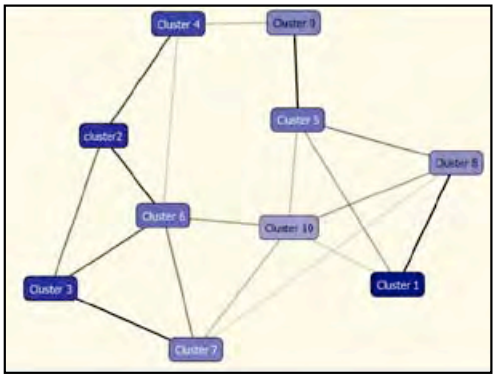

Figura 4. –Modelo de Clustering

VARIABLE	DESCRIPCIÓN
Morbilidad	Representa el estado de la madre, en este caso son dos: sanas y patológicas.
Hemorragia	Representa si las gestantes presentaron este enfermedad.
Infección en el embarazo	Representa si presentaron infección durante el embarazo.
Infección puerperal	Representa si las gestantes presentaron este tipo de mal.
Ocupación de la madre	Representa al tipo de ocupación que tienen las madres.
Fecha última de menstruación	Representa si la gestante conoce o no su fecha última de menstruación.
Aborto	Representa a qué nivel presentaron este mal.
Gestas	Cantidad que embarazos que tuvieron anteriormente sin ningún tipo de complicación.
Embarazo múltiple	Cantidad de embarazos múltiples.
Preeclampsia	Representa si tuvieron complicaciones durante el embarazo.
Eclampsia	Representa si presentaron convulsiones durante el embarazo.
Desgarros	Representa al grado de desgarros que presentaron las madres del hospital.
Parto prolongado	Representa todo aquel cuya duración del parto es mayor a 294 días o 42 semanas.

Cuadro 1. Variables para el estudio de la morbilidad

La población estudiada es de 1686 madres gestantes del hospital San Bartolomé está segmentada en 10 clústeres como se puede observar en la figura 4.

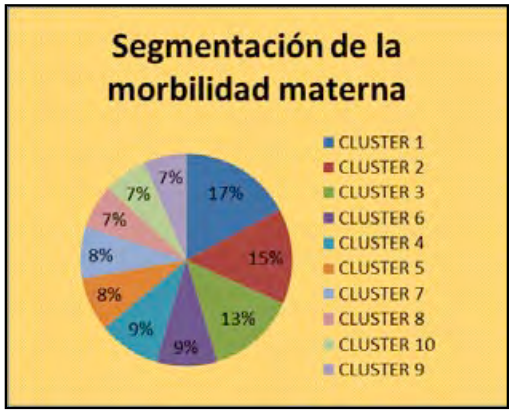

Figura 5. –Segmentación de la población de madres gestantes en función a la morbilidad

El clúster 1 está conformado por 290 madres gestantes que representa el 17.7 % de la población total.

El diagrama de clustering enfocado a la variable morbilidad muestra que el clúster 5 es el que tiene las características a presentar, mostrando un cuadro de morbilidad en las madres gestantes, por esta razón le nombramos morbilidad. Y por otro lado el clúster 2 tiene menos probabilidad de sufrir morbilidad materna.

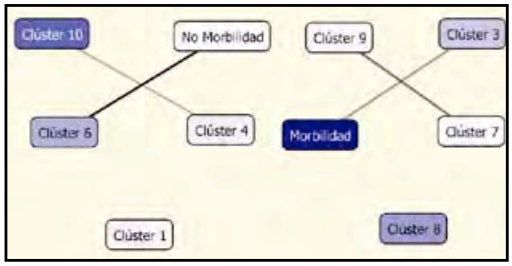

Figura 6. –Diagrama de Clustering en función al estado patológico de la morbilidad.

El atributo más importante es el clúster "Morbilidad", que de acuerdo al análisis se puede ver que el 97.2 % de la población no presenta morbilidad a diferencia del 2.8 % que sí tiene.

En relación a clústeres que presentan morbilidad, el clúster 5 es el que más morbilidad patológica presenta con un 13,4 % de 120 madres que conforman este grupo, con las siguientes características: membranas rotas, aborto incompleto, gestas cero, hemorragia H1 y la ocupación de la madre es estudiante.

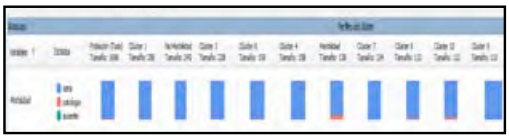

Figura 7. – Ficha Perfiles del clúster para el atributo morbilidad.

Según las fichas de características de clúster se puede ver que las madres gestantes que presentaron morbilidad tienen a compartir las mismas características tales como: membranas rotas, abortos incompletos, no tuvieron gestas, desgarros, parto prolongado.

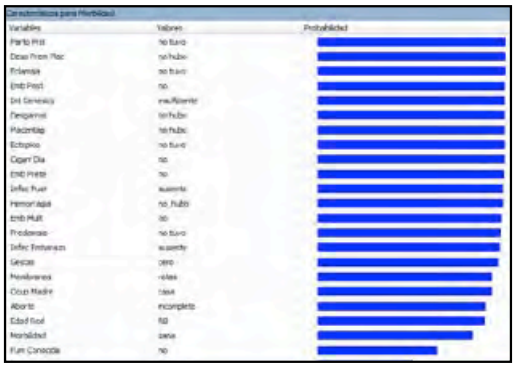

Figura 8. – Ficha de características del clúster

La ficha distinción del clúster nos permite explorar las diferencias entre los clústeres morbilidad (5) y no morbilidad (2), en la cual podemos decir que las madres que sufren morbilidad son aquellos que no conocen su última fecha de menstruación, tienen un índice de masa corporal RB, membranas rotas; la ocupación de la madre es obrera.

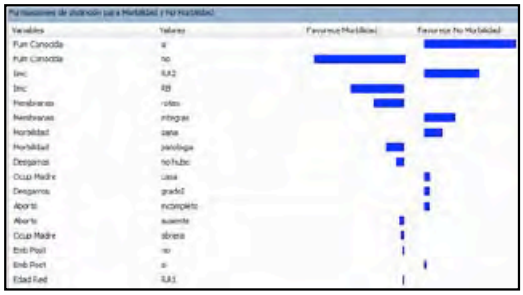

Figura 9. – Ficha distinción del clúster

1. CONCLUSIONES

En este proyecto de investigación aplicamos el algoritmo de clúster K-means de SQL Server 2008 para segmentar a las madres gestantes del hospital San Bartolomé en función a la **morbilidad**.

Para generar el modelo de clusterización hay que tener en cuenta las variables de entrada y predicción; en nuestro caso la variable de entrada fueron todas las que están descritas en la cuadro 1, y la predictora fue la **"morbilidad"**.

El diagrama de clustering generado por el Analysis Sevices de Microsoft, muestra 10 grupos; 4 de ellos tienen las características más comunes que pueden presentar un cuadro de morbilidad **materna**. Esto representa un 46,9 % de la población, el cual cada grupo tiene un perfil definido.

MORBILIDAD			
	Sana (%)	Patológica (%)	Población
CLÚSTER 1	99,7	0,3	290
CLÚSTER 2	99,9	0,1	245
CLÚSTER 3	97,4	2,6	228
CLÚSTER 6	96,1	3,9	158
CLÚSTER 4	99,5	0,5	158
CLÚSTER 5	86,6	13,4	138
CLÚSTER 7	99,9	0,1	134
CLÚSTER 8	95,1	4,9	113
CLÚSTER 10	91,9	8,1	112
CLÚSTER 9	99,7	0,3	110

Cuadro 2. Variables para el estudio de la morbilidad

A continuación describiremos los clustering que presentan un perfil de morbilidad mayor.

El Clúster 5 denominado **"morbilidad",** de acuerdo al cuadro 2, es el que presenta las características más significativas y comunes para para presentar un cuadro de morbilidad materna, esto representa el 1,1 % de la población total. Y el 13,4 % de la población estudiada (138). Esto nos indica que hay poca probabilidad de que en el hospital San Bartolomé haya una alta tendencia de morbilidad en la madres gestantes que presenten las siguientes características: membranas rotas, no presentaron hemorragia, tampoco desgarros, abortos incompletos, no tuvieron infección durante el embarazo, la ocupación de su madre es su casa, no conoce su última fecha de menstruación y no presentaron gestas mayores a cero.

El Clúster 10 es el que presenta un 8,1 % de la población del clúster teniendo las características comunes para presentar morbilidad materna con las siguientes similitudes: No presentaron abortos, han presentado embarazos múltiples, tuvieron más de una gesta, presentaron parto prolongado, no tuvieron hemorragia y la ocupación de la madre es su casa.

El Clúster 6 representa el 3.9 % de la población que presenta las características comunes para tener morbilidad materna con las siguientes similitudes: Tuvieron más de una gesta, conoce su fecha última de menstruación, no tuvieron abortos, presentaron hemorragia H1, tuvieron preeclampsia, desgarros en grado II y no hubo placenta previa.

Sin embargo, existen otros clústeres como el Clúster 2, 9, 7 que tienen otras características para no sufrir de morbilidad durante el embarazo.

En resumen, la investigación realizada aporta evidencia empírica a favor de la especificación u obtención de modelos de segmentación para el hospital San Bartolomé, con el fin de evaluar el riesgo de morbilidad en las madres gestantes.

REFERENCIAS BIBLIOGRÁFICAS

[1] Cristina García Cambronero and Irene Gómez Moreno, "Algoritmos de aprendizaje: knn & kmeans," 2006.
[2] Ing. Corso, Cynthia Lorena, "Aplicación de algoritmos de clasificación supervisada usando Weka", Universidad Tecnológica Nacional, Facultad Regional de Córdoba, Colombia, 2008.
[3] Priscila Valdiviezo Díaz, "Aplicación de técnicas de aprendizaje automático para la identificación de patrones de interacción en una experiencia virtual de aprendizaje", 2008.
[4] Ernesto Gonzáles Dias and Zady Pérez Hernández, "Obtención de patrones y reglas en el proceso académico de la Universidad de Ciencias Informáticas utilizando técnicas de minería de datos".
[5] Jorge Enrique Ugarte Humeres, "Uso de algoritmos de clustering para predecir el comportamiento de proteínas en cromatografías de interacción hidrofóbica y sistemas de dos fases acuosas", Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, departamento de Ingeniería Química y Biotecnología, Santiago de Chile, 2012.
[6] Miguel Garre, Juan José Cuadrado, Miguel A. Sicilia, Daniel Rodríguez, y Ricardo Rejas, "Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software", Alcalá de Henares, Madrid, 2007.
[7] Maribel Baltazar Domínguez, "Data Mining", Durango, 19-Mar-2010.
[8] Joel Pérez Suárez, "Desarrollo de un Data Mining para la toma de decisiones apoyado en la herramienta WEKA para el supermercado La Inmaculada", junio-2011.
[9] M.Sc. Ing. Francisco Fernández Periche, "Aproximación funcional mediante redes de funciones de base radial, una alternativa para la predicción en el proceso de reducción de mineral de la tecnología caron de producción de níquel", Universidad de Holguín, 2011.
[10] Claudio Víctor Peña Hermosilla, "Desarrollar una metodología de comportamiento dinámico de objetos con identificador", Santiago de Chile, 2007.
[11] Ramón David Lezcano "Minería de Datos", dinámico de objetos con identificador", Santiago de Chile, 2007.