II CONFERENCIA INTERNACIONAL DE PROCESAMIENTO DE LA INFORMACIÓN "CIPI 2019"

Aplicación de técnicas de minería de datos al área de Demografía de la Oficina Nacional de Estadística e Información

Application of data mining techniques to the Demographics area of the National Statistics and Information Office

Luis Mariano Rosa Castellanos[1], Yaquelín Córdova Viera[2], Miguel Ángel Lahera Rivero[3]

1- Luis Mariano Rosa Castellanos. Universidad de las Ciencias Informáticas, Cuba. E-mail: lmariano@uci.cu
2- Yaquelín Córdova Viera. Universidad de las Ciencias Informáticas, Cuba. E-mail: ycordovav@uci.cu
3- Miguel Ángel Lahera Rivero. Universidad de las Ciencias Informáticas, Cuba. E-mail: marivero@uci.cu

Resumen:

Con el aumento del volumen de los datos a lo largo de estas últimas décadas se han demandado mejoras en las tecnologías de almacenamiento y procesamiento. Las organizaciones y las empresas necesitan analizar y transformar sus datos rápidamente, para obtener conocimiento e información verdaderamente útil. De esta manera se necesitan técnica y herramientas que agilicen la extracción de conocimiento. La Oficina Nacional de Estadística e Información (ONEI) trabaja con un gran volumen de información que aumenta con el tiempo. Aunque sus sistemas informáticos permiten gestionar dicha información, actualmente necesitan procesarla en el menor tiempo posible. La presente investigación aplica a través de la Minería de datos una posible solución a este problema, obteniendo modelos de conocimiento prácticos que mejorarán y permitirán que el proceso de toma de decisiones sea más efectivo. Con estos modelos se pretende encontrar patrones y tendencias de comportamiento en el Área de Demografía de la ONEI. Para extraer el conocimiento se utilizará el mercado de datos

Demografía y la investigación será guiada por la metodología *Cross Industry Standard Process for Data Mining* (CRISP-DM). Además para el desarrollo de los modelos se seleccionó la herramienta de aprendizaje automático Weka 3.7.10.

Abstract:

With the increase of volume of data along the last few decades has been a demanded improvement in processing and storage technologies. The organizations and companies need to analyze and transform information quickly, also get the really useful knowledge. In this way require techniques and tools that speed up the extraction and processing of data. The ONEI working with a large volume of information constantly increases with time. Although their systems are updated and allow them to process their data to a large extent currently need to process large amounts of information in the shortest time possible. The current investigation applied through the Data Mining a possible solution to this problem, obtaining models practical knowledge that will improve and will allow the decision-making process more effective. With these models is to find patterns and behavioral trends in the area of Demography of the ONEI. For extract the knowledge will be used the data mart Demography of the ONEI and the investigation will be fully guided by the CRISP-DM methodology. Moreover development of the models was selected the automatic learning tool Weka 3.7.10.

Palabras Clave: Minería de datos; Weka; Demografía.

Keywords: *Datamining; Weka; Demography.*

1. Introducción

En la era actual, la informática está caracterizada por un crecimiento extraordinario de datos que se generan y almacenan en todas las áreas del desarrollo humano. Una proporción creciente de estos datos se recopilan en Bases de Datos (BD) dentro de los ordenadores, a fin de que la tecnología computacional facilite el acceso a ellos. No obstante, esta manera de proceder solo permite generar informes poco flexibles y poco escalables con grandes volúmenes de datos.

La tecnología continuó evolucionando significativamente creando una nueva forma de organizar los datos: el Almacén de Datos (AD). Se trata de "un repositorio de fuentes heterogéneas con datos integrados y organizados bajo un esquema unificado para facilitar su visualización y apoyar la toma de decisiones". [1] Esta tecnología incluye operaciones de procesamiento analítico en línea, es decir, técnicas de análisis como pueden ser el resumen, la consolidación o la agregación, así como la posibilidad de ver la información desde distintas perspectivas.

Actualmente son numerosas las naciones que han optado por aplicar estos avances de acuerdo a sus condiciones tanto económicas como políticas, con el objetivo de alcanzar un desarrollo local sustentable. Cuba no se ha quedado muy lejos de estos cambios. En la actualidad se trabaja en la ejecución de una política estatal enfocada en la orientación a los Organismos de la Administración Central del Estado para el desarrollo de sus técnicas particulares, que permitan una inserción integral a las estrategias estatales, de manera sistémica y generalizada, mediante la adopción de regulaciones y estándares generales en el desarrollo de soluciones particulares y globales. Uno de estos organismos es la Oficina Nacional de Estadísticas e Información (ONEI). [2]

La ONEI es la encargada a través de sus múltiples áreas de trabajo de gestionar datos estadísticos. Una de sus áreas es el Departamento de Población, el cual se especializa en recopilar datos demográficos del país. Estos datos se refieren, al análisis de la población distribuidos por edades, situación familiar, grupos étnicos, actividades económicas y estado civil; las modificaciones de la población tales como: nacimientos, matrimonios y fallecimientos; la esperanza de vida y estadísticas sobre migraciones junto a sus efectos sociales y económicos; grado de delincuencia; niveles de educación y otras estadísticas económicas y sociales.

Actualmente en la ONEI se cuenta con grandes volúmenes de datos referentes a dicha área. Por otra parte, es utilizado el Sistema Integral de Gestión Estadística (SIGE), una aplicación que tiene como finalidad la gestión de los datos estadísticos de forma simple y ágil, además permite el diseño de formularios y encuestas para la captura de la información estadística; a pesar de contar con las facilidades que brinda este sistema, en el momento de realizar un análisis de los datos no se le presta atención al comportamiento o las relaciones existentes entre ellos, lo cual constituye conocimiento oculto que no se espera obtener. Aprovechar este conocimiento por los compañeros de la ONEI contribuiría significativamente a facilitar la toma de decisiones del estado cubano.

En correspondencia con lo planteado anteriormente, se establece como objetivo general: Desarrollar los modelos de Minería de datos mediante las técnicas de Agrupamiento y Asociación para el descubrimiento del conocimiento oculto en los datos almacenados en el Área de Demografía de la ONEI.

La definición de la Minería de datos se puede abordar desde una triple perspectiva, en función de la amplitud de la misma. Así, se puede definir la Minería de datos desde un punto de vista estrecho como el "*descubrimiento automático de patrones o modelos interesantes y no obvios escondidos en una BD, los cuales tienen un gran potencial para contribuir en los aspectos principales del negocio*" [3]. La Minería de datos, desde un punto de vista estrecho, comprende, como sistema de extracción de relaciones; los métodos basados en la computadora, requiriendo poca intervención o apoyo por parte del analista en la obtención de información relevante. Se incluyen aquí los algoritmos de Redes neuronales, Árboles de decisión, y los Algoritmos genéticos.

Al acudir a un concepto un poco más amplio, Peacock indica que la Minería de datos también engloba, aparte de lo ya comentado, "*la confirmación o prueba de relaciones reveladas por el proceso de descubrimiento*" [3]. Se emplearían para ello métodos estadísticos clásicos y bayesianos, así como la fijación de hipótesis que se verificarán en el proceso de obtención de información, aparte de incluir la búsqueda de la confirmación de relaciones, modelos o teorías formuladas mediante la aplicación de Minería de datos desde un punto de vista estrecho. Como ejemplos están la Regresión

mínimo cuadrática[1] y el Análisis discriminante[2]. En este proceso la parte humana juega un importante papel a la hora de obtener información relevante. Se puede hablar, por tanto, de un proceso semiautomático de Minería de datos.

Por último, y como concepto más abarcador recogido en la literatura, la Minería de datos se identifica con el proceso de descubrimiento de conocimiento en bases de datos o *Knowledge Discovery in Databases* (KDD por sus siglas en inglés), incluyendo así un conjunto de actividades, entre las que se encuentra el análisis de los datos. La Minería de datos no es más que un paso hacia delante de la estadística (gracias al apoyo de la Inteligencia Artificial, que ha colaborado con la generación de nuevas técnicas). Además, la aparición de los nuevos sistemas de almacenamiento (como los almacenes de datos) es lo que permite hacer realidad la Minería de datos.

Luego de haber realizado la revisión de los principales conceptos de la Minería de datos se puede definir como un "conjunto de técnicas encaminadas a la extracción de conocimiento implícito en las BD".

Con el objetivo de apoyar la toma de decisiones la Minería de datos es aplicada en las siguientes áreas:

Negocios:

- Identificación de patrones de compra de los clientes.
- Búsqueda de asociaciones entre clientes y características demográficas.
- Predicción de respuesta a campañas de correo.
- Análisis de cestas de la compra. [4]

Fraudes:

- Detección de transacciones de lavado de dinero o de fraude en el uso de tarjetas de crédito o de servicios de telefonía móvil e, incluso, en la relación de los contribuyentes con el fisco. [4]
- Recursos humanos:
- Identificación de las características de los empleados de mayor éxito.
- Mejorar el margen de beneficios

[1] Consiste en explicar una de las variables en función de la otra a través de un determinado tipo de función (lineal, parabólica, exponencial, etc.), de forma que la función de regresión se obtiene ajustando las observaciones a la función elegida, mediante el método de Mínimos-Cuadrados (M.C.O.).

[2] Ayuda a comprender las diferencias entre grupos. Explica, en función de características métricas observadas, porqué los objetos/sujetos se encuentran asociados a distintos niveles de un factor.

- Desarrollo de planes de producción o gestión de mano de obra. [4]

Comportamiento en internet:

- Análisis del comportamiento de clientes potenciales en una página de Internet.
- Ajustar las propagandas de los productos según las necesidades específicas de los usuarios.
- Actualizar la competencia de los productos. [4]

Demografía:

- En esta área se han realizado algunos trabajos referentes al análisis de datos estadísticos tomados de los censos a determinados grupos poblacionales o a comportamientos naturales y su incidencia en distintas poblaciones, ejemplo: la fertilidad de los seres humanos y el impacto de determinadas variables en su desarrollo y evolución. [5]

2. Metodología

A continuación se realiza una breve descripción de los métodos de investigación aplicados:

Métodos Lógicos

Analítico-Sintético: División del todo en partes, con el objetivo de estudiarlas y analizarlas por separado. Reunión racional de varios elementos en una nueva totalidad. Se realiza sobre la base de los resultados del análisis. [6]

Se proponen un grupo de tareas de investigación para una metodología determinada a fin de obtener modelos de Minería de datos.

Sistémico: Estudio de los sistemas en su totalidad, complejidad y dinámica propia. [6]

Se estudian muchos temas relativos a la Minería de datos y se van aplicando según las necesidades reales del problema de investigación.

Modelación: Se crean abstracciones con el propósito de explicar la realidad. El modelo es el sustituto del objeto de investigación. [6].

Se evidencia cuando se pretende determinar la propuesta de solución, además los modelos de Minería de datos que se obtienen remplazan al objeto de la investigación.

Métodos Empíricos

Investigación – Acción: Incluye diagnóstico del problema, intervención de acción y aprendizaje reflexivo. [6].

Se realiza un estudio de los principales conceptos relacionados con la Minería de datos para dar solución al problema de investigación.

Entrevistas: La entrevista es una forma de comunicación interpersonal que tiene como objetivo proporcionar o recibir información, y en virtud de las cuales se toman determinadas decisiones. [6].

Se deben conocer las necesidades reales del cliente para poder dar solución al problema de investigación, además de que permite obtener conocimiento.

2.1 Metodología CRISP-DM

Son diversas las metodologías que han sido propuestas para el desarrollo de proyectos de Minería de datos tales como *Sample, Explore, Modify, Model, Assess*, (SEMMA por sus siglas en inglés) y *Cross Industry Standard Process for Data Mining* (CRISP-DM por sus siglas en inglés). Este último es uno de los modelos principalmente utilizados en ambientes académicos e industriales.

CRISP–DM organiza el desarrollo de un proyecto de Minería de datos, en una serie de seis fases (Ver figura 1):

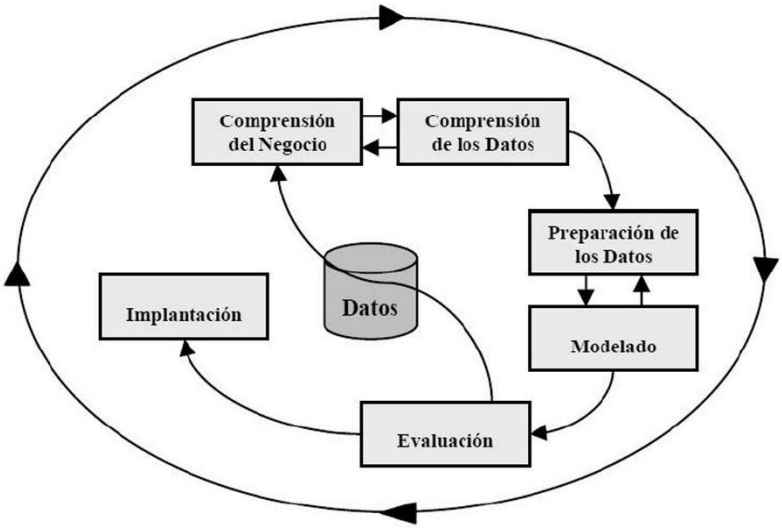

Figura 1. Fases de los proyectos de Minería de datos. [4].

La sucesión de fases no es necesariamente rígida. Cada fase es estructurada en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, donde finalmente se describen las acciones que deben ser desarrolladas para situaciones específicas, pero en ningún momento se propone como realizarlas. [7].

Se seleccionó CRISP-DM como metodología de desarrollo a utilizar en el proceso de Minería de datos, esta designación estuvo respaldada por las siguientes ventajas:

- Por ser de libre distribución.
- Constituye una de las guías de referencia más confiables y utilizadas a nivel mundial en este ámbito.
- Organiza las tareas jerárquicamente y planifica el proyecto en fases algo flexibles y cómodas.
- Las tareas generales se proyectan a tareas específicas, facilitando la planificación registro y manipulación de los trabajos mineros.

2.2 Herramientas

Después de realizado un análisis sobre las cualidades de las herramientas utilizadas mundialmente para llevar a cabo un proyecto de Minería de datos, se seleccionó para realizar el proceso de transformación de los datos la herramienta Pentaho Data Integration en su versión 5.0.1 y como entorno para la aplicación de las técnicas de Minería de datos la herramienta de aprendizaje automático Weka en su versión 3.7.1. Además se seleccionaron las técnicas de Agrupamiento y Asociación con los algoritmos Simple K-Means y Apriori respectivamente. Para el Agrupamiento el algoritmo más factible es Simple K-means pues permite procesar valores numéricos y nominales y es apto para trabajar con el gran volumen de información que contiene el Mercado de datos Demografía de la ONEI. Por su parte Apriori solo acepta valores de tipo nominal.

3. Resultados y discusión

En el presente epígrafe se define los objetivos del negocio guiados por la metodología CRISP-DM, describiendo las siguientes fases: Comprensión del negocio, Comprensión de los datos y Preparación de los datos. Es importante señalar que los datos seleccionados para la aplicación de las técnicas de Minería de datos corresponden al Mercado de datos Demografía del Departamento de Población de la ONEI.

3.1 Fase de comprensión del negocio

Determinación de los objetivos de negocio

Una vez realizadas las entrevistas con el cliente se puede llegar a la conclusión que el objetivo del negocio es analizar el comportamiento de los hechos vitales (hechos relacionados con el comienzo y fin de la vida del individuo, además de los cambios en su estado civil durante su existencia) en el movimiento natural de la población cubana

Criterios de éxito de negocio

El éxito de la investigación dependerá de una buena descripción y aplicación de los objetivos del negocio. Aquí se tendrán en cuenta los siguientes criterios:

- Aplicar las técnicas de Minería de datos con el objetivo de obtener los modelos de conocimientos.
- Hacer uso de la herramienta Weka para obtener estos modelos.
- Aplicar la Minería de datos realizando los pasos que se indican en la metodología CRISP-DM.

Determinación de los objetivos de la Minería de datos

Obtener modelos empleando las técnicas de Minería de datos, Agrupamiento y Asociación; mediante los cuales se generen un conjunto de patrones para el análisis del comportamiento de las variables demográficas en el movimiento natural de la población cubana.

3.2 Fase de comprensión de los datos

Recolección de los datos iniciales

La recolección inicial de los datos constituye el punto de partida para una buena comprensión, debido a que la salida de esta tarea es la base que sustenta las restantes tareas de la metodología. Los datos a obtener están contenidos en el Mercado de datos Demografía, el cual recoge los datos correspondientes al Departamento de Población de la ONEI, este además utiliza algunas dimensiones contenidas en el esquema dimensiones del AD de dicha entidad. En las entrevistas realizadas al cliente se logró percibir su interés con temas relacionados a los nacimientos, defunciones, matrimonios y divorcios de los individuos.

Selección de las variables

La selección se realiza mediante una exploración de los datos la cual comprende visualizar cada una de las variables con sus respectivos valores y decidir cuáles de ellas aportan mayor información al problema planteado, con este fin se utiliza la herramienta Pgadmin 1.14.1.

A partir de la estructura del mercado es posible acceder a las variables necesarias los cuales expresan características de un determinado tema. En el caso de los nacimientos (Ver tabla 1) se tiene en cuenta la provincia, el lugar donde se produce el parto, la cantidad de semanas de embarazo de la madre, el peso al nacer, el tipo de embarazo, la cantidad de nacimientos producidos por sexo y la edad de la madre.

Tabla 1. Atributos seleccionados que se relacionan con el hecho ¨Nacimientos¨. Fuente: Elaboración propia.

Nombre	Dato	Tabla	Descripción
provincia_nombre	varchar	dim_dpa	Se refiere al nombre de la provincia en que se registra el nacimiento.
lugarocurrencia_nombre	varchar	dim_lugar_ocurrencia	Se refiere al local donde se produce el nacimiento.
semanagestacion_numero	integer	dim_semanas_gestacion	Se refiere al número de la semana de gestación en que se produce el nacimiento.
rango_nombre	varchar	dim_peso_nacimiento	Se refiere al rango de pesos en el que se ubican los nacimientos.
tipoembarazo_nombre	varchar	dim_tipo_embarazo	Se refiere al tipo de embarazo.
cantidadhembrastotal	integer	hech_nacimientos	Se refiere a la cantidad de hembras que nacen.
cantidadvaronestotal	integer	hech_nacimientos	Se refiere a la cantidad de varones que nacen.

Descripción de los datos

La descripción se realiza después de adquirir los datos iniciales. Estos atributos son los que más información aportan y por lo tanto deben ser descritos con el objetivo de adecuarlos al futuro procesamiento. Este proceso involucra su identificación y el tipo de valor que poseen (Ver tabla 2) de acuerdo con la herramienta de aprendizaje automático a utilizar.

Tabla 2. Atributos recolectados y el tipo de valor que poseen. Fuente: elaboración propia.

Atributo	Tipo de valor
provincia_nombre	Nominal
lugarocurrencia_nombre	Nominal
semanagestacion_numero	Numérico
rango_nombre	Nominal
tipoembarazo_nombre	Nominal
cantidadhembrastotal	Numérico
cantidadvaronestotal	Numérico
grupoetario	Nominal

Atributo	Tipo de valor
cantidaddefuncioneshombres	Numérico
cantidaddefuncionesmujeres	Numérico
ordenmatrimonio_numero	Numérico
estadoconyugal_nombre	Nominal
edad	Numérico
numerohijos_numero	Numérico
nivel_escolaridad_nombre	Nominal

Verificación de la calidad de los datos

Comprobar la calidad de los datos garantiza en buena medida saber cuáles son los principales problemas que podrían afectar el resultado final. En el desarrollo de la investigación fue necesario realizar esta tarea de la metodología CRISP-DM sobre los datos recolectados. Con ese objetivo se emplea la herramienta DataCleaner y se realiza un estudio en cuanto a:

- Campos nulos.
- Existencia de campos vacíos.
- Representación de la realidad.
- Valores fuera de rango.

Es importante destacar que los mismos no presentan ninguna de las dificultades antes señaladas, por lo que están listos para realizar las transformaciones necesarias.

3.3 Fase de preparación de los datos

La fase de preparación de los datos cubre todas las actividades necesarias para construir el último conjunto de datos (datos que se introduce en la herramienta de aprendizaje automático) a partir de los datos brutos iniciales. Las tareas de preparación de datos probablemente se realizan varias veces y no en cualquier orden prescrito. Las tareas incluyen construir tablas, registro y selección de atributos, así como la transformación y limpieza de datos para la herramienta de aprendizaje automático.

Selección de los datos que van a ser utilizados para el análisis

En esta etapa se deciden cuáles datos serán incluidos o excluidos del proceso de análisis, apoyándose en criterios establecidos en tareas anteriores. Comprobada la calidad de los datos se determina que estos son los idóneos para continuar en el proceso. Para obtener estos datos son realizadas cuatro consultas, una de ellas devuelve la tabla ilustrada en la figura 2 relacionada con el hecho nacimientos.

provincia_nombre	lugarocur	semanages	peso_al_nacer	tipoemb	cantidadhemb	cantidadvarone	edad_de_madre
Matanzas	Hospital	39	3000-3999 grai	Sencillo	1	0	15-24
Ciudad de La Habana	Hospital	39	3000-3999 grai	Sencillo	1	0	10-14
Camagüey	Hospital	38	3000-3999 grai	Sencillo	0	1	10-14
Camagüey	Hospital	41	3000-3999 grai	Sencillo	1	0	15-24
Granma	Hospital	40	3000-3999 grai	Sencillo	0	1	15-24
Pinar del Río	Hospital	39	3000-3999 grai	Sencillo	1	0	15-24
Camagüey	Hospital	40	3000-3999 grai	Sencillo	1	0	25-34
Las Tunas	Hospital	23	2000-2999 grai	Sencillo	1	0	10-14
Granma	Hospital	37	3000-3999 grai	Sencillo	0	1	10-14
Isla de La Juventud	Hospital	40	3000-3999 grai	Sencillo	1	0	10-14

Figura 2. Datos seleccionados para obtener la vista minable Nacimientos. Fuente: elaboración propia.

Estructuración de los datos

Esta tarea incluye las operaciones de preparación de los datos tales como la generación de nuevos atributos a partir de atributos ya existentes, integración de nuevos registros o transformación de valores para atributos existentes.

Transformación de los datos para obtener las vistas minables

El atributo provincia_nombre se emplea para obtener las cuatro vistas minables (nacimientos, defunciones, matrimonios y divorcios) (Ver figura 3), además es eliminado y en su lugar se crea el atributo zona, el cual toma como valores ("Occidente", "Centro" y "Oriente").

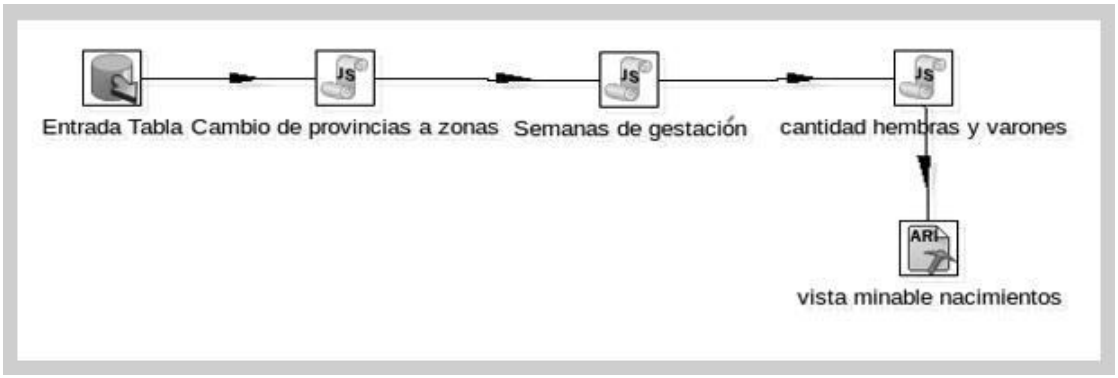

Figura 3. Transformación para obtener la vista minable ¨nacimientos¨. Fuente: elaboración propia.

Descripción de la vista minable Nacimientos

La vista está constituida por 1000 instancias, cada una de ellas está compuesta por 7 atributos (Ver figura 4) en su totalidad de tipo nominal.

lugarocurrencia_nom	peso_al_nacer	tipoembarazo	edad_de_madre	zona	edad_gesta	nacidos
Hospital	3000-3999 gra	Sencillo	15-24	Occidente	RNAT	más hembras
Hospital	3000-3999 gra	Sencillo	10-14	Occidente	RNAT	más hembras
Hospital	3000-3999 gra	Sencillo	10-14	Centro	RNAT	más varones
Hospital	3000-3999 gra	Sencillo	15-24	Centro	RNAT	más hembras
Hospital	3000-3999 gra	Sencillo	15-24	Oriente	RNAT	más varones
Hospital	3000-3999 gra	Sencillo	15-24	Occidente	RNAT	más hembras
Hospital	3000-3999 gra	Sencillo	25-34	Centro	RNAT	más hembras
Hospital	2000-2999 gra	Sencillo	10-14	Oriente	RNPT EXT	más hembras
Hospital	3000-3999 gra	Sencillo	10-14	Oriente	RNPT LIM	más varones
Hospital	3000-3999 gra	Sencillo	10-14	Occidente	RNAT	más hembras
Hospital	3000-3999 gra	Sencillo	10-14	Occidente	RNAT	más hembras
Hospital	2000-2999 gra	Sencillo	10-14	Oriente	RNPT MOI	más varones
Hospital	2000-2999 gra	Sencillo	10-14	Oriente	RNPT MOI	más varones
Hospital	3000-3999 gra	Sencillo	10-14	Oriente	RNPT LIM	más hembras
Hospital	3000-3999 gra	Sencillo	10-14	Oriente	RNPT LIM	más hembras
Hospital	3000-3999 gra	Sencillo	10-14	Oriente	RNAT	más varones
Hospital	3000-3999 gra	Sencillo	10-14	Oriente	RNAT	más varones
Hospital	3000-3999 gra	Sencillo	10-14	Centro	RNAT	más varones

Figura 4. Datos obtenidos en la vista minable Nacimientos. Fuente: elaboración propia.

Siguiendo la línea de la metodología CRISP-DM se profundizó fase a fase en varios aspectos adecuándolos a la situación real del problema de investigación. Posteriormente se hizo un detallado análisis de los atributos para seleccionar los más representativos los cuales serán utilizados en la construcción de los modelos. Se realizaron las transformaciones sobre los datos que fueron seleccionados, permitiendo obtener las vistas minables con los atributos principales para generar los modelos.

A continuación se describe la aplicación de las técnicas Agrupamiento y Asociación con los algoritmos Simple K-means y Apriori respectivamente para la generación de los modelos. Posteriormente serán interpretados para obtener un conjunto de patrones de comportamiento.

3.4 Fase de modelado

Construcción de los modelos de conocimiento

Después de seleccionadas las técnicas, se ejecutan sobre los datos previamente preparados para generar uno o más modelos. Todas las técnicas de modelado tienen un conjunto de parámetros que determinan las características de los modelos a generar.

Aplicando Simple K-means

El algoritmo Simple K-means se encuentra implementado en Weka en la clase ***weka.clusteres.SimpleKMeans***, de los parámetros esenciales para el funcionamiento de este, se empleará como función para calcular la distancia entre los centroides y el resto de la instancias: la Euclidiana, el número de semilla a utilizar estará condicionado por

seleccionar el que más minimice el error cuadrático y el número de grupos será 3 debido a que se quiere obtener un resultado orientado a las zonas geográficas del país.

Para obtener el siguiente modelo (Ver tabla 4) fue ejecutado el algoritmo Simple K-means sobre la vista minable Nacimientos con números de semilla diferentes en cada ejecución (Ver tabla 3), resultando seleccionado el 3.

Tabla 3.Selección del número de semillas para el modelo ¨nacimientos¨. Fuente: elaboración propia.

Número de semillas	Error cuadrático
1	553
2	681
3	421
4	421
5	634
6	655
7	539
8	668
9	421
10	480

Tabla 4. Modelo generado por Simple K-means para la vista minable ¨nacimientos¨. Fuente: elaboración propia.

Clusters #:			
Attribute	0 (549)	1 (159)	2 (292)
lugarocurrencia_nombre	Hospital	Hospital	Hospital
peso_al_nacer	3000-3999 gramos	2000-2999 gramos	3000-3999 gramos
tipoembarazo_nombre	Sencillo	Sencillo	Sencillo
edad_de_madre	15-24	15-24	15-24
edad_gestacional	RNAT	RNAT	RNAT
nacidos	más varones	más hembras	más hembras
Clustered Instances 0 549 (55%) 1 159 (16%) 2 292 (29%)		Cluster 0 <-- Oriente Cluster 1 <-- Centro Cluster 2 <-- Occidente	

Aplicando Apriori

El algoritmo Apriori está implementado en Weka en la clase **weka.associations.Apriori**, para su ejecución es imprescindible tener en cuenta la métrica a utilizar para la selección de las mejores reglas generadas, en este caso se usará la Confianza, además se debe especificar el criterio mínimo de confianza ya que serán seleccionadas solo las reglas cuyo valor de confianza sea superior o igual a este. A continuación se presenta el modelo obtenido tras procesar la vista minable Nacimientos:

Modelo Apriori para la vista minable Nacimientos

Para el modelo es seleccionado como criterio mínimo de confianza 1 y además es especificado que se quieren encontrar las 10 mejores reglas, las cuales son presentadas a continuación:

1. *edad_gestacional*=RNAT ==> *tipoembarazo_nombre*=Sencillo <conf:(1)>
2. *lugarocurrencia_nombre*=Hospitale *dad_gestacional*=RNAT==> *tipoembarazo_nombre*=Sencillo <conf:(1)>
3. *edad_de_madre*=15-24 *edad_gestacional*=RNAT ==> *tipoembarazo_nombre*=Sencillo <conf:(1)>
4. *lugarocurrencia_nombre*=Hospital *edad_de_madre*=15-24 *edad_gestacional*=RNAT ==> *tipoembarazo_nombre*=Sencillo <conf:(1)>
5. *peso_al_nacer*=3000-3999 gramos ==> *tipoembarazo_nombre*=Sencillo <conf:(1)>
6. *lugarocurrencia_nombre*=Hospital *peso_al_nacer*=3000-3999 gramos==> *tipoembarazo_nombre*=Sencillo <conf:(1)>

Resultados obtenidos

A continuación se presentan los patrones detectados luego de haber aplicado los correspondientes algoritmos de Minería de datos.

Patrones identificados en el Modelo generado por el algoritmo Simple K-means

Los patrones identificados en el modelo Nacimientos para cada grupo son los siguientes:

Grupo 0: Los registros de nacimientos de la zona Oriente representan el 55 %.

- En la zona oriental por lo general nacen más varones que hembras.
- Las madres con edad entre 15 y 24 años casi siempre paren recién nacidos a término.

Grupo 1: Los registros de nacimientos de la zona Centro representan el 16 %.

- Los recién nacidos a término son más hembras que varones.

Grupo 2: Los registros de nacimientos de la zona Occidente representan el 29 %.

- En la zona Occidental por lo general nacen más hembras que varones.
- Las mujeres entre 15 y 24 años tienen recién nacidos a término y de ellos son más hembras que varones.

Posibles decisiones a tomar a partir de los patrones detectados

Las siguientes decisiones equivalen a ser conocimiento que hasta el momento no se tenía:

- Incrementar la matrícula de los círculos infantiles en la zona oriental debido a ser esta la que más nacimientos registra.
- En la zona occidental se debe aumentar la producción de ropa para niñas, ya que los nacimientos de estas superan en número a los niños.
- En el centro del país brindarle una atención diferenciada a las madres con edad entre 15 y 24 años en el período de embarazo con el objetivo de incrementar el peso de los niños al nacer.

Patrones identificados en el Modelo generado por el algoritmo A priori

En el modelo Nacimientos se lograron identificar los siguientes patrones:

- R4: Si el lugar de ocurrencia es el hospital y la edad de la madre está entre 15 y 24 años y la edad gestacional es recién nacidos a término entonces el tipo de embarazo es sencillo.
- R7: Si el peso al nacer está entre 3000 y 3999 gramos y la edad gestacional es recién nacidos a término entonces el tipo de embarazo es sencillo.
- R8: Si el lugar de ocurrencia es el hospital y el peso al nacer está entre 3000 y 3999 gramos y la edad gestacional es recién nacidos a término entonces el tipo de embarazo es sencillo.
- R9: Si el peso al nacer está entre 3000 y 3999 gramos y la edad de la madre está entre 15 y 24 años entonces el tipo de embarazo es sencillo.
- R10: Si el lugar de ocurrencia es el hospital y el peso al nacer está entre 3000 y 3999 gramos y la edad de la madre está entre 15 y 24 entonces el tipo de embarazo es sencillo.

Posibles decisiones a tomar a partir de los patrones detectados

Las siguientes decisiones equivalen a ser conocimiento que hasta el momento no se tenía:

- Notificarle al Ministerio de Salud Pública que los nacimientos producidos por lo general ocurren en hospitales, provenientes de madres cuya edad oscila entre 15 y 24 años, además de que el tipo de embarazo es sencillo.

3.5 Fase de evaluación

Evaluación de los resultados obtenidos del modelo generado por el algoritmo Simple K-means

Se puede afirmar que los patrones identificados son válidos debido a que proceden de modelos los cuales fueron seleccionados en el caso en que el error cuadrático era mínimo. Es importante destacar que los patrones descubiertos se ajustan al objetivo del negocio preestablecido ya que describen en gran medida el comportamiento de las variables demográficas que inciden en el movimiento natural de las personas.

Evaluación de los resultados obtenidos del modelo generado por el algoritmo A priori

Estos resultados fueron evaluados atendiendo al indicador coeficiente de confianza el cual es utilizado para medir la precisión de las reglas, o lo que es lo mismo la probabilidad condicionada de un hecho (conclusión) con respecto a otro (condición). Al realizar una exploración de las reglas se detectó que 17 presentan como coeficiente de confianza 1 lo que representa el 53 % del total.

Con la aplicación de los algoritmos Simple K-Means y Apriori pertenecientes a las técnicas de Minería de datos Agrupamiento y Asociación respectivamente se logró obtener dos modelos de conocimientos. A partir de su interpretación se identificaron patrones de comportamientos los cuales fueron evaluados, además a partir de ellos se proponen un conjunto de decisiones que podrán ser tomadas en consideración por el personal del Departamento de Población de la ONEI.

4. Conclusiones

A partir de los resultados obtenidos a lo largo del desarrollo de la presente investigación se definen las siguientes conclusiones:

- Para guiar el proyecto de Minería de datos se estableció como metodología de desarrollo CRISP-DM, se seleccionaron las técnicas de Agrupamiento y Asociación y se utilizó como herramienta de aprendizaje automático Weka en su versión 3.7.10.

- Se seleccionaron los datos más representativos, a los cuales se les realizaron las transformaciones necesarias para obtener las vistas minables.
- Se aplicaron las técnicas Agrupamiento y Asociación mediante los algoritmos Simple K-means y Apriori los cuales están implementados en Weka.
- A partir de los modelos obtenidos, se identificaron un conjunto de patrones, los cuales constituyen información confiable y verídica que se desconocía hasta ese momento. Se propusieron decisiones a tomar.

  Ejemplo 1: Uno de los patrones obtenidos al aplicar el algoritmo Simple K-Means se refiere a que en la región Oriental ocurren el 55% de los nacimientos del país, por lo que se propuso como posible decisión a tomar, incrementar la matrícula de los círculos infantiles en la zona oriental debido a ser esta la que más nacimientos registra.

  Ejemplo 2: Uno de los patrones obtenidos al aplicar el algoritmo A Priori se refiere a que si el lugar de ocurrencia de un nacimiento es el hospital y el peso al nacer está entre 3000 y 3999 gramos y la edad de la madre está entre 15 y 24 años, entonces el tipo de embarazo es sencillo. Por lo que se propuso como posible decisión a tomar, notificarle al Ministerio de Salud Pública que los nacimientos producidos por lo general ocurren en hospitales, provenientes de madres cuya edad oscila entre 15 y 24 años, además de que el tipo de embarazo es sencillo.

5. Referencias bibliográficas

1. Hernández Orallo, J. *Introducción a la Minería de datos.* Madrid: Pearson Education S.A., 2004.
2. Vidal Ledo, M. *Alfabetización digital e informatización de la sociedad. Un reto para el presente.* [En línea] 2005. [Citado el: 16 de Noviembre de 2013.] http://www.rcim.sld.cu/revista_9/articulos_htm/alfabetizdigital.htm.
3. Peacock, P R. *Data Mining in Marketing: part 1.* s.l.: Marketing Management, 1998.
4. Bordignon, F. Aplicaciones relacionadas con la Minería de datos. [En línea] 2008. [Citado el: 16 de Noviembre de 2013.] http://ferbor.blogspot.com/2007/05/aplicaciones-relacionadas-con-minera-de.html.
5. Gans, M y Krivec, J. Demographic analysis of fertility using Data mining tools. [En línea] 12 de Marzo de 2008. [Citado el: 23 de Noviembre de 2013.] http://www.informatica.si/PDF/322/05_Gams_Demographic_Analysis_of_Fertility_Using_Data_Mining_ Tools.pdf
6. Barchini, Graciela Elisa. *Métodos "I + D" de la Informática.* [Documento] Santiago del Estero: Universidad Nacional de Santiago del Estero, 2005.
7. Mejía Giraldo, J y Jiménez Builes, J. Algunas Metodologías para Crear Proyectos de Minería de datos. [En línea] Abril de 2013. [Citado el: 28 de Noviembre de 2013.] http://www.unla.edu.ar/sistemas/redisla/ReLAIS/relais-v1-n2-TAPA.pdf. ISSN 314-2642.