PROYECTO DE GRADO

Presentado ante la ilustre UNIVERSIDAD DE LOS ANDES como requisito parcial para obtener el Título de INGENIERO DE SISTEMAS

EVALUACIÓN SEROEPIDEMIOLOGICA DEL MAL DE CHAGAS MEDIANTE REGLAS DE ASOCIACIÓN

Por

Bachiller: Juan Manuel Salcedo Suarez.

Tutor: Prof. Pablo Guillen.

Noviembre 2009

EVALUACIÓN SEROEPIDEMIOLOGICA DEL MAL DE CHAGAS MEDIANTES REGLAS DE ASOCIACION

Bachiller: Juan Manuel Salcedo Suarez
Proyecto de Grado — Sistemas Computacionales

Resumen: El objetivo de este proyecto de grado es determinar los posibles factores de riesgos epidemiológicos asociados con la transmisión de la enfermedad de Chagas en la comunidad de San Pedro, Parroquia Santa Fe del municipio Sucre, estado Sucre, para así poder tomar medidas de prevención, control y vigilancia para erradicarla.

Para resolver este problema se utiliza la minería de datos (MD), la cual es extensamente usada en la actualidad debido a su amplia aplicación en distintos sistemas. Para aplicar esta técnica se utiliza la metodología CRISP-DM, que está descrita en términos de un modelo de proceso jerárquico, la cual consistente en un conjunto de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada, e instancia de procesos.

La base de datos que se requiere para establecer los factores de riesgos epidemiológicos con la enfermedad de Chagas, es recopilada a través de encuestas que conducen a una caracterización epidemiológica de cada familia. Además, se analizaron aleatoriamente muestras sanguíneas que fueron procesadas para evaluar la serología de la enfermedad.

Para llegar a conocer los patrones existentes que asocian las variables epidemiológicas con la enfermedad, se aplico la técnica de reglas de asociación, utilizando el algoritmo *Apriori* de la biblioteca ARules, perteneciente al paquete estadístico R. Esta técnica se caracteriza por buscar patrones dentro de los datos para llegar a reglas que asocien los diferentes atributos de ella. Al evaluar el conjunto de reglas, se obtiene como resultado que las malas condiciones de las viviendas, la convivencia íntima con animales relacionados con la enfermedad, la presencia de los reservorios del *Trypanosoma cruzi* como los rabipelados y cachicamos están contribuyendo a la transmisión de la enfermedad en esta zona.

Palabras claves: Minería de datos, Reglas de asociación, Enfermedad de Chagas.

Índice

Índice de Tablas

Índice de Figuras

Capítulo 1

Aspectos Preliminares

1.1 Antecedentes

Vivimos en un mundo donde la información se duplica cada año y gran parte de esta información es histórica, es decir, habla de sucesos que han ocurrido en el pasado o se están produciendo. El problema es que, aunque la capacidad de generación y acceso a la información crece a gran velocidad no ocurre lo mismo con nuestra capacidad de asimilarla. El objetivo general de la minería de datos es el de obtener conocimiento que sirva de ayuda para mejorar un sistema estudiado a partir de su información histórica.

Las reglas de asociación son de las tareas más importantes de la minería de datos (MD), ya que nos ayuda a identificar hechos que ocurren en común en un determinado sistema. Esta técnica tiene importantes aplicaciones prácticas tales como: el análisis de la cesta de compras en un supermercado, búsqueda de patrones de páginas Web, etc. Ejemplos de algunos trabajos donde se aplicó esta técnica de minería de datos son:

En el Hospital de Ramón y Cajal de Madrid se llevo a cabo la extracción evolutiva de reglas de asociación en un servicio de urgencias psiquiátricas (Aguilera et. al., 2003). El objetivo de esta investigación fue obtener información sobre los horarios en la llegada al servicio de urgencias psiquiátricas.

Por otro lado, en la Universidad de Mondragón, se realizo un trabajo titulado "Algoritmo evolutivo de extracción de reglas de asociación aplicado a un problema de marketing (Del Jesús et. al., 2004), en donde el departamento de organización y marketing de dicha Universidad, aplicó esta herramienta para la extracción de conocimiento útil sobre certámenes feriales.

La universidad de la Rioja en conjunto con la Universidad de los Andes desarrollaron un proyecto titulado "BÚSQUEDA AUTOMÁTICA DE CONOCIMIENTO OCULTO EN SERIES TEMPORALES DE PROCESOS INDUSTRIALES VINÍCOLAS, DEL ACERO Y DE ELASTÓMEROS MEDIANTE ALGORITMOS QUE OBTIENEN REGLAS DE ASOCIACIÓN EN LÍNEA (CONOSER)", tiene como objetivo la obtención de reglas de asociación que puedan actualizarse en tiempo real y que permitan obtener conocimiento oculto de los históricos de procesos industriales que pueda ser útil en la toma de decisiones para la mejora de los mismos (Martínez et. al., 2009).

1.2 Planteamiento del Problema

La minería de datos se puede definir como el proceso de extracción de conocimiento útil y comprensible, previamente desconocido, haciendo uso de grandes cantidades de datos almacenados en distintos formatos (Witten, 2005). En los últimos años han sido muchas las aplicaciones de este conjunto de poderosas herramientas que se engloban en lo que se conoce como MD. Específicamente las reglas de asociación, derivan de un tipo de análisis que extrae la información a través de coincidencias.

Uno de los principales problemas de salud pública en nuestro país es el mal de Chagas, esta enfermedad produce alteraciones cardíacas y en una evolución avanzada de daño cardíaco los pacientes pueden presentar muerte súbita cardiaca. Actualmente la población de San Pedro, Parroquia Santa Fe del municipio Sucre, estado Sucre, constituye un foco de alta endemicidad del mal de Chagas. Se desea determinar la magnitud real de su prevalencia aplicando la Minería de Datos, particularmente con la aplicación del modelado descriptivo reglas de asociación a los datos aportados por el Laboratorio de Biología Molecular del Instituto de Investigaciones en Biomedicina y Ciencias Aplicadas de la Universidad de Oriente (IIBCA-UDO).

1.3 Justificación

Se desean obtener patrones que determinen los posibles factores de riesgo epidemiológico asociados con la transmisión de la enfermedad de Chagas en San Pedro, Parroquia Santa Fe del municipio Sucre, estado Sucre, para así poder tomar medidas de prevención, control y vigilancia para erradicarla.

1.4 Objetivos

1.4.1 Objetivo General

Aplicar y evaluar la metodología de reglas de asociación de la biblioteca ARules, para la obtención de patrones que determinen los posibles factores de riesgo epidemiológico asociados con la transmisión de la enfermedad de Chagas en San Pedro, Parroquia Santa Fe del municipio Sucre, estado Sucre.

1.4.2 Objetivo Específicos

- Crear una base de datos con las encuestas epidemiológicas sobre la enfermedad de Chagas realizada a la comunidad de San Pedro, la cual fue diseñada por especialistas de Sociología de la Salud del Centro de Investigaciones Parasitológicas "José Witremundo Torrealba" del NURR-ULA y validada por IIBCA-UDO.

- Llevar el conjunto original de datos a un nuevo conjunto más manejable y significativo, cumpliendo las siguientes etapas: Identificación y conversión de tipos, imputación (rellenar los datos inexistentes), identificación de espurios, eliminación de ruido y datos incompletos.

- Aplicar la técnica de reglas de asociación de la biblioteca ARules a los datos, a través de la utilización del paquete estadístico R.

- Evaluar las reglas obtenidas, para comprobar que los resultados sean validos y suficientemente satisfactorios.

1.5 Metodología

La metodología para minería de datos CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) (Chapman, et. al., 2002) la cual contempla un proceso jerárquico (ver figura 1.1) que se describe en las siguientes fases:

Figura 1.1: Metodología para minería de datos CRISP-DM (Chapman et. al., 2002)

- Análisis del problema: en esta fase, se determinan los objetivos que se desean alcanzar, a través del conocimiento previo de problema en estudio, y en la cual se decide la técnica de minería de datos a utilizar.

- Análisis de los datos: en esta fase, se realiza la recolección inicial de los datos para familiarizarse con ellos, se identifica la calidad de los mismos y se realiza un análisis para descubrir las relaciones más evidentes para las primeras hipótesis de relaciones ocultas entre ellos. Particularmente en esta investigación se dispone de un conjunto de datos provenientes de una encuesta epidemiológica validada por especialistas de la salud, la cual contiene información sobre las condiciones medios ambientales en que se encuentran las personas pertenecientes a la población rural de San Pedro. De un total de 293 personas censadas, se tomaron muestras sanguíneas a 194 individuos, aleatoriamente, de ambos géneros y diferentes edades.

- Preparación de los datos: en esta fase, se construye la base de datos a partir de los datos iníciales; esta tarea se realiza en varias ocasiones y no de forma estructurada. Además se incluye la selección de registros y atributos, así como también, la transformación de los datos para que puedan ser utilizados por las técnicas de minería de datos.

- Modelado: Normalmente para resolver el mismo problema existen diferentes técnicas de minería de datos, en esta fase, se seleccionan y aplican varias técnicas de modelado, además de preparar la base de datos para cada técnica en particular, ya que cada una de ella exige una entrada de datos en específico.

- Evaluación: esta parte de la metodología involucra la evaluación de los resultados obtenidos, que en el caso de las reglas de asociación consiste en determinar los factores de riesgo epidemiológico asociados con la transmisión de la enfermedad de Chagas.

Capítulo 2

Materiales y métodos

2.1 Enfermedad de Chagas

El Mal de Chagas o *tripanosomiasis* americana es una afección causada por un protozoario flagelado, denominado *Trypanosoma cruzi*. El agente etiológico de la enfermedad fue descubierto en 1907 por el doctor brasileño Carlos Chagas en muestras intestinales de insectos de la subfamilia *Triatominae* (chipos), los principales vectores de la enfermedad. Poco después hallaría el parásito en seres humanos enfermos, asociando el agente infeccioso a las patologías características de la enfermedad que lleva su nombre.

Trypanosoma cruzi es un parásito que está adaptado a una gran variedad de hospederos, incluyendo al hombre en forma accidental, principalmente en condiciones intradomiciliarias y muchos animales en condiciones peridomésticas; entre los silvestres, *Didelphis marsupialis* (rabipelado) y *Dasypus novemcinctus* (cachicamo); semidomésticos, *Rattus rattus* (rata) y *Marmosa casta* (comadreja); y entre los domésticos, *Canis familiaris* (perro) y *Felis domesticus* (gato) (Perruolo & Morales, 1987; Maekelt, 2000).

Los principales modos de transmisión son:

- A través de la deposición de heces de los triatominos en lesiones cutáneas producidas por la picadura de los propios insectos (Maekelt, 1994; Villalobos et al., 1994; Botero y Restrepo, 1998).

- Por vía transplacentaria, desde madres con parasitemia a fetos sanos, a través de la placenta. Este modo de transmisión ocurre hasta en un 11% de las mujeres infectadas y mediante transfusiones sanguíneas de individuos infectados a individuos sanos (Fragata *et al.*, 1997).

El ciclo de vida del *Trypanosoma cruzi* se divide en dos fases, una que ocurre en el hospedero invertebrado y otra en el hospedero vertebrado. En el invertebrado tiene inicio cuando el insecto se contamina al chupar sangre de los reservorios o del hombre enfermo que contiene las formas tripomastigotes sanguíneos, las cuales se transforman en epimastigotes en el intestino medio del triatomino y, posteriormente, migran y se convierten en tripomastigotes metacíclicos a nivel del recto, donde son eliminados junto con las heces y orina del vector. En el vertebrado se inicia cuando el insecto al picar, elimina en las heces y orina los tripomastigotes metacíclicos (formas infectantes), los cuales son capaces de atravesar la piel y mucosas. Dentro de las células del hospedero, los tripomastigotes comienzan su transformación a amastigotes (formas proliferativas clínicamente relevantes), iniciando un proceso de fisión binaria, convirtiéndose en formas tripomastigotas nuevamente, responsables de la ruptura de la célula por su intenso movimiento; estas formas tienen la capacidad de infectar otras células y entrar al torrente sanguíneo, distribuyéndose por todo el organismo, localizándose en el sistema nervioso central, bazo, médula ósea, corazón, ganglios linfáticos e hígado (Atias, 1991; Contreras, 1994; Brener *et al.,* 2000).

Clínicamente, se reconocen tres fases de la enfermedad de Chagas: fase aguda, fase indeterminada y fase crónica.

- La fase aguda, asintomática en el 60 % de los casos y caracterizada por elevada parasitemia, altos niveles de IgM, fiebre, linfoadenopatía y una mortalidad hasta de un 10%. Esta fase dura de 2 a 3 semanas y en algunos casos puede evolucionar en forma subaguda durante meses. Cuando la puerta de entrada se localiza en cualquier

parte del tegumento se pueden producir lesiones inflamatorias conocidas con el nombre de chagoma. Si se presenta a nivel de la conjuntiva ocular se produce el signo de Romaña caracterizado por edema periocular, unilateral o bipalpebral indoloro, hiperémico, con inflamación e infección de los ganglios linfáticos (Fragata et al., 1997; Brener et al., 2000).

- La fase indeterminada o latente es caracterizada clínicamente porque al evaluar al individuo el examen radiológico está normal, el electrocardiograma convencional es normal, pero las pruebas serológicas y parasitológicas están positivas. Este período se presenta en el 50 % de los casos, puede durar varios años o indefinidamente, existe una parasitemia subpatente y es la forma clínica más frecuente (Brener et al., 2000; Maekelt, 2000).

- La fase crónica puede aparecer de 20 a 30 años después de la fase indeterminada, donde se detecta una reducción progresiva de la función cardíaca que constituye la complicación más grave de la miocardiopatía crónica chagásica. Esta afección se desarrolla en 20% al 30% de las personas infectadas y se evidencia por manifestaciones clínicas detectadas por el examen físico, electrocardiograma y la radiografía de tórax, todo ello sumado a la serología positiva para *Trypanosoma cruzi* y el antecedente epidemiológico. Las formas digestivas y/o dilataciones del tubo digestivo conocidas como megaesófago y megacolon, que se presenta en otros países de América del Sur como Brasil, Chile, Argentina y Bolivia, son otras de las manifestaciones clínicas de esta fase (Brant et al., 1998; Brener et al., 2000; Maekelt, 2000).

Se estima que en Latinoamérica existen 11 millones de personas infectadas con *Trypanosoma cruzi* y que al menos 100 millones de personas se encuentran en zonas de riesgo. Anualmente se estima que la enfermedad de Chagas causa más de 10.000 muertes en Latinoamérica (OPS, 2008).

En Venezuela, un estudio reciente llevado a cabo sobre 3.835 personas provenientes de 10 estados en donde el mal de Chagas es considerado endémico, reveló que el 11.7% de los individuos era seropositivo para *T. cruzi*. Casi un tercio de las personas infectadas (28.5%) resultó tener 30 años o menos, indicando que la infección ha estado ocurriendo activamente durante las últimas tres décadas, época para la cual comenzó el programa de erradicación del mal de Chagas. Se cree que el mal de Chagas es responsable de 21% de las muertes asociadas a problemas cardíacos en el país (Villalobos et al., 1994).

2.2 Minería de datos

La minería de datos integra diferentes técnicas, metodologías y herramientas que nos sirven de ayuda para transformar la información histórica de un sistema o proceso, en conocimiento útil. Dentro de los conceptos que se pueden encontrar, algunos de los más significativos son los siguientes:

"*Data Mining* es el conjunto de técnicas y herramientas aplicadas al proceso trivial de extraer y presentar el conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con el objeto de predecir de forma automatizada tendencias y comportamientos y/o descubrir de forma automatizada modelos previamente desconocidos" (Piatetski-Shapiro y Frawley, 1991).

"*Data Mining* es la exploración y análisis, mediante métodos automáticos o semiautomáticos, de grandes cantidades de datos para descubrir reglas o patrones significativos" (Berry y Gordon, 1997).

"*Data Mining* es el proceso de plantear varias preguntas y extraer información útil, patrones y tendencias de grandes cantidades de datos generalmente almacenados en bases de datos." (Thuraisingham, 1999).

2.3 Arquitectura de la aplicación

Para poder realizar una búsqueda de conocimiento aplicando técnicas de minería de datos, es necesaria la obtención de una base de datos coherentes y con datos relevantes del proceso a analizar. Esto hace necesaria la disposición de un sistema de adquisición, almacenamiento y manejo de la información. A continuación se definen algunos sistemas:

2.3.1 *Data warhousing*

Es el proceso donde se organiza una gran cantidad de datos variados y almacenados, de tal forma que facilite la recuperación de información para llevar a cabo el proceso analítico (Kimball, *et al.,* 1998).

El Data warehouses genera una base de datos utilizando múltiples fuentes que se fusionan en forma coherente. Este se diseña para contener un nivel de detalle apropiado, con la intención de hacer disponible todo tipo de datos y sus características, para informar y analizar. Así un Data Warehouse se puede identificar como un recipiente de datos transaccionales que proporciona consultas operativas e información para poder llevar a cabo análisis multidimensionales. De esta forma, dentro de un almacén de datos existen dos tecnologías complementarias: una relacional para consultas y una multidimensional para análisis.

2.3.2 Sistemas OLAP

Son aplicaciones de bases de datos orientadas a realizar un análisis multidimensional de datos mediante navegación del usuario por los mismos de modo asistido. La información es vista como cubos, que contienen categorías descriptivas (dimensiones) y valores cuantitativos (medidas). El usuario puede acceder a la información bajo diferentes niveles de abstracción: desde el detalle más bajo hasta agregaciones bajo diferentes dimensiones *(*Martínez, 2003).

De esta forma, un sistema OLAP puede ser visto como la generalización de un generador de informes. Estos sistemas no tienen la necesidad de desarrollar interfaces de consulta, al mismo tiempo que ofrecen un entorno único, válido para el análisis de cualquier información histórica,

orientado a la toma de decisiones. Es necesario definir dimensiones, jerarquías y variables, organizando de esta forma los datos.

Según Berry y Linoff (Berry y Linoff, 1997) se denominan OLAP a aquellos sistemas que:

- Soportan requerimientos complejos de análisis.
- Analizan datos desde diferentes perspectivas.
- Soportan análisis complejos multidimensionales dentro de volúmenes ingentes de información.
- Permiten la posibilidad de navegar sobre la información mediante informes jerarquizados.
- Disponen de funciones como: proyectar a una tabla, expandir, colapsar, girar, rotar, etc.

2.4 Fases de un problema clásico de minería de datos

El proceso para realizar la Minería de Datos es iterativo e interactivo ya que las fases de este proceso no están claramente diferenciadas. Es iterativo debido a que en la práctica estas fases pueden cambiar de orden y frecuentemente será necesario volver a la fase anterior y repetir ciertas acciones. Interactivo ya que se necesita de un experto para la selección y análisis de los datos, así como también para la interpretación y evaluación del conocimiento adquirido. Este proceso se estructura en seis fases que serán descritas a continuación.

Figura 2.1: Fases de un proceso clásico de minería de datos.

2.4.1 Definición del alcance y los objetivos

Esta fase consiste en obtener un conocimiento del desarrollo de la aplicación, determinar el conocimiento a utilizar y establecer los objetivos del proyecto. Esta fase es de gran importancia, debido a que en ella se establece la base para la realización de las fases siguientes del proceso, por lo consiguiente el éxito o fracaso del proyecto va a depender de las decisiones tomadas en esta fase.

En esta fase se determinan los factores que son susceptibles de un procesado automático, los cuellos de botella del dominio, los conocimientos a priori que se tienen del proceso, así como cuáles son los objetivos finales que se pretenden lograr y cuáles van a ser los criterios de rendimiento exigibles. Por lo tanto, esta fase requiere cierta dependencia usuario-analista, siendo necesario el establecimiento de unos canales de comunicación entre ambas partes.

Un factor clave es el conocimiento que se tiene del sistema, muchas veces, al no tener claro el proceso a analizar, esto puede implicar perdida de tiempo en las siguientes fases, por lo que se debe conocer el proceso a analizar en todos sus detalles, para poder evitar complicaciones en las fases posteriores debido a la falta de comprensión de algunas partes del proceso.

Los siguientes aspectos nos pueden ayudar alcanzar el éxito de un proyecto:

- Identificación correcta de los problemas a resolver: En ocasiones, puede parecer que la comprensión del problema sea trivial, pero muchas veces sucede que no se comprende completamente.

- Definición con precisión de los problemas: Los problemas generalizados deberán ser divididos en componentes más pequeños que puedan ser analizados con mayor facilidad.

- Resolver las ambigüedades**:** Es conveniente resolver las ambigüedades que puedan surgir debido a que la imagen mental del problema en la mente del cliente está formada por una gran cantidad de conceptos asociados, que él tiene asumidos, pero que probablemente, pueden no ser tan claros para aquellas personas que no conocen el proceso con profundidad.

- Determinar, dentro del número de problemas, el grado de importancia y dificultad de cada uno de ellos.

- Definir qué resultados se esperan conseguir**:** Es importante definir el resultado esperado como un modelo matemático, una gráfica, unos informes, etc. Esto nos permitirá dirigir las tareas posteriores hacia el objetivo buscado.

- Implementar los resultados obtenidos: Se debe tener en cuenta la forma en que se aplicarán los resultados obtenidos.

2.4.2 Selección de los datos relevantes

La selección de los datos relevantes para una operación de Minería de Datos no puede ser realizada de forma automatizada y por ello debe ser realizada por un analista. Esta tarea consiste en crear el conjunto de datos que se utilizará para la realización de la minería.
En esta fase el analista debe trabajar de forma coordinada con el usuario, para seleccionar los datos más relevantes del proceso y la disponibilidad de los mismos. Esto implica consideraciones sobre la homogeneidad y variación a lo largo del tiempo de los datos, los grados de libertad o la estrategia de muestreo.

La obtención de los datos se puede realizar de varias maneras, directamente de un sistema transaccional, de archivos o a partir de almacenes de datos (*Data Warehouse*). Comúnmente la obtención de los datos viene dada en función de la disponibilidad de los mismos, como existencia de bases de datos, datos almacenados en archivos, necesidad de implementar un nuevo sistema de adquisición de datos, etc.

2.4.3 Preprocesamiento y limpieza de datos

El preprocesamiento de datos es una fase muy importante dentro del proceso de minería de datos, el objetivo principal es llevar el conjunto original de datos en un nuevo conjunto más manejable y significativo. Según Pyle (1999), un 60% del tiempo se dedica al preprocesado de los datos.

El preproceso de los datos incluye cuatro etapas principales: Identificación y conversión de tipos, imputación (rellenar los datos inexistentes), identificación de espurios (*outliers*) y eliminación de ruido y datos incompletos.

IDENTIFICACIÓN Y CONVERSIÓN DE ATRIBUTOS

Las primeras tareas en el preprocesado de datos son las más laboriosas, ya que debemos identificar casi manualmente, los diferentes tipos de atributos existentes en la base de datos y

convertirlos a otros tipos de variables según las necesidades posteriores. Podemos clasificar los atributos en dos grupos:

- Numéricos o Cuantitativos**:** También algunas veces llamados "continuos".

- Nominales o Cualitativos: También algunas veces llamados "discretos". corresponden a valores que tienen distintos símbolos generalmente denominados etiquetas o nombres.

Algunos atributos deben ser convertidos dependiendo de los algoritmos de la minería de datos que se van a utilizar, para que éstos puedan ser tratados convenientemente. Una vez que tenemos los tipos de atributos adaptados a nuestras necesidades, será conveniente realizar las siguientes fases:

- Detectar los espurios y eliminarlos.

- Rellenar los datos inexistentes.

- Eliminar el ruido.

2.4.4 Transformación de los datos

Esta es una de las fases críticas del proceso de minería de datos, se necesita de parte del experto un buen conocimiento y una buena intuición ya que ésto determinará el éxito o fracaso del proceso.

Por otro lado, se busca preparar la información de la mejor forma posible, para facilitar el trabajo a los algoritmos de minería de datos que se van a utilizar en la siguiente fase y además, reducir la información redundante para simplificar las tareas.

2.4.5 Uso de los algoritmos de la minería de datos

Para aplicar los algoritmos de minería de datos, se deben preparar los datos de entrada, ya que, en la mayoría de las ocasiones estos datos provienen de fuentes heterogéneas, es decir, no tienen el formato adecuado para la aplicación de los mismos.

Las herramientas de minería de datos que son utilizadas para la extracción de conocimientos se clasifican en dos grupos:

- Técnicas de verificación (en las que el sistema se limita a comprobar hipótesis suministradas por el usuario).

- Métodos de descubrimiento (en los que se han de encontrar patrones potencialmente interesantes de forma automática, incluyendo en este grupo todas las técnicas de predicción) (Daedalus, 2002).

Los métodos de descubrimiento pueden ser de carácter **descriptivo o predictivo.** Los predictivos son utilizados para predecir el comportamiento futuro de un sistema o entidad, mientras que los descriptivos son usados para la compresión de un sistema o proceso.

Los algoritmos de minería de datos pueden ser utilizados para alguna de las siguientes tareas (Westphal y Blaxton, 1998).

- Agrupamiento o segmentación: Se desea separar los datos en subgrupos o clases significativos, los miembros de un subgrupo comparten características entre sí y se diferencia de los otros subgrupos, esto permite el tratamiento particularizado de cada una de estas agrupaciones.

- Asociaciones: Establece posibles relaciones entre sucesos o acciones que son aparentemente independientes entre si. Así, se puede determinar como la ocurrencia de un suceso pude inducir la aparición de otro u otros.

- Secuenciamiento: Este concepto es similar al anterior, pero en el que se incluye un factor tiempo. Es decir, permite reconocer el tiempo en el que transcurre que un suceso induzca a otro.

- Reconocimiento de patrones: Se trata de analizar la asociación de una señal o información de entrada con aquella o aquellas con las que guarda mayor similitud, y que están ya catalogadas en el sistema. Generalmente se usan para identificar las causas de problemas o incidencias y buscar las posibles soluciones, siempre y cuando se disponga de la base de información necesaria.

- Previsión: Se busca establecer el comportamiento futuro más probable de una variable o una serie de variables a partir de la evolución pasada y presente de las mismas o de otras de las cuales dependan. Las técnicas asociadas a estas herramientas tienen ya un elevado grado de madurez.

- Simulación: Comparan la situación actual de una variable y su posible evolución futura según la variación probable de las que depende.

- Optimización: resuelve el problema de la minimización o maximización de una función que depende de una serie de variables, encontrando los valores de éstas que satisfacen la condición de máximo (típicamente beneficios), o mínimo (típicamente costes). Normalmente suele haber unas restricciones, que hacen que no todas las posibles soluciones sean aceptables, de modo que el universo de búsqueda se reduce a aquellas soluciones que satisfagan las restricciones.

- Clasificación: Agrupa a todas las herramientas que permiten asignar a un elemento la pertenencia a un determinado grupo o clase. Esto se lleva a cabo a través de la dependencia de la pertenencia a cada clase en los valores de una sede de atributos o variables. Se establece un perfil característico de cada clase y su expresión, en términos

de un algoritmo o reglas, en función de las distintas variables. Se establece también el grado de discriminación o influencia de estas últimas. Con ello, es posible clasificar un nuevo elemento una vez conocidos los valores de las variables presentes en él.

2.4.6 Interpretación de los resultados

Una vez obtenido el modelo predictivo o descriptivo, se debe proceder a su evaluación comprobando que los resultados obtenidos son válidos y suficientemente satisfactorios. La obtención de los resultados validos dependerá de varios factores, tales como; la definición de medidas de interés del tipo estadístico que permitan el filtrado de la información de forma automática, técnicas de visualización que faciliten validar los resultados o búsqueda manual de conocimiento útil entre los resultados obtenidos.

El grado de experiencia y conocimiento del analista juega un papel muy importante en esta fase. La cantidad de información obtenida dependerá del grado de conocimiento del analista que tenga sobre el problema.

Las decisiones tomadas en esta fase irán encaminadas en dos direcciones:

- Verificación de resultados: La verificación de resultados incluye determinar el grado de cumplimiento de los objetivos finales establecidos durante la primera fase del proceso de minería de datos, así como la validación de la información extraída. Durante esta fase se debe verificar la coherencia de la información obtenida con otros tipos de conocimiento ya previamente asentado y aceptado, resolviendo las posibles inconsistencias existentes. Si los objetivos finales han sido alcanzados, se procederá a la consolidación del conocimiento descubierto, incorporándolo al sistema. O simplemente documentándolo y enviándolo a la parte interesada. En caso contrario se procederá a la obtención de más información.

- Obtención de más Información: La información extraída se utilizará como información a priori para la extracción de más información. Para ello será necesario retornar a alguna de las fases anteriores del proceso de minería de datos y modificar algunas de las decisiones tomadas durante esas fases, haciendo para ello uso de la nueva información obtenida. De esta forma el proceso de minería de datos se convierte en un proceso potencialmente iterativo. Alguna de las decisiones que pueden ser tomadas para la obtención de más información son, por ejemplo: recolección de nuevos datos, separación de datos en clases, transformaciones de las variables, eliminación de datos, selección de otros algoritmos de minería de datos, cambio en los parámetros introducidos en los algoritmos, delimitación del campo de búsqueda, etc.

2.5 Técnicas y algoritmos de minería de datos

Debido a que se espera obtener relaciones entre las características de hábitat de los individuos con la enfermedad de Chagas, la técnica más apropiada para el análisis es el descubrimiento de reglas de asociación, esta técnica se caracteriza por buscar patrones para llegar a reglas que asocien los diferentes atributos contenidos en la base de datos.

2.5.1 Reglas de asociación

Las reglas de asociación es un método que se utiliza para descubrir interesantes relaciones entre variables en una gran base de datos. Agrawal, Imielinski, y Swami (1993) definen el problema de minería de reglas de asociación de la siguiente manera:

Sea $I = \{I_1, I_2, \ldots, I_m\}$ un conjunto de atributos binarios llamados ítems. Sea T una base de datos de transacciones, donde cada transacción t es representada por un vector binario, con $t[k] = 1$, si en t aparece el ítem I_k, $t[k] = 0$ en caso contrario. Sea X un conjunto de algunos elementos de I, decimos que una transacción t satisface X si para todos los elementos I_k de X, $t[k] = 1$. Una regla de asociación está definida implícitamente de la forma $X \Rightarrow Y$, donde

$X, Y \subseteq I \; y \; X \cap Y = \emptyset$. Los conjuntos de ítems $X \, e \, Y$ son llamados antecedente o predecesor (lado izquierdo de la regla) y sucesor o consecuente (lado derecho de la regla).

Para ilustrar el concepto, utilizamos un pequeño ejemplo de análisis de la cesta de compras de un supermercado. Se tiene el conjunto de ítems $I = \{carne, carbón, cerveza, pan\}$ y una base de datos que contienen los ítems que se muestran en la figura 2.2. Un ejemplo de una regla para el supermercado podría ser $\{carne, carbón\} \Rightarrow \{cerveza\}$, lo que significa que si un cliente compra carne y carbón también compra cerveza.

Transacción ID	ítems
1	carne, carbón, cerveza
2	carbón, cerveza
3	pan, carne
4	carne, carbón, cerveza
5	carne, carbón

Figura 2.2: Ejemplo de una base de datos de un supermercado con cinco transacciones

Para seleccionar ciertas reglas de interés de todo el conjunto de reglas, se pueden utilizar limitaciones o medidas de interés. Las medidas más conocidas son soporte (o cobertura) y confianza (o eficiencia).

El soporte $supp(X)$, se define como la proporción de transacciones que contiene el conjunto X. En la base de datos del ejemplo de la figura 2.2 el conjunto de ítems $\{carne, carbón\}$ tienen un soporte de $3/5 = 0.6$, ya que se produce en el 60% de todas las transacciones.

El soporte de una regla de asociación es la proporción de transacciones que contienen tanto al predecesor como al consecuente. La confianza de una regla de asociación se define como la probabilidad condicional de que una transacción que contenga X y que también contenga Y. Así para una asociación $X \Rightarrow Y$:

$$supp\left(X \Rightarrow Y\right) = supp(X \cup Y)$$

$$conf\left(X \Rightarrow Y\right) = \frac{supp(X \cup Y)}{supp(X)}$$

A continuación se ilustra el cálculo del soporte y la confianza para la regla $\{carne, carbón\} \Rightarrow \{cerveza\}$ de la base de datos del ejemplo, esta regla tiene un soporte de $2/5 = 0.4$, ya que se produce en el 40% de todas las transacciones y una confianza del $0.4/\ 0.6 = 0.6666$, lo que significa que el 66.6% de las transacciones que contengan carne y carbón son correctas. La confianza puede ser interpretada como la probabilidad de encontrar el consecuente dado el predecesor de la regla.

La búsqueda inicial de reglas de asociaciones permite encontrar todas las asociaciones que satisfaga las restricciones iníciales de soporte y confianza, esto puede llevar a obtener un gran número de de reglas.

Una solución práctica al problema de encontrar muchas reglas de asociación que satisfagan las restricciones de cobertura y confianza, es filtrando el resultado usando medidas de interés adicionales. Una medida popular es el lift (Brin, Motwani, Ullman, y Tsur, 1997).

El lift de una regla es definido como $lift(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)supp(Y)}$. Esta medida compara un subconjunto de los datos contra todos, dando un resultado más generalizado que el soporte y la confianza, los cuales solo nos provee resultados evaluados en un subconjunto de todos los datos. Valores de lift mayores a 1 indican que el consecuente es más frecuente en transacciones que contienen también el antecedente, que en transacciones que no la contienen.

Por ejemplo, considerando la asociación $\{carbón\} \Rightarrow \{cerveza\}$. Si $supp(\{cerveza\})$ = 0.6 y $conf\left(\{carbón\} \Rightarrow \{cerveza\}\right) = 0.75$. Entonces $lift\left(\{carbón\} \Rightarrow \{cerveza\}\right)$ = 1.25

Como contraste, consideramos otra asociación con la misma confianza,$\{carbón\} \Rightarrow \{carne\}$, donde si $supp(\{carne\})$ = 0.8, entonces: $lift\left(\{carbón\} \Rightarrow \{carne\}\right)$ = 0.9375. Estos valores relativos de lift indican que el carbón tiene un mayor impacto en la frecuencia de la cerveza que en la frecuencia de la carne.

Algoritmo Apriori:

El algoritmo *Apriori* (Agrawal 1994), ataca el problema reduciendo el número de conjuntos considerados. El usuario define una cobertura mínima, *Apriori* genera todos los conjuntos que cumplan con la condición de tener una cobertura mayor o igual a la introducida por el usuario. Para cada conjunto frecuente X se generan todas las reglas de asociación $A \Rightarrow C$ tales que $A \cup C = X \; y \; A \cap C = \emptyset$. Cualquier regla que no satisfaga la condiciones impuestas por el usuario, como por ejemplo la confianza mínima, se desechan y las reglas que si cumplan se conservan. El algoritmo *Apriori* se presenta a continuación (ver figura 2.3).

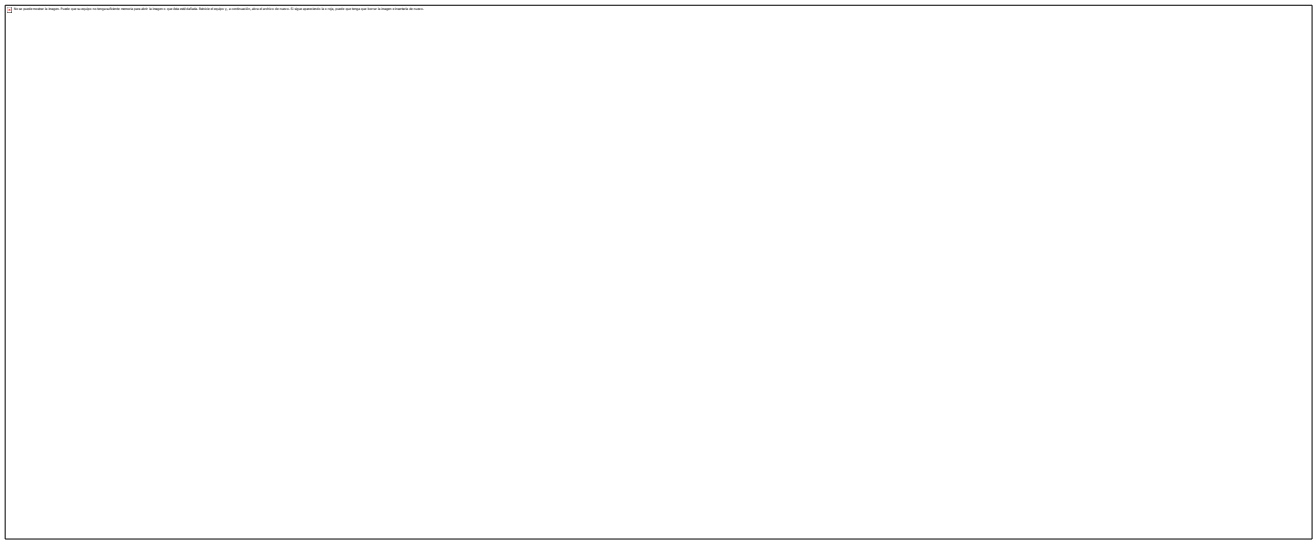


Figura 2.3 Algoritmo *Apriori*

Como vemos, comienza calculando en F1 todos los conjuntos frecuentes formados por un solo ítem y mientras se puedan combinar elementos en el último Fk-1 formado (en la primera iteración k = 2), genera mediante una llamada a *AprioriGen* todos los posibles conjuntos candidatos de k elementos, recorre todas las transacciones de la tabla para calcular el soporte de cada conjunto candidato y en Fk permanecen sólo aquellos conjuntos cuyo soporte sea mayor que MinSop. Devuelve al terminar todos los conjuntos frecuentes Fk, $k \geq 1$ (ver figura 2.4).

Algoritmo AprioriGen
Entrada: F_{k-1}: Conjunto frecuente (k-1)-itemsets
Salida: C_k: Conjunto candidato k-itemsets
Método:
$C_k = \varnothing$
Insertar en C_k, todos los conjuntos ordenados de ítems, estos son:
$\{ Z[1], ..., Z[k-1], Y[k-1] \mid Y, Z \in F_{k-1}, Z[1] = Y[1], ..., Z[k-2] = Y[k-2] \wedge Z[k-1] < Y[k-1] \}$
Borrar todos los $Z \in C_k$, tal que algún (k-1)-subconjunto de Z no esté en F_{k-1}
Return C_k

Figura 2.4: Algoritmo *AprioriGen*

Notemos que este algoritmo forma todos los conjuntos ordenados de k elementos tal que sus subconjuntos de k-1 elementos coincidan con algún conjunto frecuente de k-1 elementos.

2.6 Herramienta utilizada para la realización de la minería de datos

En este proyecto se seleccionó el paquete estadístico R, por su condición de ser software libre y su alto rendimiento en la realización de tareas estadísticas que nos servirán de apoyo para llegar al éxito en este proyecto. R es un lenguaje y un entorno de código abierto para realizar cómputos y gráficos estadísticos. Es un proyecto GNU, que fue desarrollado en los laboratorios BELL (Ihaka y Gentleman, 1996). R ofrece una amplia variedad estadística (modelos lineales y no lineales, pruebas estadísticas clásicas, análisis de series de tiempos, clasificación, clustering, etc.) y técnicas de graficación. Además, es altamente extensible por su condición de código abierto que

ofrece una ruta para la participación en el desarrollo de este lenguaje. Una de las fortalezas de R es la facilidad que tiene al adquirir información de calidad sobre el manejo de esta herramienta. R contiene una suite integrada para la manipulación de datos, cálculo y visualización de gráficos, que incluye:

- Un efectivo manejo de datos y facilidad de almacenamiento.
- Un conjunto de operadores para los cálculos de arreglos, en particular matrices.
- Una colección de instrumentos, coherentes e integrados para el análisis de los datos.
- Facilidades gráficas para el análisis de los datos.
- Un lenguaje de programación bien desarrollado, simple y eficaz que incluye estructuras condicionales, bucles, funciones recursivas, etc.

Para realizar la minería de reglas de asociación se utilizará una extensión al paquete R, incluyendo la biblioteca ARules**.** La cual proporciona la infraestructura necesaria para crear y manipular conjuntos de datos de entrada a los algoritmos de minería y para facilitar el análisis de los resultado obtenidas.

2.7 Base de datos

Los datos que se requieren para establecer los factores de riesgos epidemiológicos asociados con la enfermedad de Chagas, en las localidades rurales de la parroquia Santa Fe, municipio Sucre del estado Sucre, fueron recopilados a través de una encuesta epidemiológica (Aza, 2003) que conduce a una caracterización epidemiológica de cada familia, la cual fue diseñada y validada por especialistas en sociología de la salud. Además, se tomaron muestras sanguíneas de individuos, de ambos géneros y diferentes edades, que fueron procesadas para evaluar la serología de la enfermedad de Chagas.

Las encuestas y diagnósticos serológicos fueron realizadas por el personal del Laboratorio de Biología Molecular del Instituto de Investigaciones en Biomedicina y Ciencias Aplicadas de la Universidad de Oriente (IIBCA-UDO) y fue facilitada una copia de los originales que consistió en 293 encuestas y resultados de 88 pruebas serológicas a la enfermedad de Chagas.

Los datos procedentes de las encuestas están estructurados en cuatro secciones:

- La primera sección define las características socioeconómicas expresadas por el jefe del grupo familiar, esta contiene la siguiente información: nombre y apellido, lugar de nacimiento, edad, grado de instrucción y ocupación.

- La segunda sección hace referencia al nivel de conocimiento que tienen las personas sobre el modo en que se transmite la enfermedad de Chagas y las consecuencias que esta acarrea. Además, se especifica si las personas han sido picadas por el insecto (vector) transmisor.

- La tercera sección describe las características de la vivienda donde habita el grupo familiar, especificando el tipo de construcción, distribución de ambientes, servicios públicos existentes, deposición de excretas, el tiempo que ha vivido en la zona, etc.

- La última sección hace referencia a la presencia de animales domésticos relacionados con la transmisión de la enfermedad de Chagas, reservorios naturales del parásito *Trypanosoma cruzi*, como lo son el rabipelado o cachicamo, y palmas cerca de la vivienda, hábitat por excelencia del insecto vector.

Por otra parte, se tiene el análisis de la serología de las muestras sanguíneas que fueron tomadas aleatoriamente a 88 individuos, lo cual correspondió a una muestra estadísticamente representativa (30% del total de los individuos censados). Dicho análisis determinó la presencia

(seropositividad) o ausencia de anticuerpos totales (IgM, IgA e IgG) *anti-Trypanosoma cruzi*, agente causal de la enfermedad de Chagas.

De estas encuestas, se seleccionaron los datos correspondientes a las características epidemiológicas más relevantes que están asociadas frecuentemente con la transmisión de la enfermedad de Chagas ("Casa enferma"), como son las socioeconómicas, conocimiento de la enfermedad, características de la vivienda, presencia de animales domésticos relacionados con la transmisión de la enfermedad, reservorios naturales, presencia de palmas y diagnósticos serológicos.

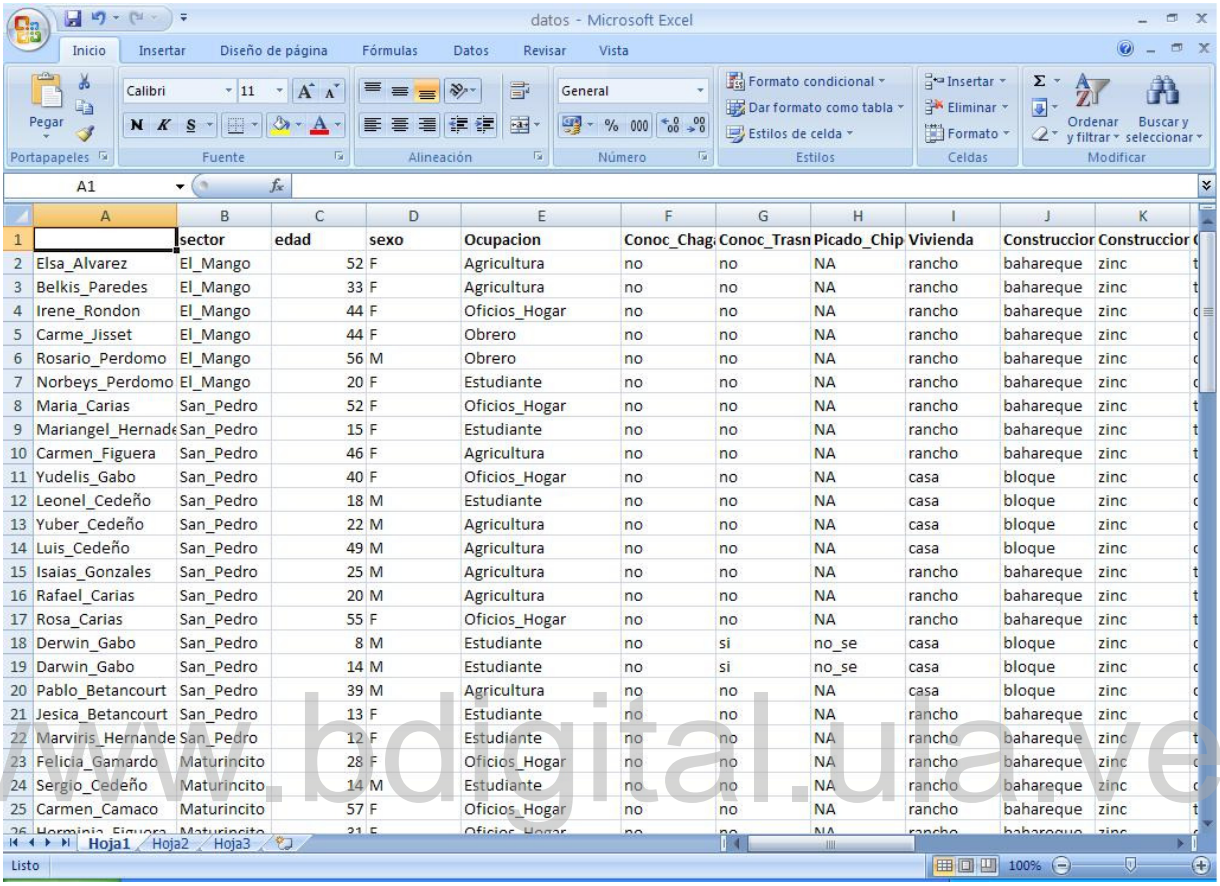
Se utilizó una hoja de cálculo (EXCEL) para representar los datos obtenidos, donde cada columna contiene las características epidemiológicas y el resultado del diagnóstico serológico de cada individuo (ver figura 2.5), con el fin de preparar el archivo que será utilizado por la herramienta estadística R, para el análisis y aplicación de la minería de datos.

2.7.1 Descripción de los datos

Se seleccionaron las características o atributos útiles de las variables, debido a que existían datos que se consideraron innecesarios o irrelevantes para la investigación que se lleva a cabo. A continuación se describen los atributos que permanecieron en cada fuente de datos y cuales fueron eliminados.

Características socioeconómicas

En esta sección de la encuesta, se consideraron como variables sin relevancia el nombre, lugar de nacimiento y grado de instrucción, debido a la naturaleza de la enfermedad de Chagas y no poseer relación directa con la misma. En la tabla 2.1 se presentan las variables relevantes de los datos que pertenecen a la sección características socioeconómicas.



	A	B	C	D	E	F	G	H	I	J	K
1		sector	edad	sexo	Ocupacion	Conoc_Chag	Conoc_Trasn	Picado_Chip	Vivienda	Construccior	Construccior
2	Elsa_Alvarez	El_Mango	52	F	Agricultura	no	no	NA	rancho	bahareque	zinc
3	Belkis_Paredes	El_Mango	33	F	Agricultura	no	no	NA	rancho	bahareque	zinc
4	Irene_Rondon	El_Mango	44	F	Oficios_Hogar	no	no	NA	rancho	bahareque	zinc
5	Carme_Jisset	El_Mango	44	F	Obrero	no	no	NA	rancho	bahareque	zinc
6	Rosario_Perdomo	El_Mango	56	M	Obrero	no	no	NA	rancho	bahareque	zinc
7	Norbeys_Perdomo	El_Mango	20	F	Estudiante	no	no	NA	rancho	bahareque	zinc
8	Maria_Carias	San_Pedro	52	F	Oficios_Hogar	no	no	NA	rancho	bahareque	zinc
9	Mariangel_Hernade	San_Pedro	15	F	Estudiante	no	no	NA	rancho	bahareque	zinc
10	Carmen_Figuera	San_Pedro	46	F	Agricultura	no	no	NA	rancho	bahareque	zinc
11	Yudelis_Gabo	San_Pedro	40	F	Oficios_Hogar	no	no	NA	casa	bloque	zinc
12	Leonel_Cedeño	San_Pedro	18	M	Estudiante	no	no	NA	casa	bloque	zinc
13	Yuber_Cedeño	San_Pedro	22	M	Agricultura	no	no	NA	casa	bloque	zinc
14	Luis_Cedeño	San_Pedro	49	M	Agricultura	no	no	NA	casa	bloque	zinc
15	Isaias_Gonzales	San_Pedro	25	M	Agricultura	no	no	NA	rancho	bahareque	zinc
16	Rafael_Carias	San_Pedro	20	M	Agricultura	no	no	NA	rancho	bahareque	zinc
17	Rosa_Carias	San_Pedro	55	F	Oficios_Hogar	no	no	NA	rancho	bahareque	zinc
18	Derwin_Gabo	San_Pedro	8	M	Estudiante	no	si	no_se	casa	bloque	zinc
19	Darwin_Gabo	San_Pedro	14	M	Estudiante	no	si	no_se	casa	bloque	zinc
20	Pablo_Betancourt	San_Pedro	39	M	Agricultura	no	no	NA	casa	bloque	zinc
21	Jesica_Betancourt	San_Pedro	13	F	Estudiante	no	no	NA	rancho	bahareque	zinc
22	Marviris_Hernande	San_Pedro	12	F	Estudiante	no	no	NA	rancho	bahareque	zinc
23	Felicia_Gamardo	Maturincito	28	F	Oficios_Hogar	no	no	NA	rancho	bahareque	zinc
24	Sergio_Cedeño	Maturincito	14	M	Estudiante	no	no	NA	rancho	bahareque	zinc
25	Carmen_Camaco	Maturincito	57	F	Oficios_Hogar	no	no	NA	rancho	bahareque	zinc

Figura 2.5: Muestra de la base de datos de transacciones.

Nombre de la variable	Descripción
Sector	Representa el sector donde habita el individuo
Edad	Menores de 20 años indican transmisión activa.
Sexo	Representa el género del individuo
Ocupación	Indica con qué frecuencia está en riesgo el individuo de contraer la enfermedad

Tabla.2.1: Atributos relevantes de la sección Características Socioeconómicas.

Características referentes a la enfermedad

En esta sección de la encuesta, se consideró como consultas sin relevancias: la gravedad y el riesgo que la persona encuestada tiene como opinión de contraer la enfermedad, ya que estos datos en su mayoría no fueron respondidos en las encuestas. En la tabla 2.2 se presentan las variables relevantes de los datos que pertenecen a la sección características referentes a la enfermedad.

Nombre de la variable	Descripción
Conocimiento de la existencia del Chagas	Indica si la persona tiene conocimiento de la existencia de la enfermedad
Conocimiento del transmisor o Vector	Indica si la persona tiene conocimiento del insecto transmisor de la enfermedad
Contacto con el vector	Indica si la persona ha sido picada por el insecto transmisor (vector) de la enfermedad.

Tabla 2.2: Atributos relevantes de la sección Conocimiento sobre la enfermedad.

Características de la vivienda

Las variables consideradas sin relevancia en esta sección de la encuesta son: distribución del ambiente, servicios públicos existentes, debido a que se conocen con exactitud los servicios públicos que existen en la zona. En la tabla 2.3 se presentan las variables relevantes de los datos que pertenecen a la sección características de la vivienda.

Nombre de la variable	Descripción
Tipo de Vivienda	Indica el tipo de vivienda donde habita la persona, por ejemplo: rancho, casa o quinta.
Material de Paredes	Indica el tipo de material con el que fueron construida las paredes de la vivienda que favorece o no la colonización del chipo
Material de Techo	Indica el tipo de material con el que fueron construido el techo de la vivienda que favorece o no la colonización del chipo
Material de Piso	Indica el tipo de material con el que fueron construido el piso de la vivienda que favorece o no la colonización del chipo
Deposición de Excretas	Indica la exposición que tiene el individuo a contraer la enfermedad en el momento de la deposición de las excretas.
Tiempo en la Zona	Indica si la enfermedad es endémica de la zona.
Lugar Previo	Indica si la enfermedad fue contraída fuera de la zona.
Visitado poblados	Indica en que población pudo la persona contraer la enfermedad
Vivienda Fumigada	Medida de organismos sanitarios de protección con el vector.
Usan Insecticidas	Medidas de protección familiar contra el vector.

Tabla 2.3: Atributos relevantes de la sección características de la vivienda.

Presencia de animales relacionados con la transmisión de la enfermedad

En esta sección de la encuesta la única variable eliminada fue la presencia de cafetales en la zona, debido a que esta variable no tiene relevancia con la investigación en curso. En la tabla 2.4 se presentan las variables relevantes de los datos que pertenecen a la sección presencia de animales.

Nombre de la variable	Descripción
Perros	Indica la presencia de perros (hospederos intermediarios del parásito) en la vivienda del individuo
Gatos	Indica la presencia de gatos (hospederos intermediarios del parásito) en la vivienda del individuo.
Aves de Corral	Indica la presencia de aves de corral (atractores del insectos vectores) en la vivienda del individuo
Burros	Indica la presencia de burros (hospederos intermediarios del parásito) en la vivienda del individuo
Rabipelados-Cachicamos	Indica la presencia de rabipelados o cachicamos en la cercanía de la vivienda, reservorios naturales del parásito
Consumo RC	Indica si el individuo ha consumido rabipelados o cachicamos
Palmas	Indica la presencia de palmas cerca de la vivienda, hábitat por excelencia del insecto vector.

Tabla 2.4: Atributos relevantes de la sección presencia de animales.

Diagnóstico serológico

De las pruebas sanguíneas se eliminaron los atributos siguientes: nombre de la persona, edad y sector, ya que estos son reflejados en la sección de características socioeconómicas. Se utilizaron para relacionar las encuestas con las pruebas sanguíneas y así poder crear el almacén de datos, en

la tabla 2.5 se presentan las variables relevantes de los datos que pertenecen al diagnóstico serológico.

Nombre de la variable	Descripción
serología	Determina la presencia(seropositividad) o ausencia de anticuerpos totales (IgM, IgA e IgG) *anti-Trypanosoma cruzi*, agente causal de la enfermedad de Chagas

Tabla 2.5: Atributos relevantes de la pruebas sanguíneas.

Capítulo 3

Resultados

En este capitulo se lleva a cabo el desarrollo de las fases: preparación de los datos, modelado y evaluación de la metodología CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) (Chapman, Kerber, Khabaza, Reinartz, Shearer, & Wirth, 2002).

3.1 Exploración de los datos

En esta fase se desarrollaran una serie de actividades, que permitirán tener una mayor comprensión de los datos, y describir los primeros conocimientos o subconjuntos interesantes para formular hipótesis, sobre las condiciones socio-económicas y medio-ambientales de los individuos que por su hábitat tienen riesgo de contraer la enfermedad de Chagas.

Se aplicaron algunas técnicas de estadística básica para analizar con mayor detalle las propiedades de las variables más relevantes de cada fuente de datos. Se utilizaron procedimientos descriptivos y de frecuencia que permitieron obtener información tanto de las variables categóricas como de las numéricas. Las distribuciones de frecuencia dan a conocer los valores que adoptan cada variable, el número de veces que ésta se repite y su porcentaje correspondiente. Además, se utilizaron gráficos estadísticos para tener una mejor compresión, los gráficos utilizados fueron los histogramas. Para aplicar estas técnicas de estadística básica se utilizó la herramienta estadística R, con la cual fue posible obtener una serie de resultados descritos a continuación.

Características socioeconómicas

En esta sección se muestran los valores correspondientes a cada variable con su respectivo histograma de las características socioeconómicas de los individuos pertenecientes a la comunidad de San Pedro.

Resultados de la variable Sector

Los valores correspondientes de la variable sector se muestran en la tabla 3.1 los cuales se encuentran graficados en la figura 3.1.

Valores	Numero de incidencia	Porcentaje
Cambural	5	5.68%
El Maco	30	34,10%
El Mango	6	6,82%
El Puente	2	2,27%
La Sabana	18	20,45%
Maturincito	6	6,82%
San Pedro	21	23,87%

Tabla 3.1: Valores correspondientes a la variable sector

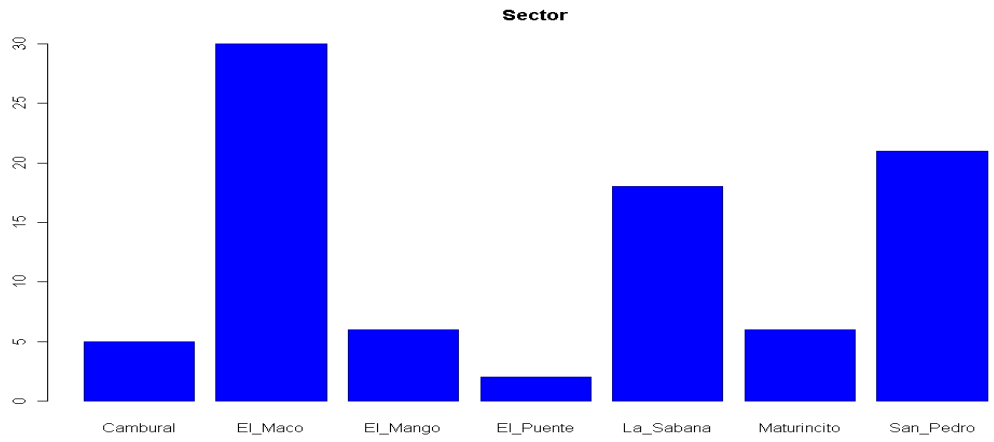


Figura 3.1: Histograma correspondiente de la variable sector

Resultados de la variable edad

En esta sección se muestran los valores correspondientes a la variable edad (ver tabla 3.2) y a continuación se describe el significado de la notación empleada.

- *Valor mínimo:* representa el valor más pequeño.
- *Primer cuartil:* representa el valor de grupo de datos en el cual está el 25% de todos los valores de la distribución.
- *Mediana:* representa el valor central de los datos ordenados de forma que el 50% de los datos están por debajo de la misma.
- *Media:* Corresponde al valor medio de los datos.
- *Tercer cuartil:* Valor del grupo de datos en el cual está el 75% de todos los valores de la distribución.
- *Valor máximo:* Representa el valor más grande.

Valor mínimo	Primer cuartil	Mediana	Media	Tercer cuartil	Valor máximo
2	13	24	26	40	74

Tabla 3.2: Valores correspondientes a la variable edad

Resultados de la variable sexo

En esta sección se muestran los valores correspondientes a la variable sexo (ver tabla 3.3), los cuales son representados mediante un histograma (ver figura 3.2).

Valores	Numero de instancias	Porcentaje
Femenino	49	55,68%
Masculino	39	44,32%

Tabla 3.3: Valores correspondientes a la variable sexo

Figura 3.2: Histograma correspondiente a la variable sexo

Resultados de la variable ocupación

En esta sección se muestran los valores correspondientes a la variable ocupación (ver tabla 3.4), los cuales son representados con un histograma (ver figura 3.3).

Valores	Numero de instancias	Porcentaje
Agricultura	17	19,32%
Estudiante	38	43,18%
Oficio del hogar	27	30,68%
Otras Ocupaciones	6	6,82%

Tabla 3.4: Valores correspondientes a la variable ocupación

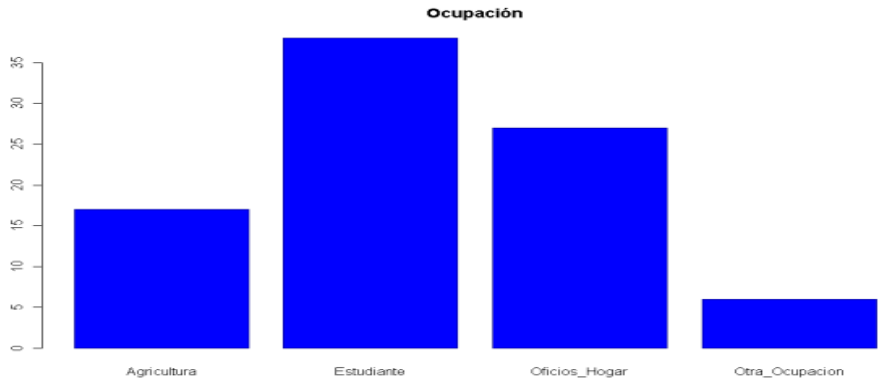


Figura 3.3: Histograma correspondiente de la variable ocupación

Características referentes a la enfermedad

En esta sección se especifican cada uno de los datos correspondientes a la enfermedad de Chagas. La tabla 3.5 contiene el número de incidencias y el porcentaje de la población que conoce la enfermedad, además estos valores son representados con un histograma (ver figura 3.4).

Valores	Numero de instancias	Porcentaje
Si	26	29.55%
No	62	70.45%

Tabla 3.5: Valores correspondientes a la variable conoce la enfermedad

Figura 3.4: Histograma correspondiente a la variable conoce la enfermedad

La tabla 3.6 contiene los valores correspondientes que determinan el conocimiento por la población del insecto transmisor de la enfermedad, representados con un histograma (ver figura 3.5).

Valores	Numero de instancias	Porcentaje
Si	26	29.55%
No	62	70.45%

Tabla 3.6: Valores correspondientes que determinan el conocimiento de la población del insecto transmisor de la enfermedad

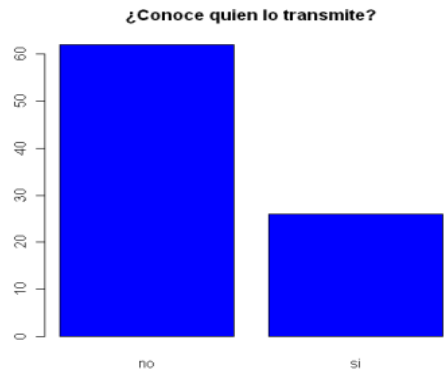


Figura 3.5: Histograma de valores correspondientes que determinan el conocimiento de la población del insecto transmisor de la enfermedad

La tabla 3.7 contiene los valores correspondientes a la variable contacto con el vector, representados con un histograma (ver figura 3.6).

Valores	Numero de instancias	Porcentaje
No	21	23.86%
No se	2	2.27%
Si	5	5.68%
NA(no respuesta)	60	68.19%

Tabla 3.7: Valores correspondientes contacto con el vector

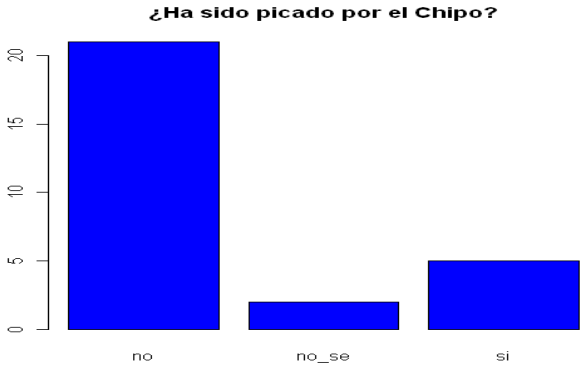


Figura 3.6: Histograma correspondiente a la variable contacto con el vector

Características de la vivienda

En esta sección se muestran los valores correspondientes a cada variable con su respectivo histograma de las características de la vivienda de los individuos pertenecientes a la comunidad de San Pedro.

La tabla 3.8 contiene los valores correspondientes a la variable tipo de vivienda, representados con un histograma (ver figura 3.7)

Valores	Numero de instancias	Porcentaje
Casa	45	51.14%
Rancho	43	48.86%

Tabla 3.8: Valores correspondientes a la variable tipo de vivienda

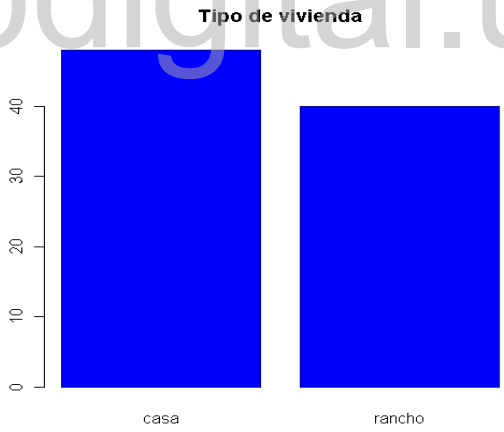


Figura 3.7: Histograma correspondiente a la variable tipo de vivienda

La tabla 3.9 contiene los valores correspondientes a la variable construcción de paredes, representados con un histograma (ver figura 3.8)

Valores	Numero de instancias	Porcentaje
bahareque	41	46,59%
bloque	45	51,14%
zinc	2	2,27%

Tabla 3.9: Valores correspondientes a la variable construcción de paredes

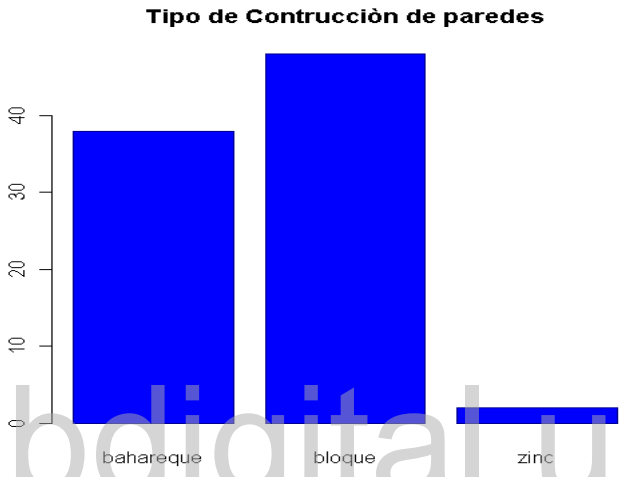


Figura 3.8: Histograma correspondiente a la variable construcción de paredes

La tabla 3.10 contiene los valores correspondientes a la variable construcción de techos, representados con un histograma (ver figura 3.9)

Valores	Numero de instancias	Porcentaje
Concreto	1	1,14%
Zinc	87	98,86%

Tabla 3.10: Valores correspondientes a la variable construcción de techos

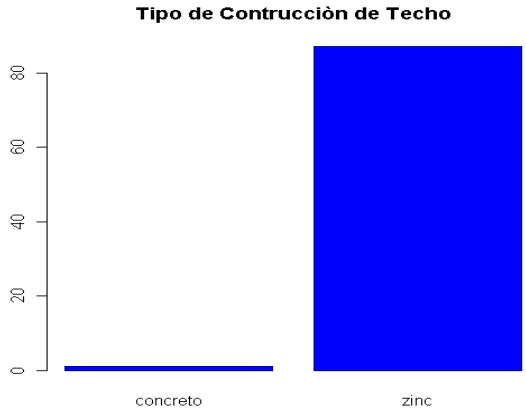


Figura 3.9: Histograma correspondiente a la variable construcción de techos

La tabla 3.11 contiene los valores correspondientes a la variable construcción de piso, representados con un histograma (ver figura 3.10)

Valores	Numero de instancias	Porcentaje
Cemento	76	86,36%
Tierra	12	13,64%

Tabla 3.11: Valores correspondiente a la variable construcción de pisos

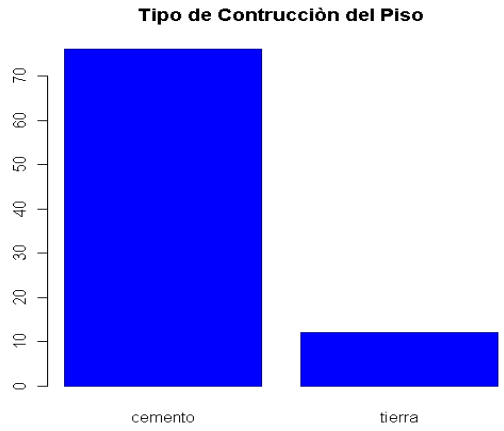


Figura 3.10: Histograma correspondiente a la variable construcción de piso

La tabla 3.12 contiene los valores correspondientes a la variable deposición de excretas, representados con un histograma (ver figura 3.11)

Valores	Numero de instancias	Porcentaje
Campo Abierto	26	29,55%
Pozo Séptico	62	70,45%

Tabla 3.12: Valores correspondientes a la variable deposición de excretas

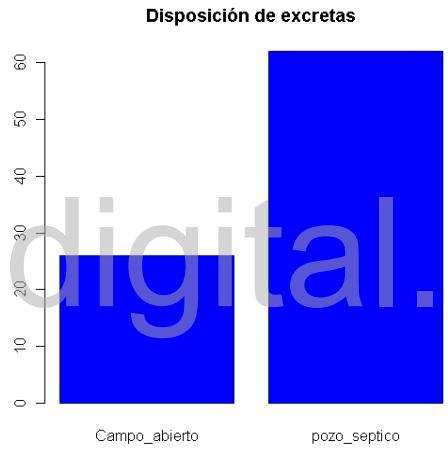


Figura 3.11: Histograma correspondiente a la variable deposición de excretas

La tabla 3.13 contiene los valores correspondientes a la variable tiempo en la zona, representados con un histograma (ver figura 3.12).

Valores	Numero de instancias	Porcentaje
Años	10	11,36%
Toda la vida	78	88,64%

Tabla 3.13: Valores correspondientes a la variable valores tiempo en la zona

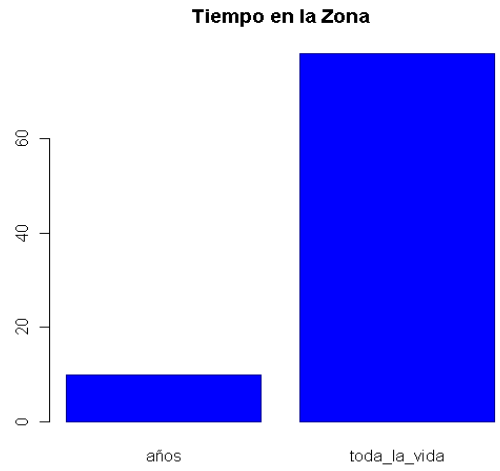


Figura 3.12: Histograma correspondiente a la variable tiempo en la zona

La tabla 3.14 muestra el porcentaje de personas que aseguran que la vivienda donde habitan ha sido fumigada por organismos sanitarios y cuantas de ellas usan insecticidas para protegerse de los insectos.

Valores	Numero de instancias/n	Porcentaje
Vivienda fumigada	68/88	77,27%
Usan Insecticidas	47/88	53,41%

Tabla 3.14: Valores correspondientes a la variable vivienda fumigada de los 88 individuos evaluados.

Presencia de Animales relacionados con la transmisión de la enfermedad

La tabla 3.15 presenta información sobre la presencia de animales domésticos relacionados con la enfermedad de Chagas, reservorios naturales del parásito *Trypanosoma cruzi*, como lo son el rabipelado y cachicamo, y palmas cerca de la vivienda.

Valores	Numero de instancias/n	Porcentaje
Perros	56/88	63,64%
Gatos	23/88	26,14%
Aves de Corral	73/88	82,95%
Burros	20/88	22,73%
Rabipelados o Cachicamos	70/88	79,55%
Consumo de rabipelados o cachicamos	25/88	28,41%
Palmas	35/88	39,77%

Tabla 3.15: Valores correspondientes al porcentaje de los tipos de animales de los 88 individuos evaluados.

Diagnóstico serológico

La tabla 3.16 contiene los valores correspondientes a la variable serología, representados con un histograma (ver figura 3.13)

Valores	Numero de instancias	Porcentaje
Positiva	35	39,77%
Negativa	53	60,23%

Tabla 3.16: Valores correspondientes a la variable serología

Al analizar las estadísticas realizadas a cada una de las variables, surgieron hipótesis sobre los factores epidemiológicos que pueden tener relación con el ciclo de transmisión de la enfermedad del Chagas en la comunidad de estudio. Las hipótesis fueron las siguientes:

- El desconocimiento de la enfermedad y su transmisor sugieren la falta de información útil para el individuo afectado, ser el responsable de su propia protección para evitar contraer la enfermedad.

Figura 3.13: Histograma correspondiente a la variable serología

- Las viviendas rurales de construcción de materiales naturales propios de la zona favorecen a la colonización del vector dentro del hogar.

- La presencia de animales domésticos relacionados con la transmisión de la enfermedad los convierten en potenciales hospederos intermediarios o atractores del vector (aves de corral) poniendo en riesgo a las personas que habitan la vivienda.

- Los rabipelados y cachicamos son el principal portador de parasito *Trypanosoma cruzi*.

- La existencia de palmas indica la presencia del vector.

3.2 Limpieza y Preparación de los datos

Antes de transformar el banco de datos originales al tipo de estructura que es capaz de leer el paquete estadístico R, para aplicar la técnica de reglas de asociación, se procedió a eliminar de la base de datos las variables que se consideraron con información incompleta o redundante.

Se procedió a la eliminación de la variable contacto con el vector, ya que se observó que un 68.19% de los individuos no contestaron esta pregunta de la encuesta (ver tabla 3.7), lo cual representa una información vaga de esta característica.

Se consideró la variable tipo de vivienda como redundante, ya que para obtener un resultado mas descriptivo de la influencia que presentan las características de la vivienda con el ciclo de transmisión de la enfermedad de Chagas, se consideró analizar las variables relacionadas con el tipo de construcción de las distintas partes de la vivienda y no de forma general como esta variable la representa.

Se procedió reagrupar la variable sector en dos regiones (ver figura 3.14). Esta agrupación se realizo bajo los siguientes criterios:

- Por su ubicación geográfica.
- Por las características epidemiológicas.

De esta manera se favorece la búsqueda de asociaciones específicas a cada región y tener una mejor comprensión de las causas que hacen favorables la transmisión de la enfermedad en estas comunidades.

La región occidental, considerando la proximidad que existe entre los sectores abarca: Cambural, El Mango, San Pedro y La Sabana, mientras que la región oriental abarca los siguientes sectores: El Maco, El Puente, Maturicito y Barrio Torció. En la tabla 3.17 se encuentran los datos de las instancias clasificadas según las regiones anteriormente descritas.

Valores	Numero de instancias	Porcentaje
región occidental	38	43.18%
región oriental	50	56.82%

Tabla 3.17: Datos de las instancias clasificadas según las regiones

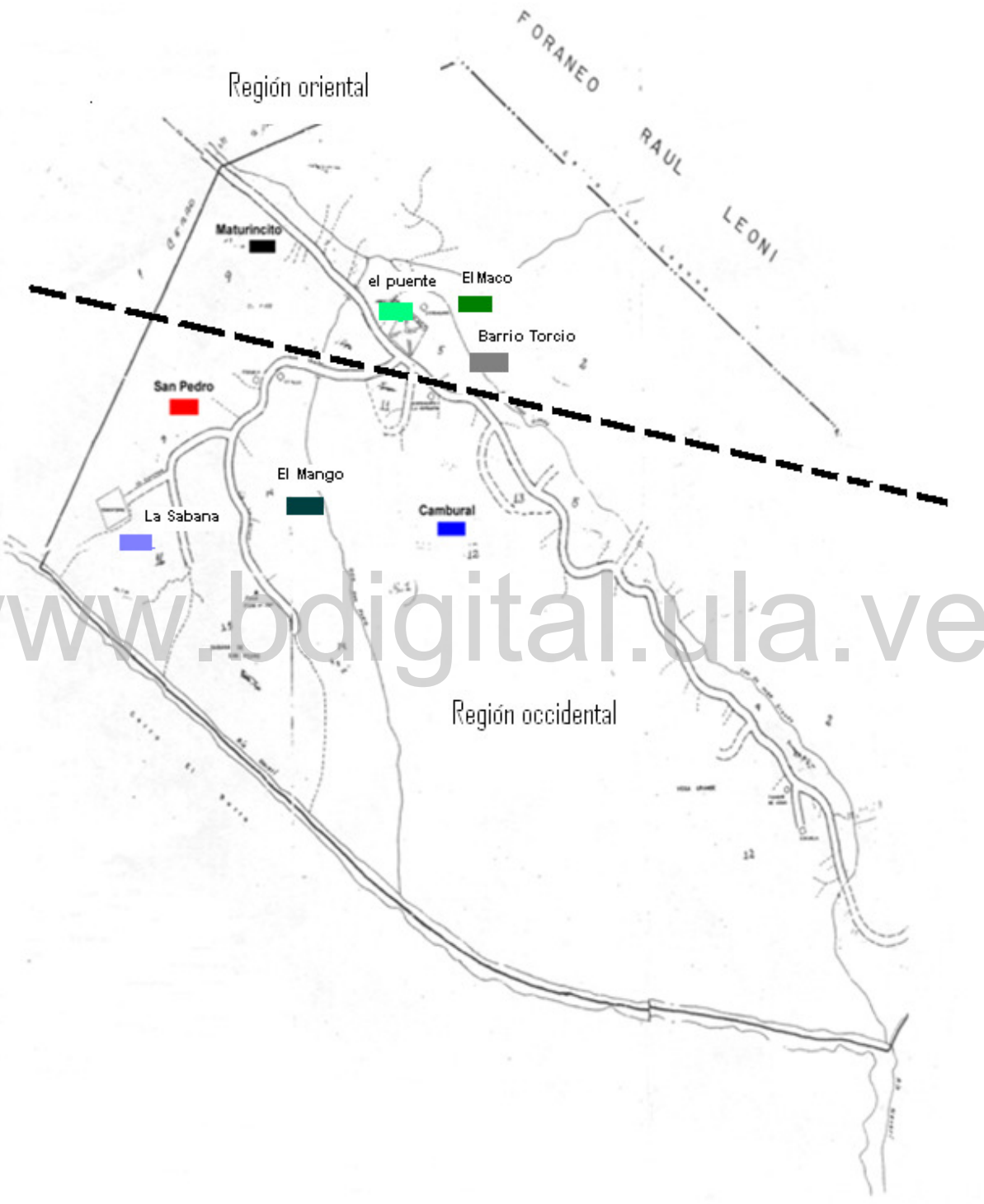


Figura 3.14: Localidades estudiadas en la parroquia Santa Fe, municipio Sucre, estado Sucre.

Se procedió también, a discretizar la variable edad, para obtener una descripción de los tipos de transmisión que se puedan presentar en esta comunidad, dividiéndola en tres grupos. El primer grupo llamado jóvenes, reflejan transmisión activa de la enfermedad en las personas seropositivas que son menores de 20 años de edad. El segundo grupo o adultos activos, comprende a los predispuestos a incapacitación para el trabajo al contraer la enfermedad, y son los individuos que van desde 21 hasta 40 años. El tercer grupo, los adultos mayores pasivos, son aquellos cuya edad es mayor a los 41 años, que denotan una transmisión antigua de la enfermedad. La tabla 3.18 muestra la frecuencia de cada uno de los grupos:

Intervalo	Frecuencia	Clase
[2,20]	41	Jóvenes
[21,40]	28	Adulto Activo
[41,100]	19	Adulto Pasivo

Tabla 3.18: Discretización de la variable edad

Para poder representar y trabajar con los datos, la biblioteca ARules provee una estructura bien diseñada, capaz de hacer frente de manera eficaz al manejo de grandes cantidades de datos binarios. Para la entrada de datos es provista la clase "*transactions*" que representa los datos con una matriz de incidencia binaria, donde las columnas representan a los ítems y las filas corresponden a las transacciones. Las entradas de la matriz son representadas con uno (1), que indica la presencia de un determinado ítem y cero (0) representa la ausencia de un ítem en particular.

Originalmente los datos están representados como un objeto de la estructura data.frame, esta estructura es la que utiliza el paquete estadístico R para representar un banco de datos típicos, rectangulares, de individuos por variables. Para llevar a cabo la conversión se utilizó la función as() de la biblioteca ARules, con esta función los datos son automáticamente codificados como una matriz de incidencia binaria.

3.3 Modelado

En esta fase de la minería de datos, se utilizó la técnica de reglas de asociación, la cual es un tipo de modelizado descriptivo, que permitirá tener una descripción total de los datos y descubrir correlaciones o concurrencias que determinen la presencia de la enfermedad del Chagas y se formaliza en la obtención de reglas del tipo SI… ENTONCES.

Para la obtención de las reglas se utilizó el algoritmo *Apriori* implantado en la biblioteca ARules, del paquete estadístico R. Es necesario proveer los parámetros adecuados para obtener una buena cantidad de reglas de asociación. Los parámetros provistos son:

- Soporte mínimo = 4%: Es el soporte mínimo a tener en cuenta para que las reglas que asocien la seropositividad con las variables epidemiológicas sean consideradas. Este soporte es muy pequeño debido a la relación entre la cantidad de reglas y la cantidad de atributos que se tiene. Dado que la matriz de transacciones de esta base de datos es una matriz dispersa, necesitamos utilizar un valor de soporte bajo para obtener reglas desde nuestro archivo. Es por esto que dentro del algoritmo *Apriori* definimos como soporte mínimo 4%.

- Tipo de métrica = Confianza: Las opciones disponibles para esta opción son los tres tipos de métrica explicadas en el capítulo 2 de este proyecto: soporte, confianza y lift. En este caso se indica que se consideran las reglas con confianza mínima de 60%.

Al aplicar el algoritmo *Apriori* se encontraron un total de 225.428 reglas. A continuación se presente un resumen del soporte y confianza de todo el conjunto de reglas encontradas (ver tabla 3.19).

	Soporte	Confianza
Valor mínimo	0.04545	0.6000
Primer cuartil	0.05682	0.7692
Mediana	0.07955	0.9000
Media	0.10172	0.9722
Tercer cuartil	0.12500	1.0000
Valor máximo	0.98860	1.0000

Tabla 3.19: Resumen del soporte y confianza de todo el conjunto de reglas obtenidas

En la primera inspección a las reglas, encontramos reglas que son predecibles y que no aportan mayor información a la investigación, tales como:

- Si la construcción de paredes es igual a bahareque, entonces la vivienda es de tipo rancho. Soporte = 46.59%, Confianza = 100%.

- Si la construcción de paredes es igual a bahareque y construcción de piso es igual a tierra, entonces la vivienda es de tipo rancho. Soporte = 13.64%, Confianza = 100%.

- Si Deposición de excretas es igual a campo abierto, entonces la vivienda es de tipo rancho. Soporte = 29.55%, Confianza = 100%.

- Si Deposición de excretas es igual a campo abierto, entonces la vivienda no contiene baño. Soporte = 10.23%, Confianza = 100%.

- Si la ocupación es igual a oficios del hogar, entonces el sexo es igual a femenino. Soporte = 30.68%, Confianza = 100%.

Debido a la gran cantidad de reglas obtenidas, se procedió a extraer subconjuntos de reglas que contengan ciertas variables epidemiológicas consideradas relevantes en la transmisión de la enfermedad de Chagas, para su posterior análisis.

3.3.1 Reglas seropositivas

El primer subconjunto de reglas que fue extraído, contiene todas aquellas reglas donde su sucesor o consecuente (lado derecho de la regla) esté presente la variable serología con valor positivo. Este subconjunto de reglas dió información sobre las variables que pueden llevar a un individuo a contraer la enfermedad.

El subconjunto contiene un total de 3.526 reglas, donde el soporte, la confianza y el lift varían en los intervalos [4.55%, 29.54%], [60%, 100%] y [1.509, 2.514], respectivamente. Estas reglas fueron agrupadas según las variables presentes en el lado izquierdo de la regla y ordenadas de forma descendente por la medida lift, para hacer un análisis detallado de los casos que están asociados a la enfermedad.

Reservorios

Este subconjunto de reglas representa a todas aquellas reglas donde aparece la variable Rabipelados-Cahicamos, que indica la presencia de estos animales en el sector, siendo estos los reservorios naturales del parásito *Trypanosoma cruzi*.

- SI Ocupación = Oficios del Hogar, Perros = si, Rabipelados-Cahicamos = si y Palmas = si, Entonces serología = positiva. Soporte = 13.64%, Confianza = 92.31%, Lift = 2.320879.

- SI Ocupación = Oficios del Hogar, Vivienda fumigada = no, Perros = si y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 11.36%, Confianza = 90.91%, Lift = 2.285714.

- SI Ocupación = Oficios del Hogar, Construcción de paredes = bahareque, Rabipelados-Cahicamos = si y Palmas = si, Entonces serología = positiva. Soporte = 11.36%, Confianza = 90.91%, Lift = 2.285714.

- SI Ocupación = Oficios del Hogar, Vivienda fumigada = no, Aves de corral = si y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 10.22%, Confianza = 90%, Lift = 2.262857.

- SI Vivienda fumigada = no, Perros = si y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 14.77%, Confianza = 86.67%, Lift = 2.179048.

- SI Conoce la enfermedad = no, Vivienda fumigada = no, Perros = si y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 13.63%, Confianza = 85.71%, Lift = 2.155102.

- SI Tiempo en la zona = toda la vida, Vivienda fumigada = no, Perros = si y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 13.63%, Confianza = 85.71%, Lift = 2.155102.

- SI Vivienda fumigada = no, Aves de corral = si y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 12.50%, Confianza = 84.62%, Lift = 2.127473.

- SI Vivienda fumigada = no, Perros = si, Aves de corral = si y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 12.50%, Confianza = 84.62%, Lift = 2.127473.

- SI Conoce la enfermedad = no, Vivienda fumigada = no, Aves de corral = si y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 12.50%, Confianza = 84.62%, Lift = 2.127473.

Conocimiento de la enfermedad

De este grupo de reglas se observó, las personas que no tienen conocimiento de la enfermedad o de que la transmite.

- SI Ocupación = Oficios del Hogar, Conoce la enfermedad = no, Perros = si y Palmas = si, Entonces serología = positiva. Soporte = 10.22%, Confianza = 90%, Lift = 2.262857.

- SI Conoce la enfermedad = no, Vivienda fumigada = no, Perros = si y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 13.63%, Confianza = 85.71%, Lift = 2.155102.

- SI Conoce la enfermedad = no, Tiempo en la zona = toda la vida, Vivienda fumigada = no y Perros = si, Entonces serología = positiva. Soporte = 12.5%, Confianza = 84.62%, Lift = 2.127473.

- SI Conoce la enfermedad = no, Vivienda fumigada = no, Aves de corral = si y Rabipelados-Cahicamos = si, Entonces serología=positiva. Soporte = 12.5%, Confianza = 84.62%, Lift = 2.127473.

- SI Ocupación = Oficios del Hogar, Conoce la enfermedad = no, Vivienda fumigada = no y Rabipelados-Cahicamos = si,, Entonces serología = positiva. Soporte = 11.36%, Confianza = 83.33%, Lift = 2.095238.

- SI Conoce la enfermedad = no, Tiempo en la zona = toda la vida, Vivienda fumigada = no y Aves de corral = si, Entonces serología = positiva. Soporte = 11.36%, Confianza = 83.33%, Lift = 2.095238.

- SI Conoce la enfermedad = no, Vivienda fumigada = no y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 14.77%, Confianza = 81.25%, Lift = 2.042857.

Palmas

Las siguientes reglas permiten identificar las variables involucradas que conllevan a la seropositividad, cuando exista la presencia de palmas en el sector.

- SI Ocupación = Oficios del Hogar, Perros = si y Palmas = si, Entonces serología = positiva. Soporte = 13.63%, Confianza = 92.31%, Lift = 2.320879.

- SI Ocupación = Oficios del Hogar, Perros = si, Rabipelados-Cahicamos = si y Palmas = si, Entonces serología = positiva. Soporte = 13.64%, Confianza = 92.31%, Lift = 2.320879.

- SI Ocupación = Oficios del Hogar, Tiempo en la zona = toda la vida, Perros = si y Palmas = si, Entonces serología = positiva. Soporte = 13.63%, Confianza = 92.31%, Lift = 2.320879.

- SI Ocupación = Oficios del Hogar, Perros = si, Aves de corral = si y Palmas = si, Entonces serología = positiva. Soporte = 12.5%, Confianza = 91.67%, Lift = 2.304762.

- SI Ocupación = Oficios del Hogar, Construcción de paredes = bahareque, Rabipelados-Cahicamos = si y Palmas = si, Entonces serología = positiva. Soporte = 11.36%, Confianza = 90.91%, Lift = 2.285714.

- SI Ocupación = Oficios del Hogar, Construcción de paredes = bahareque, Aves de corral = si y Palmas = si, Entonces serología = positiva. Soporte = 11.36%, Confianza = 90.91%, Lift = 2.285714.

- SI Ocupación = Oficios del Hogar, Construcción de paredes = bahareque y Palmas = si, Entonces serología = positiva. Soporte = 13.64%, Confianza = 85.71%, Lift = 2.155102.

Características de la vivienda y presencia de animales domésticos relacionados con la transmisión de la enfermedad

El siguiente subconjunto de reglas representa las variables que tienen relación con el hogar, considerando las características de la vivienda y la cría de animales domésticos de aquellas personas que resultaron positivas en la prueba sanguínea.

- SI Ocupación = Oficios del Hogar, Construcción de paredes = bahareque, Rabipelados-Cahicamos = si y Palmas = si, Entonces serología = positiva. Soporte = 11.36%, Confianza = 90.91%, Lift = 2.285714.

- SI Ocupación = Oficios del Hogar, Construcción de paredes = bahareque, Aves de corral = si y Palmas = si, Entonces serología = positiva. Soporte = 11.36%, Confianza = 90.91%, Lift = 2.285714.

- SI Ocupación = Oficios del Hogar, Construcción de paredes = bahareque y Palmas = si, Entonces serología = positiva. Soporte = 13.63%, Confianza = 85.71%, Lift = 2.155102.

- SI Ocupación = Oficios del Hogar, Construcción de paredes = bahareque, Aves de corral = si y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 18.18%, Confianza = 84.21%, Lift = 2.117293.

- SI Construcción de paredes = bahareque, Tiempo en la zona = toda la vida, Vivienda fumigada = no, Perros = si, Entonces serología = positiva. Soporte = 10.23%, Confianza = 81.82%, Lift = 2.057143.

- SI Construcción de paredes = bahareque, Vivienda fumigada = no, Perros = si y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 10.23%, Confianza = 81.82%, Lift = 2.057143.

Tiempo en la zona

A continuación se presentan las reglas, que nos describen la relación directa entre el tiempo de residencia en el área y la seropositividad, es decir, la endemicidad de la enfermedad de Chagas en la zona de estudio.

- SI Ocupación = Oficios del Hogar, Tiempo en la zona = toda la vida, Perros = si y Palmas = si, Entonces serología = positiva. Soporte = 13.63%, Confianza = 92.31%, Lift = 2.320879.

- SI Ocupación = Oficios del Hogar, Tiempo en la zona = toda la vida, Vivienda fumigada = no y Perros = si, Entonces serología = positiva. Soporte = 13.63%, Confianza = 92.31%, Lift = 2.320879.

- SI Ocupación = Oficios del Hogar, Tiempo en la zona = toda la vida, Vivienda fumigada = no y Aves de corral = si, Entonces serología = positiva. Soporte = 10.23%, Confianza = 90%, Lift = 2.262857.

- SI Tiempo en la zona = toda la vida, Vivienda fumigada = no y Perros = si, Entonces serología = positiva. Soporte = 14.77%, Confianza = 86.67%, Lift = 2.179048.

- SI Tiempo en la zona = toda la vida, Vivienda fumigada = no, Perros = si y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 13.64%, Confianza = 85.71%, Lift = 2.155102.

- SI Tiempo en la zona = toda la vida, Vivienda fumigada = no y Aves de Corral = si, Entonces serología = positiva. Soporte = 12.5%, Confianza = 84.62%, Lift = 2.127475.

- SI Conoce la enfermedad = no, Tiempo en la zona = toda la vida, Vivienda fumigada = no y Perros = si, Entonces serología = positiva. Soporte = 12.5%, Confianza = 84.62%, Lift = 2.127473.

- SI Tiempo en la zona = toda la vida, Vivienda fumigada = no, Perros = si y Aves de corral = si, Entonces serología = positiva. Soporte = 12.5%, Confianza = 84.62%, Lift = 2.127473.

- SI Conoce la enfermedad = no, Tiempo en la zona = toda la vida, Vivienda fumigada = no y Aves de corral = si, Entonces serología = positiva. Soporte = 11.36%, Confianza = 83.33%, Lift = 2.095238.

Grupo de edades

En esta sección, se seleccionaron las reglas que contenían la variable edad. A continuación se presentan las reglas más relevantes clasificadas por los intervalos de edad a la que corresponden.

Jóvenes (edad menor a 20 años): su seropositividad implica transmisión activa de la enfermedad en la zona, asociada con los factores epidemiológicos descritos en las reglas.

- Si edad = Joven, Ocupación = Estudiante, Construcción de paredes = bahareque y Tiempo en la zona = toda la vida, Entonces serología = positiva. Soporte = 11.36%, Confianza = 83.33%, Lift = 2.095238.

- Si edad = Joven, Construcción de paredes = bahareque y Tiempo en la zona = toda la vida, Entonces serología = positiva. Soporte = 11.36%, Confianza = 76.92%, Lift = 1.934066.
- Si edad = Joven, Ocupación = Estudiante, Construcción de paredes = bahareque, Tiempo en la zona = toda la vida y Perros = si, Entonces serología = positiva. Soporte = 11.36%, Confianza = 76.92%, Lift = 1.934066.

Adultos activos (edad mayor a 20 y menor de 41 años): su seropositividad implica incapacitación física por la enfermedad en la población de edad productiva para el trabajo, asociada con los factores epidemiológicos descritos en las reglas.

- Si edad = Adulto Activo, Ocupación = Oficios del Hogar, Construcción de paredes = bahareque y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 6.82%, Confianza = 85.71%, Lift = 2.155102.

- Si edad = Adulto Activo, Perro = si, Construcción de paredes = bahareque y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 9.09%, Confianza = 80%, Lift = 2.011429.

- Si edad = Adulto Activo, Aves de Corral = si, Construcción de paredes = bahareque y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 9.09%, Confianza = 72.73%, Lift = 1.828571.

Adultos pasivos (edad mayor a 41 años): su seropositividad implica transmisión de la enfermedad de vieja data, asociada con los factores epidemiológicos descritos en las reglas.

Al realizar la selección de las reglas, en este grupo de edad, no se obtuvieron resultados, ya que el soporte que se le introdujo al algoritmo *Apriori*, es mayor al soporte que contiene estas reglas.

Regiones

Se extrajeron las reglas que contienen las dos regiones (oriental y occidental) definidas anteriormente, que reagrupan los sectores de esta comunidad, para poder evaluar los factores de riesgo epidemiológico asociados a la seropositividad en cada región.

Región oriental

- Si Región = oriental, Ocupación = Oficios del Hogar, Vivienda fumigada = no y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 10.23%, Confianza = 90%, Lift = 2.262857.

- Si Región = oriental, Ocupación = Oficios del Hogar, Construcción de paredes = bahareque y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 10.23%, Confianza = 90%, Lift = 2.262857.

- Si Región = oriental, Vivienda fumigada = no y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 11.36%, Confianza = 83.33%, Lift = 2.095238.
- Si Región = oriental, Ocupación = Oficios del Hogar, Construcción de paredes = bahareque y Usan Insecticidas = no, Entonces serología = positiva. Soporte = 11.36%, Confianza = 83.33%, Lift = 2.095238.

- Si Región = oriental, Tiempo en la zona = toda la vida, Perros = si y Aves de corral = si, Entonces serología = positiva. Soporte = 11.36%, Confianza = 83.33%, Lift = 2.095238.

- Si Región = oriental, Conoce la enfermedad = no, Vivienda fumigada = no y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 11.36%, Confianza = 83.33%, Lift = 2.095238.

- Si Región = oriental, Tiempo en la zona = toda la vida y Vivienda fumigada = no, Entonces serología = positiva. Soporte = 10.23%, Confianza = 81.82%, Lift = 2.057143.

- Si Región = oriental, Vivienda fumigada = no, Perros = si y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 10.23%, Confianza = 81.82%, Lift = 2.057143.

Región occidental

- Si Región = occidental, Vivienda fumigada = no, Rabipelados-Cahicamos = si y Palmas = si, Entonces serología = positiva. Soporte = 4.55%, Confianza =100%, Lift = 2.514286.

- Si Región = occidental, Ocupación = Oficios del Hogar, Perros = si, Palmas = si, Entonces serología = positiva. Soporte = 7.95%, Confianza = 87.5%, Lift = 2.200000.

- Si Región = occidental, Ocupación = Oficios del Hogar, Construcción de paredes = bahareque, Palmas = si, Entonces serología = positiva. Soporte = 9.09%, Confianza = 80%, Lift = 2.011429.

- Si Región = occidental, Tiempo en la zona = toda la vida, Rabipelados-Cahicamos = si y Vivienda fumigada = no, Entonces serología = positiva. Soporte = 4.55%, Confianza =80%, Lift = 2.011429.

- Si Región = occidental, Rabipelados-Cahicamos = si y Vivienda fumigada = no, Entonces serología=positiva. Soporte = 4.55%, Confianza =80%, Lift = 2.011429.

Deposición de excretas

A continuación se presentan las asociaciones de la variable deposición de excretas con la seropositividad, para poder determinar una relación directa entre ellas.

- SI Edad = Joven, Conoce la enfermedad = no, Deposición de Excretas = Campo abierto, Palmas = si Entonces serología = positiva. Soporte = 5.68%, Confianza = 100%, Lift = 2.514286.

- SI Edad = Joven, Deposición de Excretas = Campo abierto, Palmas = si Entonces serología = positiva. Soporte = 5.68%, Confianza = 100%, Lift = 2.514286.

- SI Edad = Joven, Tiempo en la zona = toda la vida, Deposición de Excretas = Campo abierto, Palmas = si Entonces serología = positiva. Soporte = 5.68%, Confianza = 100%, Lift = 2.514286.

- SI Sexo = masculino, Deposición de Excretas = Campo abierto, Rabipelados-Cahicamos = si y Palmas = no, Entonces serología = positiva. Soporte = 5.68%, Confianza = 83.33%, Lift = 2.095238.

- SI Sexo = masculino, Deposición de Excretas = Campo abierto, Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 9.09%, Confianza = 80%, Lift = 2.011429.

3.3.2 Reglas generales

En esta sección se describen las reglas, donde en el predecesor (lado izquierda de la regla) está presente la variable serología con valor positivo; este subconjunto de reglas describe las variables que están más relacionadas con la enfermedad.

- Si serología =positiva, Entonces Perros = si, Soporte = 38.64%, Confianza = 97.14%.

- Si serología = positiva, Entonces Tiempo la zona = toda la vida, Soporte = 37.5%, Confianza = 94.28%.

- Si serología = positiva, Entonces Rabipelados-Cahicamos = si, Soporte = 36.36%, Confianza = 91.43%.

- Si serología = positiva, Entonces Construcción de paredes = bahareque, Soporte = 29.54%, Confianza = 74.29%.

- Si serología = positiva, Entonces Palmas = si, Soporte = 23.86%, Confianza = 60%.

3.3.3 Reglas seronegativas

Este subconjunto de reglas que fue extraído comprende aquellas reglas en donde su sucesor o consecuente (lado derecho de la regla), esté presente la variable serología con valor negativo. Este subconjunto de reglas arroja información sobre las condiciones en la que habitan los individuos que no han tenido ningún tipo de contacto con el parásito.

El subconjunto contiene un total de 7.914 reglas, donde el soporte, la confianza y el lift varían en los intervalo [4.55%, 60.23%], [60%, 100%] y [0.9962, 1.6604] respectivamente. Estas reglas fueron agrupadas según las variables presentes en el lado izquierdo de la regla y ordenadas de forma descendente por la medida lift, para hacer un análisis detallado de los casos que conllevan a la enfermedad.

Reservorios

El siguiente subconjunto de regla, muestra las reglas donde no exista la presencia de los rabipelados y cachicamos, y si existen que medidas toman los individuos para protegerse de la enfermedad.

- SI construcción de paredes = bloque y Rabipelados-Cahicamos = no, Entonces serología = negativo, Soporte = 11.36%, Confianza = 100%, Lift = 1.660377.

- Si usan insecticidas y Rabipelados-Cahicamos = no, Entonces serología = negativo, Soporte = 11.36%, Confianza = 100%, Lift = 1.660377.

- Si Deposición de excretas = pozo séptico y Rabipelados-Cahicamos = no, Entonces serología = negativo, Soporte = 11.36%, Confianza = 100%, Lift = 1.660377.

- Si construcción de paredes = bloque, Deposición de excretas = pozo séptico y Rabipelados-Cahicamos = no, Entonces serología = negativo, Soporte = 11.36%, Confianza = 100%, Lift = 1.660377.

- Si construcción de paredes = bloque, vivienda fumigada = si y Rabipelados-Cahicamos = si, Entonces serología = negativo, Soporte = 10.22%, Confianza = 100%, Lift = 1.660377.

- Si no conoce la enfermedad, usan insecticidas, Perro = si y Rabipelados-Cahicamos = si, Entonces serología = negativo, Soporte = 10.22%, Confianza = 100%, Lift = 1.660377.

Conocimiento de la enfermedad

En el siguiente subconjunto se seleccionaron las reglas que contengan la variable conocimiento de la enfermedad, para obtener un análisis de la influencia que tiene, que un individuo seronegativo conozca o no la enfermedad.

- Si conoce la enfermedad, Construcción de paredes = bloque, Burros = si y usan insecticidas = si, Entonces serología = negativo, Soporte = 10.22%, Confianza = 100%, Lift = 1.660377.

- Si conoce la enfermedad, usan insecticidas = si y palmas = no, Entonces serología = negativo, Soporte = 10.22%, Confianza = 100%, Lift = 1.660377.

- Si conoce la enfermedad, Deposición de Excretas = Campo abierto y Rabipelados-Cahicamos = no, Entonces serología = negativo, Soporte = 10.22%, Confianza = 100%, Lift = 1.660377.

- Si no conoce la enfermedad, Construcción de paredes = bloque, vivienda fumigada = si y Burros = si, Entonces serología = negativo, Soporte = 10.22%, Confianza = 100%, Lift = 1.660377.

- Si no conoce la enfermedad, vivienda fumigada = si, Burros = si y palmas = no, Entonces serología = negativo, Soporte = 10.22%, Confianza = 100%, Lift = 1.660377.

Palmas

En el siguiente subconjunto se seleccionaron las reglas donde existe la presencia de la variable palmas cerca del hogar, a continuación se presentan las reglas más relevantes.

- Si Construcción de paredes = bloque, palmas = no, Rabipelados-Cahicamos = no, entonces serología = negativo, Soporte = 11.36%, Confianza = 100%, Lift = 1.660377.

- Si Construcción de paredes = bloque, conoce la enfermedad y palmas = no, entonces serología = negativo, Soporte = 11.36%, Confianza = 100%, Lift = 1.660377.

- Si insecticidas = si, animales en la vivienda = si, Rabipelados-Cahicamos = no, y palmas = no, entonces serología = negativo, Soporte = 11.36%, Confianza = 100%, Lift = 1.660377.

Características de la vivienda

En el siguiente subconjunto de reglas se seleccionaron las variables que tienen relación con el hogar, considerando las características de la vivienda de las personas que resultaron negativas en el análisis serológico.

- Si Construcción de paredes = bloque y Rabipelados-Cahicamos = no, Entonces serología = negativo, Soporte = 11.36%, Confianza = 100%, Lift = 1.660377.

- Si Construcción de paredes = bloque, Rabipelados-Cahicamos = no y Palmas = no, Entonces serología = negativo, Soporte = 11.36%, Confianza = 100%, Lift = 1.660377.

- Si Construcción de paredes = bloque, Rabipelados-Cahicamos = no y Baños = si, Entonces seropositividad = negativo, Soporte = 11.36%, Confianza = 100%, Lift = 1.660377.

- Si Construcción de paredes = bloque, Rabipelados-Cahicamos = no y Vivienda fumigada = si, Entonces seropositividad = negativo, Soporte = 11.36%, Confianza = 100%, Lift = 1.660377.

Animales domésticos relacionados con la transmisión de la enfermedad

En el siguiente subconjunto de reglas, se seleccionaron aquellas donde existan la presencia de animales domestico dentro del hogar, de las personas que resultaron negativas en el análisis serológico.

- Si Conoce la enfermedad = si, Vivienda fumigada = si, Perros = si y Construcción de paredes = bloque, Entonces serología = negativo, Soporte = 10.23%, Confianza = 100%, Lift = 1.6603774.

- Si insecticidas = si, Perros= si, Palmas = no, y palmas = no, entonces serología = negativo, Soporte = 10.23%, Confianza = 100%, Lift = 1.6603774.

- Si Vivienda fumigada = si, Aves de corral = si, Burros =si y Construcción de paredes = bloque, Entonces serología = negativo, Soporte = 10.23%, Confianza = 100%, Lift = 1.6603774.

3.4 Discusión de los resultados

Esta discusión de los resultados se centra en la evaluación de posibles relaciones entre los niveles de seropositividad *anti-T. cruzi* detectados y las variables epidemiológicas (factores de riesgo) asociadas con la enfermedad de Chagas. En este sentido los resultados están basados en las reglas obtenidas y corroborados por estudios realizados a otras comunidades.

Los animales silvestres como los rabipelados o cachicamos son los reservorios naturales del parásito *Trypanosoma cruzi,* agente causal de la enfermedad de Chagas (Perruolo y Morales, 1987; Maekelt, 2000).

- SI Ocupación = Oficios del Hogar, Perros = si, Rabipelados-Cahicamos = si y Palmas = si, Entonces serología = positiva. Soporte = 13.64%, Confianza = 92.31%, Lift = 2.320879.

Esta regla relaciona la seropositividad presente en los individuos pertenecientes a la comunidad en estudio con la presencia de estos animales a los alrededores de las viviendas, ligado a la existencia de palmas y animales domésticos como el perro, siendo más afectadas las personas que permanecen con más frecuencia dentro de las viviendas, sometiéndose a un mayor contacto y

riesgo de ser picados por el vector de hábitos intradomiciliarios. Esta regla tiene un nivel de confianza del 92.31% lo que indica que en la mayor parte de las transacciones, esta regla es correcta.

Referente a las palmas presentes a la cercanía de la vivienda, es un factor de riesgo de gran influencia en la epidemiología de la enfermedad de Chagas, ya que éstas son los hábitats naturales por excelencia del vector. Figuera (2002) y Abreu (2003) en el estado Sucre y Jiménez (1995) en el estado Anzoátegui, consiguieron resultados similares a los reportados en el presente estudio.

Por otro lado, la cría de animales domésticos como el perro y aves de corral son importantes en el ciclo de transmisión de la enfermedad de Chagas, debido a que estos influyen en un aumento de la infestación de los vectores y las aves domésticas permiten mantener mayores densidades poblacionales del vector (Sanmartino y Crocco, 2000).

Otro factor de riesgo de gran importancia en el mantenimiento de la transmisión en forma endémica es el tipo de vivienda del medio rural, representada por construcciones con techos de paja o palma, paredes de barro (bahareque o tablas mal ajustadas), lo cual condiciona que el vector colonice las grietas de las paredes o los techos. Se puede decir que la enfermedad de Chagas es un reflejo de las precarias condiciones de la habitación rural, aunque, algunos investigadores han referido la presencia de insectos infestados por *Trypanosoma cruzi* en áreas urbanas y suburbanas (Torres, 1992; Caso de escuela Bolivariana de Chacao, diciembre 2007). Sin embargo, estudios realizados en décadas pasadas por Pifano (1973), demostraron los cambios de hábitos que experimenta el insecto vector, adaptado a las habitaciones humanas, especialmente al rancho.

- SI Construcción de paredes = bahareque, Tiempo en la zona = toda la vida, Vivienda fumigada = no, Perros = si, Entonces serología = positiva. Soporte = 10.23%, Confianza = 81.82%, Lift = 2.057143.

Esta regla describe la asociación de la seropositividad con viviendas construidas con bahareque, la cría de animales relacionados con la transmisión y sin ninguna medida de protección contra el insecto vector. Además, es importante resaltar que estos individuos seropositivos tienen toda la vida habitando en la zona, sugiriendo esto una relación directa entre el tiempo de permanecía y la seropositividad. La baja migración podría indicar la presencia de infecciones autóctonas (endémicas). Estos resultados coinciden con los reportados por Figuera (2002) y Abreu (2003) en el estado Sucre.

Es necesario destacar la importancia que tiene que los individuos de la comunidad en estudio tengan conocimiento que les ayude a comprender el problema y habilidades prácticas que les permitan reconocer y protegerse contra esta enfermedad, debido a que los resultados obtenidos demuestran una asociación significativa entre la seropositividad y la falta de conocimiento que presentan los individuos cuando habitan en zonas con las condiciones favorables para la transmisión de la misma. Sanmartino y Crocco (2000) le dan mucha importancia al papel de la educación, como una herramienta fundamental en la lucha contra la enfermedad de Chagas, por ser el medio más indicado para promover en las personas cambios permanentes.

Las siguientes reglas muestran la relación que tiene el desconocimiento de la enfermedad con la seropositividad.

- SI Conoce la enfermedad = no, Vivienda fumigada = no, Perros = si y Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 13.63%, Confianza = 85.71%, Lift = 2.155102.

- SI Conoce la enfermedad = no, Vivienda fumigada = no, Aves de corral = si y Rabipelados-Cahicamos = si, Entonces serología=positiva. Soporte = 12.5%, Confianza = 84.62%, Lift = 2.127473.

Basados en las reglas obtenidas referentes a las dos regiones observadas, en la sección 3.3.1, se puede inferir que las localidades evaluadas presentaron factores de riesgo epidemiológicos asociados similares, posiblemente, por la cercanía en que se encuentran. En cuanto a los grupos

etarios, se encontró que existe una transmisión activa de la enfermedad de Chagas en la zona en estudio, asociada con viviendas de paredes de bahareque y permanencia toda la vida de los jóvenes seropositivos, es decir, endemicidad de la enfermedad. Asimismo, se detecta la presencia de una población activa seropositiva en edad plena apta para el trabajo, potencialmente discapacitados físicamente, habitando viviendas tipo rancho con perros y reservorios naturales presente en la zona.

De la regla, SI Sexo = masculino, Deposición de Excretas = Campo abierto, Rabipelados-Cahicamos = si, Entonces serología = positiva. Soporte = 9.09%, Confianza = 80%, Lift = 2.011429, se desprende que la deposición a campo abierto, es decir, frecuente contacto con los reservorios naturales y animales domésticos relacionados con la transmisión presentes en la zona, aunado a la presencia toda la vida en la misma, constituyen factores de riesgo para contraer la enfermedad en la población de San Pedro. Debemos destacar que aunque en las reglas no se detecta el vector, necesario para la transmisión, debemos considerar que su hábitat natural es precisamente el silvestre donde la mayoría de los individuos hacen sus necesidades fisiológicas.

Al estudiar el grupo de reglas que asocian las variables con la seronegatividad se encontró, la ausencia de los reservorios naturales del parásito, aunado a viviendas de construcción de paredes de bloques, condiciones que no favorecen la transmisión de la enfermedad. También se observó que las que habitan en zonas donde exista la presencia de los reservorios naturales del parásito, utilizan medidas de protección contra la enfermedad como el uso de insecticidas y fumigaciones realizadas por las autoridades sanitarias. Las siguientes reglas reflejan lo expuesto:

- SI construcción de paredes = bloque y Rabipelados-Cahicamos = no, Entonces serología = negativo, Soporte = 11.36%, Confianza = 100%, Lift = 1.660377.

- Si construcción de paredes = bloque, vivienda fumigada = si y Rabipelados-Cahicamos = si, Entonces serología = negativo, Soporte = 10.22%, Confianza = 100%, Lift = 1.66037.

En síntesis, en las localidades estudiadas están presentes todos los factores de riesgos clásicos asociados con la transmisión del mal de Chagas, resaltando que la transmisión activa se manifiesta en la alta seropositividad observada en los grupos etarios menores de 20 años. Ello contrasta considerablemente con la afirmación realizada por los investigadores Aché y Matos (2001) quienes sostuvieron que fue interrumpida la transmisión de la enfermedad de Chagas en Venezuela.

Capítulo 4

Conclusiones y Recomendaciones

En este proyecto de grado se mostró la aplicación de reglas de asociación para evaluar los distintos factores de riesgo epidemiológicos que favorecen la transmisión de la enfermedad de Chagas en la comunidad de San Pedro, Parroquia Santa Fe del municipio Sucre, estado Sucre.

Dada la naturaleza del problema planteado, mediante el análisis de reglas de asociación se llegan a obtener resultados que de otra manera hubiesen sido difícil de conocer, ya que el análisis manual de los datos no es tarea fácil y obtener reglas por medios empíricos, como la experiencia, es poco fiable.

En cuanto a los resultados obtenidos podemos resaltar los siguientes:

- La alta seropositividad encontrada en los individuos menores de 20 años, en la comunidad estudiada, indica una transmisión activa en esta comunidad.

- Los individuos seropositivos con exclusiva permanecía en la zona evaluada demuestran la endemicidad de la enfermedad.

- Las personas que crían animales domésticos relacionados con la transmisión de la enfermedad y habitan en áreas donde existan los reservorios naturales del parásito *Trypanosoma cruzi*, y palmas, están en mayor riesgo de ser picados por el vector, siendo más afectadas las personas que permanecen con más frecuencia dentro de la vivienda.

- Un factor de riesgo de gran importancia en el mantenimiento de la infección en forma endémica en esta comunidad es el tipo de vivienda rural, representada por construcción de paredes de bahareque y la cría de animales domésticos como perros y aves de corral. La presencia de los perros influyen en el aumento de infectividad de los vectores y las aves domésticas permite mantener mayor densidad de población del insecto dentro o cerca del hogar.

- Los individuos seropositivos no tienen conocimiento alguno de la enfermedad y su forma de transmisión. El estudio realizado indica la importancia que tiene que los individuos de la zona conozcan perfectamente su situación para poder protegerse a sí mismo y a sus familiares, y poder reducir el grado de infección en la zona.

Hasta el momento no existe una vacuna que haga posible prevenir la enfermedad. Eliminar completamente la enfermedad es muy difícil, ya que ésta se encuentra en la naturaleza entre vectores y reservorios naturales, sin embargo, es posible reducir el grado de infección humana, adoptando ciertas medidas para evitar ser infectados, entre ellas tenemos:

- Conocer la enfermedad y aprender a identificar el vector y diferenciarlo de otros insectos.

- Fumigar las viviendas periódicamente, así como las aéreas verdes y árboles propensos a albergar insectos.

- No dormir con animales domésticos dentro de la habitación.

- Uso de mosquiteros y repelentes para evitar la picada del insecto

- Eliminar las grietas de las paredes para que no sean colonizadas por el vector.

- Mejoramiento de la calidad de las viviendas para reducir el número de nichos disponibles para el establecimiento de los vectores.

Se recomienda que este estudio sea aplicado en otras comunidades del estado Sucre y otras comunidades del país, donde hayan evidencias de la presencia de la enfermedad de Chagas, para identificar los factores de riesgo epidemiológicos que conllevan a la enfermedad, ya que ésto puede variar debido que el insecto transmisor se adapta a distintos ambientes, dependiendo si las condiciones se hacen favorables para su supervivencia.

Bibliografía

Abreu, L. 2003. "Evaluación Seroepidemiológica de la Enfermedad de Chagas en la Población de los Altos de Sucre del Municipio Sucre, Estado Sucre". Trabajo de Pregrado. Departamento de Bioanálisis, Universidad de Oriente, Cumaná, Venezuela.

Aché, A. 1993. "Programa de control de la enfermedad de Chagas en Venezuela". Bol. Dir. Malariol. y San. Amb.

Aché, A. & Matos, A. 2001. "*Interrupting Chagas disease transmisión in Venezuela*". Rev. Inst. Med. Trop. S. Paulo.

Agrawal, R.; Imielinski, T.; and Swami, A. (1993) "*Mining association rules between sets of items in large Databases*". In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pages 207{216. ACM Press, 1993. URL http://doi.acm.org/10.1145/170035.170072.

Agrawal, R.; Srikant, R. (1994). "*Fast algorithms for mining association rules*". In Proceedings of the 20th International Conference on Very Large Databases, pages 487--499, Santiago, Chile.

Aguilera, J. J.; Del Jesus, M. J.; Gonzales, P.; Herrera, F.; Navio M.; Sainz J. (2002). "Extracción Evolutiva de Reglas de Asociación en un Servicio de Urgencias Psiquiátricas".

Atias, A. (1991). "Parasitología Clínica". 3ra. edición. Publicaciones técnicas Mediterráneo. Santiago, Chile.

Aza, T. 2003. "Evaluación seroepidemiológica del mal de Chagas en la población de San Pedro, parroquia de Santa Fe del municipio Sucre, estado Sucre", Trabajo de Grado. Licenciatura en Bioanálisis, Escuela de ciencias, Núcleo de Sucre, UDO.

Berry, J.A.; Linoff, G. (1997). "Data *Mining Techniques For Marketing, Sales and Customer Support*". Editorial Wiley.

Brant, R.; Troncon, L.; Oliveira, R. y Meneghelli, U. (1998). "*Gastrointestinal manifestations of Chagas disease*". *Am. Coll. Gastroenterol*.

Brener, Z.; Souza, W.; Andrade, Z. & Barral-Netto, M. (2000). "*Trypanosoma cruzi* e Doença de Chagas". Segunda edición. Editora Guanabara Koogan, S.A. Brasil.

Brin, S.; Motwani, R.; Ullman, J. D. and Tsur, S. "Dynamic itemset counting and implication rules for market basket data". In SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, pages 255{264, Tucson, Arizona, USA, May 1997.

Botero, D. y Restrepo, M. (1998). "Parasitología Humana". Tercera edición. Corporación para Investigaciones Biológicas. Medellín, Colombia.

Chapman, P. C.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C. y Wirth, R. (2002). "*CRISP-DM 1.0. Step-by-step data mining guide*". Obtenido de http://www.crisp-dm.org.

Contreras, V. 1994. "Elementos de apoyo para trabajar en la enfermedad de Chagas". Clemente Editores C.A, Valencia. Venezuela.

Daedalus (2002). "Minería de Datos: Conceptos y Objetivos". Dirección Web: http://www.daedalus.es.

Del Jesus, M.J.; Gonzales, P.; Herrera, F. y Mesonero M. (2004) "Algoritmo Evolutivo de Extracción de Reglas de Asociación aplicado a un Problema de Marketing".

Fragata, A.; Ostermayer, A.; Prata, A; Dias, E.; de Oliveira, H.; Romeu, J.; Rodríguez, J.; Gumes, S.; Macedo, V.; Amato, V.; de Oliveira, W. & Brener, Z. (1997). *"Etiological treatment for Chagas Disease". The National Health Foundation of Brazil.*

Kimball, R.; Reeves, L.; Ross, M.; Thornthwaite, W. (1998). "*The Data Warehouse Lifecycle Toolkit*". Wiley Computer Publishing.USA.

Maekelt, A. 1994. "Publicaciones de Medicina Tropical". Facultad de Medicina. Universidad Central de Venezuela.

Maekelt, A. (2000). "Programa de enseñanza. La enfermedad de Chagas". Tomo II. Medicina Tropical, Facultad de Medicina UCV.

Martínez, F. (2003) "Optimización Mediante Técnicas de Minería de Datos del Ciclo de Recocido de una Línea de Galvanizado". Servicio de Publicaciones de la Universidad de La Rioja.

Martínez ,F., Ordieres J. (2004). "Apuntes de la asignatura Mineria de Datos". Servicio de Publicaciones de la Universidad de La Rioja.

Martínez, F.; Pernia, A.; Fernández, R.; Escribano, R.; Guillén, P.; Conti, D. (2009). "Proyecto CONOCER". Obtenido de http://api.unirioja.es/conoser.

OPS (2008). "Enfermedad de Chagas". DDNNI (Iniciativa para el desarrollo de drogas para los países olvidados)

Perruolo, G. y Morales, O. 1987. "Reservorios de la enfermedad de Chagas en el estado Táchira", Venezuela (zona norte). Bol. Dir. Malariol. y San Amb.

Pifano, Felix. 1973. La dinámica epidemiológica de la Enfermedad de Chagas en el Valle de los Naranjos, Estado Carabobo, Venezuela. Arch. Ven. Med. Trop. y Parasitol. Med.

Piatetsky-Shapiro, G. y Frawley, W. J. (1991). "Discovery, analysis, and presentation of strong rules". editors, Knowledge Discovery in Databases. AAAI/MIT Press, Cambridge, MA.

Pyle, Dorian. "*Data Preparation For Data Mining*". Morgan Kaufmann Publishers. San Francisco, California (1999).

Sanmartino, M. y Crocco L. 2000. "Conocimientos sobre la enfermedad de Chagas y factores de riesgo en comunidades epidemiológicamente diferentes de Argentina". Rev. Panam.

Thuraisingham, B. (1999). "*Data Mining. Technologies, Techniques, Tools and Trends*". Ed. CRC Press LLC.

Torres, R. 1992. "La infección chagasica en el barrio cerros de Marín de la ciudad de Maracaibo, Venezuela: Estudio Serológico". Kasmera

Villalobos, L.; De Sequeda, M. y De Aponte, M. (1994). "Enfermedad de Chagas: Transmisión Vectorial y su Control en Venezuela".

Westphal, C.; Blaxton, T. (1998). "*Data Mining Solutions. Methods and Tools for Solving Real-World Problems*". John Wiley & Sons. USA.

Witten, I. y Frank, E. "Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations. Second Edition". Morgan Kaufmann Publishers. San Francisco, California (2005).

Ihaka R. y Gentleman R. 1996. "*R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics* ".