UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

PROGRAMA DE MAESTRÍA

MAESTRÍA EN INFORMÁTICA

TESIS

PATRONES PARA LA ESTIMACIÓN DE CONSUMO DE MEDICAMENTOS CON MINERÍA DE DATOS REDES PUNO

PRESENTADA POR:

ALCIDES DEMETRIO MELO CHURA

PARA OPTAR EL GRADO ACADÉMICO DE:

MAGISTER SCIENTIAE EN INFORMÁTICA

MENCIÓN EN GERENCIA DE TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIONES

PUNO, PERÚ

2018

UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

PROGRAMA DE MAESTRÍA

MAESTRÍA EN INFORMÁTICA

TESIS

PATRONES PARA LA ESTIMACIÓN DE CONSUMO DE MEDICAMENTOS CON MINERÍA DE DATOS REDES PUNO

PRESENTADA POR:

ALCIDES DEMETRIO MELO CHURA

PARA OPTAR EL GRADO ACADÉMICO DE:

MAGISTER SCIENTIAE EN INFORMÁTICA

MENCIÓN EN GERENCIA DE TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIONES

APROBADA POR EL SIGUIENTE JURADO:

PRESIDENTE:

.....

Dra. MARIA MAURA SALAS PILCO

PRIMER MIEMBRO:

.....

Mg. **EMMA ORFELINDA AZAÑERO DE AGUIRRE**

SEGUNDO MIEMBRO:

.....

MC. CESAR AUGUSTO LLUEN VALLEJOS

ASESOR DE TESIS:

.....

M. Sc. REMO CHOQUEJAGUA ACERO

Puno, 13 de abril del 2018

ÁREA: Inteligencia Artificial
TEMA: Minería de Datos

DEDICATORIA

A mis padres por ser mi pilar fundamental en todo lo que he conseguido hasta estas instancias de mi vida, por su apoyo incondicional, por demostrarme siempre que el que persevera alcanza.

AGRADECIMIENTOS

A la Universidad Nacional del Altiplano, ser mi Alma Mater.

A los docentes de la Maestría en Informática de la UNA Puno, por su dedicación y sus enseñanzas para alcanzar este objetivo

ÍNDICE GENERAL

Pág.

CAPÍTULO II

MARCO TEÓRICO

CAPÍTULO III

METODOLOGÍA

ÍNDICE DE CUADROS

Pág.

ÍNDICE DE FIGURAS

Pág.

ÍNDICE DE ANEXOS

Pág.

RESUMEN

La investigación plantea como objetivo determinar los patrones de predicción mediante el uso de minería de datos del consumo de medicamentos de Redes Puno, mediante los árboles de decisión, para asi tener la información disponible, mediante el proceso KDD, finalmente aplicar WEKA como algoritmos de minería de datos, para obtener los arboles de decisión y clasificación, para extraer conocimiento. La metodología se basa en las etapas del Knowledge Discovery in Databases (KDD), implementadas de acuerdo a la metodología CRISP-DM, la cual es usada para el desarrollo de proyectos de minería de datos. Inicialmente se hizo un diagnóstico de la situación actual de los Establecimientos de Salud. Luego, se seleccionaron y procesaron las fuentes de datos para el estudio de forma automática, almacenando las variables en un repositorio multidimensional (DataMart), reduciendo el tiempo de cálculo de indicadores y recursos humanos. Se aplicaron técnicas de clustering para obtener grupos de elementos con datos de clientes cuyas características fueran similares, según información histórica de consumo. Finalmente se generó un modelo de clasificación que asignara, de acuerdo a una medida de similitud, elementos que no habían sido evidenciados, y de esta manera estimar la necesidad de los medicamentos. Con esto se logró diseñar nuevas métricas para el proceso e identificar a los medicamentos más críticos, lo que permite llegar a valores más exactos de los ingresos perdidos en cada segmento. Con este estudio, se puede tomar decisiones informadas y mejorar su capacidad de control de provisión de medicamentos.

Palabras Claves: Arboles de Decisión, CRISP-DM, KDD, Minería de datos, y WEKA.

ABSTRACT

The research aims to determine the prediction patterns through the use of data mining of drug consumption of Puno Networks, through the decision trees, in order to have the information available, through the KDD process, finally apply WEKA as mining algorithms of data, to obtain the trees of decision and classification, to extract knowledge. The methodology is based on the Knowledge Discovery in Databases (KDD) stages, implemented according to the CRISP-DM methodology, which is used for the development of data mining projects. Initially a diagnosis was made of the current situation of the Health Establishments. Then, the data sources for the study were selected and processed automatically, storing the variables in a multidimensional repository (DataMart), reducing the calculation time of indicators and human resources. Clustering techniques were applied to obtain groups of elements with data from clients whose characteristics were similar, according to historical consumption information. Finally, a classification model was generated that assigned, according to a measure of similarity, elements that had not been evidenced, and in this way estimated the need for medications. With this, it was possible to design new metrics for the process and identify the most critical medications, which allows reaching more accurate values of the lost revenues in each segment. With this study, you can make informed decisions and improve your ability to control the supply of medications.

Keywords: CRISP-DM , Data Mining, Decision Trees, , KDD and WEKA.

INTRODUCCIÓN

El estudio de consumo de medicamentos se ha convertido en un problema social, para el tratamiento de diversas enfermedades que afecta a muchas sectores de la población en todo el mundo, reducir el riesgo de las enfermedades es un tema que tienen muy presente todos los profesionales de salud, la misma que para implementar un plan estratégico para reducir el morbilidad, por diferentes circunstancias.

Con este análisis de la información, que se cuenta de los sistemas informáticos de la Red Puno, se podrá crear un modelo de análisis de datos que permita obtener patrones de comportamiento consumo de medicamentos. La creación del presente modelo se lo realiza a través del análisis de la información: y de la interacción en el entorno de almacenamiento en la Base de datos

Para contribuir con la solución al problema del desabastecimiento o sobrestock de medicamentos se plantea la aplicación de técnicas de minería de datos, con el objeto de “Comprender cuáles son las posibles causas

De acuerdo a Hand, Mannila & Smyth (2011) “la Minería de datos es un proceso que reúne un conjunto de herramientas de diversas ciencias (Estadística, Informática, Matemáticas, Ingeniería, entre otras)” que persigue extraer conocimiento oculto o información no trivial de grandes volúmenes de datos, con la finalidad de dar soluciones a problemas específicos en las empresas.

CRISM-DM fue la metodología utilizada para la creación del modelo, la misma que es una de las más usadas en la actualidad para la generación de proyectos

de Minería de datos, con ella se pretende obtener un modelo de análisis de datos, que con la ayuda de la implementación de algoritmos de Inteligencia Artificial, ya incorporados en la herramienta de pre-procesamiento de datos Weka,

La investigación esta estructurado de la siguiente manera:

En el capítulo 1 se describe el estado del arte del presente proyecto definiendo en el mismo el análisis del aprendizaje y algunos conceptos que engloban la minería de datos como son las técnicas y tareas de la minería de datos para extraer conocimiento, además de analizar las diferentes herramientas y librerías que se utilizan para poder analizar el conocimiento extraído. Del mismo modo se describen en el estado de arte las diferentes áreas en la actualidad en donde se están empleando técnicas de minería de datos; se empleó además un estudio de proyectos similares acerca de la aplicación de técnicas de minería de datos para predecir la deserción de estudiantes de carreras universitarias.

En el capítulo 2 se describe el análisis del problema que se requiere modelar, así mismo se establece el diseño de la solución para el mismo.

En el capítulo 3 se detalla, la implementación de la Metodología CRISP-DM, para la elaboración del modelo de minería de datos.

En el capítulo 4 se presentan las conclusiones y recomendaciones del proyecto, respecto a los resultados encontrados en la minería de datos.

En la parte final se presenta las conclusiones, sugerencias, bibliografía y anexos correspondientes al trabajo de investigación.

CAPÍTULO I

PROBLEMÁTICA DE INVESTIGACIÓN

1.1. PLANTEAMIENTO DE LA INVESTIGACIÓN

La gestión de compra de medicamentos es la acción de buscar mejoras permanentes, al realizar compras utilizando los recursos de que se dispone en forma eficaz y efectiva con el propósito de conseguir aquellos bienes y servicios que necesita la institución para su funcionamiento es una de las actividades más complejas dentro de la Red de Salud de Puno. Esta labor se hace de vital importancia a la hora de hacer los requerimientos de compra de medicamentos, y se hace necesaria de una herramienta que permita en la ayuda de la toma de decisiones de manera oportuna. Es por esto que tanto investigadores como personal de salud, han trabajado por años en la solución a dicha complejidad, apoyándose en técnicas matemáticas y estadísticas, desarrollando modelos para la correcta gestión del abastecimiento, de los mismos.

Dicho proceso es administrado mediante la Oficina de Informatica, servicio realizado por las diferentes ventanillas de Farmacia, Admision y triaje. La generación de requerimientos de medicamentos demanda tiempo y esfuerzo por parte de los diferentes trabajadores de Salud, debido a que es un trabajo realizado en forma manual.

Sin embargo se evidencia que la mayor parte de las instituciones de salud, toman sus decisiones en base a criterios o experiencia y no están fundamentando esas decisiones en datos reales, actualmente el mundo institucional pública y privada es tan competitivo que no se puede dar el lujo de equivocarse en una decisión por más pequeña que ésta sea.

Es por ello que se propone el desarrollo de patrones de minería de datos, que ayude a los encargados de ésta área a hacer un mejor uso de los datos almacenados que generalmente no se aprovecha, pero que son de vital importancia como apoyo al proceso de toma de decisiones.

1.2 DEFINICIÓN DEL PROBLEMA

1.2.1. Problema Principal

La generación de requerimientos de medicamentos demanda tiempo y esfuerzo por parte de los diferentes trabajadores de Salud, debido a que es un trabajo realizado en forma manual por lo que es indispensable: “Determinar los patrones de predicción mediante el uso de minería de datos del consumo de medicamentos de Redes Puno, y mejorar la oportunidad de la información a través del uso de los arboles de decisión.”

1.3 JUSTIFICACIÓN

El Optimo abastecimiento de medicamentos para una optimo inversión quienes se benefician tanto la población, al encontrar medicamentos que se necesitan para su atención de manera oportuno, para el tratamiento de las diversas enfermedades asi como a los estudiantes los profesionales, entre otros

El resultado de la investigación aporta en el área de administración de la información, un modelo computacional de minería de datos práctico enfocado a inventarios y un análisis comparativo de técnicas de minera de datos aplicadas a este enfoque, que servirá en investigaciones futuras.

El modelo de minería de datos permitirá mejorar la eficiencia del proceso de toma de decisiones en la gestión del inventario, de manera que esta sea más racional y menos intuitiva con menos posibilidades de equivocación, con el fin de facilitar el logro de ventajas competitivas.

1.4 OBJETIVOS

1.4.1 Objetivo General

Determinar los patrones de predicción mediante el uso de minería de datos del consumo de medicamentos de Redes Puno, y mejorar la oportunidad de la información a través del uso de los arboles de decisión.

1.4.2 Objetivos Específicos

a. Diseñar y documentar el estado del arte en técnicas de minería de datos y en modelamiento multidimensional de datos transaccionales operativas

de consumo de medicamentos de Redes Puno, mediante el modelo CRISP-DM.

b. Diseñar e implementar un Data Mart para el área de Farmacia, Admisión y triaje, como repositorio único y consolidado de datos e indicadores, que permita tener la información disponible para su consulta, mediante el proceso KDD.

c. Aplicar WEKA como algoritmos de minería de datos para obtener los arboles de decisión y clasificación, para extraer conocimiento desde los datos de Farmacia, Admisión y triaje de consumo de medicamentos de Redes Puno.

1.5 HIPÓTESIS

La minería de datos mejorar la oportunidad de la información y optimizaran el abastecimiento de medicamentos para una mejor toma de decisiones.

CAPÍTULO II

MARCO TEÓRICO

2.1. ANTECEDENTES

Martinez (2012) la metodología de este trabajo se basa principalmente en las etapas del proceso conocido como Knowledge Discovery in Databases (KDD), implementadas de acuerdo a la metodología CRISP-DM, la cual es usada para el desarrollo de proyectos de minería de datos. Para comenzar, se hizo un levantamiento de las métricas existentes para la gestión de la provisión de servicios. Luego, se seleccionaron y procesaron las fuentes de datos para el estudio de forma automática, almacenando las variables más relevantes en un repositorio multidimensional (Data Mart), reduciendo drásticamente el tiempo de cálculo de indicadores y liberando recursos humanos altamente calificados. A partir de lo anterior, se aplicaron técnicas de clustering para obtener grupos de elementos con datos de clientes y servicios cuyas características fueran similares, asociándoles un valor de precio según información histórica de consumo. Por último, se generó un modelo de clasificación que asignara, de

acuerdo a una medida de similitud, elementos que no habían sido facturados a los grupos previamente definidos, y de esta manera estimar los ingresos no percibidos.

Gildo Tapia Rivas (2006) en su tesis Minería de datos, sectorizo en el consumo de medicamentos, para descubrir y enumerar patrones presentes en los datos, utilizando algoritmos de segmentación o clasificación, para evaluar la forma con la que se consumen los medicamentos en un Hospital en el Perú y poder identificar alginas realidades o características no observables que producirían desabastecimiento o insatisfacción del paciente, y para que sirva como una herramienta en la toma de decisión sobre el abastecimiento de medicamentos en el hospital.

Banet (2001) sostiene que la minería de datos es, en realidad, una prolongación de una práctica estadística de larga tradición, la de Análisis de Datos. Existe, además, una aportación propia de técnicas específicas de Inteligencia Artificial, en particular sobre la integración de los algoritmos, la automatización del proceso y la optimización del coste. Por otro lado, en el mundo estadístico más académico, la minería de datos ha sido considerada en su inicio como una moda más, aparecida después de los sistemas expertos, conocida desde hacía tiempo bajo el nombre de "data fishing".

El trabajo de García, López,, Moreno, Abad, & Blasco. (2009) pretende principalmente acercar a los investigadores del campo de las drogodependencias una metodología de análisis de datos orientada al descubrimiento de conocimiento en bases de datos (KDD). El KDD es un proceso que consta de una serie de fases, la más característica de las cuales

se denomina Data Mining (DM), en la que se aplican diferentes técnicas de modelado para detectar patrones y relaciones en los datos. Se analizan los factores comunes y diferenciadores de las técnicas DM más ampliamente utilizadas, desde una visión principalmente metodológica, y ejemplificando su uso con datos provenientes del consumo de alcohol en adolescentes y su posible relación con variables de personalidad (N=7030). Aunque la precisión global obtenida (% de predicciones correctas) es muy similar en los tres modelos analizados, las redes neuronales generan el modelo más preciso (64.1%), seguidas de los árboles de decisión (62.3%) y Naive Bayes (59.9%).

Martínez Álvarez (2012) el estudio se enfoca en el análisis de ingresos no percibidos en la empresa de telecomunicaciones ENTEL, dentro del proceso de provisión de servicios privados de telefonía, internet y comunicaciones a los clientes de mercados no residenciales. Dicho proceso es controlado mediante indicadores de gestión, obtenidos a partir de la transformación de datos de clientes y servicios. La generación de estos indicadores demanda tiempo y esfuerzo por parte de los analistas de la empresa, debido a que es un trabajo realizado en forma manual. El objetivo principal de esta tesis consiste en reducir el tiempo de cálculo de los indicadores de servicios privados de ENTEL, para lo cual se aplicó modelamiento multidimensional, técnicas de minería de datos y automatización de procesos, y de este modo poder entregar información más oportunamente. La metodología de este trabajo se basa principalmente en las etapas del proceso conocido como Knowledge Discovery in Databases (KDD), implementadas de acuerdo a la metodología CRISP-DM, la cual es usada para el desarrollo de proyectos de minería de datos.

Siccha Vega, Hober Willy (2012) Minería de datos aplicados a las ventas con tarjeta de crédito clásica realizados en las tiendas Saga Falabella en la ciudad de Lima. El interés de esta investigación es determinar el comportamiento a futuro y la naturaleza de los datos históricos de ventas con tarjeta de crédito clásica en las tiendas de Saga Falabella de la ciudad de Lima a través de la explotación de las técnicas de minería de datos, con la finalidad de ayudar a los miembros de la alta dirección a analizar los hábitos de los clientes a fin de satisfacer mejor su demanda, mejorar la administración de los inventarios de los productos que están asociados a las transacciones de ventas y mejorar los volúmenes de ventas

Monserrat Sergio(2013) Minería de Datos en Base de Datos de Servicios de Salud, que ha tomado como caso de estudio una empresa mutual que brinda servicios de salud. La empresa canaliza la atención de los afiliados a través de convenios con entidades de profesionales de la medicina e instituciones sanitarias como hospitales, sanatorios y clínicas. Tiene cobertura nacional con administración centralizada en Santa Fe. Las prestaciones incluyen prácticas médicas, internaciones, óptica, ortopedia, rehabilitación y farmacia. Con la finalidad de Los resultados obtenidos del proceso de minería de datos, realizado sobre la base de datos de una empresa de servicios de salud, permitieron alcanzar el objetivo propuesto de buscar patrones de consumo de medicamentos por franja etaria, por sexo y por estaciones del año.

Jorge Luis Ruge Leiva y Camilo Andrés Sierra Niño (2015) quien Construyo un sistema prototipo que complemente al sistema actual (SAHI) en la toma de decisiones sobre el control de los medicamentos que tiene la farmacia general del Hospital Universitario San Ignacio, teniendo en cuenta la demanda y la

fecha de vencimiento de los productos. Quien Se logró cumplir con el objetivo general planteado en este trabajo, ya que se desarrolló un sistema que complementa al sistema actual (SAHI) para apoyar la toma de decisiones en el manejo de inventario de medicamentos de la farmacia general del H.U.S.I. Se enfatizó en la generación de pronósticos de la demanda de medicamentos y en la utilización de alertas para informar sobre el vencimiento de medicamentos.

2.2. MARCO CONCEPTUAL

2.2.1 Datawarehouse

Un Data Warehouse (DW) es un gran repositorio lógico de datos que permite el acceso y la manipulación flexible de grandes volúmenes de información provenientes tanto de transacciones detalladas como datos agregados de fuentes de distinta naturaleza . Los sistemas de administración de DW integran información procedente de diversos sistemas operacionales, la seleccionan, la historizan y la almacenan para proporcionar la base para la planeación, control y toma de decisiones a un alto nivel.

2.2.2. Arquitectura General

La arquitectura general de un DW es la que se muestra en la Figura 1. Este diagrama muestra como primer componente dentro de la arquitectura DW a las fuentes desde las cuales se extrae la información necesaria para poblar la base de datos. Conectada a cada una de las fuentes se encuentran los siguientes componentes básicos de la arquitectura, los wrappers o extractores, los cuales extraen y transforman la información de las fuentes. Posteriormente través de un integrador

dicha información se carga a la base de datos, la cual constituye el siguiente componente básico de la arquitectura. Este proceso de cargado de la información ejecuta las tareas siguientes:

- Transforma los datos de acuerdo al modelo de datos del warehouse.
- Limpia dichos datos para corregir y depurar errores que pueden contener las fuentes (por lo general se generan en la captura de los datos en los sistemas de transacción diaria).
- Integra todos los datos para formar la base de datos en la cual se encontrará la información.

De igual manera, los meta datos deben ser refrescados dentro de este proceso. Dicho proceso es crítico para asegurar la calidad de la información y soportar una adecuada toma de decisiones con datos correctos y previamente verificados. Una vez que los datos han sido cargados se encuentran disponibles para un sistema que soporte decisiones. Sin embargo, las aplicaciones no accesan directamente el warehouse debido a que es demasiado grande, además de poseer un esquema genérico no óptimo para el usuario final. Por consiguiente, vistas especializadas más pequeñas del DW son cargadas en los data marts, éstos son repositorios más pequeños con vistas materializadas para facilitar la consulta de los datos Esta carga se realiza a través de un segundo proceso más simple debido a que los datos ya se encuentran ordenados y verificados dentro del DW. Únicamente se seleccionan las vistas requeridas y a través de una serie de

transformaciones necesarias quedan establecidas para facilitar y acelerar el proceso de consulta del usuario. Finalmente los datamarts son accesados a través de las herramientas para el usuario final (OLAP o ambientes de consultas analíticas, generalmente), las cuales permiten analizar la información disponible en el warehouse para la generación de consultas especializadas, reportes, nuevas clasificaciones y tendencias que sirvan de apoyo a la toma de decisiones.

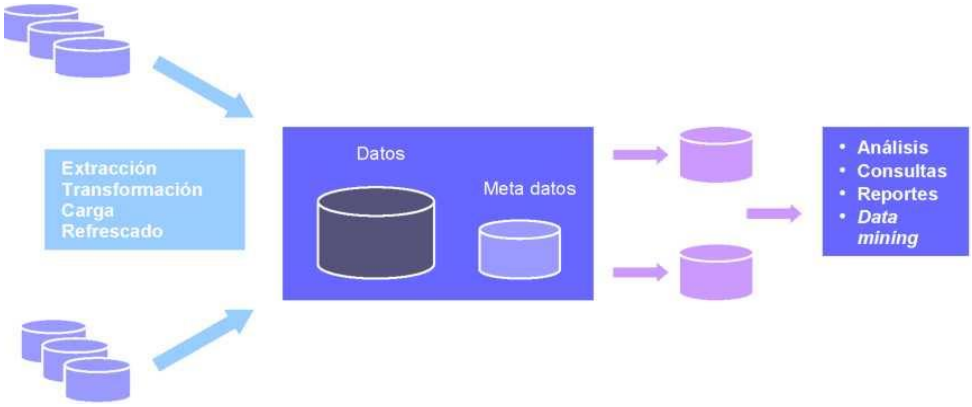


Figura 1. Arquitectura básica de un DW.

2.2.3 Modelo multidimensional

Para facilitar el análisis de los datos, un Data Warehouse representa los datos que contiene usando modelos multidimensionales. De manera general, un modelo multidimensional provee dos conceptos principales: medida y dimensión. Una medida es un valor en un espacio multidimensional definido por dimensiones ortogonales. Así, el cubo es el concepto central del modelado de datos multidimensional, donde se muestra una instancia del modelo multidimensional: un esquema del mismo tipo.

Dentro del modelo de datos multidimensional las medidas o atributos numéricos describen un cierto proceso del mundo real el cual va a ser objeto de un análisis. Estos atributos dependen de ciertas dimensiones las cuales proveen el contexto a través del cual van a ser interpretadas las medidas. Dichas dimensiones regularmente se encuentran en orden jerárquico (ejemplo: tiempo-í-día ^mes año). Las medidas pueden ser agregadas a lo largo de las dimensiones lo cual resulta en un cubo el cual es la base para el uso de las operaciones OLAP, estas operaciones serán explicadas más adelante en otra sección.

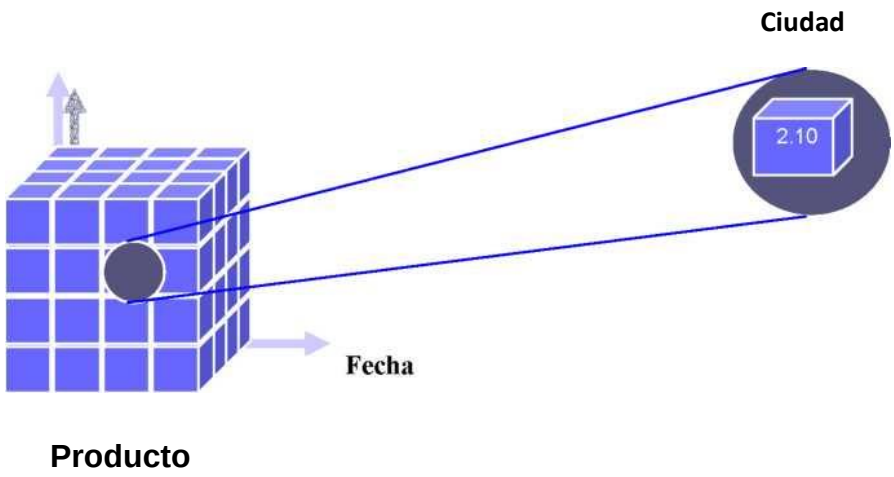


Figura 2. Esquema multidimensional de bases de datos

El esquema multidimensional presentado puede ser implementado directamente a través de un servidor MOLAP (Multidimensional OLAP). Dichos servidores soportan vistas multidimensionales de los datos a través de un repositorio multidimensional. Esto hace posible implementar consultas multidimensionales a la base a través de un mapeo directo. Otra alternativa para implementar el modelo multidimensional es a través de la tecnología ROLAP (Relational OLAP).

Esta tecnología utiliza un esquema de bases de datos relacional para representar la información (datos y medidas) del esquema multidimensional. Ambas tecnologías son útiles y tienen sus méritos. El esquema relacional puede manejar grandes cantidades de datos y los nuevos avances en esta tecnología han mejorado para el manejo de Data Warehouses. Los sistemas MOLAP debido a la representación de los datos pueden responder rápidamente a consultas muy complejas y permitir así un análisis rápido de la información. Sin embargo siguen teniendo problemas para bases con grandes cantidades de datos.

La tecnología mayormente usada para representar los esquemas multidimensionales en el manejo de DW es la ROLAP. Cuando un servidor relacional es utilizado, el modelo multidimensional y sus operaciones deben ser mapeadas a relaciones y las consultas basadas en SQL (Structured Query Language). La mayoría de los DW utilizan el esquema en estrella para representar el modelo multidimensional de bases de datos. se muestra un esquema de este tipo. La base consiste de una tabla simple de hechos que contiene un apuntador a cada una de las dimensiones que proveen las coordenadas del esquema multidimensional y guarda las medidas numéricas para esas coordenadas. Cada tabla de dimensión consiste de columnas que corresponden a los atributos de cada dimensión.

Los esquemas en estrella generalmente no proveen un soporte explícito para la jerarquía de cada una de las dimensiones. Por lo cual generalmente después de generar un esquema en estrella se realiza una normalización del mismo generando un esquema copo de nieve Los

esquemas copo de nieve se basan en el esquema en estrella para realizar una normalización del mismo y obtener un esquema que representa de mejor manera el modelo multidimensional del DW.

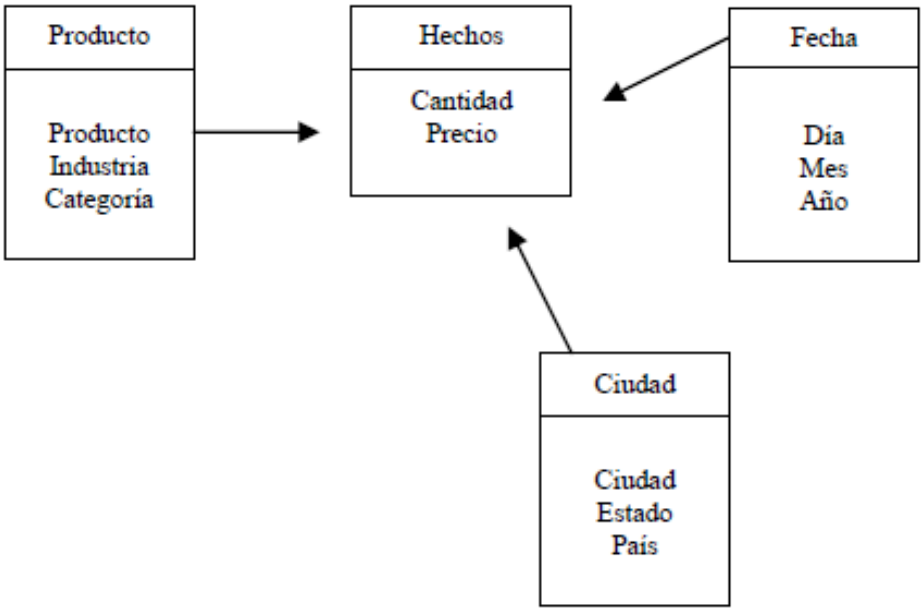


Figura 3. Esquema copo de nieve

2.2.4 Minería de Datos

Se denomina Minería de Datos al conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con objeto de predecir de forma automatizada tendencias y comportamientos (Perichinsky, G., M. Servente, A. Servetto, R. García-Martínez, R. Orellana, A. Plastino, 2003); para descubrir y enumerar patrones presentes en los datos, se desarrollaron un gran número de métodos de análisis de datos basados en la estadística (Michalski et al, 2003). En la

medida en que se incrementaba la cantidad de información almacenada en las bases de datos, estos métodos empezaron a enfrentar problemas de eficiencia y escalabilidad y es aquí donde aparece el concepto de minería de datos. Una de las diferencias entre al análisis de datos tradicional y la minería de datos es que el primero supone que las hipótesis ya están construidas para validadar, mientras que el segundo supone que los patrones e hipótesis son automáticamente extraídos de los datos (Hernández, 2000).

La Minería de Datos es un proceso completo de descubrimiento de conocimiento que involucra varios pasos (Morales, 2003):

i. Entendimiento del dominio de aplicación, el conocimiento relevante a utilizar y las metas del usuario.
ii. Seleccionar un conjunto de datos en donde realizar el proceso de descubrimiento.
iii. Limpieza y pre procesamiento de los datos, diseñando una estrategia adecuada para manejar ruido, valores incompletos, valores fuera de rango, valores inconsistentes, etc.
iv. Selección de la tarea de descubrimiento a realizar, por ejemplo, clasificación, agrupamiento o clustering, reglas de asociación, etc.
v. Selección de los algoritmos a utilizar.
vi. Transformación de los datos al formato requerido por el algoritmo especifico de explotación de datos, hallando los atributos útiles, reduciendo las dimensiones de los datos, etc.
vii. Llevar a cabo el proceso de minería de datos para encontrar patrones interesantes.

viii. Evaluación de los patrones descubiertos y presentación de los mismos mediante técnicas de visualización. Quizás sea necesario eliminar patrones redundantes o no interesantes, o se necesite repetir algún paso anterior con otros datos, con otros algoritmos, con otras metas o con otras estrategias.

ix. Utilización del conocimiento descubierto, ya sea incorporándolo dentro de un sistema o simplemente para almacenarlo y reportarlo a las personas interesadas.

Es muy importante la etapa del pre-procesamiento de los datos y su transformación al formato requerido por el algoritmo, ya que dependiendo de cómo se realicen estas tareas, va a depender la calidad final de los patrones descubiertos. Un patrón es interesante si es fácilmente entendible por las personas, potencialmente útil, novedoso o valida alguna hipótesis que el usuario busca confirmar. Un patrón interesante representa conocimiento (Ale, 2005).

Las principales técnicas de minería de datos se suelen clasificar según su tarea de descubrimiento en: Agrupación o clustering, Clasificación, Asociación y otros.

A continuación se realiza una breve descripción de cada una de estas técnicas y los algoritmos más utilizados.

2.2.5 Agrupación de datos

La agrupación o clustering consiste en agrupar un conjunto de datos basándose en la similitud de los valores de sus atributos. De esta

manera se busca maximizar la similitud de las instancias en cada cluster y minimizar la similitud entre clusters (Han & Kamber, 2001).

La técnica de clustering ha sido estudiada en las áreas de la estadística (Cheeseman & Stutz, 1996; Jain & Dubes, 1988), machine learning (Fisher, 1996), base de datos espaciales y minería de datos (Cheeseman & Stutz, 1996; Ester et al., 1995; Ng & Han, 1994; Zhang et al., 1996).

Dos de los algoritmos de clustering más utilizados son Self Organizing Maps (SOM) y K-means. SOM, también denominado redes de Kohonen, fue creado por Teuvo Kohonen en 1982. Se trata de un modelo de red neuronal con capacidad para formar mapas de características de manera similar a como ocurre en el cerebro.

SOM está basado en el aprendizaje no supervisado y competitivo, lo cual quiere decir que no se necesita intervención humana durante el mismo y que se necesita saber muy poco sobre las características de la información de entrada. SOM provee un mapa topológico de datos, que se representan en varias dimensiones, utilizando unidades de mapa (las neuronas) para simplificar la representación (Kohonen, 1995).

Las neuronas usualmente forman un mapa bidimensional, por lo que el mapeo transforma un problema de muchas dimensiones en el espacio, a un plano. La propiedad de preservar la topología significa que el mapeo preserva las distancias relativas entre puntos. Los puntos que están cerca unos de los otros en el espacio original de entrada son mapeados a neuronas cercanas en SOM.

Por esta razón, SOM es muy útil como herramienta de análisis de clases de datos de muchas dimensiones (Vesanto & Alhoniemi, 2000), y además tiene la capacidad de generalizar (Essenreiter et al., 1999), lo que implica que la red puede reconocer o caracterizar entradas que nunca antes ha encontrado. K-means es un método iterativo que busca formar k clusters, con k predeterminado antes del inicio del proceso. K-means comienza particionando los datos en k subconjuntos no vacíos, calcula el centroide de cada partición como el punto medio del cluster y asigna cada dato al cluster cuyo centroide sea el más próximo. Luego vuelve a particionar los datos iterativamente, hasta que no haya más datos que cambien de cluster de una iteración a la otra.

2.2.7 Clasificación de datos

Clasificar un conjunto de datos basado en los valores de sus atributos. Por ejemplo, clasificar a distintas personas para la otorgación de un préstamo en riesgo bajo, medio y alto, teniendo en cuenta información histórica de las mismas. La clasificación encuentra las propiedades comunes entre un conjunto de objetos y los clasifica en diferentes clases, de acuerdo a un modelo de clasificación.

Para construir este modelo, se utiliza un conjunto de entrenamiento, en el que cada instancia consiste en un conjunto de atributos y el valor de la clase a la cual pertenece para analizar los datos de entrenamiento y, mediante un método supervisado, desarrollar una descripción o un modelo para cada clase utilizando las características disponibles en los datos. Esta descripción o modelo permite clasificar otras instancias, cuya clase es

desconocida. El método se conoce como supervisado debido a que, para el conjunto de entrenamiento, se conoce la clase de pertenencia y se le indica al modelo si la clasificación que realiza es correcta o no. La construcción del modelo se realimenta de estas indicaciones del supervisor (Chen *et al*., 1996).

Los algoritmos mayormente utilizados para las tareas de clasificación son los algoritmos de inducción que generan árboles de decisión. La clasificación basada en árboles de decisión es un método de aprendizaje supervisado que construye árboles de decisión a partir de un conjunto de entrenamiento.

Un sistema típico de construcción de árboles de decisión es ID3, que utiliza para minimizar la cantidad de pruebas para clasificar un objeto. Al utilizar métodos heurísticos, ID3 garantiza un árbol simple, pero no necesariamente el más simple. Una extensión de ID3 es C4.5 (Quinlan, 1993ª), que extiende el dominio de clasificación de atributos categóricos a numéricos. Un paso importante en la construcción del árbol de decisión es la poda, la cual elimina las ramas no necesarias, resultando en una clasificación más rápida y una mejora en la precisión de la clasificación de datos (Han & Kamber, 2001).

Existen muchos otros algoritmos de clasificación de datos, incluyendo métodos estadísticos (Cheeseman & Stutz, 1996), como el análisis de regresión lineal (Elder IV & Pregibon, 1996); algoritmos de machine learning (Cheeseman & Stutz, 1996); redes neuronales (Lu et al., 1995), algoritmos genéticos y lógica difusa.

2.2.8 Reglas de asociación

La minería de reglas de asociación consiste en encontrar reglas de la forma (A_1 y A_2 y ... y A_m) => (B_1 y B_2 y ... y B_n), donde A_i y B_j son valores de atributos del conjunto de datos (Chen et al., 1996). Por ejemplo, se podría encontrar en un gran repositorio de datos de compras en un supermercado, la regla de asociación correspondiente a que si un cliente compra leche, entonces compra pan. Una regla de asociación es una sentencia probabilística acerca de la co-ocurrencia de ciertos eventos en una base de datos, y es particularmente aplicable a grandes conjuntos de datos (Hand *et a*l., 2001).

Existen varios algoritmos que realizan el descubrimiento de reglas de asociación, uno de los más utilizados es A priori según (Ibermática, 2007), una data mart es: una base de datos especializada, departamental, orientada a satisfacer las necesidades específicas de un grupo particular de usuarios (en otras palabras, una data mart normalmente en un subconjunto del total corporativo con transformaciones específicas para el área a la que va dirigido).

Bill Inmon: Data Warehouse es una parte del todo que conforma un sistema de inteligencia de negocios. Una empresa tiene una Data Warehouse, y los data marts tienen como fuente de información ese Data Warehouse. Ésta aproximación también es conocida como "Top-Down"

Ralph Kimball: bajo este paradigma, el Data Warehouse se compone por el conglomerado de todos los Data Marts generados en una empresa. La

información siempre se almacena en un modelo dimensional. Otra forma de denominar ésta aproximación es como "Bottom-up".

Almacenar recursos, analizar e interpretar la información que se genera y se acumulan para su análisis con el fin de tomar decisiones críticas que permitan su existencia pero sobre todo que maximicen su prosperidad; por lo que se vuelve prioritario crear sistemas de análisis y retroalimentación para comprender su información (Data Warehouse) y de esta manera contar con los elementos adecuados para la toma de decisiones.

Estandarización de la información: Estando disponibles los datos en bruto en el data Warehouse se llevan a cabo los procesos de transformación: normalización y limpieza de datos. De esta forma que los datos almacenados guarden una coherencia de formato, cambios de unidad, operaciones entre campos, etc.

Limpieza de datos: Generalmente, tras la extracción en bruto de la información, hay datos que no interesan mantener, o son datos duplicados. Es frecuente realizar procesos de limpieza o de filtrado para eliminar información innecesaria, redundante o errónea.

Carga de datos: Tras aplicar todos los procesos de transformación, se lleva a cabo la carga consolidada de los datos. Es habitual disponer de dos bases de datos separadas físicamente una para la preparación de los datos y otra para el data Warehouse en sí. El proceso de volcado sería pues el paso de la primera de estas bases de datos (llamada staging area o interfaz) al data Warehouse. Es habitual que este proceso

requiera elborrado de algunos datos del data Warehouse que van a ser refrescados.

PKI Mide lo que es importante designar las métricas importantes, se denomina indicadores de gestión, KPI (Key Performance Indicators). Los sistemas de Business Intelligence están específicamente diseñados para asimilar grandes cantidades de datos complejos de diferentes fuentes y combinar estos datos utilizando algoritmos complejos con el fin de asignar, agregar y, en definitiva, jugar con la información. El resultado es la obtención sistemática de informes con las métricas, ratios e indicadores del negocio; los auténticos KPI que los gerentes necesitan identificar, analizar y utilizar para tomar decisiones de forma frecuente. (Vitt, Luckevich y Misner. 2002).

2.3. HERRAMIENTAS DE MINERÍA DE DATOS.

2.3.1 Weka

WEKA (Waikato Environment for Knowledge Analysis) es una herramienta visual de libre distribución (Licencia GNU) desarrollada por un equipo de investigadores de la 25 universidad de Waikato (Nueva Zelanda). La herramienta está implementada en Java. Como entorno de Minería de Datos conviene destacar:

- **Acceso a datos:** Los datos son cargados desde un archivo en formato ARFF (Archivo plano organizado en filas y columnas). El usuario puede observar en los diferentes componentes gráficos, información de interés sobre el conjunto de muestras (talla del conjunto, número de atributos, tipo de datos, medias y varianzas de

los atributos numéricos, distribución de frecuencias en los atributos nominales, etc.)

- **Pre-procesado de datos:** Selección de atributos, discretización, tratamiento de valores desconocidos, transformación de atributos numéricos.
- **Modelo de aprendizaje:** Árboles de decisión, tablas de decisión, vecinos más próximos, máquinas de vectores soporte, reglas de asociación, métodos de agrupamiento (K medias, EM y Cobweb), modelos combinados.
- **Visualización:** La interfaz gráfica se compone de diversos entornos. El entorno Explorer permite controlar todas las operaciones anteriores (filtrado, selección y especificación del modelo, diseño de experimentos, etc). El entorno consola (CLI) posibilita la invocación textual de las operaciones anteriores. El entorno Experimenter facilita el diseño y la realización de experimentos complejos. El proceso global de Minería de Datos en Weka se acelera considerablemente gracias al entorno KnowledgeFlow que, de una forma gráfica y a modo de flujo de operaciones, permite definir la totalidad del proceso.

CAPÍTULO III

METODOLOGÍA

3.1. Tipo y Diseño de Investigación

La presente investigación implica la producción de un nuevo conocimiento para la solución de problemas prácticos(Arias,2004), es de tipo Investigación Aplicada y la metodología es CRISP-DM por su uso frecuente en proyectos de minería de datos.

3.2. Población

Todos los datos de consumo de medicamentos en el ámbito de la Red Puno.

3.3. Muestra

Todos los datos de Consumo de Medicamentos en el periodo 2016 en el ámbito de la Red Puno.

3.4. Ubicación y Descripción de la Población

Se investigó en los Establecimientos de la Red Puno.

3.5. La Metodología CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining), es un modelo de proceso de minería de datos que describe una manera en la que los expertos en esta materia abordan el problema.

Para implementar una tecnología en un negocio es necesaria una metodología. Estos métodos suelen venir de las experiencias propias y también de los procedimientos estándar más conocidos. En el caso de los proyectos de implementación de minería de datos una de las metodologías que ha tenido más apoyo de las empresas privadas y organismos públicos es CRISP-DM, como se puede observar en la siguiente gráfica (figura 3), publicada en kdnuggets.com, y que representa el grado de utilización de las principales guías de desarrollo de proyectos de minería de datos según las encuestas realizadas. Como se puede observar CRISP-DM ha experimentado un ligero descenso en los últimos años, pero sigue siendo la más empleada de las distintas metodologías.

CRISP-DM incluye un modelo y una guía, estructurados en seis fases, algunas

de las cuales son bidireccionales, es decir que de una fase en concreto se puede volver a una fase anterior para poder revisarla, por lo que la sucesión de fases no tiene porqué ser ordenada desde la primera hasta la última. se puede observar las fases en las que se divide CRISP-DM y las posibles secuencias a seguir entre ellas.

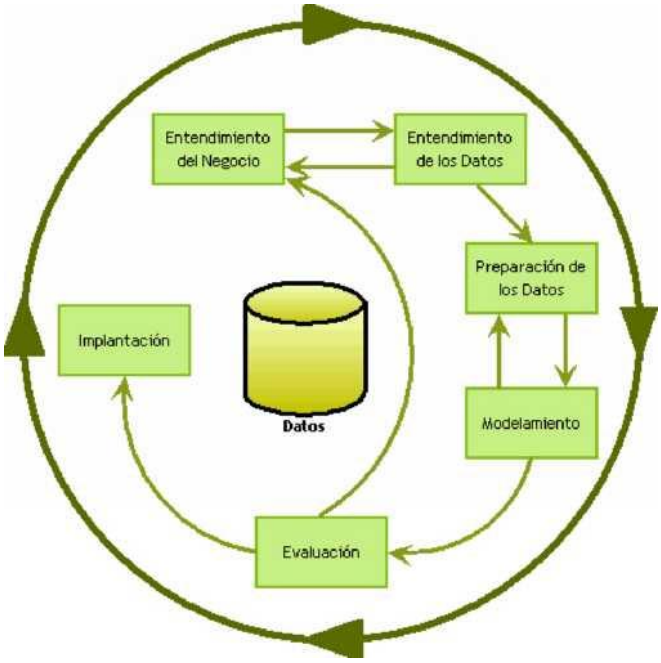


Figura 4. Secuencia del proceso CRISP - DM

A continuación, se explica cada una de estas fases (Rodríguez, 2010):

A. Comprensión del negocio.

Esta primera fase es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva de negocio, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables. Para obtener el mejor provecho de

la minería de datos, es necesario entender de la manera más completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados. En esta fase, es muy importante la capacidad de poder convertir el conocimiento adquirido del negocio en un problema de minería de datos y en un plan preliminar cuya meta sea el alcanzar los objetivos del negocio. A continuación vemos una descripción de cada una de las principales tareas que componen esta fase, figura 5 (CRISP- DM, 2000).

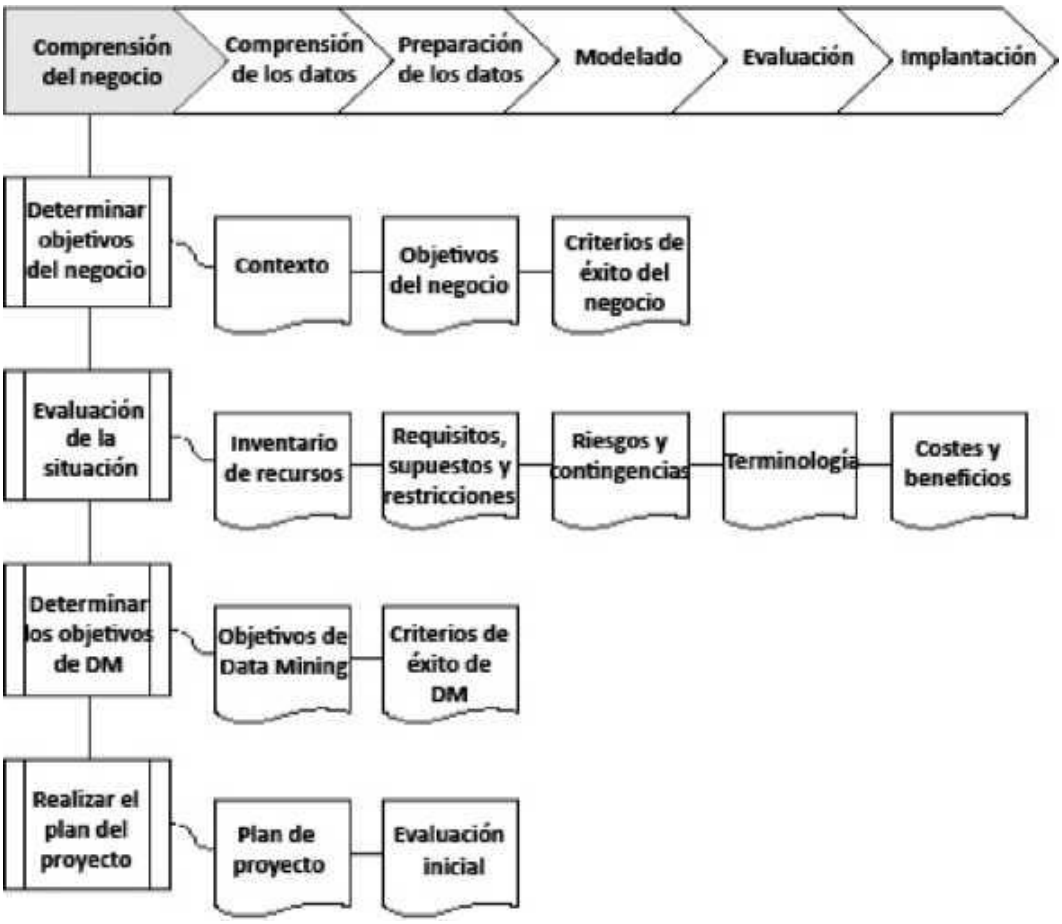


Figura 5. Fase de comprensión del negocio

Determinar los objetivos del negocio.

Esta es la primera tarea a desarrollar y tiene como metas

determinar cuál es el problema que se desea resolver, por qué la necesidad de utilizar la minería de datos y definir los criterios de éxito. Los problemas pueden ser diversos, como por ejemplo, detectar fraude en el uso de tarjetas de crédito, detección de intentos de ingreso indebido a un sistema, asegurar el éxito de una determinada campaña publicitaria, etc. En cuanto a los criterios de éxito, estos pueden ser de tipo cualitativo, en cuyo caso un experto en el área de dominio califica el resultado del proceso de minería de datos, o bien de tipo cuantitativo, por ejemplo, el número de detecciones de fraude o la respuesta de clientes ante una campaña publicitaria.

Evaluación de la situación.

En esta tarea se debe calificar el estado de la situación antes de iniciar el proceso de minería de datos, considerando aspectos tales como: ¿cuál es el conocimiento previo disponible acerca del problema?, ¿se cuenta con la cantidad de datos requerida para resolver el problema?, ¿cuál es la relación coste beneficio de la aplicación de minería de datos?, etc. En esta fase se definen los requisitos del problema, tanto en términos de negocio como en términos de minería de datos.

Determinar los objetivos de la minería de datos.

Esta tarea tiene como objetivo representar los objetivos del negocio en términos de las metas del proyecto de minería de datos, como por ejemplo, si el objetivo del negocio es el

desarrollo de una campaña publicitaria para incrementar la asignación de créditos hipotecarios, la meta de minería de datos será por ejemplo determinar el perfil de los clientes respecto de su capacidad de endeudamiento.

Realizar el plan del proyecto.

Esta última tarea de la primera fase de CRISP-DM tiene como meta desarrollar un plan para el proyecto, que describa los pasos a seguir y las técnicas a emplear en cada uno de ellos.

B. Comprensión de los datos.

Esta segunda fase comprende la recolección inicial de los datos con el objetivo de establecer un primer contacto con el problema, familiarizarse con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las dos siguientes fases son las que demandan el mayor esfuerzo y tiempo en un proyecto de minería de datos. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos específica para el proyecto de DM (Data Mining), ya que durante el desarrollo del proyecto es posible que se generen frecuentes y abundantes accesos a la base de datos con el fin de realizar consultas y probablemente se produzcan modificaciones, lo cual podría generar muchos problemas. Vemos las tareas que componen esta fase, figura 6 (CRISP- DM, 2000).

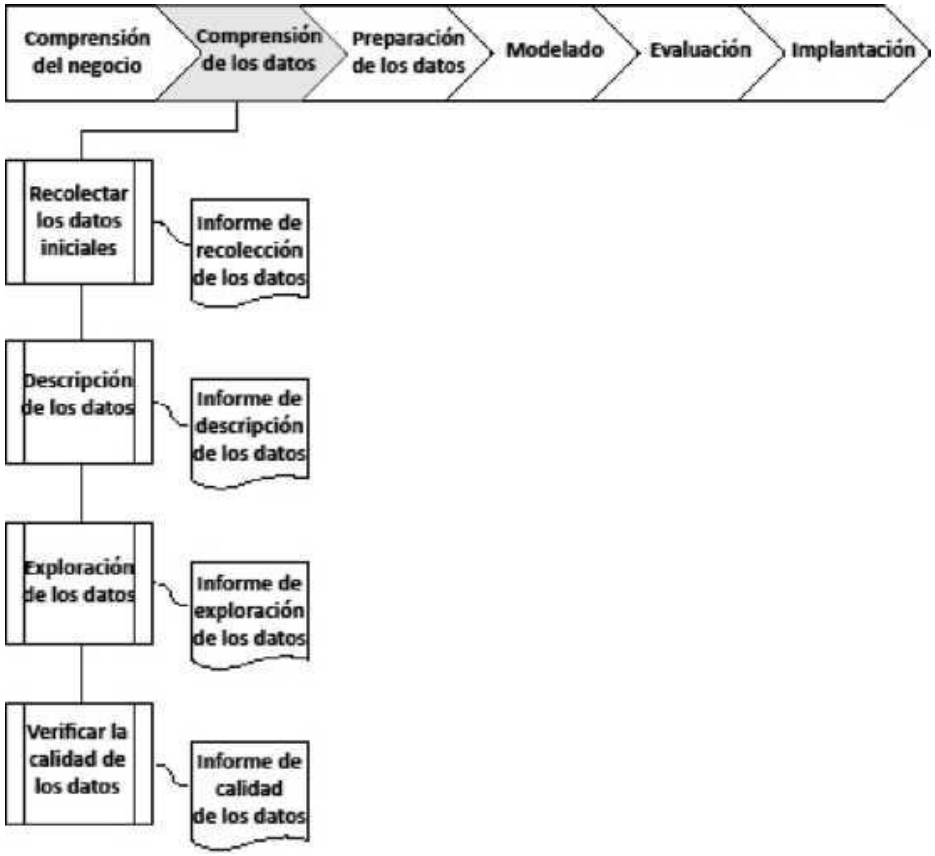


Figura 6. Fase de comprensión de los datos

- **Recolectar los datos iniciales.**

La primera tarea en esta segunda fase del proceso de CRISP-DM es la recolección de los datos iniciales y su adecuación para el futuro procesamiento. Esta tarea tiene como objetivo elaborar informes con una lista de los datos adquiridos, su localización, las técnicas utilizadas en su recolección y los problemas y soluciones inherentes a este proceso.

- **Descripción de los datos.**

Después de adquiridos los datos iniciales, estos deben ser descritos. Este proceso implica establecer volúmenes de datos (número de registros y campos por registro), su identificación, el significado de cada campo y la descripción del formato inicial.

- **Exploración de los datos.**

Una vez realizada la descripción de los datos, se procede a su exploración, cuyo fin es encontrar una estructura general para los datos. Esto implica la aplicación de pruebas estadísticas básicas que revelen propiedades en los datos recién adquiridos, se crean tablas de frecuencia y se construyen gráficos de distribución. La salida de esta tarea es un informe de exploración de los datos.

- **Verificar la calidad de los datos.**

En esta tarea se efectúan verificaciones sobre los datos para determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los valores nulos, y para encontrar valores fuera de rango, los cuales pueden constituirse en ruido para el proceso. La idea una vez llegados a este punto es poder garantizar la completitud y corrección de los datos.

C. Preparación de los datos.

En esta fase y una vez efectuada la recolección inicial de los datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se van a utilizar posteriormente, éstas pueden ser técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para explotación de los datos. La preparación de los datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

Esta fase se encuentra relacionada con la fase de modelado, ya que en función de la técnica de modelado elegida, los datos requieren ser procesados de una manera o de otra, por esta razón las fases de preparación y de modelado interactúan de forma permanente. En la figura 7 (CRISP-DM, 2000) se pueden ver cada una de las tareas de las que se compone esta fase así como las salidas de cada una de ellas.

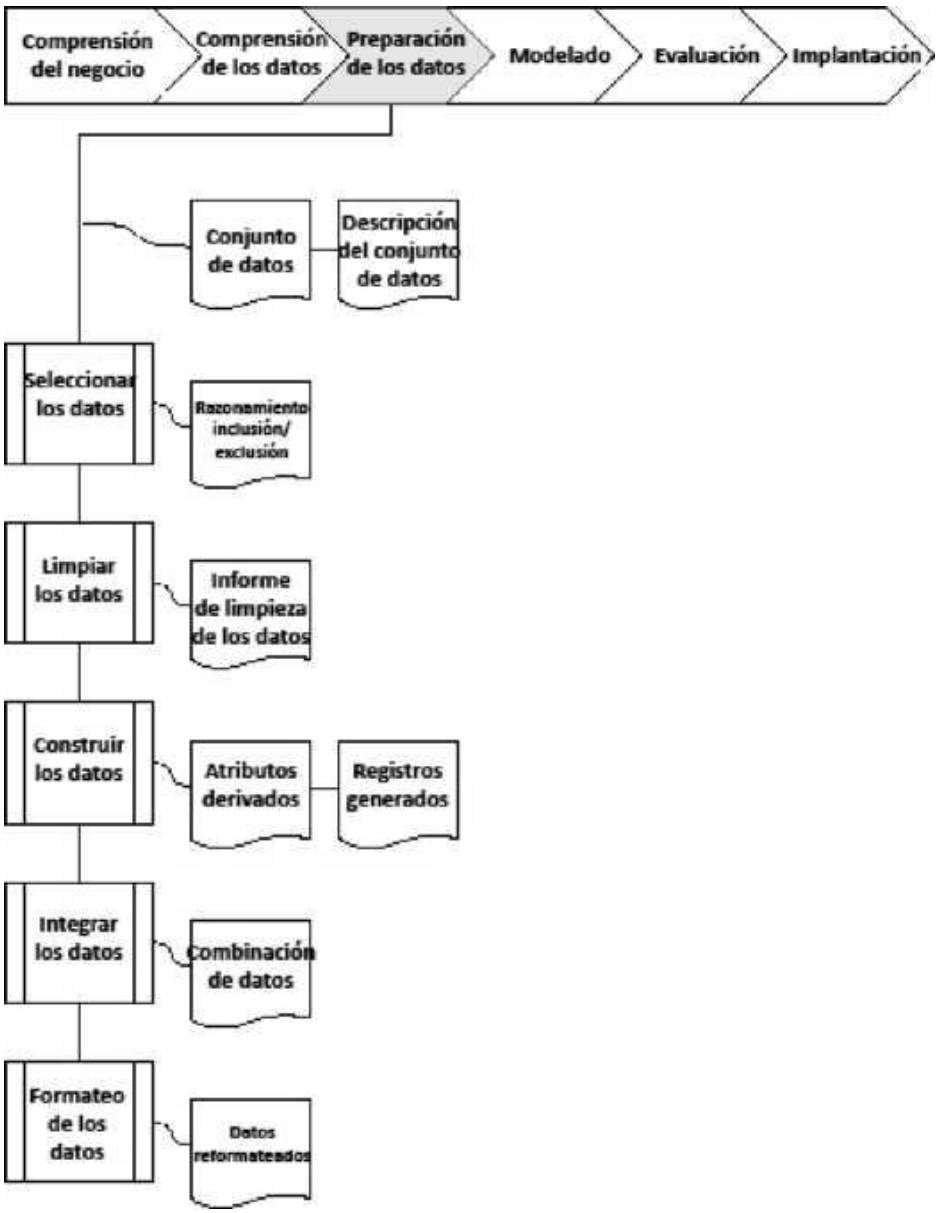


Figura 7. Fase de preparación de los datos

Seleccionar los datos.

En esta etapa se selecciona un subconjunto de los datos adquiridos anteriormente apoyándose en criterios previamente definidos en las fases anteriores como la calidad de los datos en cuanto a su completitud, corrección de los datos y limitaciones en el volumen o en los tipos de datos que están relacionados con las técnicas de minería de datos seleccionadas. Limpiar los datos.

Esta tarea complementa a la anterior y es una de las que más tiempo y esfuerzo consume debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos a objeto de prepararlos para la fase de modelación. Algunas de las técnicas a utilizar para este propósito son la normalización de los datos, discretización de campos numéricos, tratamiento de valores faltantes, reducción del volumen de datos, etc.

Construir los datos.

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

Integrar los datos.

La integración de los datos implica la creación de nuevas estructuras a partir de los datos seleccionados, por ejemplo, generación de nuevos campos a partir de otros existentes, creación de nuevos

registros, fusión de tablas campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen.

Formateo de los datos.

Esta tarea consiste principalmente en la realización de transformaciones sintácticas de los datos sin modificar su significado de tal forma que se permita y se facilite utilizar alguna técnica de minería de datos en concreto, como por ejemplo la reordenación de los campos y/o de los registros de la tabla o el ajuste de los valores de los campos a las limitaciones de las herramientas de modelación (eliminar comas, tabuladores, caracteres especiales, máximos y mínimos para las cadenas de caracteres, etc.).

D. Modelado.

En esta fase de CRISP-DM se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos específico. Las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios: o Ser apropiada para el problema. o Disponer de los datos adecuados. o Cumplir los requisitos del problema. o Tiempo adecuado para obtener un modelo. o Conocimiento de la técnica.

Previamente al modelado de los datos se debe determinar un método de evaluación de los modelos que permita establecer el grado de adecuación de cada uno de ellos. Después de concluir estas tareas genéricas se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo dependen de las

características de los datos y de las características de precisión que se quieran lograr con el modelo. La figura 8 muestra las tareas y las salidas que se obtienen en esta fase, a continuación describimos las tareas principales de esta fase.

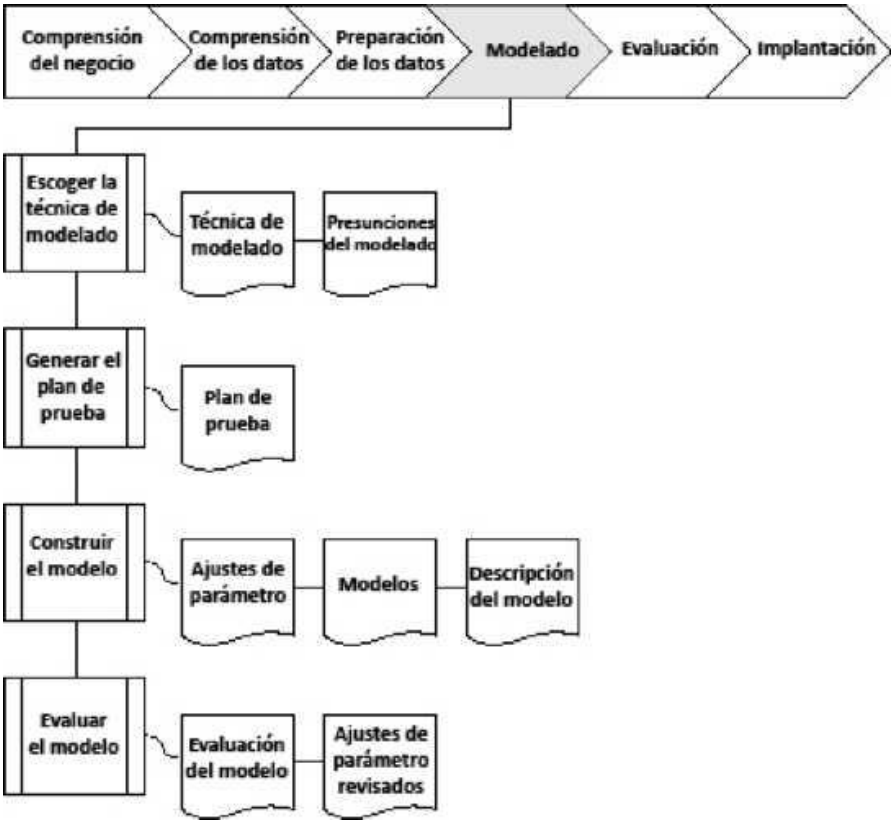


Figura 8. Fase de Modelado

Escoger la técnica de modelado.

Esta tarea consiste en la selección de la técnica de minería de datos más apropiada al tipo de problema que se quiere resolver. Para esta selección, se debe considerar el objetivo principal del proyecto y la relación con las herramientas de minería de datos existentes. Por ejemplo, si el problema es de clasificación, se podrá elegir de entre árboles de decisión, k-nearest neighbors o razonamiento basado en casos (CBR), si el problema

es de predicción, análisis de regresión o redes neuronales, o si el problema es de segmentación, redes neuronales, técnicas de visualización, etc.

Generar el plan de prueba.

Se debe generar un procedimiento destinado a probar la calidad y validez del modelo elegido una vez que éste esté construido. Por ejemplo, en una tarea supervisada de minería de datos como la clasificación, es común usar la razón de error como medida de la calidad. Entonces, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba.

Construir el modelo.

A continuación se ejecuta la técnica seleccionada sobre los datos previamente preparados para generar uno o más modelos. Todas las técnicas de modelado tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.

Evaluar el modelo.

En esta última tarea de esta fase de modelado los ingenieros de DM interpretan los modelos de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos. Expertos en el dominio del

problema juzgan los modelos dentro del contexto del dominio y expertos en minería de datos aplican sus propios criterios (seguridad del conjunto de prueba, pérdida o ganancia de tablas, etc.).

E. Evaluación.

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se pueda haber cometido algún error. Considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados. Si el modelo generado es válido en función de los criterios de éxito establecidos en la fase anterior, se procede a la explotación del modelo. La figura 9 detalla las tareas que componen esta fase y los resultados que se deben obtener. Las tareas involucradas en esta fase del proceso son las siguientes:

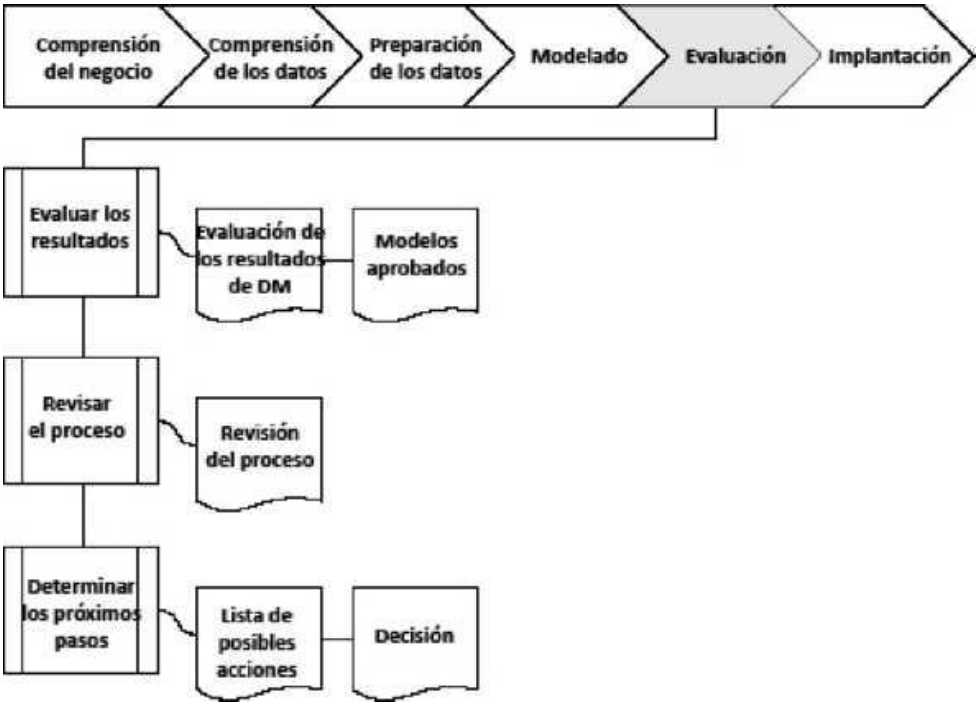


Figura 9. Fase de evaluación

Evaluar los resultados.

En los pasos de evaluación anteriores se trataron factores tales como la exactitud y generalidad del modelo generado. Esta tarea involucra la evaluación del modelo en relación a los objetivos del negocio y busca determinar si hay alguna razón de negocio para la cual el modelo sea deficiente, o si es aconsejable probar el modelo en un problema real si el tiempo y las restricciones lo permiten. Además de los resultados directamente relacionados con el objetivo del proyecto, ¿es aconsejable evaluar el modelo en relación a otros objetivos distintos a los originales?, esto podría revelar información adicional.

Revisar el proceso.

Este proceso se refiere a calificar al proceso entero de minería de datos a objeto de identificar elementos que pudieran ser mejorados.

Determinar los próximos pasos.

Si se ha determinado que las fases hasta este momento han generado resultados satisfactorios podría pasarse a la siguiente fase, en caso contrario podría decidirse por hacer otra iteración desde la fase de preparación de los datos o de modelado con distintos parámetros. Podría incluso darse el caso de que en esta fase se decida empezar desde cero con un nuevo proyecto de minería de datos.

F. Despliegue o implantación.

En esta fase, y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, esto puede hacerse por ejemplo cuando el analista recomienda acciones basadas en la observación del modelo y sus resultados, o por ejemplo aplicando el modelo a diferentes conjuntos de datos o como parte del proceso (en análisis de riesgo de créditos, detección de fraudes, etc.). Generalmente un proyecto de minería de datos no concluye en la implantación del modelo, ya que se deben documentar y presentar los resultados de manera comprensible para el usuario con el objetivo de lograr un incremento del conocimiento. Por otra parte, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados. Las tareas que componen esta fase (figura 10) son:

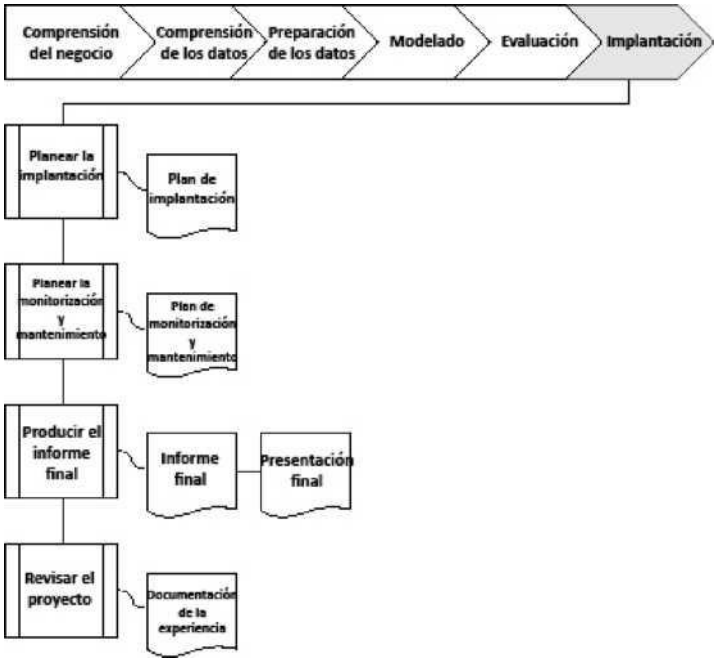


Figura 10. Fase de implantación

Planear la implantación.

Para implementar el resultado de la minería de datos en la organización, esta tarea toma los resultados de la evaluación y concluye una estrategia para su implementación. Si un procedimiento general se ha identificado para crear el modelo, este procedimiento debe ser documentado para su posterior implementación.

Planear la monitorización y mantenimiento.

Si los modelos resultantes del proceso de minería de datos son implementados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitorización y mantenimiento para ser aplicadas sobre los modelos. La retroalimentación generada por la monitorización y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente.

Producir el informe final.

Es la conclusión del proyecto de minería de datos realizado. Dependiendo del plan de implementación, este informe puede ser sólo un resumen de los puntos importantes del proyecto y la experiencia adquirida o puede ser una presentación final que incluya y explique los resultados logrados con el proyecto.

Revisar el proyecto.

En esta tarea se evalúa que cosas se hicieron correctamente y cuales fueron incorrectas, así como aquellos puntos que se podrían mejorar en el proyecto.

El proceso de minería de datos permite obtener conocimiento para que sean explotados con la finalidad de obtener información de por si oculta. Por lo que se realizara por medio de la metodología CRISP DM

3.6. DESARROLLO DE LA METODOLOGÍA

Según la metodología CRISP – DM, se tienen los siguientes procedimientos.

3.6.1. Comprensión del negocio

Según esta metodología es necesario determinar los objetivos de la metodología CRISP -DM

Establecimiento de los objetivos del negocio

El En la metodología CRISP – DM es necesario especificar los objetivos que se buscaran siendo este.

- Determinación de la morbilidad que se tiene en el ámbito de Red Puno, así como el uso de los medicamentos en base a los datos existentes

3.6.2 Objetivos del Data Mining.

- Los objetivos de la minería de datos es el estudio de la morbilidad así como el uso de medicamentos en el ámbito de la ciudad de Puno

3.6.3. Recopilación inicial de datos

Para el proceso de Minería de Datos son los que se encuentran en hojas de Excel de los diferentes servicios que cuentan los Establecimientos de Salud de la Red Puno, ya sea de Farmacia , Triaje, Hospitalización , entre otros.

Las fuentes de información son las Fichas His, Informes de Consumo de Medicamentos por parte de farmacia.

Para lo cual se realizara un proceso de normalización de la información, para luego ser llevado a una Base de Datos. En SQL SERVER 2008.

Para tal fin, se empezó con el proceso de normalización de los datos, en una hoja de Excel, para luego llevarlos a una Base de datos denominada DATABASE_MEDIC, que contienen información de pacientes, Catálogo de Productos, atenciones, diagnósticos, siendo este como sigue.

(ver Anexo Nº 1)

Posteriormente se seleccionar los registros que son utilizados en el en grafico anterior para luego crear una tabla “MORBILIDAD” que serán utilizados en el proceso de Minería de Datos.

Estos datos comprenden información de un 1 año, es decir de los meses de enero a diciembre del 2017.

3.6.4. Descripción de los datos.

Dada la cantidad de campos con que se cuenta, el siguiente paso será determinar las variables mas importantes, que nos permita darle solución al problema plateado.

Descripción de las entidades influyentes en el modelo.

PACIENTES.

Dado que en esta tabla se cuenta con información que podría resultar sensible, como son DNI, nombre, edad, Lugar de Nacimiento, Fecha de Nacimiento se procedido a procedido a Obtener solamente información como Sexo, Edad, profesión.

FACTCATALOGOBIENESINSUMOS,

La cual contiene todos los medicamentos como son Nombre, Presentación, Concentración, Nombre Comercial, Forma Farmacéutica.

Las tablas secundarias son:

ATENCIONES

Que contiene todas las atenciones realizadas por los pacientes durante un periodo de tiempo tales como dia, mes, año, el tipo de enfermedad es decir el código CIE10

FARMMOVIMIENTOVENTAS.

Contiene la información la enfermedad diagnosticada por el Especialista, además de la fecha en que dispenso sus medicamentos.

FARMMOVIMIENTOVENTADETALLE.

Contiene la información sobre que medicamentos que recogió del servicio de Farmacia, además de la fecha.

Diagnostico.

Esta la clasificación de todas las enfermedades que contiene el código CIE10 y la descripción de la enfermedad.

Servicio.

Contiene la información de todos los servicios con los que cuenta el hospital

3.6.5.. Exploración de los datos

En esta sección se realizará gráficos estadísticos que se utilizara en el proceso de Minería de datos

3.6.6. Verificación de calidad de datos

Otro punto importante es el proceso eliminar información innecesaria, inconsistente, redundante o errónea en el diseño de Data Mart, para lo cual se pudo observar que existen información con valores nulos, que puede ser debido a fallas en el Momento de recolección de información por parte de los trabajadores de salud, o incoherencias o omisión en el llenado de la información, para lo cual es importante un mecanismo para el tratamiento de dichos valores.

Eliminación de los valores nulos

Para la eliminación de inconsistencias en el manejo de información se debe de tener en cuanta que todos lo campos tengan valores completos, debido a que a falta de uno valores en los campos distorsionaría la información a ser analizada.

Por lo que se considera como información perdida, por lo que no podrá ser utilizada en el proceso de Minería de datos

3.6.7. Selección de los datos

A partir de ahora solo se consideran los datos que se utilizaran en la Minería de Datos, es decir, todas las variables que se utilizan en la determinación de la morbilidad y del uso de medicamento., por lo que unas de las variables es la morbilidad, esta esta relacionado, con el año, mes, medicamento, sexo del paciente, cantidad por meses,

Ver Anexo Nª 2

3.6.8. Limpieza de datos

En esta sección se determinó algunas fallas para la aplicación de la Minería de datos, por lo que se encontró información incompleta en un total de 20 registros, y algunos valores nulos, por lo que se procedió a eliminar dicha información

3.6.9. Construcción de datos

Para la construcción de determinar la morbilidad en el ámbito de la Red Puno, se procedió a seleccionar información de 1 año, para luego agruparlos meses, por diagnostico, por sexo, medicamentos,

3.6.10. Formateo de datos

Para el formateo de los datos se tuvieron que darle, en cuanto al nombre del medicamento al ser un tipo de dato string o cadena y se procedió a cambiarle por un equivalente como se muestra en la siguiente tabla

Cuadro 1. Tabla de Medicamentos utilizados en el Análisis

idProducto	medicamento	cuenta
med1	DEXAMETASONA FOSFATO 2 mL 4 mg/2mL INY 2 mL 4 mg/2 mL INY	87
med2	DICLOFENACO 3 mL 25 mg/mL INY 3 mL 25 mg/mL INY	23
med3	DEXTROSA 1 L 5 g/100mL (5%) INY 1 L 5 g/100 mL (5 %) INY	51

idProducto	medicamento	cuenta
med4	METAMIZOL SODICO 2 mL 1 g INY 2 mL 1 g INY	109
med5	OMEPRAZOL 20 mg TAB 20 mg CAP_LM	22
med6	PARACETAMOL 10 mL 100 mg/ mL SOL 10 mL 100 mg/mL SOL	40
med7	PARACETAMOL 60 mL 120 mg/5 mL JBE 60 mL 120 mg/5 mL JBE	51
med8	PARACETAMOL 500 mg TAB 500 mg TAB	32
med9	POTASIO CLORURO 10 mL 20 g/100mL INY 10 mL 20 g/100 mL (20 %) INY	45
med10	SALES DE REHIDRATACION ORAL 27.9 g PLV 27.9 g PLV	27
med11	RANITIDINA 2 mL 25 mg/mL INY 2 mL 25 mg/mL INY	33
med12	SODIO CLORURO 0.9 X 1L, INYECTABLE (SINONIMIA: CLORURO DE SODIO 9% X 1L INYECTABLE) 1 L 900 mg/100 mL (0.9 %) INY	94
med13	SODIO CLORURO 20 % X 20 ML INYECTABLE (SINONIMIA: CLORURO DE SODIO 20 % X 20 ML, INYECTABLE) 20 mL 20 g/100 mL (20 %) INY	59
med14	OMEPRAZOL 40 mg INY 40 mg INY	32
med15	CEFTRIAXONA SODICA (Con diluyente) 1 g INY 1 g INY	71

3.6.11. Selección de la técnica de modelado

En esta parte se selecciona la técnica para el modelado,

Para el modelado se utilizara la técnica J48, teniendo en cuenta que el periodo es anual y los 10 diagnósticos mas utilizados en la Red Puno.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1 Evaluación del modelo

De la matriz de confusión se puede determinar

Cuadro 2. Matriz de confusión

Han sido clasificados correctamente de un total de 709 instancias un 65.95.73%,		
No se ha podido clasificar 240 instancias que representan un 32.0427%		
Correctly Classified Instances	509	67.9573 %
Incorrectly Classified Instances	240	32.0427 %
Kappa statistic	0.522	
Mean absolute error	0.0992	
Root mean squared error	0.2227	
Relative absolute error	60.2982 %	
Root relative squared error	77.7448 %	
Total Number of Instances	749	

Además, existe una gran cantidad de pacientes que vienen a los Establecimientos de salud por motivos de fiebre no especificada, que es más

recurrente durante todos los meses del año en hombres y en mujeres durante todo el año 2016 de los diagnósticos, siendo este un patrón de las enfermedades más recurrente del estudio.

De los resultados obtenido en la tabla, se puede interpretar de la siguiente forma.

Hasta el mes de abril (4) es recurrente en personas solteras, y con edad menor o igual a 11 años 13 de 42 personas presentan fiebre no especificada; y si la edad es mayor a 11 años y es menor de 14 años y el mes es antes de marzo 1 persona de 5 presenta dolor no especificado;

Si la edad es mayor de 14 años y menor de 19 presentan otras convulsiones y las no especificadas en un numero de 5 de 19

Y si es mayor de 19 años, presentan 10 de 37 personas fiebre no especificada.

Si el mes es posterior a marzo y la edad es menor o igual a 18 años presentan 9 personas dolor no especificado.

Si el mes el posterior a marzo y es mayor a 18 años presentan 2 personas fiebre no especificada.

Si el estado civil es conviviente.

Y la edad es menor o igual a 24 años y mes en los meses de enero, febrero y marzo 2 personas presentan neumonía no especificada

Si es posterior al mes de marzo, presentan fiebre no especificada

En mayores de 24 años y en el mes de enero se presentan otras convulsiones y las no especificadas

Y en meses posteriores a enero y el sexo es masculino, otras convulsiones y las no especificada

Y en sexo femenino dolor no especificada 1 de 8 personas

Si es casado y es de sexo masculino y tiene seguro dolor no especificada 1 de 15 personas

Y si no tiene seguro fiebre no especificada

Si es casada y es de sexo femenino y tiene menos de 48 años tiene un diagnóstico de faringitis aguda

Y si es mayor de de 48 años fiebre no especificada

En las personas viudas presentan un síntoma de dolor no especificado

Y en personas mayores de 75 y menores iguales de 86 otras convulsiones y las no especificadas

Y en personas mayores a 86 años neumonía no especificada

En meses posteriores a abril y personas solteras con edad menor o igual a 4 años hasta los meses de julio y tienen seguro 46 de 123 personas presentan fiebre no especificada

Y si no tienen seguro en los meses anteriores a junio fiebre no especificada y posteriores a julio fiebre con escalofrió

Si la edad es mayor a 4 años y tiene seguro y menor igual a 5 fiebre no especificada 2 de 8

Y mayor de 5 neumonía no especificada

Y si no tiene seguro fiebre no especificada

En personas mayores a 6 años 6 de 111 presentan náuseas y vómitos

En recién caídos 20 de 33 pacientes presentan faringitis aguda

Y en personas de sexo masculino presentan fiebre no especificada

Y en sexo femenino 8 de 16 personas fiebre con escalofrió

Y en los meses de octubre noviembre y diciembre

Fiebre no especificada

En el mes de diciembre

Los menores de edad presentan fiebre no especificada 2 de 8

Y mayores de 0 años fiebre con escalofrío 6 de 17

Si la edad es mayor a 3 años y antes de diciembre fiebre no especificada

Y en mayores de 11 años

Y la edad es menor e igual a 7 fiebre no especificada

Y la edad mayor a 7 años náuseas y vómitos

Si la edad es mayor a 11 años tiene seguro y menor a 40 37 de 68 presentan fiebre no especificada

Con edad mayor a 59 años y sexo masculino en los meses antes de julio neumonía no especificada

En los meses después de julio faringitis aguda

En sexo femenino 1 de 4 personas otras convulsiones y las no especificadas

Y en mayor de 59 años faringitis aguda

En los meses posteriores a octubre fiebre no especificada

Y mayores de mayor y antes de julio y edad menor o igual a 59 años sexo masculino neumonía no especificada

Y en los meses después de julio faringitis aguda

Las personas que no tiene seguro y menores de 18 4 de 7 personas náuseas y vómitos

Y en mayores de 18 antes de agosto fiebre no especificada

En personas mayores de 7 fiebre con escalofrió

En los convivientes, fiebre no especificada 23 de 44

En personas casadas y mayores de 63 tiene que seguro de sexo masculino fiebre no especificada

En sexo femenino gastritis no especificada

No tienen seguro particular nausea y vomito

En personas mayores de 63 y enantes del mes de julio fiebre no especificada

Y en los meses posteriores de julio y hasta septiembre neumonía no especificada 3 de 10 personas

Y e b los meses posteriores a octubre 10 de 16 personas co fiebre no especificada

En personas viudas presenta un diagnóstico de dolor agudo.-

Cuadro 3. Matriz de confusión de enfermedades

```
=== Confusion Matrix ===

  a   b   c   d   e   f   g   h   i   <-- classified as

 321   0   3   8  12   4   4   0   1 |   a = Fiebre_no_especificada

 13  39   0   0   0   0   0   0   0 |   b = Dolor_no _especificado

 26   2  25   0   0   1   0   0   0 |   c =
Otras_convulsiones_y_las_no_especificadas

 20   0   2  18   8   9   1   0   3 |   d = Náusea_y_vómito

 24   0   0   0  41   7   0   0   0 |   e = Fiebre_con_escalofrío

 29   0   0   4   7  28   0   0   0 |   f = Faringitis_aguda

 18   1   1   1   0   0  17   0   0 |   g = Gastritis_no_especificada

 10   0   0   0   1   0   0   1   0 |   h = Dolor_agudo

 18   0   0   0   0   0   2   0  19 |   i = Neumonía_no_especificada
```

Cuadro 4. Resultado del proceso de Minería de Datos WEKA de medicamentos

```
mes <= 4

|   edad <= 75

|   |   Descripcion = Soltero

|   |   |   edad <= 11: Fiebre_no_especificada (42.0/13.0)

|   |   |   edad > 11

|   |   |   |   mes <= 3

|   |   |   |   |   edad <= 14: Dolor_no _especificado (5.0/1.0)

|   |   |   |   |   edad > 14

|   |   |   |   |   |   edad <= 19: Otras_convulsiones_y_las_no_especificadas
(10.0/5.0)

|   |   |   |   |   |   edad > 19: Fiebre_no_especificada (37.0/10.0)

|   |   |   |   mes > 3

|   |   |   |   |   edad <= 18: Dolor_no _especificado (9.0)

|   |   |   |   |   edad > 18: Fiebre_no_especificada (2.0)

|   |   Descripcion = Conviviente

|   |   |   edad <= 24

|   |   |   |   mes <= 3: Neumonía_no_especificada (2.0)

|   |   |   |   mes > 3: Fiebre_no_especificada (7.0)
```

| | | edad > 24

| | | | mes <= 1: Otras_convulsiones_y_las_no_especificadas (5.0)

| | | | mes > 1

| | | | | sexo = Masculino: Otras_convulsiones_y_las_no_especificadas (2.0)

| | | | | sexo = Femenino: Dolor_no _especificado (8.0/1.0)

| | Descripcion = Casado

| | | sexo = Masculino

| | | | seguro = SIS: Dolor_no _especificado (15.0/1.0)

| | | | seguro = Particular: Fiebre_no_especificada (2.0)

| | | sexo = Femenino

| | | | edad <= 48: Faringitis_aguda (3.0/1.0)

| | | | edad > 48: Fiebre_no_especificada (11.0)

| | Descripcion = Viudo: Dolor_no _especificado (5.0)

| edad > 75

| | edad <= 86: Otras_convulsiones_y_las_no_especificadas (10.0)

| | edad > 86: Neumonía_no_especificada (4.0)

mes > 4

```
|   Descripcion = Soltero
|   |   edad <= 11
|   |   |   mes <= 7
|   |   |   |   edad <= 6
|   |   |   |   |   edad <= 4
|   |   |   |   |   |   seguro = SIS: Fiebre_no_especificada (123.0/46.0)
|   |   |   |   |   |   seguro = Particular
|   |   |   |   |   |   |   mes <= 6: Fiebre_no_especificada (17.0/5.0)
|   |   |   |   |   |   |   mes > 6: Fiebre_con_escalofrío (5.0)
|   |   |   |   |   edad > 4
|   |   |   |   |   |   seguro = SIS
|   |   |   |   |   |   |   edad <= 5: Fiebre_no_especificada (8.0/2.0)
|   |   |   |   |   |   |   edad > 5: Neumonía_no_especificada (5.0/1.0)
|   |   |   |   |   |   seguro = Particular: Fiebre_no_especificada (3.0)
|   |   |   |   edad > 6: Náusea_y_vómito (11.0/6.0)
|   |   |   mes > 7
|   |   |   |   edad <= 3
|   |   |   |   |   mes <= 11
```

```
|  |  |  |  |  |  edad <= 0: Faringitis_aguda (33.0/20.0)
|  |  |  |  |  |  edad > 0
|  |  |  |  |  |  |  mes <= 10
|  |  |  |  |  |  |  |  sexo = Masculino: Fiebre_no_especificada (4.0)
|  |  |  |  |  |  |  |  sexo = Femenino: Fiebre_con_escalofrío (16.0/8.0)
|  |  |  |  |  |  |  mes > 10: Fiebre_no_especificada (19.0/3.0)
|  |  |  |  |  mes > 11
|  |  |  |  |  |  edad <= 0: Fiebre_no_especificada (8.0/2.0)
|  |  |  |  |  |  edad > 0: Fiebre_con_escalofrío (17.0/6.0)
|  |  |  |  edad > 3
|  |  |  |  |  mes <= 11: Fiebre_no_especificada (20.0/3.0)
|  |  |  |  |  mes > 11
|  |  |  |  |  |  edad <= 7: Fiebre_no_especificada (2.0)
|  |  |  |  |  |  edad > 7: Náusea_y_vómito (5.0)
|  |  edad > 11
|  |  |  seguro = SIS
|  |  |  |  edad <= 40: Fiebre_no_especificada (68.0/37.0)
|  |  |  |  edad > 40
```

| | | | | mes <= 10

| | | | | | edad <= 59

| | | | | | | sexo = Masculino

| | | | | | | | mes <= 7: Neumonía_no_especificada (2.0)

| | | | | | | | mes > 7: Faringitis_aguda (2.0)

| | | | | | | sexo = Femenino: Otras_convulsiones_y_las_no_especificadas (4.0/1.0)

| | | | | | edad > 59: Faringitis_aguda (11.0)

| | | | | mes > 10: Fiebre_no_especificada (6.0/1.0)

| | | seguro = Particular

| | | | edad <= 18: Náusea_y_vómito (7.0/4.0)

| | | | edad > 18

| | | | | mes <= 7: Fiebre_no_especificada (7.0/3.0)

| | | | | mes > 7: Fiebre_con_escalofrío (31.0/14.0)

| Descripcion = Conviviente: Fiebre_no_especificada (44.0/23.0)

| Descripcion = Casado

| | edad <= 63

| | | seguro = SIS

| | | | sexo = Masculino: Fiebre_no_especificada (8.0)

| | | | sexo = Femenino: Gastritis_no_especificada (24.0/7.0)

| | | seguro = Particular: Náusea_y_vómito (8.0/3.0)

| | edad > 63

| | | mes <= 7: Fiebre_no_especificada (25.0)

| | | mes > 7

| | | | mes <= 9: Neumonía_no_especificada (10.0/3.0)

| | | | mes > 9: Fiebre_no_especificada (16.0/10.0)

| Descripcion = Viudo: Dolor_agudo (1.0)

Cuadro 5. Tabla de Medicamentos utilizados en el Análisis

idProducto	medicamento	Cuenta
med1	DEXAMETASONA FOSFATO 2 mL 4 mg/2mL INY 2 mL 4 mg/2 mL INY	87
med2	DICLOFENACO 3 mL 25 mg/mL INY 3 mL 25 mg/mL INY	23
med3	DEXTROSA 1 L 5 g/100mL (5%) INY 1 L 5 g/100 mL (5 %) INY	51
med4	METAMIZOL SODICO 2 mL 1 g INY 2 mL 1 g INY	109
med5	OMEPRAZOL 20 mg TAB 20 mg CAP_LM	22

idProducto	medicamento	Cuenta
med6	PARACETAMOL 10 mL 100 mg/ mL SOL 10 mL 100 mg/mL SOL	40
med7	PARACETAMOL 60 mL 120 mg/5 mL JBE 60 mL 120 mg/5 mL JBE	51
med8	PARACETAMOL 500 mg TAB 500 mg TAB	32
med9	POTASIO CLORURO 10 mL 20 g/100mL INY 10 mL 20 g/100 mL (20 %) INY	45
med10	SALES DE REHIDRATACION ORAL 27.9 g PLV 27.9 g PLV	27
med11	RANITIDINA 2 mL 25 mg/mL INY 2 mL 25 mg/mL INY	33
med12	SODIO CLORURO 0.9 X 1L, INYECTABLE (SINONIMIA: CLORURO DE SODIO 9% X 1L INYECTABLE) 1 L 900 mg/100 mL (0.9 %) INY	94

Además de la tabla de medicamentos, se puede determinar que no es posible clasificar dado que

En relación al tipo de enfermedades se puede determinar Fiebre_no_especificada: se utiliza DEXAMETASONA FOSFATO 2 mL 4 mg/2mL INY 2 mL 4 mg/2 mL INY

(353.0/284.0)

Para el diagnostico Dolor_no _especificado en sexo = Masculino:

DEXTROSA 1 L 5 g/100mL (5%) INY 1 L 5 g/100 mL (5 %)INY

(34.0/27.0)

En el sexo = Femenino en los meses de enero y febrero DEXAMETASONA FOSFATO 2 mL 4 mg/2mL INY 2 mL 4 mg/2 mL INY (7.0/4.0)

| | | mes > 1: SODIO CLORURO 0.9 X 1L, INYECTABLE (SINONIMIA: CLORURO DE SODIO 9% X 1L INYECTABLE) 1 L 900 mg/100 mL (0.9 %) INY (2.0)

| | mes > 2: POTASIO CLORURO 10 mL 20 g/100mL INY 10 mL 20 g/100 mL (20 %) INY (9.0/4.0)

enfermedad = Otras_convulsiones_y_las_no_especificadas

| seguro = SIS

| | sexo = Masculino: POTASIO CLORURO 10 mL 20 g/100mL INY 10 mL 20 g/100 mL (20 %) INY (19.0/15.0)

| | sexo = Femenino: SODIO CLORURO 0.9 X 1L, INYECTABLE (SINONIMIA: CLORURO DE SODIO 9% X 1L INYECTABLE) 1 L 900 mg/100 mL (0.9 %) INY

(30.0/24.0)

| seguro = Particular

| | mes <= 2: POTASIO CLORURO 10 mL 20 g/100mL INY 10 mL 20 g/100 mL (20 %) INY (3.0/1.0)

| | mes > 2: DEXAMETASONA FOSFATO 2 mL 4 mg/2mL INY 2 mL 4 mg/2 mL INY (2.0/1.0)

enfermedad = Náusea_y_vómito

| seguro = SIS

| | sexo = Masculino: RANITIDINA 2 mL 25 mg/mL INY 2 mL 25 mg/mL INY (20.0/15.0)

| | sexo = Femenino

| | | mes <= 9: PARACETAMOL 500 mg TAB 500 mg TAB (18.0/14.0)

| | | mes > 9: OMEPRAZOL 20 mg TAB 20 mg CAP_LM (3.0/2.0)

| seguro = Particular

| | mes <= 5

| | | mes <= 3: POTASIO CLORURO 10 mL 20 g/100mL INY 10 mL 20 g/100 mL (20 %) INY (2.0/1.0)

| | | mes > 3: DEXAMETASONA FOSFATO 2 mL 4 mg/2mL INY 2 mL 4 mg/2 mL INY (4.0/2.0)

| | mes > 5: DICLOFENACO 3 mL 25 mg/mL INY 3 mL 25 mg/mL INY (14.0/9.0)

enfermedad = Fiebre_con_escalofrío

| mes <= 6

| | mes <= 5: PARACETAMOL 60 mL 120 mg/5 mL JBE 60 mL 120 mg/5 mL JBE

(4.0/2.0)

| | mes > 5: DEXAMETASONA FOSFATO 2 mL 4 mg/2mL INY 2 mL 4 mg/2 mL INY (4.0/2.0)

| mes > 6

| | seguro = SIS: OMEPRAZOL 20 mg TAB 20 mg CAP_LM (39.0/30.0)

| | seguro = Particular: DEXAMETASONA FOSFATO 2 mL 4 mg/2mL INY 2 mL 4 mg/2 mL INY (25.0/19.0)

enfermedad = Faringitis_aguda: DEXAMETASONA FOSFATO 2 mL 4 mg/2mL INY 2 mL 4 mg/2 mL INY (68.0/47.0)

enfermedad = Gastritis_no_especificada

| mes <= 5: POTASIO CLORURO 10 mL 20 g/100mL INY 10 mL 20 g/100 mL (20 %) INY (10.0/7.0)

| mes > 5: DICLOFENACO 3 mL 25 mg/mL INY 3 mL 25 mg/mL INY (28.0/21.0)

enfermedad = Dolor_agudo

| mes <= 10: POTASIO CLORURO 10 mL 20 g/100mL INY 10 mL 20 g/100 mL (20 %) INY (4.0/3.0)

| mes > 10: PARACETAMOL 60 mL 120 mg/5 mL JBE 60 mL 120 mg/5 mL JBE (8.0/5.0)

enfermedad = Neumonía_no_especificada

| mes <= 7: DEXAMETASONA FOSFATO 2 mL 4 mg/2mL INY 2 mL 4 mg/2 mL INY (25.0/20.0)

| mes > 7

| | mes <= 8: DICLOFENACO 3 mL 25 mg/mL INY 3 mL 25 mg/mL INY (5.0/3.0)

| | mes > 8: PARACETAMOL 500 mg TAB 500 mg TAB (9.0/7.0)

Ver Anexo Nª 3

Ver Anexo Nº 4

CONCLUSIONES

- Mediante el uso de los arboles de decisión se llego a determinar que los patrones de consumo están dadas por: las enfermedades mas comunes y la medicinas que están requieran para su tratamiento. De esta manera se pudo determinar los pacientes con mayor afluencia al centro de salud tienen los síntomas de fiebre con escalofríos, neumonía, dolor agudo. Así mismo las medicinas que mayor demanda se tiene es: Paracetamol, Metamizol Sódico, Dexametasona Fosfato y Sodio Cloruro 0.9 X 1L, Inyectable. Así mismo permite comprender el comportamiento del consumo de medicamentos, de acuerdo a las enfermedades mas recurrentes.
- El modelo CRISP – DM es adecuado para el uso de minería de datos y sumamente importante en el la realización del proceso de extracción,

normalización, limpieza y carga de datos existentes, como también para eliminar información innecesaria, inconsistente, redundante o errónea en el diseño del Data Mart.

- Se demostró que la implementación de un Data Mart es muy útil para el proceso de minería de datos, de igual manera que el proceso KDD permite la descripción de la información que se obtuvo de la oficina de Informática respecto a la demanda de medicamentos con relación a los síntomas del área de Farmacia, Admisión y Triaje.
- La aplicación de WEKA y el algoritmo J48 específicamente fue ideal para el estudio de la obtención del conocimiento y así determinar las medicinas mas demandadas y las enfermedades mas comunes por las que las personas acuden al establecimiento de salud. Redes Puno.

RECOMENDACIONES

- Se recomienda el uso de minería de datos con Cubos OLAP, de esta manera comparar resultados y poder determinar ventajas frente a WEKA.
- A futuro la oficina de informática tenga un datamart como política y como necesidad ya que esto permitirá una toma de decisiones de manera eficaz y rápida.
- Se recomienda para futuras investigaciones tomar en cuenta otras áreas como son: hospitalización, logística y adquisiciones entren otros.
- Se recomienda una reingeniería en la base de datos de la RED PUNO, de tal manera que permita y facilite la aplicación de minería de datos mas exhaustivamente.

BIBLIOGRAFÍA

Ale, J., (2005). Análisis de Clusters.

Charaja Cutipa, Francisco. (2009). *El MAPIC en la Metodología de la Investigación*. Peru

Chen, H., W. Chung, J. Xu, G. Wang, Y. Qin, M. Chau. (2004*). Crime Data Mining: A General Framework and Some Examples. IEEE Computer Society, vol. 37, no. 4*. Páginas 50-56.

Chen, M., J. Han, (1996). *Data mining: An overview from database perspective. IEEE Transactions on Knowledge and Data* Eng.

Evangelos, S., J. Han, (1996). *Proceeding of the Second International Conference on Knowledge Discovery and Data Mining*. Portland, EEUU.

Fisher, D. (1996). *Iterative optimization and simplification of hierarchical clusterings. Departament of Computer Science. Vanderbilt* University, Nashville, EEUU.

Hand, D. J., (1997). *Data Mining: Statistics and More?. The American Statistician*.

Ibermática. (2007) *Business Intelligence, el conocimiento compartido*. Disponible en:

http://www.ibermatica.com/ibermatica/publicaciones/BusinessIntelligence.pdf

Jain, A. K., R. C. Dubes, (1988). *Algorithms for Clustering Data. Prentice Hall*.

Kantardzic, M. (2002). *Data Mining: Concepts, models, methods and algorithms. Wiley- IEEE Press*. ISBN 0-471-22852-4.

Kaplan R, Norton D.(1996) *The Balanced Scorecard, Boston, MA: Harvard. Business School Press*.

Kohonen, T. (1995*). Self-Organizing Maps. Springer-Verlag.*

Mannila, H. (1997). *Methods and problems in data mining. In Proc. of International Conference on Database Theory, Delphi, Greece*.

Michalski R., A. Baskin, K. Spackman. (1982). *A Logic-Based Approach to Conceptual*

Morales, E., (2003). *Descubrimiento de Conocimiento en Bases de Datos*.

Servente, M., R. García-Martínez, (2002). *Tesis Doctoral Algoritmos TDIDT aplicados a la minería de datos inteligente*. Universidad de Buenos Aires, Argentina.

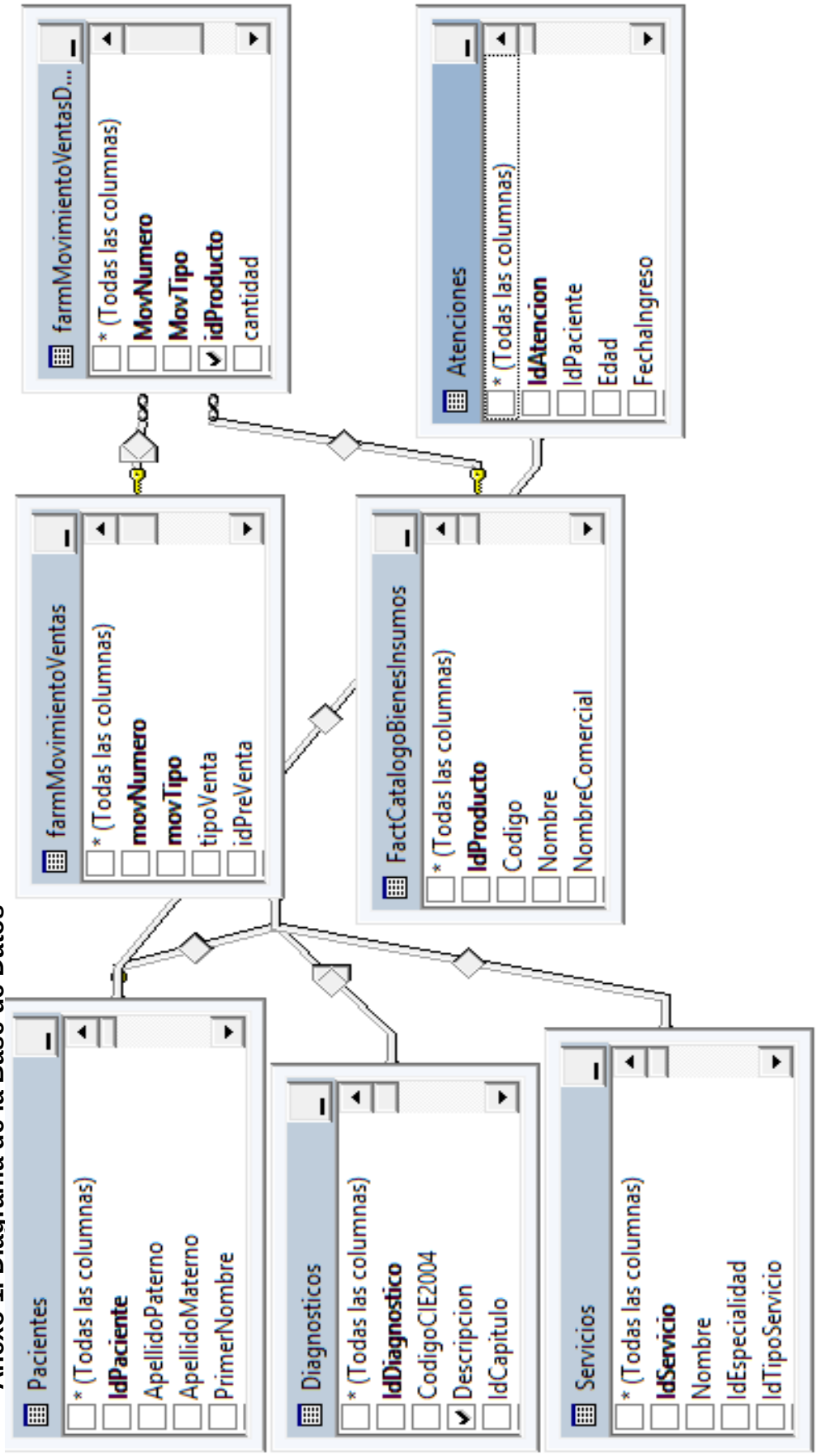
Suarez J.C, Gomez A.(2003) *Sistemas de Información Herramientas Prácticas para la Gestión Empresarial*, Ra-Ma. Madrid.

Vesanto J., E. Alhoniemi, (2000). *Clustering of the Self-Organizing Map. IEEE transactions on neural networks,* Vol 11, No. 3.

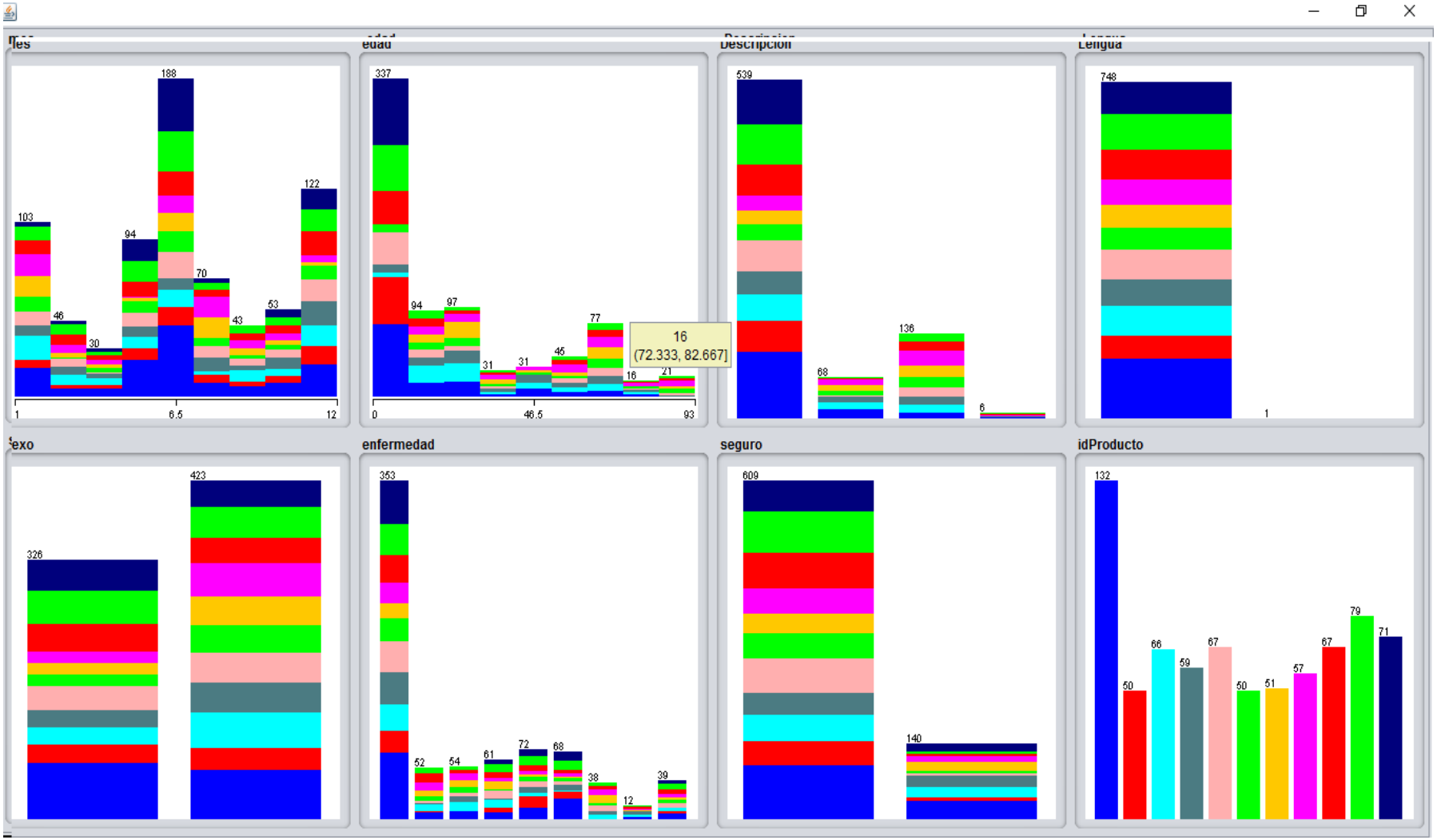
Vitt E, Luckevich M, Misner S. (2002) *Busines*.

ANEXOS:

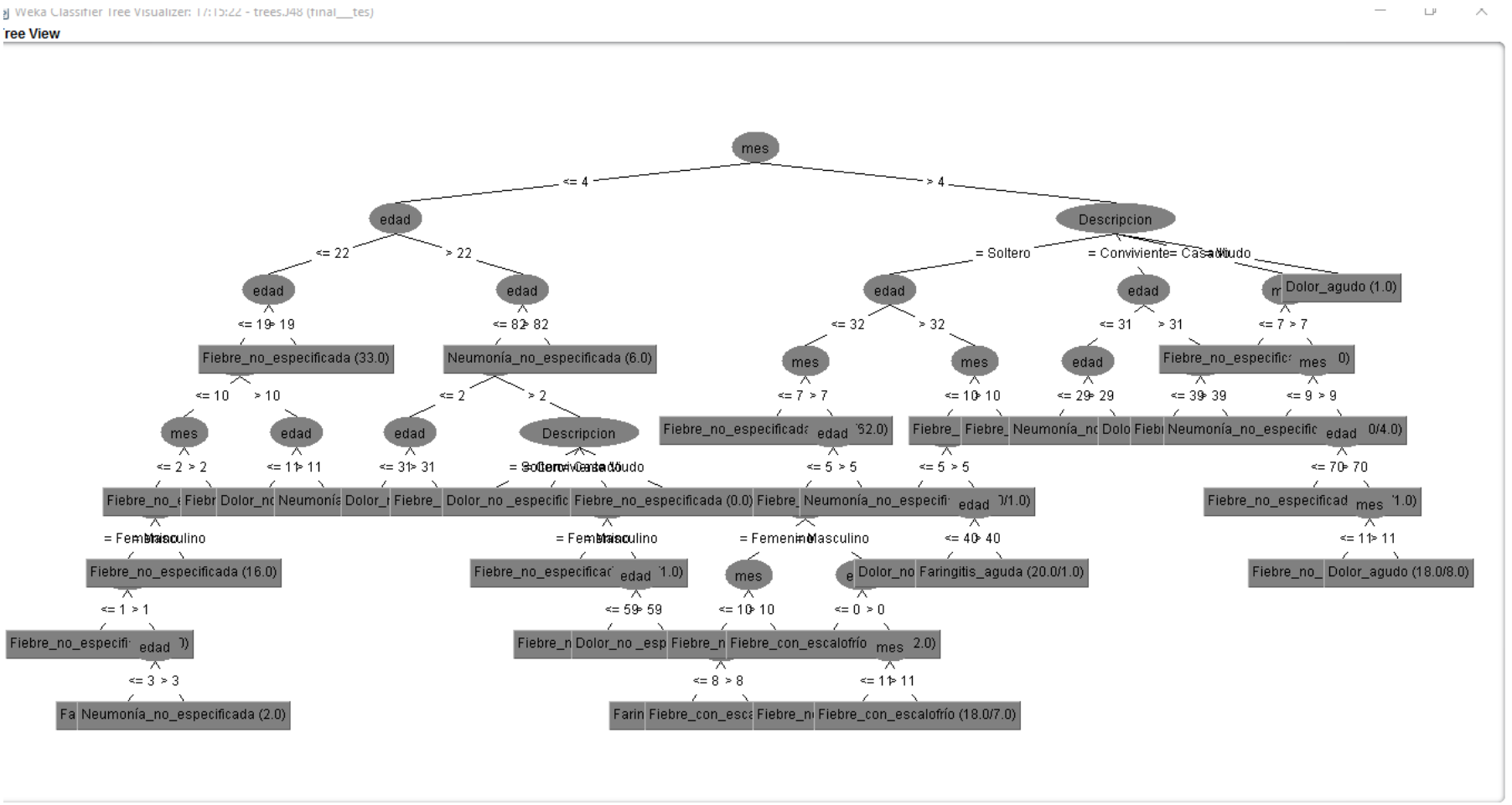
Anexo 1. Diagrama de la Base de Datos



Anexo 2. Estadística de los Datos a Utilizar



Anexo 3. Árbol J48 de enfermedades



Anexo 4. Árbol J48 de medicamentos