FACULTAD DE POSTGRADO

TESIS DE POSTGRADO
DETECCIÓN DE ANOMALÍAS EN EL RPI APLICANDO MINERÍA DE DATOS

SUSTENTADO POR:

KATHLEEN SASKIA ZAVALA VARELA

MANUEL SALVADOR GARCÍA LACAYO

PREVIA INVESTIDURA AL TÍTULO DE

MÁSTER EN GESTIÓN DE TECNOLOGÍAS DE INFORMACIÓN

TEGUCIGALPA, F.M., HONDURAS, C.A.

ABRIL, 2013

UNIVERSIDAD TECNOLÓGICA CENTROAMERICANA

UNITEC

FACULTAD DE POSTGRADO

AUTORIDADES UNIVERSITARIAS

RECTOR

LUIS ORLANDO ZELAYA MEDRANO

SECRETARIO GENERAL

JOSÉ LÉSTER LÓPEZ

VICERRECTOR ACADÉMICO

MARLON ANTONIO BREVÉ REYES

DECANO DE LA FACULTAD DE POSTGRADO

JEFFREY LANSDALE

DETECCIÓN DE ANOMALÍAS EN EL RPI APLICANDO MINERÍA DE DATOS

TRABAJO PRESENTADO EN CUMPLIMIENTO DE LOS REQUISITOS EXIGIDOS PARA OPTAR AL TÍTULO DE MÁSTER EN
GESTIÓN DE TECNOLOGÍAS DE INFORMACIÓN

ASESOR METODOLÓGICO
JUAN MARTÍN HERNÁNDEZ

ASESOR TEMÁTICO
FRANCISCO ARMANDO PÉREZ ORDOÑEZ

MIEMBROS DE LA TERNA:
CINTHIA CANO
CLAUDIO ARCHILA
JUAN SOLANO

DEDICATORIA

A mis padres, quienes son mi ejemplo de superación y entrega. A ellos que con su apoyo constante, preocupación y paciencia han estado siempre presentes y han sido un incentivo para seguir adelante y culminar otra etapa en mi vida.

A mis hermanos quienes han sido parte fundamental de mi superación personal.

Kathleen Saskia Zavala Varela

DEDICATORIA

A Dios que me ha regalado la vida, que día a día me manifiesta su infinita misericordia y me permite descubrir que con perseverancia todo es posible.

A mi esposa y mis hijas que son mi mayor motivación para seguir superándome en la vida.

A mis padres que han estado siempre presentes apoyándome, aconsejándome y velando por mi bienestar.

Manuel García Lacayo

AGRADECIMIENTO

En primer lugar agradezco a Dios quien siempre estuvo a mi lado y me dio fortaleza y salud para alcanzar esta meta.

A mis padres, quienes permanentemente me alentaron y a mi familia en general por su apoyo incondicional, lo que me ha permitido culminar con éxito este proyecto.

A los docentes y a nuestros asesores de tesis, cuyas aportaciones y orientaciones fueron indispensables para el desarrollo de esta investigación.

Al Programa de Administración de Tierras de Honduras (PATH), y a su Coordinador Nacional por brindarnos la oportunidad de realizar esta investigación. Asimismo al personal del Área de Infotecnología, por proporcionarnos las facilidades y colaboración necesarias para completar este estudio.

A todas aquellas personas que han contribuido de manera directa o indirecta a la realización de esta tesis.

Kathleen Saskia Zavala Varela

AGRADECIMIENTO

A Dios que ha estado conmigo en cada paso que doy, guiándome, cuidándome y dándome la fe, la fortaleza y la salud para terminar esta investigación.

A mi esposa por su paciencia, comprensión y apoyo incondicional en todo momento y en especial durante el desarrollo de mis estudios de maestría.

A las autoridades del Programa de Administración de Tierras de Honduras (PATH) por permitirnos realizar esta investigación.

Al personal Técnico Programa de Administración de Tierras de Honduras (PATH) por habernos proporcionado las facilidades y colaboración necesarias para completar esta investigación.

A nuestros asesores de tesis por su ayuda y orientación durante la realización de la investigación y cuyos comentarios y sugerencias nos permitieron mejorar la presente tesis.

Manuel García Lacayo

FACULTAD DE POSTGRADO

DETECCIÓN DE ANOMALÍAS EN EL RPI APLICANDO MINERÍA DE DATOS

AUTOR:

Kathleen Saskia Zavala Varela

Manuel Salvador García Lacayo

RESUMEN

Con esta investigación se buscó formalizar el descubrimiento de patrones anómalos en las transacciones del Registro de la Propiedad Inmueble (RPI) que se encuentran almacenadas en el Sistema Unificado de Registros (SURE), basado en los métodos y técnicas de minería de datos. Se realizó la recopilación de las transacciones y casos que han motivado quejas o denuncias y de esta manera se definió un catálogo de esquemas de anomalías. Posteriormente se realizo una prueba de funcionalidad del modelo utilizando un software gratuito a través del cual se aplicaron diversos algoritmos para detectar anomalías en conjuntos de datos de prueba, demostrando que los métodos de minería de datos pueden identificar registros maliciosos. Se propuso la implementación de un modelo de detección de anomalías, que pueda ser integrado a futuro en el SURE, el cual actúe como un mecanismo de alarma sobre registros irregulares que requieran de un mayor análisis por parte de las autoridades.

Palabras Claves: Minería de Datos, Descubrimiento de Conocimiento en Base de Datos, Inteligencia de Negocios, Detección de Anomalías, Sistemas Registrales

GRADUATE SCHOOL

ANOMALY DETECTION IN RPI APPLYING DATA MINING

AUTHOR:

Kathleen Saskia Zavala Varela

Manuel Salvador García Lacayo

ABSTRACT

This study aimed to formalize the discovery of anomalous patterns in the Registro de la Propiedad Inmueble (RPI) transactions which are stored in the Sistema Unificado de Registros (SURE), based on data mining the methods and techniques. We performed a compilation of transactions and cases that have motivated complaints and reports and thus defined a catalog of patterns of anomalies. Subsequently a functional test of the model using an open source software through which various algorithms were applied to detect anomalies in the test datasets, showing that data mining methods can identify malicious records. We proposed the implementation of an anomaly detection model that can be integrated to SURE in the future, which act as an alarm mechanism on irregular records that will require further analysis by authorities.

Keywords: Data Mining, Knowledge Discovery in Database, Business Intelligence

ÍNDICE

CAPÍTULO I. PLANTEAMIENTO DE LA INVESTIGACIÓN

1.1 INTRODUCCIÓN

La seguridad jurídica es fundamentalmente la garantía que se otorga a los individuos sobre el hecho de que su persona, sus derechos y bienes no serán violentados, por ende esta representa una obligación del Estado, ya que este tiene el papel de regular las relaciones en la sociedad y es quien crea las disposiciones legales a seguir.

Esta protección o seguridad jurídica aplica sobre los bienes inmuebles (terrenos urbanizados o no, etc.) y la misma ha sido una preocupación histórica para los gobiernos, porque son de importancia económica, ya que permiten la generación de ingresos, la obtención de crédito y pueden ser sujeto de impuestos.

Es así que el Registro de la Propiedad se vuelve relevante, al ser el medio por el cual se provee seguridad jurídica, ya que permite la publicación de los derechos sobre los bienes inmuebles.

El Registro de la Propiedad al igual que cualquier organización en la actualidad, hace uso de sistemas de información, para la automatización de sus procesos y un mejor desarrollo de sus actividades diarias. Asimismo busca proporcionar integridad, confidencialidad y alta disponibilidad de la información que se gestiona dentro de su sistema de información.

En ese sentido, el objetivo del presente estudio es contribuir a reforzar la seguridad de la información del sistema de propiedad y tenencia de la tierra del país, a través de la creación un modelo de detección de anomalías (registros inusuales) basado en los fundamentos y técnicas de la minería de datos, que pueda ser incorporado a futuro en el Sistema Unificado de Registros (SURE), como sistema de alerta y monitoreo de la información registral perteneciente al Registro de la Propiedad Inmueble (RPI).

En el capítulo I de este estudio se presenta el planteamiento de la investigación y sus objetivos, fundamentando su origen, definiendo claramente el problema del que se ocupa así como cuales son las interrogantes que se dieron respuesta, justificando su validez e importancia.

En el capítulo II se presenta el marco teórico, que es la teoría que fundamenta los conceptos que se han manejado durante la investigación y que han servido para comprender el proceso y técnicas registrales, el rol del sistema SURE y la metodología y las técnicas de minería de datos en la identificación de anomalías.

En el capítulo III se presenta la metodología, que describe la naturaleza y el diseño de la investigación así como las técnicas e instrumentos utilizados para recolectar datos.

En el capítulo IV comprende un resumen de la información obtenida durante la investigación donde se describen los hallazgos más relevantes que son fundamentados mediante el análisis de los investigadores.

En el capítulo V se presentan las conclusiones sobre los resultados obtenidos en el capitulo anterior orientadas a dar respuestas a la problemática que origino la investigación. Igualmente incluye las recomendaciones que son sugerencias y son parte de los aportes de los investigadores.

En el capítulo VI se presenta la aplicabilidad que incluye una guía para la implementación del modelo de detección de anomalías.

En general el presente aporte busca demostrar la aplicabilidad de la minería de datos como fuente de conocimiento, que genera apoyo en la investigación de actos irregulares o inusuales que se presente en el SURE y de esta manera fortalecer la seguridad del mismo y por ende otorgar mayor seguridad jurídica a los poseedores de bienes inmuebles.

1.2 ANTECEDENTES

La promulgación de la Ley de la Propiedad en el año 2004 permitió crear un marco legal e institucional para el ordenamiento de la tenencia de la tierra de forma integrada. De esta manera se creó el Instituto de la Propiedad (IP) para fiscalizar el catastro nacional y el registro de bienes inmuebles, empleando para ello el Sistema Nacional de Administración de la Propiedad (SINAP), el cual es la plataforma tecnológica que contempla módulos para el Sistema Unificado de Registros (SURE), el Registro Nacional de Normativas de Ordenamiento Territorial (RENOT) y el Sistema Nacional de Información Territorial (SINIT).

Aunque se han obtenido varios logros aun hay camino por recorrer ya que de acuerdo al Banco Mundial (2011), Honduras aun enfrenta grandes retos para poder consolidar la gobernanza de la propiedad en el país: el marco legal y político requiere de mayor coherencia; las agencias claves necesitan mayor fortalecimiento institucional y coordinación e independencia de las turbulencias políticas; …. mejorar los mecanismos alternativos de resolución de conflictos y promover la cultura de registro de bienes inmuebles.

Esto motivó al Gobierno de la República a firmar un convenio para la ejecución de la segunda fase del Programa de Administración de Tierras de Honduras (PATH); cuyo objetivo consiste en beneficiar a la población del área de influencia del proyecto con servicios descentralizados y mejorados de administración de tierras, entre ellos mayor acceso e información más completa y fidedigna de los registros y transacciones de propiedades (Congreso Nacional, 2011).

Dentro de las opciones y propuestas de acción para las mejoras en la seguridad y calidad del SURE, Proenza (2007) propone que se deberán fortalecer los controles de seguridad del SURE para garantizar un servicio electrónico que goce de la confianza de los usuarios y la Corte. En ese sentido el PATH ha sido esencial para desarrollar el

SINAP y sus subsistemas, así como también para formular y ejecutar las reformas legales e institucionales.

Entre los esfuerzos recientes del PATH por mejorar la seguridad y control de las transacciones en el SURE, se pueden mencionar:

- Sistema de Seguridad y Auditoria SURE: Tiene por objetivo administrar las actividades de seguridad y auditoría de las aplicaciones o módulos que integran el SINAP. (PATH , 2006)
- Sistema de quejas y denuncias: Es un sistema para recibir, rastrear y monitorear quejas y denuncias. Los usuarios tienen la opción de recibir notificación vía convencional o electrónica de los resultados de las quejas o gestiones interpuestas. (Banco Mundial, 2011)
- Creación de roles y reporte de seguimiento: Esto con el fin de apoyar las funciones de los supervisores del IP, quienes realizan el seguimiento a las transacciones, de manera que se pueda optimizar los tiempos de entrega de los documentos.
- Reportes a demanda: Envío de reportes electrónicos o impresos a solicitud de las autoridades que contienen información específica de un periodo.

Estos esfuerzos en general se basan en la fiscalización y reconstrucción de la ocurrencia de los hechos, pero también se debe tomar en cuenta que existen métodos preventivos para satisfacer la necesidad de seguridad de los derechos de las personas que poseen bienes inmuebles.

En esta misma línea hemos encontrado que en Costa Rica, Jiménez Calderón y Mc Hugh en 2011 desarrollaron un estudio y análisis de varios mecanismos tecnológicos de protección de la información registral, enfocándose en la implementación de un mecanismo virtual.

Por su parte la Oficina de Auditoría General de Alberta, Canadá, realizó un análisis del sistema de registro de tierras con el fin de formular recomendaciones para hacer frente al creciente fraude hipotecario en dicha provincia. Su hallazgo fue el siguiente: El Departamento puede mejorar sus actividades de detección de fraudes y disuasión mediante el uso de técnicas de minería de datos. Se identificaron patrones específicos en los datos del sistema de títulos de tierra que concluimos podrían indicar una actividad potencial fraude hipotecario, los cuales requieren seguimiento e investigación por la Unidad Especial de Investigaciones del Departamento (Auditor General of Alberta, 2010).

1.3 DEFINICIÓN DEL PROBLEMA

1.3.1 ENUNCIADO DEL PROBLEMA

El SURE permite al IP automatizar el proceso de trámite de presentación de escrituras públicas, que involucra la recepción, revisión, calificación, inscripción y vinculación a información catastral donde esta exista. Asimismo el SURE ofrece un servicio de consultas públicas gratuitas en su sitio web, sobre el trámite de presentaciones de escrituras públicas, imágenes digitalizadas de las presentaciones y planos topográficos de zonas que han sido catastradas.

De acuerdo a datos proporcionados por el Área de Infotecnología del PATH, en el año 2012 se recibieron y registraron el en SURE 41,570 presentaciones de escrituras públicas pertenecientes a las oficinas de Francisco Morazán del IP. El trámite individual de cada presentación requiere de un tiempo promedio de 10 a 12 días en el caso de una compra-venta de un inmueble, pero este tiempo varía según el acto expresado en la escritura pública.

Es durante este proceso de trámite de presentación de escrituras públicas, donde se presenta la posibilidad de errores humanos, omisiones e incluso irregularidades, que

causan inconsistencias en los datos registrados y por lo tanto puede afectar los derechos sobre los inmuebles de las personas que hacen uso de los servicios del IP.

El SURE ofrece al IP reportes financieros que reflejan cifras, rubros y clasificaciones; reportes operativos donde se resume la carga diaria de trabajo y reportes de transacciones históricas donde se detalla los usuarios que han creado y modificado transacciones en el sistema según criterios específicos. Pero cuando se requiere realizar una investigación, el personal encargado recurre al PATH solicitando informes de apoyo a investigaciones, los cuales son procesados posteriormente por los solicitantes. Estos informes así como una serie de reportes a demanda se entregan en medios digitales a las autoridades del IP que así los soliciten.

Los análisis de estos reportes se realizan mediante herramientas estadísticas descriptivas básicas, como el Excel. Sin embargo, muchas veces estos análisis no reflejan el problema real, debido a los grandes volúmenes de datos y puesto que se desconoce que buscar y donde buscar, por lo tanto se requiere un tratamiento más complejo para los datos.

1.3.2 PLANTEAMIENTO DEL PROBLEMA

El IP carece de un mecanismo para extraer conocimiento de las grandes cantidades de datos almacenados en el subsistema SURE, el cual identifique registros de datos inusuales, que de ser relevantes requieran mayor investigación por parte de los funcionarios de la institución.

1.3.3 PREGUNTAS DE INVESTIGACIÓN

- ¿Cómo definir esquemas de anomalías en las transacciones de bienes inmuebles dentro del RPI?

- ¿Cómo determinar criterios de aplicación de las técnicas y algoritmos de minería de datos?

- ¿Qué técnica de minería de datos es apropiada para determinar patrones de comportamiento transaccional irregular o anómalo en la base de datos del SURE?

- ¿Cuáles son los elementos necesarios para implementar una herramienta que permita identificar transacciones inusuales o patrones de comportamiento anómalos?

1.4 OBJETIVOS DEL PROYECTO

1.4.1 OBJETIVO GENERAL

Contribuir a reforzar la seguridad de la información del sistema automatizado del IP, a través de la creación un modelo de detección de anomalías (registros inusuales) basado en los fundamentos y técnicas de la minería de datos, que pueda ser incorporado a futuro en el SURE, como sistema de alerta y monitoreo de la información registral.

1.4.2 OBJETIVOS ESPECÍFICOS

- Identificar dentro del contexto del subsistema SURE los escenarios que han motivado quejas o denuncias por irregularidades en el proceso registral.

- Analizar un conjunto de datos históricos del sistema para determinar los criterios de aplicación de las técnicas de minería de datos.

- Determinar la técnica de minería de datos y el software a utilizar en el diseño del modelo de detección de anomalías.

- Proponer una guía para la implementación del modelo de identificación de anomalías para el RPI.

1.5 HIPÓTESIS Y/O VARIABLES DE ESTUDIO

Las variables o características que fueron sujetas a estudio durante la investigación se muestran a continuación:

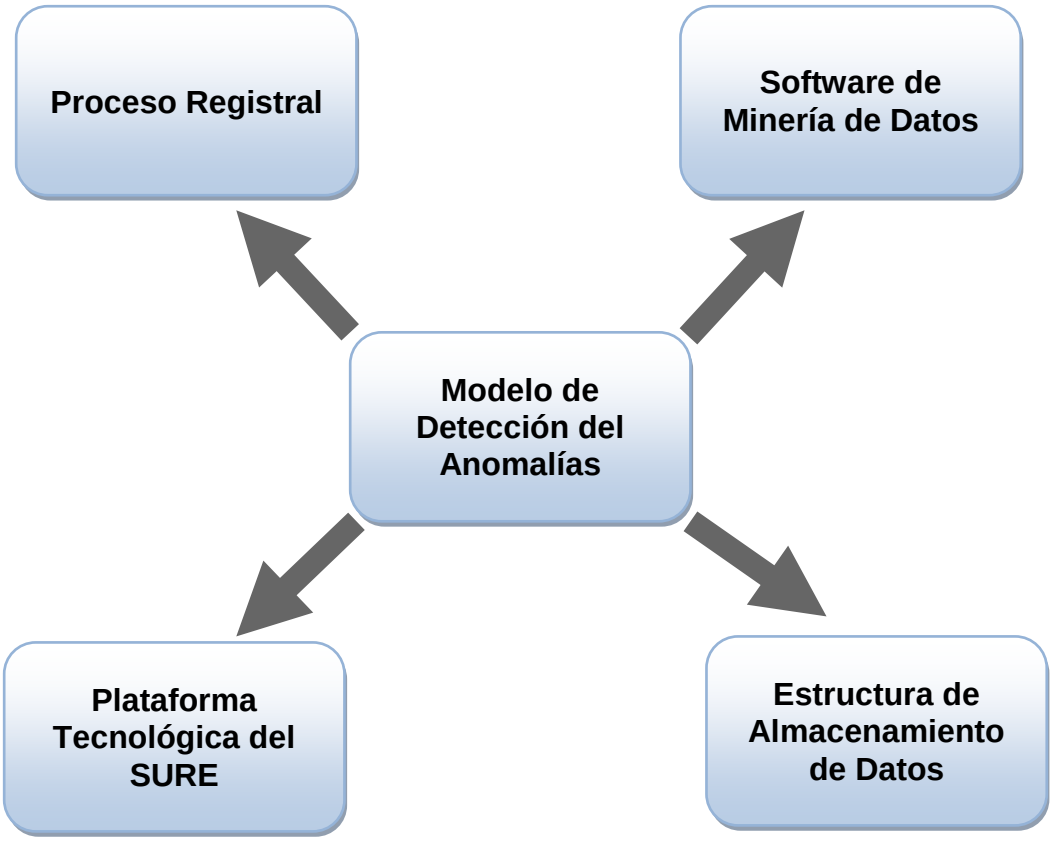


Figura 1. Variables de Estudio

V1 Proceso Registral: Se refiere al conjunto de actuaciones realizadas desde el momento en que se presenta la solicitud de registro de un título hasta el momento en que el registrador resuelve esta petición, suspendiendo, denegando o practicando el asiento solicitado.

V2 Estructuras de Almacenamiento de Datos: Se refiere al dominio y tipo de datos de cada atributo disponibles dentro de las bases de datos.

V3 Software Base de Minería de Datos: Se refiere a las herramientas de software que pueden ser utilizadas para llevar a cabo tareas de minería de datos mediante la aplicación de técnicas y algoritmos de minería de datos

V4 Plataforma Tecnológica del SURE: Se refiere al conjunto de elementos y recursos tecnológicos que conforman la base sobre la cual la institución puede construir sistemas de información específicos. Incluye: hardware, software, redes, equipo de telecomunicaciones, sistemas operativos, gestores de bases de datos, lenguajes de programación, aplicativos de oficina.

1.6 JUSTIFICACIÓN

El presente proyecto de investigación pretende proporcionar una propuesta de un modelo de minería de datos, mediante el cual se puedan detectar anomalías. El modelo busca identificar aquellas transacciones que se encuentren fuera de los lineamientos establecidos mediantes sus rasgos característicos.

Este mecanismo mejorará la forma de analizar las grandes cantidades de datos almacenadas en el SURE provenientes de las transacciones realizadas por el RPI y de esta manera realizar un estudio más profundo de los casos que así lo ameriten y que son identificados a partir de los análisis.

A su vez permitirá a las autoridades, canalizar mejor los esfuerzos de investigación y evitar irregularidades dentro del contexto del SURE, antes de que estas afecten a los poseedores de los bienes inmuebles. Al mismo tiempo esta herramienta tendrá efectos positivos en la imagen y calidad de los servicios prestados por el IP.

CAPÍTULO II. MARCO TEÓRICO

2.1 CONCEPTUALIZACIONES Y DEFINICIONES

Acto Jurídico: Todo acontecimiento voluntario al que el ordenamiento legal ya le ha señalado las consecuencias a actualizarse por su verificación.(«Diccionario Jurídico», s. f.)

Anomalía: Es un valor atípico, es básicamente un dato que normalmente se califica de extremo ya que su valor se aleja mucho de la tendencia (media o moda) de los demás datos que se estén analizando (Han & Kamber, 2006)

Asiento Registral: El asiento es un extracto del contenido del título, expresivo de sus datos fundamentales, extracto que determina el Registrador, con amplias facultades para la calificación.(«Diccionario Jurídico», s. f.)

Arrendar: Ceder o adquirir por precio el goce o aprovechamiento temporal de cosas, obras o servicios. («Diccionario de la lengua española - Vigésima segunda edición», s. f.)

Bien Mueble: Los que, por oposición a los inmuebles, se caracterizan por su movilidad y posibilidad de traslación, y ciertos derechos a los que las leyes otorgan esta condición. («Diccionario de la lengua española - Vigésima segunda edición», s. f.)

Bien Inmueble: Tierras, edificios, caminos, construcciones y minas, junto con los adornos o artefactos incorporados, así como los derechos a los cuales atribuye la ley esta consideración.(«Diccionario de la lengua española - Vigésima segunda edición», s. f.)

Cognoscibilidad: Posibilidad de ser conocido. («Diccionario de la lengua española - Vigésima segunda edición», s. f.)

Dar fe: Seguridad, aseveración de que algo es cierto. («Diccionario de la lengua española - Vigésima segunda edición», s. f.)

Derecho de Propiedad: Poder jurídico que una persona ejerce en forma directa e inmediata sobre una cosa, que le permite su aprovechamiento total o parcial en sentido jurídico y que es oponible a terceros.(«Diccionario Jurídico», s. f.)

Derecho Registral: aquella rama del derecho, formada por el conjunto de normas jurídicas y principios regístrales que regulan la organización y funcionamiento de los registros públicos, los derechos inscribibles y medidas precautorias en los diversos registros, en relación con terceros. (Guevara Manrique, 1988)

Dominio: en teoría de base de datos se refiere al conjunto de valores que se pueden asignar a un atributo de una entidad (Elmasri & Navathe, 2007)

Escritura Pública: Es aquel documento que se otorga ante un notario público y por este mismo funcionario, en el que asienta lo que ante él sucede y refieren las partes, además de autorizarlo. («Diccionario Jurídico», s. f.)

Hiperplano: Es una línea divisora que sirve como frontera de decisión entre los registros de diferentes tipos. (Han & Kamber, 2006)

Inscribir: Tomar razón, en algún registro, de los documentos o las declaraciones que han de asentarse en él según las leyes. («Diccionario de la lengua española - Vigésima segunda edición», s. f.)

Impugnar: Interponer un recurso contra una resolución judicial. («Diccionario de la lengua española - Vigésima segunda edición», s. f.)

Minería de Datos: Conjunto de técnicas matemáticas, estadísticas y computacionales que, junto a un enfoque de las ciencias de la conducta, permite el análisis de datos y la

elaboración de modelos matemáticos descriptivos y predictivos. (Palma, Palma, & Perez, 2009)

Presunción iuris tantum: presunción salvo prueba en contrario o iuris tantum, ésta podrá dirigirse tanto a probar la inexistencia del hecho presunto como a demostrar que no existe. («Diccionario jurídico: Iuris tantum», s. f.)

Publicidad Jurídica: Consiste en la exteriorización o divulgación de una situación jurídica para producir cognoscibilidad general. (Cornejo, 1994)

Registrador: Persona que tiene a su cargo algún registro público, especialmente el de la propiedad. . («Diccionario de la lengua española - Vigésima segunda edición», s. f.)

Sistema Registral: Conjunto de normas que en un país determinado regulan las formas de publicidad de los derechos reales sobre bienes inmuebles mediante el registro de la propiedad, así como el régimen y organización de esa institución. (Lacruz Berdejo, 2011).

Usufructo: Derecho a disfrutar bienes ajenos con la obligación de conservarlos, salvo que la ley autorice otra cosa.. («Diccionario de la lengua española - Vigésima segunda edición», s. f.)

2.2 LA PUBLICIDAD JURÍDICA Y EL REGISTRO COMO INSTRUMENTO DE CERTEZA

A través de la historia ha sido una preocupación básica la propiedad de la tierra, su modalidad de ocupación y la transferencia de la posesión de la misma, puesto que la tierra es un medio para la producción de alimentos, la generación de ingresos, así como también permite el otorgamiento de crédito y recaudación de impuestos. El problema jurídico más relevante de la propiedad es la cuestión de sus orígenes inciertos (Álvarez

Caperochipi, 2006) por la dificultad de determinar el primer propietario, si la propiedad se ha vendido antes y del contenido material del objeto de la propiedad.

Es a través de la publicidad de la posesión, que se ha buscado resolver la incertidumbre con respecto a la propiedad. La publicidad, consiste en la exteriorización o divulgación de una situación jurídica para producir cognoscibilidad general. (Cornejo, 1994). Por su parte, López de Zavalía (1983) define la publicidad como la cognoscibilidad de hechos en base a una declaración señalativa del órgano competente, puesta a disposición del público por los medios previstos por la ley. De acuerdo con este concepto para concretar la publicidad, se requiere de un mecanismo y procedimientos establecidos.

Hernández Gil (1983) opina que la publicidad - en sentido técnico jurídico - solo se consigue por medio de órganos públicos registrales dispuestos para ese fin específico. En este caso el mecanismo de publicidad es el Registro, el cual asegura las relaciones jurídicas y salvaguarda los intereses de los participantes en los actos.

2.2.1 EL REGISTRO DE LA PROPIEDAD Y EL SISTEMA REGISTRAL INMOBILIARIO

Se encontró que el Registro de la Propiedad se puede definir desde 4 puntos de vista distintos:

- Como institución jurídica
- Como servicio público
- Como oficina
- Como conjunto de libros

Para los fines de esta investigación la definición desde el punto de vista de oficina es la más adecuada. Desde esta perspectiva, el Registro de la Propiedad es el centro público o lugar donde se hallan los libros en los que se practican los asientos regístrales; en el que el Registrador presta el servicio registral a través de un conjunto de medios materiales y humanos por él organizados (Faudos, et al, 2008). En concordancia con

esta definición el sistema legislativo hondureño lo contempla como el lugar en el que se lleva a cabo la inscripción de las escrituras, con la finalidad primordial de dar seguridad y certeza al comercio jurídico.

Existen diferentes formas en las que se pueden organizar los registros de la propiedad o inmobiliarios. La vía a través de la cual se regula la publicidad registral así como el registro, es el Sistema Registral. De ahí se entiende por Sistema Registral: el conjunto de normas que en un país determinado regulan las formas de publicidad de los derechos reales sobre bienes inmuebles mediante el registro de la propiedad, así como el régimen y organización de esa institución (Lacruz Berdejo, 2011). Éste es en esencia la pieza fundamental de los entes administradores de la propiedad.

En resumen, este conjunto de normas son reguladoras de los entes organizados por el Estado que tienen a su cargo la función de registrar personas, sus derechos y los actos que deban ser inscritos. Un sistema registral debe tener por objetivo el de otorgar seguridad jurídica a los particulares en el trafico inmobiliario, permitiendo saber quiénes son los titulares y cuál es el estado jurídico de esos bienes que aseguran y dan eficacia a ese tipo de transacciones, de tal forma que la persona que figure registralmente como propietario o titular de un derecho real, en verdad lo sea, salvo que judicialmente se impugne el contenido de la inscripción. («Revista del Colegio de Notarios de Jalisco», s.f.). Los mismos deben brindar seguridad en todo el proceso, para cumplir el deber de proteger el derecho de propiedad.

2.2.1 PRINCIPIOS DEL DERECHO REGISTRAL EN HONDURAS

En Honduras se emplean varios principios registrales, que permiten conocer cómo funciona el registro, estos se definen como: orientaciones capitales, líneas directrices del sistema, la serie sistemática de bases fundamentales y el resultado de la sintetización o condensación del ordenamiento jurídico registral. (IP, 2012). Dichas directrices se encuentran estrechamente relacionadas entre sí.

En ese sentido, la aplicación de estos principios, permite al Registro de la Propiedad ser ágil, transparente e imparcial. El Registro tiene por objeto y finalidad garantizar a los usuarios y terceros que las inscripciones y servicios registrales se efectúen bajo los principios de organización, eficacia registral, legalidad, prioridad, rogación, obligatoriedad, publicidad, tracto sucesivo, especialidad, celeridad, universalidad y fe pública registral (Congreso Nacional, 2004). Según el momento en el que se llevan a cabo, se pueden clasificar en los que se dan previo a que se realice la inscripción, durante la misma y posteriores a la inscripción.

Dentro de aquellos que operan antes de la inscripción están el principio de rogación y el principio de legalidad. Por el principio de rogación el Registrador, con carácter general, no puede actuar de oficio en la práctica de los asientos regístrales, sino que debe de mediar una solicitud o petición de parte interesada o un mandato de la autoridad judicial o administrativa con los que se iniciará el procedimiento registral. (Faudos, et al, 2008). El principio de legalidad se refiere que todo título que pretenda su inscripción y/o anotación preventiva, sin excepción, debe estar sometido a una previa calificación registral, a fin de que en los asientos correspondientes solamente tenga acceso los títulos válidos y perfectos. («Principios Registrales - Ilustrados!», s.f.).

Entre los principios que se dan simultáneos a la inscripción esta el Principio de Prioridad. Es el principio hipotecario en virtud del cual el acto registrable que primeramente ingrese en el Registro de la Propiedad, se antepone con preferencia excluyente o superioridad de rango a cualquier otro acto registral que siéndole incompatible o perjudicial, no hubiere sido presentado al registro o lo hubiere sido con posterioridad, aunque dicho acto fuese de fecha anterior. (Faudos, et al, 2008) También se incluye el de especialidad, que observa que los derechos inscritos en el Registro, se hagan con precisión en cuanto a la persona, el bien inmueble y el derecho. Y el de Tracto sucesivo que tiene por objetivo la continuidad en orden cronológico de los actos de manera que ello refleje el historial sucesivo de cada propiedad.

Finalmente están aquellos principios que se dan posterior a la inscripción, como el principio de Legitimación: se presume *iuris tantum* que el contenido del Registro es exacto y corresponde a la realidad. (Faudos, et al, 2008). También se encuentra el de Fe Pública que presume que el contenido de los libros donde se lleva el registro es exacto; el principio de publicidad indica que el registro puede ser consultado por cualquier persona y por último el principio de Oponibilidad cuyo objetivo es impedir que se inscriban derechos que puedan resultar incompatibles con otro derecho que se haya inscrito anteriormente.

2.2.2 TÉCNICAS REGISTRALES

Las técnicas registrales más conocidas son dos: Folio Personal y Folio Real. Se encontró que el Folio Personal se basa en la generación de asientos regístrales efectuados en libros, los cuales se ordenan por el propietario del bien inmueble, mientras que el Folio Real consiste en otorgar un número único por propiedad al cual se le denomina matricula.

La siguiente tabla muestra las características más relevantes de ambas técnicas.

Tabla 1: Características de las Técnicas Registrales

Folio Personal	Folio Real
En cuanto a los derechos de propiedad, los cambios que afectan un inmueble se registran por separado, en libros o tomos distintos.	La primera inscripción hace referencia a la matrícula o número único que identifica el inmueble y los subsiguientes registros se hacen bajo este número.
Los números de inscripción, se asignan de manera cronológica a cada uno de los actos jurídicos que son presentados por las partes interesadas para su inscripción.	El registro de los instrumentos o documentos se realiza en base a cada lote y no del propietario de este, lo que permite darle seguimiento a los cambios que se dan en los derechos, por lo cual hay un mejor control
Se da lugar a la formación de “Libros” o Tomos que contienen información dispersa de los inmuebles.	Tomando como base este número de matrícula, se anexan a esta, notas con la descripción del inmueble, dirección, naturaleza, área, colindantes, clave catastral y planos si se cuenta con uno. También se anexa los datos de los derechos de los propietarios.

2.3 EL INSTITUTO DE LA PROPIEDAD EN HONDURAS

Con la promulgación de la Ley de la Propiedad en el año 2004, se sientan las bases para modernizar las instituciones del sector público que requerían con urgencia cambios profundos, a través de un ordenamiento jurídico y abriendo las posibilidades al empleo de la tecnología como herramienta para automatizar y mejorar los procesos, así como también para garantizar la seguridad jurídica.

Se crea así el Instituto de la Propiedad (IP), el cual es un ente desconcentrado de la Presidencia de la República, con su propia personalidad jurídica y patrimonio propio. Este ente funciona en todo el territorio hondureño. Fue creado para administrar la información territorial así como para otorgar, garantizar y mantener la seguridad jurídica a los titulares de derechos constituidos sobre bienes materiales e inmateriales mediante la aplicación de la constitución de la República y demás leyes de la nación. («Instituto de la Propiedad», s. f.). Con esto se pretende que el país cuente con un sistema de la propiedad incluyente, donde se busca que un alto porcentaje de la población pueda participar en las actividades económicas del Estado.

La Ley de la Propiedad es enfática en cuanto a las atribuciones y derechos del IP, de entre los cuales se encuentran los siguientes:

- ....Coordinar la creación y operación de un sistema integrado de información de la propiedad;
- Crear, administrar y poner a disposición del público por cualquier medio electrónico o físico la información sobe los derechos y registros que son de su competencia con las limitaciones señaladas por la ley; (Congreso Nacional, 2004)

Para el logro de estas actividades el IP cuenta con su propia estructura administrativa y marco de competencias.

En el Reglamento de la Ley de Propiedad, Capitulo II se detallan los diferentes órganos y unidades que componen el IP. Algunos de ellos son:

- Consejo Directivo
- Secretaria Ejecutiva
- Direcciones Generales
- Inspectoría General
- Superintendencia de Recursos
- Secretaria General
- Programa de Administración de Tierras en Honduras (PATH).

Actualmente el IP cuenta con las siguientes Direcciones Generales: la Dirección General de Registro (DGR), Dirección General de Catastro y Geografía (DGCG), Dirección General de Regularización Predial (DGRP). Las Direcciones Generales son los órganos responsables de atender, coordinar y ejecutar a nivel nacional los asuntos que sean competencias especificas y especializadas del Instituto de la Propiedad (Instituto de la Propiedad, 2010). Cada una de ellas está a cargo de un Director General y todos los Registros que dependan del IP atañen a estas Direcciones. En base a esta estructura el IP tiene como parte de sus funciones, lograr a cabalidad los objetivos y finalidades del Registro de la Propiedad Inmueble (RPI) tal como se menciona en el Articulo 28 de la Ley de la Propiedad.

El RPI comprende el registro de todos los derechos e intereses pertenecientes a los bienes inmuebles. Es decir, el RPI inscribe instrumentos que crean, transfieren, modifican, o cancelan la pertenencia o derechos de posesión. Similarmente, documentos relacionados a herencias, usufructos, servidumbre, y arrendamiento, cuando tales derechos tienen la intención de tener efectos contra terceras personas deben ser inscritos en el RPI. (PATH, 2005). Este Registro forma parte de los que por ley se denomina el Registro Unificado de la Propiedad, que además de la Propiedad Inmueble incluye la Propiedad Mueble, Mercantil, Intelectual y los registros especiales (personas jurídicas, información cartográfica, geográfica, áreas protegidas y las demás que el IP desee incorporar).

2.3.1 PROGRAMA DE ADMINISTRACIÓN DE TIERRAS DE HONDURAS (PATH)

El Programa de Administración de Tierras de Honduras (PATH) tuvo sus inicios en el Proyecto de Administración de Áreas Rurales (PAAR) en el año 1995. Posteriormente bajo una nueva iniciativa denominada Proyecto de Modernización de los Sistemas de Registro de Propiedad Inmueble y Mercantil en Honduras (PMSR) cambia de denominación y se enfoca en la automatización y reforma estructural del Registro de la Propiedad en Honduras.(PATH, 2005). Actualmente es un programa adscrito al IP, el cual es financiado por el Banco Mundial, que busca establecer un sistema de administración de tierras integrado y descentralizado.

Como parte de sus compromisos el PATH deberá institucionalizar el Sistema de Administración de la Propiedad, otorgando apoyo técnico a las instituciones públicas y privadas que integren el mismo. Entre los objetivos y propósitos de mayor relevancia del PATH podemos mencionar:

- Generar, registrar, vincular y administrar, información veraz y actualizada, sobre las transacciones características y normativas de la propiedad.
- Propiciar el fortalecimiento de los mercados financieros secundarios, mediante la seguridad en las transacciones de bienes inmobiliarios, bienes muebles, propiedad intelectual y otros. (PATH, 2012)

Con estos objetivos se busca como resultado que el IP se transforme en una institución fortalecida que sea capaz de asumir un papel protagónico en la administración de tierras en el país.

La estructura organizacional del PATH corresponde a la definición de objetivos, metas y la definición de la estrategia de ejecución que fue planteada para el mismo. En la figura 2 se muestra el organigrama del PATH disponible en su página web http://www.path.hn/path/index.php/pathii/organigrama.

Según se observa en dicho organigrama, éste se subdivide en componentes y áreas, los cuales responden a objetivos específicos y han sido subdivididos en “áreas de

acción temáticas", cada una de ellas atiende los proyectos que colaboran con las metas que han sido establecidas.

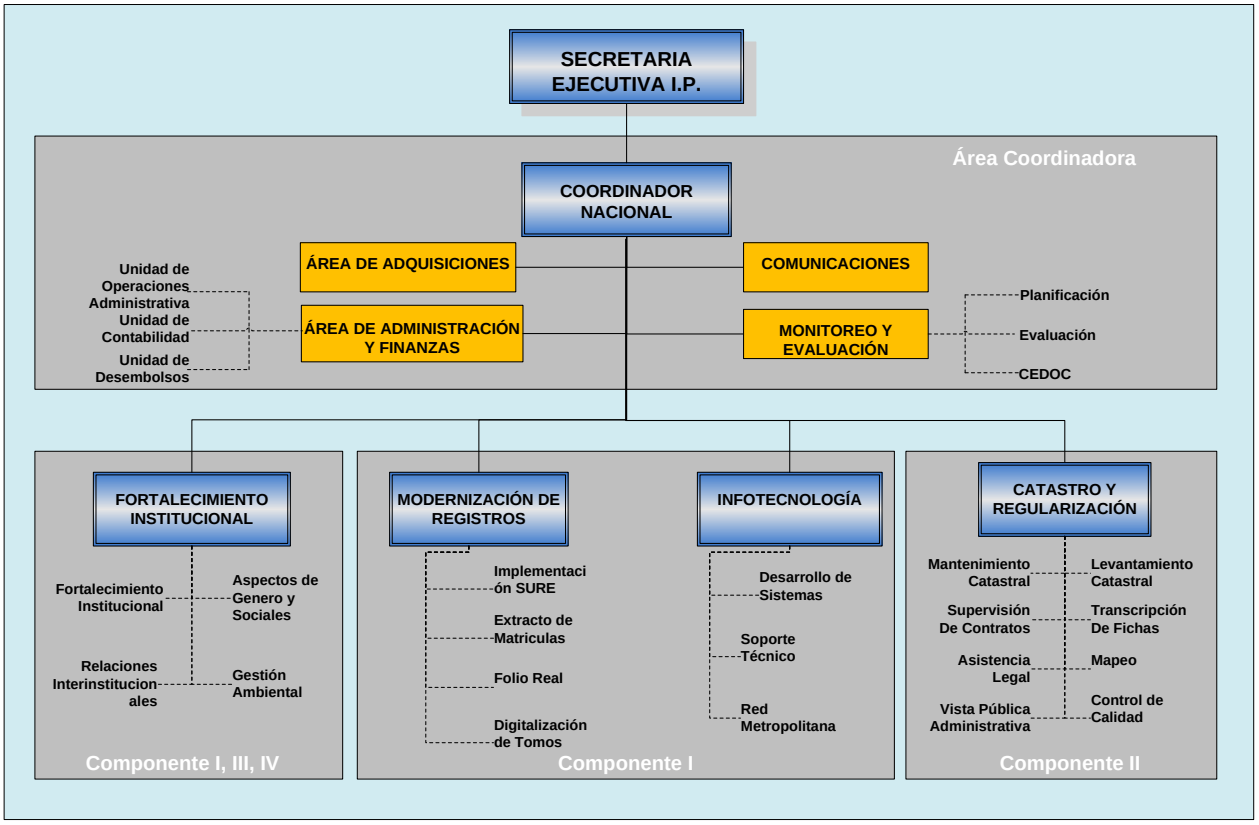


Figura 2. Organigrama del PATH

Fuente: www.path.hn (2012)

2.4 SISTEMA UNIFICADO DE REGISTROS (SURE)

De acuerdo a lo dispuesto en el Artículo 5, Numeral 3, de la Ley de Propiedad, el IP generará, operará y administrará un sistema de información, utilizando medios electrónicos y las nuevas tecnologías.

El Sistema Nacional de Administración de Propiedad (SINAP) operará como un sistema integrado de Información del territorio en materia de geografía y propiedad, e integrará, registrara y publicará la información contenida en los subsistemas que los constituyen: el Sistema Nacional de Información Territorial (SINIT), el Registro Nacional de Normativas de Ordenamiento Territorial (RENOT), Sistema Unificado de Registros

(SURE) y el Sistema Nacional de Información Geográfica de Honduras (SINIGH) (Instituto de la Propiedad, 2010). El desarrollo de esta plataforma tecnológica ha sido realizado por el PATH.

El SURE es el aplicativo medular del SINAP, y es fundamental para la seguridad del sistema de propiedad y tenencia del país. El SURE es el subsistema a través del cual es posible vincular la información registral y catastral de cada predio, así como los datos generales de la o las personas que se declaran con derecho sobre él.(PATH, s. f.) El SURE es una herramienta de apoyo al trabajo diario registral del IP, pero la tutela del sistema la tiene el PATH.

Figura 3. Pantalla Principal del SURE

Fuente: www.sure.sinap.hn (2012)

La figura 3 muestra la pantalla principal del SURE el cual es un sistema web que cuenta con 30 módulos con propósitos específicos entre ellos el Modulo de Registro de Propiedad Inmueble. Dicho sistema, permite registrar y consultar vía una página web, actualmente es utilizado en 9 registros que han sido modernizados para emplear esta técnica automatizada.

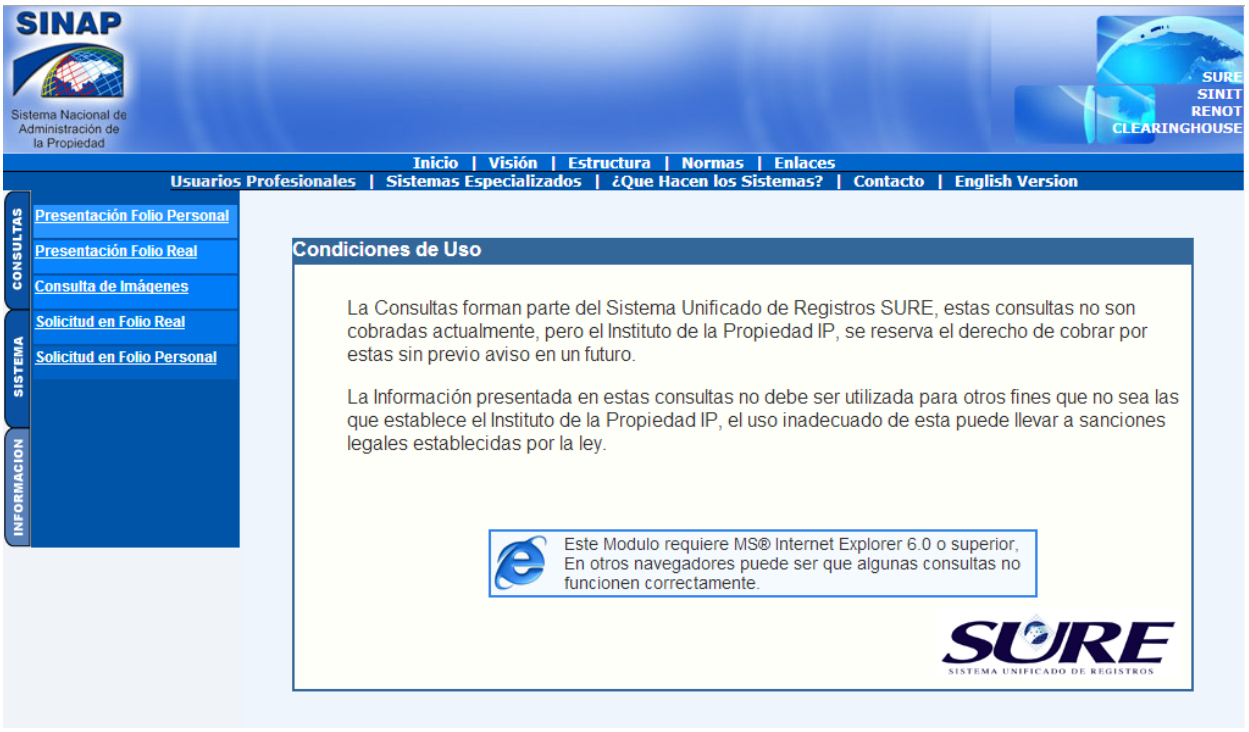


Figura 4. Consultas Públicas del SURE

Fuente: Portal SINAP (2012)

En la figura 4 se muestra el sitio de Consultas Públicas del SURE, que pueden ser accedidas por entidades públicas y privadas como los bancos e instituciones de crédito en este enlace: http://www.sinap.hn/portal/page? _pageid=52,1&_dad= portal&_schema =PORTAL

2.4.1 PROCESO REGISTRAL - INSCRIPCIÓN EN FOLIO REAL USANDO SURE

El proceso de inscripción de una escritura pública involucra una serie de pasos que sigue un documento desde que el Usuario lo presenta hasta su inscripción o en su

defecto la Resolución de Denegatoria Provisional o Definitiva, según se muestra en la Figura 5.

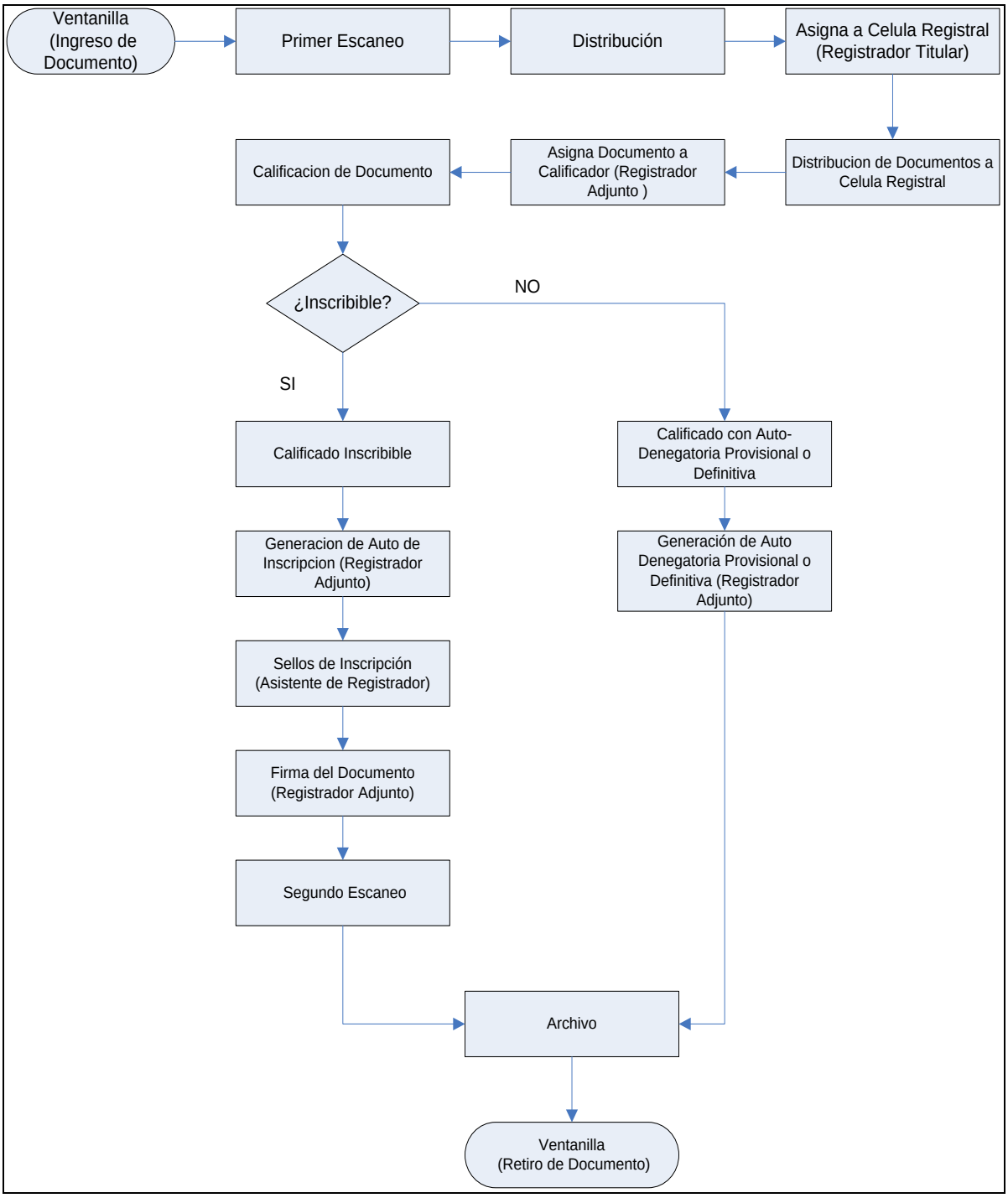


Figura 5. Proceso de Inscripción bajo la técnica del Folio Real en el RPI

Fuente: PATH (2011)

En la figura se observa que el proceso empieza con la presentación al registro de una escritura pública en ventanilla.

Es importante mencionar que los datos generales de ese documento son digitados en el SURE, incluyendo la fecha y hora de recibido. Adicionalmente, a la presentación se le asigna un número de transacción único, el cual puede ser usado más adelante para determinar cuando el documento fue ingresado en relación a los demás.

Se realiza un primer escaneo y pasan a una oficina de distribución. Aquí, los documentos se distribuyen a una Célula Registral, que no es más que un equipo de abogados liderados por un Registrador Adjunto. Los Registradores Adjuntos están adscritos a un Registrados Titular. Existe un Registrador Titular por cada Registro en el país.

El Registrador Adjunto distribuye dentro de su Célula Registral los documentos. Al miembro de la célula al que se le asigno el documento, empieza por revisar si cumple con las formalidades legales establecidas, esta etapa se denomina “calificación”. Si se encuentran errores o deficiencias en el documento y la deficiencia puede ser corregida, se elabora una nota describiendo cual es el cambio requerido dando lugar a un Auto Denegatoria Provisional y el documento es devuelto al área de archivo para que sea recogida por la persona que lo presento. Cuando un error o falla no puede ser arreglado, la presentación se deniega, (Auto Denegatoria Definitiva) y el documento es devuelto a la oficina de archivo.

Si no se encuentran errores en el documento, o si los errores, una vez encontrados, han sido corregidos, la aplicación pasa nuevamente al Registrador Adjunto para la Generación del Auto de Inscripción. El documento es sellado y firmado. Luego es escaneado por segunda vez pasando finalmente al área de archivo para que sea retirado por el interesado.

2.5 MINERÍA DE DATOS

La Minería de datos es un término utilizado para describir el descubrimiento de conocimiento dentro de grandes cantidades de datos. Técnicamente se puede definir como un proceso que usa técnicas estadísticas, matemáticas y de inteligencia artificial para extraer e identificar información útil y posteriormente conocimiento (o patrones) de grandes conjuntos de datos (Turban, Sharda, & Delen, 2009). Esta definición ubica a la minería de datos como un proceso donde se interceptan técnicas de diferentes disciplinas, y aunque no se requiere ser experto en ellas es necesario un conocimiento básico.

Según autores como Prabhu & Venatesan, (2007) la minería de datos, conocida también como descubrimiento de conocimiento en bases de datos, es la extracción no trivial de información implícita, previamente desconocida y potencialmente útil a partir de los datos. Estos elementos hacen de la minería de datos un proceso iterativo destinado al análisis de grandes conjuntos de datos que permitan realizar inferencias a individuos involucrados en la toma de decisiones y solución de problemas en diferentes áreas como ser: ventas, mercadeo y finanzas, por mencionar algunas.

2.5.1 HISTORIA DE LA MINERÍA DE DATOS

El termino minería de datos puede parecer nuevo, pero el concepto como tal y las ideas detrás del mismo no lo son, ya desde los años sesenta se manejaban términos como Data Fishing con la idea de encontrar correlaciones dando como resultado que muchas de las técnicas utilizadas tuvieran sus raíces en otras disciplinas. Sin embargo, La industria de base de datos ha sido testigo del camino evolutivo en el desarrollo de funcionalidades como (ver Figura 6): recolección de datos y creación de base de datos, administración de datos y análisis de datos avanzados.(Han & Kamber, 2006). Está claro que la minería de datos puede ser vista como el resultado de la evolución natural de la tecnología de información y a lo largo de los años su evolución está estrechamente relacionada con la evolución de las bases de datos.

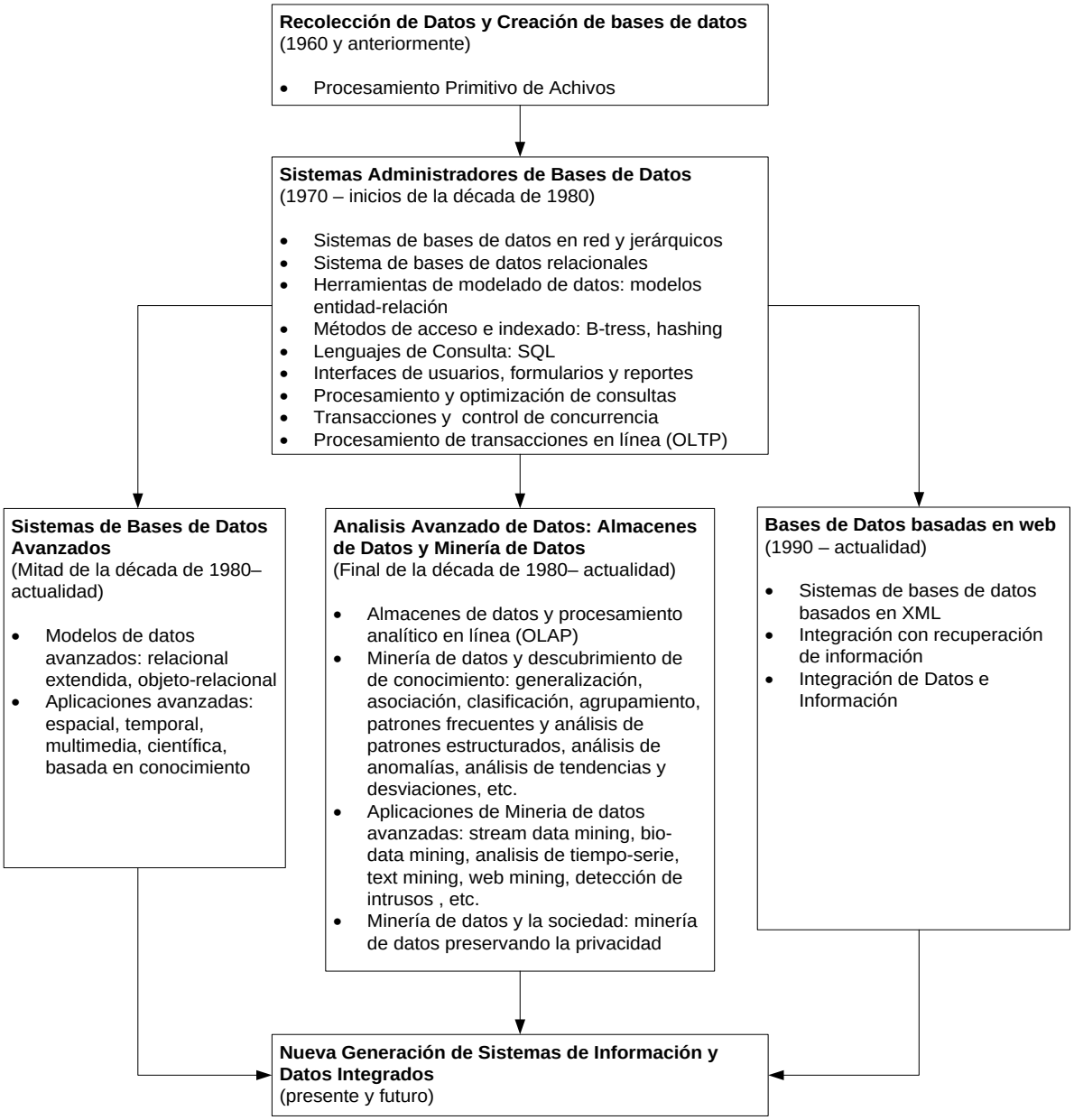


Figura 6. La evolución de los sistemas de Base de Datos

Fuente: Han & Kamber, (2006)

2.6 TAXONOMÍA DE MINERÍA DE DATOS

Taxonomía en su sentido más general es la ciencia de la clasificación. Es importante que antes de revisar las diferentes técnicas y algoritmos disponibles para la minería de

datos hagamos un esfuerzo por dar una clasificación que nos permitan ubicarlas dentro de un marco más amplio.

Durante la investigación se pudo constatar que no existe una clasificación generalmente aceptada y que su clasificación depende básicamente de tres factores: del conjunto de datos que se dispone, del problema a resolver y de las herramientas disponibles. La siguiente imagen resume la clasificación de tareas, técnicas y algoritmos que a nuestro criterio es la más clara y fácil de comprender.

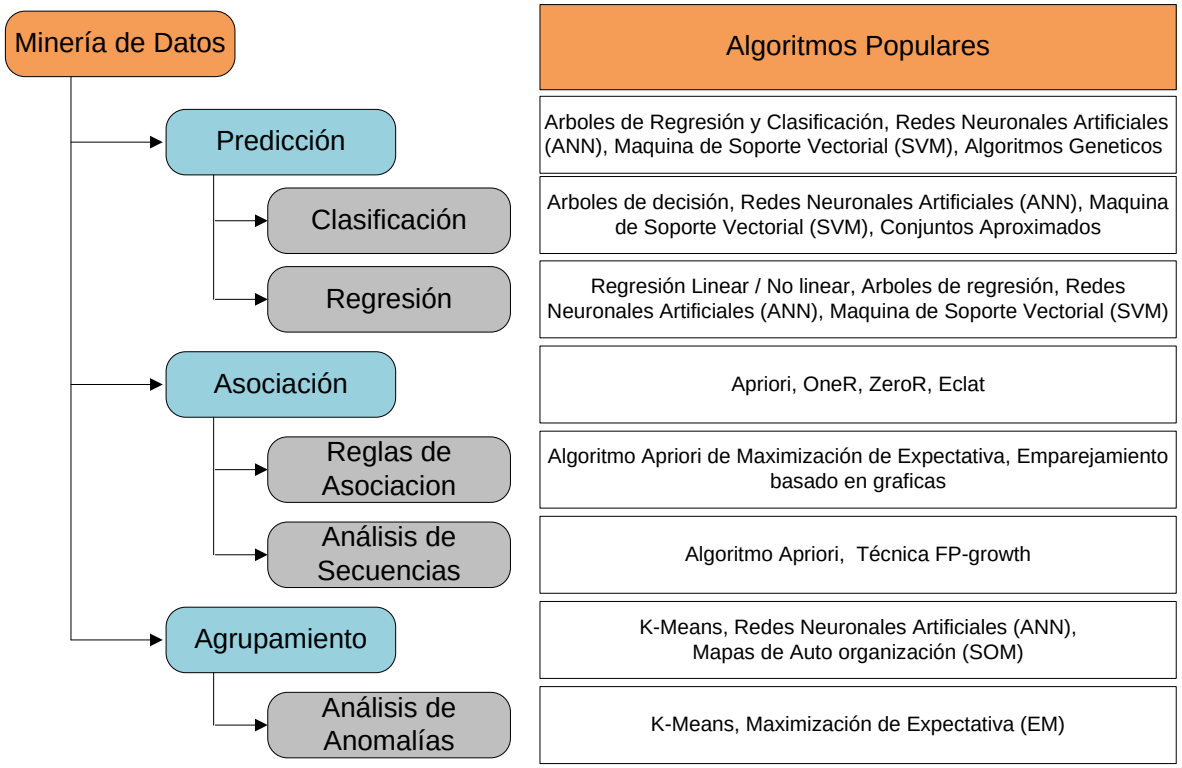


Figura 7. Una Taxonomía simple de las tareas de la Minería de Datos

Fuente: Turban et al., (2009)

2.7 TÉCNICAS DE MINERÍA DE DATOS

El uso de minería de datos es cada vez más importante en las organizaciones que mantienen grandes repositorios de datos y necesitan extraer el conocimiento que se encuentra dentro de ellos de manera que puede ayudarlos a mejorar. Con la minería de

datos pueden llevarse a cabo principalmente tres tareas, denominadas también técnicas: predicción, asociación y agrupamiento.

2.7.1 PREDICCIÓN

Es una de las principales técnicas por las cuales se usa minería de datos y es en ocasiones conocida como estimación y a veces asociada con el término pronostico. Predicción tiene como propósito anticipar el valor que una variable aleatoria asumirá en el futuro o estimar la probabilidad de futuros eventos (Vercelli, 2009). Normalmente hacemos uso de esta técnica, aunque de manera indirecta, al revisar los pronósticos del tiempo o para algunos interesados en el mundo financiero al revisar la predicción del alza o baja de acciones, por mencionar algunos de sus principales usos hoy en día.

2.7.2 ASOCIACIÓN

Es una de las principales técnicas mediante las cuales se pueden descubrir relaciones interesantes en nuestros conjuntos de datos. “Asociación es el trabajo de encontrar que atributos “van juntos”…. “busca descubrir reglas para cuantificar la relación entre dos o más atributos” (Larose, 2005). Este tipo de técnica es principalmente utilizada en análisis del “carrito de compras” donde se analiza los artículos adquiridos por cada cliente para determinar patrones significativos de compra.

2.7.3 AGRUPAMIENTO (CONGLOMERADOS)

Es otra de las principales técnicas de minería de datos cuyo fin es identificar conglomerados con características similares. Una definición de agrupamiento podría ser “el proceso de organizar objetos dentro de conglomerados cuyos miembros son similares de alguna manera”. Un conglomerado es por lo tanto una colección de objetos que son “similares” entre ellos y son “no similares” a los objetos que pertenecen a otros conglomerados.(Prabhu & Venatesan, 2007). Esta técnica nos ayudará a encontrar a conglomerados homogéneos que nos permitan identificar, según nuestros intereses, ya

sea un nuevo nicho de mercado o detectar valores atípicos que no son incluidos en ninguno de los conglomerados resultantes.

2.8 MÉTODOS DE APRENDIZAJE

Antes de describir los principales algoritmos de minería de datos, es importante conocer los diferentes métodos de aprendizaje que un algoritmo de minería de datos puede utilizar. En minería de datos los algoritmos pueden ser clasificados por su método o proceso de aprendizaje en tres categorías que son: supervisados, semi-supervisados y no supervisados.

2.8.1 SUPERVISADO

En un método de aprendizaje supervisado (o directo), un atributo objetivo representa la clase a la que cada registro pertenece… el proceso de aprendizaje supervisado es por lo tanto orientado a la predicción e interpretación con respecto a un atributo objetivo. (Vercelli, 2009) Los métodos supervisados aunque son altamente efectivos y pueden tener un alto grado de exactitud, sin embargo lograr un alto nivel de efectividad y exactitud puede ser excesivamente caro.

2.8.2 SEMISUPERVISADO

Es un método de aprendizaje que aprende de un pequeño conjunto de registros etiquetados (con el atributo objetivo) y un gran conjunto de datos no etiquetados (sin el atributo objetivo)(Liu, 2007). Los métodos semi-supervisados tienen la opción de preparar un conjunto de datos que incluya una etiqueta (atributo objetivo) que identifique a que clase pertenece, y cuando hablamos de detectar anomalías identificamos básicamente dos clases: normal y anomalía. Y lo más común es que ese conjunto de datos se haga con la clase normal que es relativamente fácil de obtener; por el contrario un conjunto de datos que represente todos los escenarios de anomalías existentes es caro y difícil de obtener.

2.8.3 NO SUPERVISADO

En un método de aprendizaje no supervisado (o indirecto) no se guía por un atributo objetivo. Por lo tanto, las tareas de minería de datos en este caso son dirigidas a descubrir patrones y afinidades en el conjunto de datos. (Vercelli, 2009). Este método al contrario de los anteriores no presupone existe un atributo objetivo sino que hace otro tipo de suposiciones sobre los datos, que dependerán del algoritmo seleccionado. Aunque es uno de los métodos más populares de aprendizaje es víctima de un alto porcentaje de falsas alarmas, eso debido a que frecuentemente las suposiciones realizadas no son ciertas.

2.9 ALGORITMOS DE MINERÍA DE DATOS

Hay muchas maneras diferentes de desempeñar las tareas de minería de datos. Estas tareas no solo requieren ciertos tipos o estructuras de datos sino también que implican algún tipo de técnica o algoritmo matemático. Existe una amplia lista de algoritmos disponibles para llevar a cabo las principales tareas de minería de datos, sin embargo a continuación mencionamos los más importantes.

2.9.1 REDES NEURONALES ARTIFICIALES (ARTIFICIAL NEURAL NETWORK)

Las redes neuronales son modelos matemáticos que se pueden adecuar para casi todas las técnicas de minería de datos, con especial énfasis en predicción. Las Redes Neuronales son técnicas analíticas que siguen el modelo del proceso de aprendizaje en el sistema cognitivo y las funciones neurológicas del cerebro, capaz de predecir nuevas observaciones (sobre las variables específicas) de otras observaciones (en las misma u otras variables) después de ejecutar un proceso de llamado aprendizaje a partir de los datos existentes. (Prabhu & Venatesan, 2007). Este proceso de aprendizaje implica un “entrenamiento” -que puede ser supervisado (guiado) o no supervisado- el cual una vez concluido da como resultado una red neuronal lista para ser aplicada y utilizada para realizar predicciones.

2.9.2 REGRESIÓN LINEAR (LINEAR REGRESSION)

Se usa principalmente en taras de estimación (predicción). La regresión linear simple nos permite relacionar una variable dependiente continua Y a una continua e independiente variable X. Es comúnmente aceptado que los valores para X sean controlados y no sujetos a errores de medidas, y los correspondientes valores de Y son observados (Tufféry, 2011). Este algoritmo tiene algunos requerimientos que se debe tener en cuenta, por ejemplo, todos los predictores sean numéricos y preferiblemente continuos; tiene dificultades para manejar valores omitidos

2.9.3 MÁQUINA DE SOPORTE VECTORIAL (SUPPORT VECTOR MACHINE)

Según Han & Kamber, (2006) la Máquina de Soporte Vectorial es un algoritmo que utiliza un mapeo no linear para transformar los datos de entrenamiento originales a una dimensión superior. Dentro de esta nueva dimensión, busca el hiperplano óptimo de separación lineal (es decir, una "frontera de decisión" que separa los registros de una clase de otra) Este algoritmo encuentra esa frontera de decisión utilizando vectores de soporte, posiblemente sea uno de los algoritmos más eficaces pero a su vez de los más lentos por lo que la velocidad de entrenamiento del conjunto de datos y de las pruebas pueden convertirse en poco factibles cuando se trata de grandes conjuntos de datos.

2.9.4 ÁRBOLES DE DECISIÓN (DECISION TREES)

Los árboles de decisión son una de los algoritmos de minería de datos directas más poderosos, porque pueden usarse en una amplia gama de problemas y ellos producen modelos que explican cómo estos problemas trabajan. Un árbol de decisión es una colección jerárquica de reglas que describe cómo dividir una larga colección de de registros dentro de grupos cada vez más pequeños de registros. Con cada sucesiva división los miembros del segmento resultante se vuelven más y más similares entre ellos.(Linoff & Berry, 2011).

2.9.5 APRIORI

El algoritmo Apriori es un método clásico de reglas de asociación que funciona de manera iterativa y genera conjuntos de ítems frecuentes basado en un precepto teórico llamado el principio de Apriori. El algoritmo Apriori es un método eficiente de extraer reglas rigurosas contenidas en un conjunto de transacciones. En su primera fase el algoritmo genera los conjuntos de ítems frecuentes en una forma sistemática, sin explorar el espacio de todos los ítems candidatos, mientras que en la segunda fase extrae las reglas fuertes… El principio de Apriori: si un conjunto de ítems es frecuente entonces todos sus subconjuntos son también frecuentes (Vercelli, 2009) La determinación de esas "reglas fuertes" se basa en que deben cumplir con el requerimiento de no exceder el umbral mínimo predefinido de confianza.

2.9.6 K-MEANS

El algoritmo de agrupamiento K-means es un algoritmo sencillo y efectivo que define un parámetro K que indica cuantos conglomerados se buscan. Los puntos K son elegidos aleatoriamente como centros de un conglomerado. Todas las instancias son asignadas a su centro de conglomerado más cercano de acuerdo a una distancia métrica. Luego la media de cada instancia es calculada…y estas son tomadas para ser los nuevos valores centrales. El proceso se repite con los nuevos valores centrales hasta que los mismos puntos son asignados a cada conglomerado en iteraciones consecutivas, etapa en la que el centro del conglomerado se ha estabilizado (Witten, Frank, & Hall, 2011). Hay que tener en cuenta que este algoritmo es sensible valores atípicos ya que estos pueden afectar el cálculo de la media.

Se pueden lograr mejores resultados si se tiene claridad con antelación el número de conglomerados que se desean generar y si se asigna un valor a la media de estos de manera que no sea aleatorio.

2.9.7 K VECINO MÁS CERCANO (K NEAREST NEIGHBOR)

Es un algoritmo que busca patrones usando similitud basado en un concepto de cercanía que realmente se traduce en una distancia numérica, por lo que la distancia de cada punto hacia todos sus vecinos más cercanos es determinada. KNN funciona de la siguiente manera: siendo D nuestro conjunto de datos de entrenamiento. Nada es realizado con los registros de entrenamiento. Cuando una instancia de prueba se presenta, el algoritmo compara d con cada instancia del conjunto de datos de entrenamiento para calcular la similitud o distancia entre ellos. Las instancias k mas similares (cercanas) en D son seleccionadas. Este set de instancias es llamada los vecinos más cercanos de d (k nearest neighbors of d). Luego d toma la clase más frecuente entre los vecinos más cercanos. (Liu, 2007)

Es un algoritmo que hace gran cantidad de cálculos y a medida que el conjunto de datos expuesto aumenta su cantidad de registros así lo hace el costo de cálculo del mismo.

2.9.8 FACTOR DE ANOMALÍA LOCAL (LOCAL OUTLIER FACTOR)

LOF es uno de los primeros algoritmos con enfoque de vecinos más cercano y basado en densidad, comparando la densidad de un objeto con los demás en su vecindario. Este algoritmo considera un objeto como una anomalía local si se encuentra aislado en relación a sus vecinos locales… no considera la propiedad de ser una anomalía como una propiedad binaria. En vez de eso evalúa el grado en el que un objeto es una anomalía. Este grado de anormalidad es calculado como el factor de anomalía local de un objeto. Es local en el sentido que el grado depende de cuan aislado el objeto este respecto a su vecindario circundante.(Han & Kamber, 2006) Aunque se han implementado variantes de este algoritmo (como COF, INFLO, LOCI) y a pesar de ser un algoritmo tiene un alto costo de cálculo sigue siendo uno de los más utilizados.

2.10 MODELOS DE MINERÍA DE DATOS

Un modelo, por lo menos en minería de datos, es una representación computarizada de observaciones del mundo real. Modelos son la aplicación de algoritmos para buscar, identificar y mostrar cualquier patrón o mensaje dentro de los datos (North, 2012).

Un modelo de minería de datos se crea mediante la aplicación de un algoritmo a los datos, pero es algo más que un algoritmo o un contenedor de metadatos: es un conjunto de datos, estadísticas y patrones que se pueden aplicar a los nuevos datos para generar predicciones y deducir relaciones.(«Modelos de minería de datos», s. f.)

De las dos definiciones anteriores podemos resumir que un modelo de minería de datos se compone de varios elementos, como: la definición de un conjunto de atributos que deben extraerse de una base de datos para el conjunto de datos a minar, la aplicación de una técnica o algoritmo de minería de datos al conjunto de datos seleccionado y la generación de nuevos atributos que permitan obtener conocimiento. Este último en nuestro caso sería que nos permita identificar ejemplos (registros) considerados como anomalías.

2.11 DETECCIÓN DE ANOMALÍAS

2.11.1 ANOMALÍA

Una anomalía o valor atípico es básicamente un dato que normalmente se califica de extremo ya que su valor se aleja mucho de la tendencia (media o moda) de los demás datos que se estén analizando. Los valores atípicos son puntos de datos que están muy lejos de otros puntos de datos. Los valores atípicos pueden ser errores en los datos almacenados o puntos de datos especiales con valores muy diferentes (Liu, 2007).Por lo tanto algunos valor atípicos pueden ser validos y simplemente resultado de la variabilidad de algunos atributos en ciertos conjunto de datos, por ejemplo, el salario de

los ejecutivos de alto nivel si estos se analizaran en conjunto con los salarios del resto de empleados de la compañía

2.11.2 MÉTODOS PARA IDENTIFICAR ANOMALÍAS

En la literatura hay una amplia gama de enfoques relacionados con la detección de anomalías siendo los de más amplio uso los siguientes: método estadístico, método basado en distancia, método basado en densidad y método basado en conglomerados.

2.11.3 MÉTODO ESTADÍSTICO

Según Maimon & Rokach, (2010) en el enfoque estadístico los valores atípicos para campos específicos son identificados basados en estadísticas automáticamente calculadas. Para cada campo, la media y desviación estándar son utilizadas, y basadas en el teorema de Chebyshev, aquellos registros que tienen ciertos valores en un campo determinado fuera de un numero de desviaciones estándares de la media son identificados como atípicos. Este enfoque se puede decir que consta de dos etapas, primero la creación del modelo probabilístico que represente los registros “normales” y luego la determinación de un umbral que utilizando las medidas de tendencia central del conjunto de datos se use como referencia para determinar aquellos registros atípicos.

2.11.4 MÉTODO BASADO EN DISTANCIA

Estos algoritmos se basan generalmente en las medidas de distancia locales y son capaces de manejar grandes bases de datos La mayoría de algoritmos basados en distancia son definidos basados en el concepto de vecindario local o k vecinos más cercanos de los datos. La noción de anomalías basadas en distancia no asume ninguna distribución de datos subyacente y generaliza muchos conceptos de los algoritmos basados en distribución (Zhang, 2013). Estos algoritmos son más eficientes por lo que requieren menores capacidades de cálculo.

2.11.5 MÉTODO BASADO EN DENSIDAD

Estos algoritmos utilizan mecanismos más complejos para determinar el grado de anormalidad que los basados en distancia. Los algoritmos basados en densidad asumen que los puntos que pertenecen a cada conglomerado provienen de una distribución probabilística específica. La distribución global de los datos se supone que es una mezcla de muchas distribuciones. El objetivo de estos algoritmos es identificar los conglomerados y sus parámetros de distribución. (Maimon & Rokach, 2010) Los algoritmos basados en densidad cuentan con un modelado más fuerte de valores atípicos pero a la vez requieren de mayores capacidades de cálculo.

2.11.6 MÉTODO BASADO EN CONGLOMERADOS

Muchos algoritmos de minería de datos en la literatura encuentran anomalías como consecuencia de algoritmos de agrupamiento y definen anomalías como puntos que no encuentran dentro de un conglomerado o están localizados muy lejos de uno. Así los algoritmos de agrupamiento implícitamente definen anomalías como el ruido de fondo de los conglomerados… siendo las principales categorías de algoritmos de agrupación: particionamiento de conglomerados, jerarquización de conglomerados, conglomerados basados en densidad.(Zhang, 2013) Originalmente los algoritmos de agrupamiento no estaban destinados a identificar anomalías estos se han adaptado en base al hecho que las anomalías son agrupadas en conglomerados pequeños o no pertenecen a ninguno.

2.12 METODOLOGÍAS DE APLICACIÓN DE LA MINERÍA DE DATOS

Para poder llevar a cabo un proyecto de minería de datos de forma ordenada y eficiente, se han desarrollando varias metodologías, las cuales están basadas en las mejores prácticas y con el tiempo se han convertido en un estándar que permiten maximizar las posibilidades de éxito en dichos proyectos. Dentro de están se encuentran principalmente dos: la metodología CRISP-DM y la metodología SEMMA.

2.12.1 METODOLOGÍA CRISP-DM

Uno de los más conocidos es *Cross-Industry Standard Process for Data Mining*, que fue concebido por un consorcio Europeo a finales de los años 1990s.

Esta guía se considera una metodología, pues es descrita en términos de modelo de procesos jerárquico, consistente en grupos de tareas detalladas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada, e instancia de procesos (Chapman, Clinton, Kerber, Khabaza, Reinartz, , Shearer, & Wirth, 2000), tal y como se puede observar en la Figura 8.

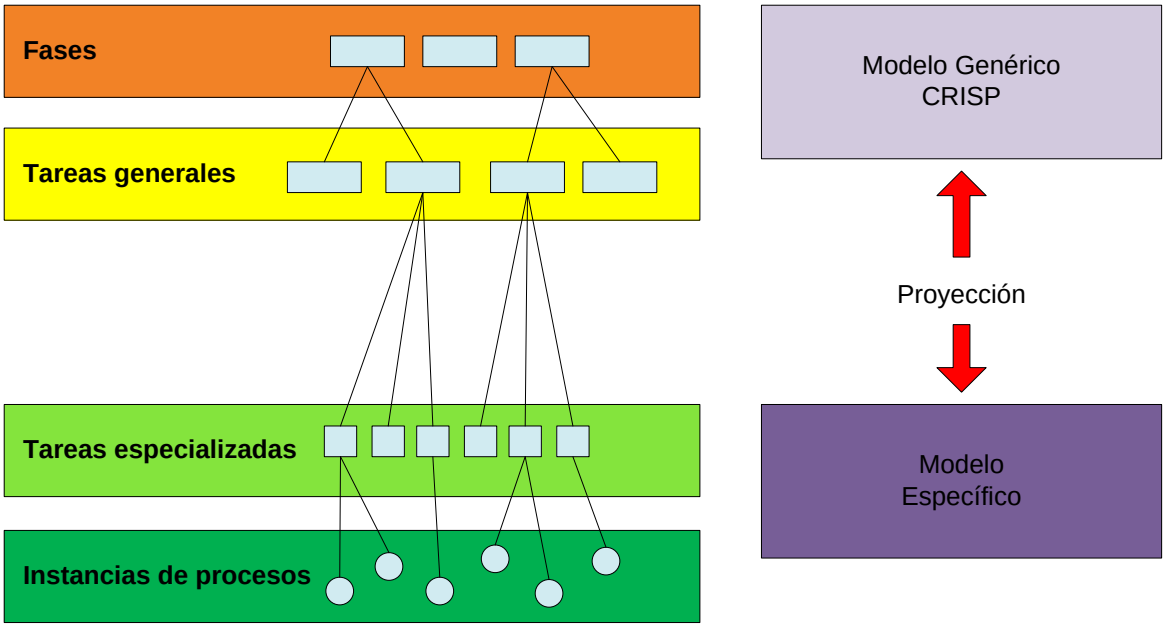


Figura 8. Cuatro niveles de abstracción de la metodología CRIPS-DM

Fuente: IBM (2011)

En el nivel superior, el proceso de minería de datos consta de fases; y estas fases contienen tareas genéricas ubicadas en el segundo nivel denominado “genérico”, llamado así ya que trata de cubrir todas las posibles situaciones de la minería de datos.

En el tercer nivel están las tareas especializadas, este incluye la descripción de las acciones de cada una de las tareas genéricas y como deberán ser llevadas a cabo.

El cuarto nivel, contempla un registro de todas las acciones, decisiones, y de los resultados obtenidos de una implementación de minería de datos. Representa lo que en realidad pasó en una implementación en particular.

Así mismo CRISP-DM se considera un modelo de procesos, puesto que provee una visión general del ciclo de vida de proyecto de minería de datos. Este contiene las fases de un proyecto, sus respectivas tareas, y la relación entre esas tareas. A este nivel de descripción, no es posible identificar todas las relaciones. Pueden existir relaciones entre cualquiera de las tareas de minería de datos dependiendo de las metas, el antecedente y el interés de los usuarios – y lo más importante – de los datos. (IBM, 2011).

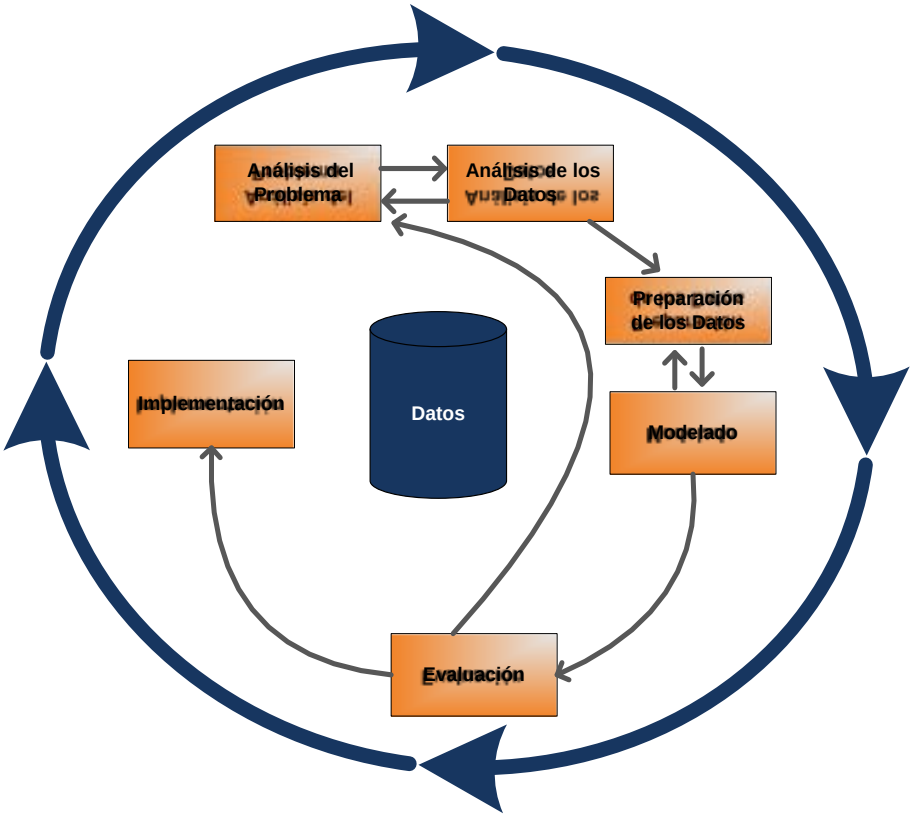


Figura 9. Fases del modelo de referencia CRIPS-DM

Fuente: CRISP-DM (2000)

Como se observa en la figura el ciclo de vida del proyecto de minería de datos consta de 6 fases, donde el paso de una a otra o secuencia de las mismas no es rígido, por lo cual se puede avanzar y regresar a una fase o tarea de una fase según se considere necesario.

2.12.2 METODOLOGÍA SEMMA

Es el proceso de selección, exploración, modelado y evaluación de grandes cantidades de datos para descubrir patrones de negocio desconocidos. El nombre es un acrónimo que corresponde a las nombres, en inglés, de los pasos básicos del proceso de conducir una minería de datos: Seleccionar (Sample), Explorar (Explore), Modificar (Modify), Modelar (Model) y Evaluar (Asses).

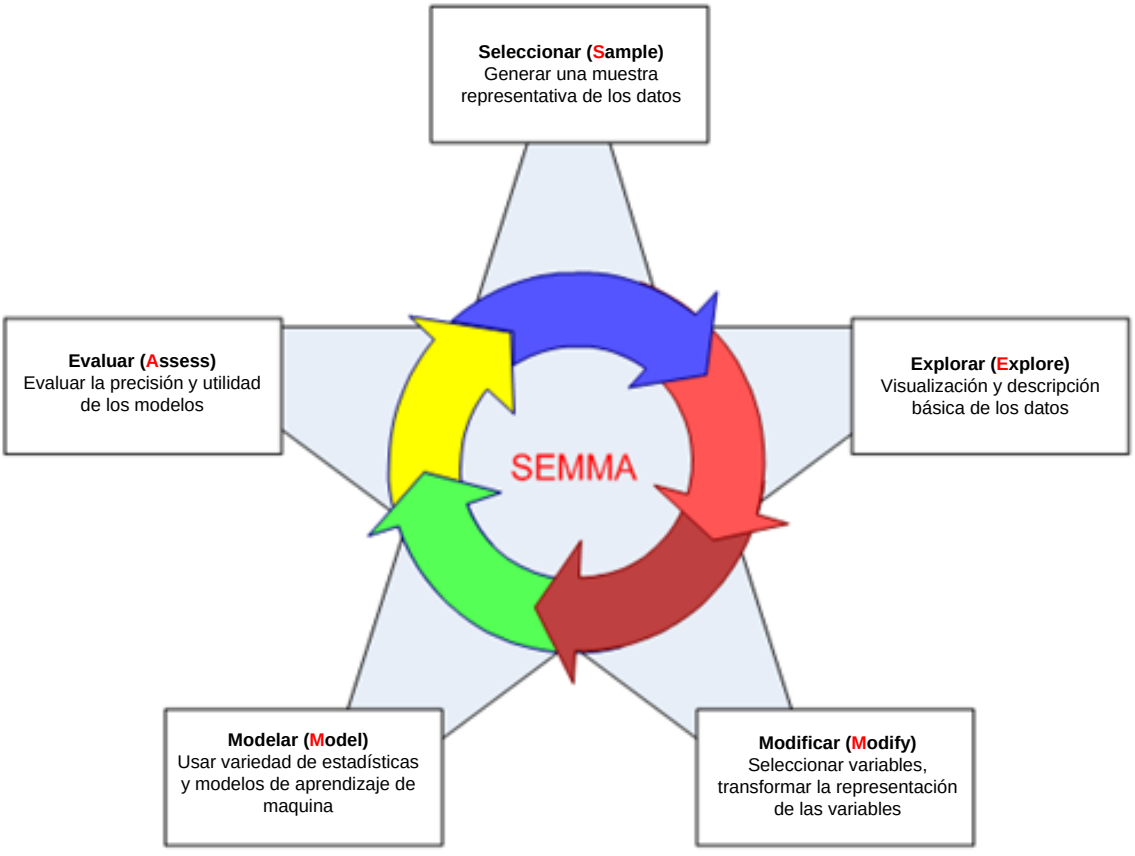


Figura 10. Pasos básicos metodología SEMMA

Fuente: http://www.sas.com

Como se ve en la figura, estos pasos se realizan de manera iterativa, aunque no necesariamente en el orden presentado.

A continuación se detalla en qué consiste cada unos de los pasos:

- Seleccionar: Elementos relevantes de los datos son extraídos de una base de datos o de un almacén de datos dentro de una tabla que contiene todas las variables necesarias para el modelado.(Rokach & Maimon, 2008). Se selecciona una muestra de los datos extrayendo una porción de un conjunto de datos lo suficientemente grande para contener información significativa, pero lo suficientemente pequeño que se pueda manipular rápidamente. («SAS Enterprise Miner - SEMMA», s. f.)

- Explorar: un conjunto de pasos de exploración son realizados para obtener un mejor entendimiento acerca de las relaciones entre los datos. También se realiza una búsqueda de tendencias inesperadas y anomalías con el propósito de mejorar el entendimiento y obtener ideas.(Refaat, 2007) Explorar el conjunto de datos gráfica y estadísticamente. Buscando por pistas inesperadas u observaciones inusuales y también familiarizándose con los datos (graficar los datos, obtener estadísticas descriptivas, identificar variables importantes para el modelo) (Matignon, 2005)

- Modificar: en base al descubrimiento en la fase de exploración, puede ser necesario manipular los datos para incluir información. Se puede también buscar valores atípicos y reducir el número de variables para limitarlas a las más significantes. Finalmente se trata de crear, seleccionar y transformar las variables para enfocar el proceso de selección del modelo y tomar ventaja de las relaciones entre las variables. (Turban et al., 2009)

- Modelar: en este paso es donde es donde el usuario busca por una combinación de variables que de manera confiable predice un resultado esperado. Una vez que has preparado tus datos estás listo para construir un modelo que explique patrones en tus datos… Cada tipo de modelo tiene una fortaleza particular y es apropiada dentro de cada situación de minería de datos dependiendo de los datos.(Olson & Delen, 2008)

- Evaluar: se deben evaluar los datos mediante la valoración de la utilidad y confiabilidad de los resultados del proceso de minería de datos y estimar qué tan bien se lleva a cabo. Una manera común de evaluar un modelo es aplicarla a una porción de datos, la cual fue creada durante la primera etapa de seleccionar. Si el modelo es válido, deberá funcionar con esta muestra de igual manera que con la muestra usada durante la construcción del modelo. (Turban et al., 2009)

En el proceso de minería de datos los pasos de extraer y preparar los datos ocupa hasta un 80% del tiempo del proyecto, por eso se debe asegurar que se ha extraído los datos correctos y más relevantes.

SEMMA ha sido defendida por el SAS Institute, quien ha lanzado un software de minería de datos que implementa esta metodología. (Refaat, 2007)

CAPÍTULO III. METODOLOGÍA

A continuación se presenta el diseño metodológico, el cual tiene como propósito describir el proceso llevado a cabo para responder las preguntas de investigación y cumplir con los objetivos establecidos.

3.1 ENFOQUE Y MÉTODO

El proceso de investigación se realizó utilizando un enfoque cualitativo. Este fue seleccionado porque el propósito de la misma era comprender un fenómeno o realidad, por ejemplo: el proceso registral. Se recolectó información de naturaleza cualitativa, a la cual no se aplicó mediciones numéricas y se siguió una lógica inductiva durante todo el proceso. Por lo anterior, se determino que mediante un enfoque cualitativo se lograría comprensión del negocio así como también comprensión de los datos tal como la investigación lo requería.

3.2 DISEÑO

El proceso de investigación se realizó de manera no experimental, ya que no se manipuló deliberadamente ninguna de las variables de la investigación. Debido a que no existen estudios previos sobre detección de anomalías en el RPI en Honduras y que la recolección de datos fue en un solo momento (en un tiempo único) se trató de una investigación de tipo exploratoria y transversal.

3.2.1 ESQUEMA Y PROCEDIMIENTO

- Método de Aprendizaje de Algoritmos de Minería de Datos: De los métodos disponibles de aprendizaje se eligió el método no supervisado, básicamente porque no se dispone de un conjunto de datos “etiquetados” que cubra al menos uno de los escenarios posibles (registros considerados normales o registros considerados anomalías) y que sirvan para entrenar el modelo. Además los métodos de

aprendizaje no supervisado permiten identificar escenarios de anomalías desconocidas.

- Criterio de Éxito de Minería de Datos: Se definió como exitoso aquel algoritmo cuyos resultados de detección de anomalías logren un nivel de precisión de al menos 50%. Este criterio varía de un proyecto de minería de datos a otro, para este estudio dicho valor se determino considerando dos factores: se utiliza un método de aprendizaje no supervisado y se trata de un estudio exploratorio. El modelo seleccionado será el que tenga mayor nivel de precisión.

- Metodología de Minería de Datos: Se aplicó CRISP-DM la cual es la metodología generalmente utilizada en los proyectos de minería de datos. Su ciclo de vida se divide en seis etapas: Compresión del Negocio, Comprensión de los Datos, Preparación de los datos, Modelado, Evaluación e Implementación. El proceso de investigación se realizó siguiendo las primeras cinco etapas de la metodología y la ultima se ve reflejada en la sección de aplicabilidad como una guía para la implementación del modelo de minería de datos.

  La primera fase es la **comprensión del negocio**, para lograr esto fue necesario realizar una serie de entrevistas con los usuarios expertos en el manejo del sistema, conocidos como Implementadores; así como con el oficial legal y los inspectores generales. También se realizó una revisión de los informes de apoyo de investigación del PATH y de cierta documentación como ser: manuales de procedimiento, ley de la propiedad y su respectivo reglamento.

  Como resultado de esta fase se tiene: objetivos de minería de datos que en este estudio se plasma en el planteamiento del problema, inventario de recursos que por razones de confidencialidad no puede ser publicado, sin embargo se incluye características relevantes de los recursos y definición del criterio de éxito del proyecto de minería de datos.

La segunda fase es la **comprensión de los datos**, para lograr esto fue necesario analizar el modelo lógico (entidad relación) y físico de la base de datos relacional del SURE, principalmente aquellas entidades que tienen relación con el proceso registral. Se realizaron diferentes actividades, incluyendo: una recolección inicial de datos, análisis y revisión de los objetos (tablas, atributos y sus relaciones), exploración de datos mediante consultas para identificar información adicional sobre atributos de interés, identificación de si faltan atributos o existen atributos con valores nulos o en blanco. Como resultado de esta fase se tienen: reporte inicial de recolección de datos, descripción de los datos y reporte de exploración de los datos. Por motivos de confidencialidad estos no serán incluidos, sin embargo en la sección de análisis y resultado se incluye un diagrama de abstracción de la base de datos incluyendo las entidades relacionadas con el proceso registral.

La tercera fase es la **preparación de los datos**, para lograr esto fue necesario el análisis de datos y crear una estructura de datos que se alimente de la base de datos transaccional a través de un proceso de extracción, transformación y carga de datos conocido como ETL, éste no fue un proceso trivial, por lo que fue necesario el apoyo de personal técnico experto, como ser el administrador de la base de datos. Los datos generados fueron importados a un repositorio dentro del software base de minería de datos y se le aplicaron diferentes operaciones para realizar: filtrado de datos, reducción de datos y conversión de datos. Como resultado de esta fase se tienen: listado de los datos a ser utilizados y listado de atributos derivados, listado de registros generados (no fue necesario su uso en este estudio) y catalogo de esquemas de anomalías.

La cuarta fase es el **modelado**, para lograr esto fue necesario, dentro del software de minería de datos, realizar pruebas con diferentes algoritmos de detección de anomalías basadas en distancia como: KNN, LOF, COF. Se hicieron múltiples pruebas realizando cambios en los parámetros de los algoritmos en busca de una configuración que permitiera descubrir registros anómalos o inusuales. Adicionalmente se probaron variaciones de estos algoritmos como ser CBLOF y

LDCOF, haciendo uso del mismo procedimiento repetitivo de probar la ejecución de los mismos con diferentes parámetros. Como resultado de esta fase se tienen: diseño de las pruebas que se explica más adelante en este capítulo, y modelos generados que se incluyen en el capítulo de análisis y resultados.

La quinta fase es la **evaluación**, para lograr esto fue necesaria la revisión de los puntos identificados como anómalos o inusuales contra el sistema SURE y/o contra los informes de apoyo de investigación y de esta manera determinar de forma manual si estos podían considerarse como anomalías. En esta etapa se determino el nivel de precisión de cada algoritmo, se identifico aquellos que cumplen el criterio de éxito y se selecciono el algoritmo con mayor nivel de precisión. Como resultado de esta fase se tienen: Evaluación de los resultados de minería de datos respecto al criterio de éxito definido (nivel de precisión) y el Modelo Final Seleccionado.

La sexta, y última, fase es la **implementación**, la investigación contempla las primeras cinco fases de la metodología CRISP-DM, por lo que la implementación esta fuera de su alcance, sin embargo en el capítulo de aplicabilidad se incluyo una guía para la implementación del modelo de detección de anomalías.

- Diseño de las pruebas. Las pruebas deben realizarse siguiendo los siguientes pasos:
    - Seleccionar un único conjunto de datos y exportarlo a un archivo CSV (ambos resultado de la fase de preparación de los datos) el cual posteriormente se debe importar al software de minería de datos
    - Aplicar al conjunto de datos importados diferentes algoritmos de detección de anomalías (kNN, LOF, COF, k-Means, CBLOF)
    - Realizar con cada algoritmo pruebas iníciales sobre un subconjunto (muestra aleatoria simple) de datos en los siguientes porcentajes: 5%, 10% y 25% para tener una noción de los tiempos de ejecución requeridos. En estas pruebas no será evaluado el nivel de precisión.

- Realizar con el 100% del conjunto de datos todas las pruebas de los diferentes algoritmos que serán sometidas a evaluación.
    - Realizar para cada algoritmo varias pruebas modificando los parámetros de configuración. La primera prueba será con los parámetros predeterminados.
    - Visualizar para cada prueba en un grafico de dispersión el atributo evaluado versus el grado de anormalidad (atributo “outlierness”)
    - Identificar visualmente las anomalías o mediante el atributo anormalidad aquellas que tengan los valores más altos

3.2.2 POBLACIÓN Y MUESTRA

En el marco del enfoque cualitativo la población fue el Personal Experto en el manejo del SURE, Personal Técnico Experto y Personal de Inspectoría General del IP; ambos grupos asignados en las oficinas de Francisco Morazán, Honduras. En total quince personas, siete personas del PATH: oficial legal, coordinador de modernización de registros, implementador del sistema SURE, administrador de base de datos, oficial de infraestructura, oficial de seguridad de la información y oficial gestor de operaciones; y ocho personas del IP: todos inspectores generales.

El tipo de muestra en ambos casos fue no probabilista o dirigida de tipo homogénea que permita describir un subgrupo en profundidad.

En el marco de la metodología CRISP-DM se hizo uso de un conjunto de datos, cuyas características poblacionales son (ver Figura 11):

Figura 11. Detalle del conjunto de datos a utilizar

Para este estudio se permitió aplicar únicamente sobre los datos históricos del año 2008, por restricciones de la institución. Una consideración adicional a este respecto, es que el uso de datos históricos permitió comprobar casos detectados por el modelo, puesto que ya existe un registro de algunas de las irregularidades que fueron reportados para ese año.

La muestra que fue sometida al proceso de modelado fue igual a la totalidad de registros, es decir igual a la población. Lo anterior se justifica debido a que según Stéphane Tufféry, (2011) el muestreo es aplicable sólo si, por un lado, podemos controlar la representatividad de la muestra, y, por otro lado, si no estamos buscando fenómenos excesivamente poco comunes. Para este caso, sobre la detección de anomalías se requiere tomar todo el conjunto de datos disponibles.

3.3 TÉCNICAS E INSTRUMENTOS APLICADOS

Un instrumento de medición adecuado es aquél que registra datos observables que representan verdaderamente los conceptos o las variables que el investigador tiene en mente. (Hernández, et al, 2010)

En este proceso de investigación se utilizaron las siguientes técnicas e instrumentos: Entrevista semiestructuradas a Expertos con preguntas abiertas (anexo 1 y 2), Cuestionarios con preguntas abiertas y autoadministrado (anexo 3), Observación, Revisión Documental y Datos Secundarios.

3.4 FUENTES DE INFORMACIÓN

3.4.1 PRIMARIAS

Las fuentes primarias son aquellas que proveen un testimonio o evidencia directa sobre el tema de investigación. Las fuentes primarias son escritas durante el tiempo que se está estudiando o por la persona directamente envuelta en el evento. A continuación listamos las fuentes primarias de para este estudio:

- Personal experto del PATH
- Personal Técnico Experto del PATH
- Funcionarios de Inspectoría General del IP
- Informes de apoyo de Investigación de Denuncias
- Manuales de Usuario y Técnicos del SURE
- Documentación Técnica de la base de datos
- Leyes y Reglamentos

3.4.2 SECUNDARIAS

Son fuentes secundarias aquellas que contienen información de fuentes primarias pero de manera sintetizada, reorganiza o reprocesada. Éstas pueden estar disponibles físicamente o de manera digital. A continuación se listan las fuentes primarias para este estudio:

- Libros de Texto
- Artículos
- Encuesta de software base de minería

CAPÍTULO IV. RESULTADOS Y ANÁLISIS

Este capítulo del trabajo de investigación se centra en la presentación de los datos y hallazgos obtenidos mediantes los instrumentos diseñados y técnicas usadas en el estudio. Estos se muestran de acuerdo al orden de las variables de estudio.

La primera sección trata de la Plataforma Tecnológica del SURE, donde se exponen las respuestas del personal técnico expertos en la misma.

En la segunda sección se aborda la Estructura de Almacenamiento de Datos, donde se describe en forma general la misma en base a las respuestas proporcionadas por el Administrador de Base de Datos y en base a lo observado en esta investigación.

El Proceso Registral se aborda en la tercera sección, donde se detallan los hallazgos sobre las clases de anomalías que se pueden presentar dentro del contexto del SURE, las cuales fueron extraídas de las entrevistas con el personal experto tanto de PATH como del IP.

La cuarta sección refleja la comparación de características de cuatro productos o software de minería de datos.

Finalmente, en la quinta sección se muestra la funcionalidad del software de minería de datos, así como el uso de los algoritmos de detección de anomalías aplicados en la identificación de un tipo de anomalía que se da en el RPI.

Es necesario mencionar que se siguieron lineamientos de privacidad y confidencialidad solicitados por el PATH, con respecto a la no divulgación de información restringida.

4.1 PLATAFORMA TECNOLÓGICA DEL SURE

Como parte de la etapa de Comprensión del Negocio de la metodología CRISP-DM, es necesario conocer la situación actual, para ello fue necesario conocer la Infraestructura y la Base de Datos del SURE.

4.1.1 INFRAESTRUCTURA DEL SURE

La infraestructura agrupa y organiza un conjunto de elementos tecnológicos que integran el SURE, entre ellos los firewalls, servidores de publicación, servidores de aplicación, de manejo de información geográfica así como de base de datos. Todos ellos soportan las operaciones del mismo.

A continuación mencionamos lo aspectos relevantes de la misma:

- Actualmente cuenta con más de 25 servidores de diferentes marcas para las funciones de publicación del portal SINAP - SURE- SINIT-RENOT, manejo de dominio, base de datos y almacenamiento de imágenes escaneadas, entre otras.
- La infraestructura de servidores se ha organizado por bloques, es decir servidores de base de datos, de aplicaciones, etc.
- Cuenta con una arquitectura modular, que facilite el crecimiento futuro de la planta de servidores.
- Posee sistemas contingentes contra fallas de energía eléctrica y facilidad para restauración inmediata del servicio.
- Sistemas de alta eficiencia energética que permitan una operación con bajos costos de energía para el PATH.
- Para aprovechar los recursos de hardware el PATH recientemente ha iniciado un proceso de virtualización de varios de sus equipos.
- Los sistemas operativos empleados son: Windows y Red Hat Linux.

- La planta de equipos es administrada por 5 personas, donde 3 de ellas se dedican al soporte y mantenimiento de los mismos y 2 se dedican a la administración y configuración.

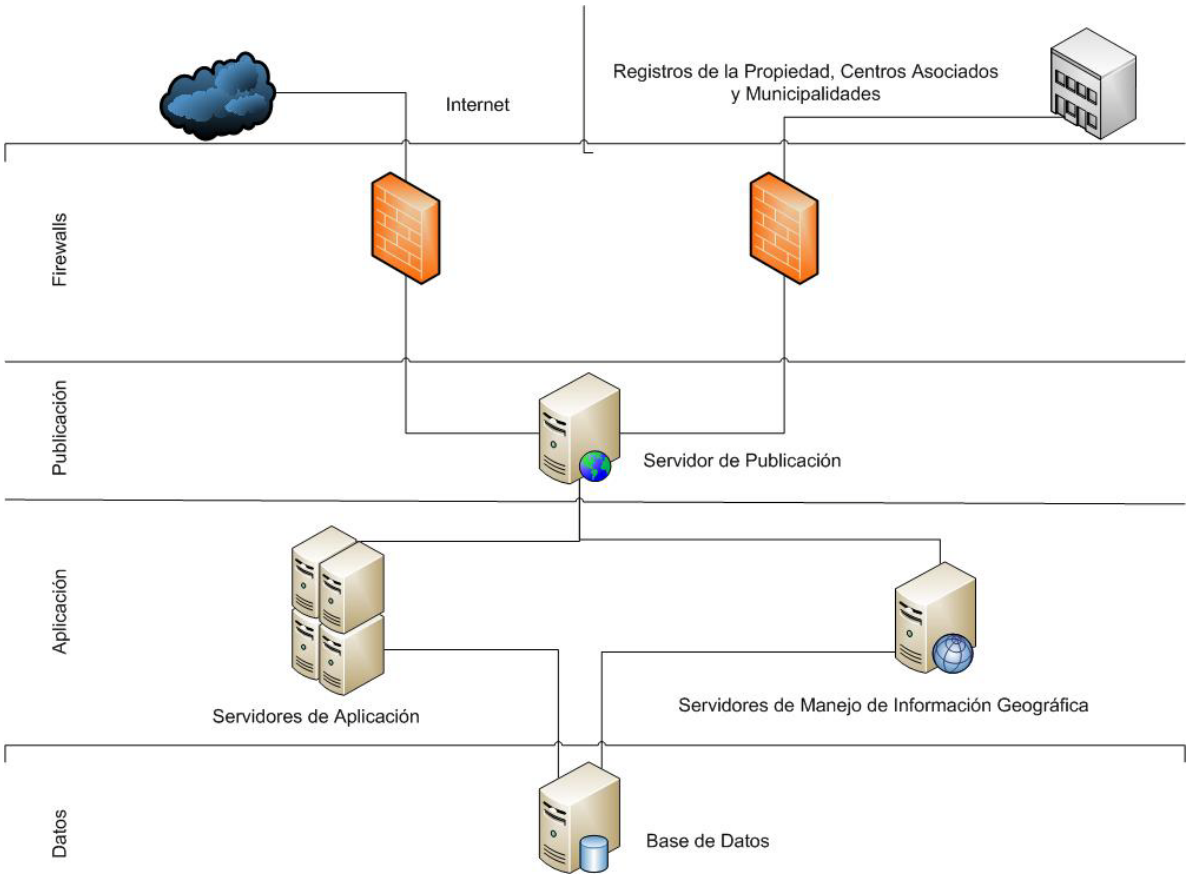


Figura 12. Esquema básico de la Plataforma Tecnológica del SURE

Se encontró que el SURE opera en ambiente web por medio de navegadores. Su infraestructura está basada en tecnologías ampliamente reconocidas y probadas en el mercado, lo que la hace robusta, solida y le permite operar de manera eficiente.

Esta fue diseñada para soportar las demandas actuales y futuras requeridas para la función crítica de este sistema con niveles altos de servicios y prestaciones.

4.1.2 BASE DE DATOS DEL SURE

Con respecto a la Base de datos del SURE, se encontró que se trata de una base de datos relacional, es decir que esta permite establecer relaciones entre los datos a través de conexiones previamente establecidas.

El desempeño de la misma esta estrictamente vigilado por un Administrador de Base de Datos.

Se encontró que a la fecha esta base de datos tiene registradas más de 900,000 presentaciones de documento del RPI.

4.2 ESTRUCTURAS DE ALMACENAMIENTO DE LOS DATOS

La fase de Comprensión de los Datos permitió conocer la estructura de almacenamiento de datos. En ese sentido se realizó una exploración del diagrama entidad relación del SURE, así como la inspección de las tablas

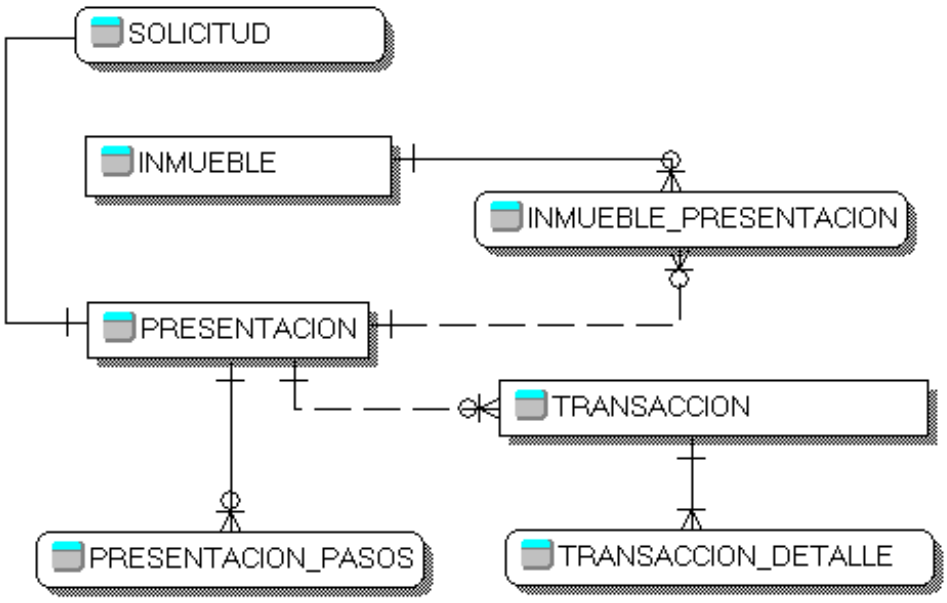


Figura 13. Abstracción de la Base de Datos del SURE

La figura muestra algunas de las entidades que componen la Base de Datos del SURE, las cuales se encuentran íntimamente relacionados a las funciones del RPI. No presentamos más detalle de los componentes de la misma por razones de confidencialidad.

Se encontró en su mayoría tipos de datos primitivos, es decir aquellos que almacenan directamente el valor. Entre ellos se encuentran:

- Carácter: es el tipo de datos que puede contener letras (a-z), números (0-9) y símbolos (#, $,&, etc.)
- Numéricos: es el tipo de datos que puede contener números reales o enteros.
- Booleano: es el tipo de datos que puede contener valores lógicos (verdadero o falso).
- Fecha: tipos de datos que pueden contener fechas.

4.3 PROCESO REGISTRAL

Como se menciono en el Marco Teórico este proceso comienza con la presentación de la escritura en las oficinas del RPI y finaliza hasta que esta procede a ser inscrito o denegado provisionalmente o definitivamente.

Durante este proceso se pueden dar errores, omisiones e irregularidades que de alguna forma afecten a los propietarios de bienes inmuebles.

4.3.1 ESQUEMAS DE ANOMALÍAS

Esta parte del estudio consiste precisamente en la observación y clasificación de aquellas situaciones atípicas que se ha presentado dentro del contexto del SURE y representan una irregularidad en el proceso registral.

Por lo tanto se procedió a la recopilación "esquemas de anomalías", las cuales no son más que una noción más abstracta y general por la cual la irregularidad es reconocida.

Determinar estos esquemas de anomalías fue una de las mayores dificultades que afrontamos en el estudio, puesto que requiere de mucha experiencia en materia del Proceso Registral, para determinar qué aspectos se encuentran dentro de la legalidad y cuáles no, es por ellos que recurrimos a expertos tanto del PATH como del IP.

Durante las entrevistas los expertos del PATH, nos indicaron que no existe una categorización documentada de estos esquemas. Ahondando mas en el tema los expertos nos revelaron diferentes situaciones que han observado durante el cumplimiento de sus funciones y que representan un mal uso o abuso del sistema y que finalmente se convierte errores o irregularidades que afectan a los propietarios de bienes inmuebles.

Ante la interrogante de los tipos o categorías de situaciones irregulares que se presentan en mayor grado nos manifestaron los casos comunes, los cuales se resumieron y se describen en la Tabla 2.

Tabla 2: Esquemas de Anomalía más frecuentes descritos por Expertos PATH

Anomalía	Descripción
Rompimiento del Orden de Prelación	Lo que por ley se conoce primero en tiempo, primero en derecho. Se trata acerca de que se debe dar trámite a la presentación que llego primero a la ventanilla, en caso de que haya más de una presentación que afecte el mismo inmueble.
Transacciones fuera de Horario Laboral	Transacciones realizadas fuera del horario laboral (8:00am-4:00pm) o en días feriados no competen a las actividades normales de la institución, por lo tanto se consideran irregulares.
Tramite inmediato de presentaciones	Se trata de que la presentación sea inscrita en pocos días o incluso se le dé trámite el mismo día.

Otras formas de irregularidades que se presentan pero en menor grado son las siguientes:

Tabla 3: Esquemas de Anomalía menos frecuentes descritos por Expertos PATH

Anomalía	Descripción
Ausencia de bases cartográficas	Inmuebles sin referencia catastral, lo cual da lugar a descripción de la ubicación y aéreas del inmueble ambiguas, y que afecta a los inmuebles contiguos.
Afectaciones de antecedentes erróneas	Que se haga referencia a documentos anteriores cuyos inmueble no coincide con el inmueble que se eta afectando en la presentación actual. Por lo cual se presentan cambios de antecedentes.
Asignación errónea de los derechos de propiedad	Que se asigne los derechos de posesión a otra persona.
Uso inadecuado de claves de usuario	Uso de Claves de usuario de personal retirado de su cargo, la cuales no fue solicitada su cancelación.

En cuanto a los funcionarios de Inspectoría General (IG), nos describieron situaciones que han causado quejas o denuncias por parte de las personas afectadas, ya que ellos no cuentan con un procedimiento formal (documentado) para la búsqueda de situaciones que representen una situación fraudulenta.

Se elaboró un catalogo que resume los casos detallados agrupando aquellos que fueron mencionados de manera coincidente por los Inspectores Generales del IP. El catalogo se muestra en la Tabla 4:

Tabla 4: Catalogo de Esquemas de Anomalías reportadas por IG

Esquema	Descripción
Presentaciones sin primer escaneo	Presentaciones que no cuentan con imagen del primer escaneo que permita la verificación de los medios digitales y físicos.
Reingreso de presentaciones que cuentan con múltiples denegatorias	Presentaciones que se deniegan por no coincidir con lo registrado o no cumplir con los requisitos. Cuando esto se da de forma consecutiva indica una irregularidad.
Presentaciones inscritas de forma expedita o inmediatas	Presentaciones que se inscriben en un tiempo menor al tiempo normal establecido para cada servicio registral. A esta se les denomina comúnmente "rapidonas".
Uso del mismo instrumento (escritura) para varias presentaciones	Detección de reingreso de documentos denegados múltiples veces con otro número de presentación utilizando los siguientes criterios para su verificación: numero de instrumento, abogado y fecha.
Comparecen participantes fallecidos	Que en la presentación se detalle persona que ha fallecido y esta no se verifique en el Registro de Personas.
Da fe notario fallecido	Que en la presentación se detalle un notario que ha fallecido y esta no se verifique en el Registro de Notarios.
Transacciones fuera de horario laboral	Aquellas transacciones que se realizan fuera de las 8:00 am que abre el diario automatizado de registro y después de las 5:00pm.
Cambio del área de un inmueble sin transacción q justifique.	Cambios de área injustificados, es decir que no se presente una remedida, partición, reunión u otro servicio registral que afecte el área del mismo.
Doble registro de un inmueble, tanto bajo la técnica de folio real como de folio personal	Inmuebles cuyo tracto ya fue cerrado en folio personal y se traslado a folio real, pero que aun sigue siendo afectado por transacciones en folio personal, haciendo caso omiso al traslado

Asimismo Inspectoría General del IP solicita al PATH apoyo para realizar investigaciones sobre casos de irregularidades, los cuales han sido reportados por las personas afectadas mediante una queja o denuncia a este ente. Durante el año 2012 el PATH recibió 16 solicitudes de apoyo a investigaciones.

A través de la revisión de los informes elaboramos una clasificación de los mismos y a continuación se muestra la tabla resumen.

Tabla 5: Esquemas de Anomalías de los informes de Apoyo de Investigación

Descripción	Cantidad
Tramite inmediato	3
Varias irregularidades en el mismo caso	3
Problemas con los derechos del inmueble	2
Imágenes no escaneadas	2
No corresponden al uso del sistema	2
Se rompe el Orden de prelación	2
Antecedentes incorrectos	1
Cambio de los datos generales del inmueble (área y ubicación)	1
Total casos	**16**

Como se puede observar las presentaciones de trámite inmediato es una de las anomalías que mas presentan. Los expertos nos mencionaron que estas situaciones se presentan cuando hay conflictos derechos de un bien inmueble. Asimismo como se observa en la tabla hay casos en los que varias anomalías se presentan juntas.

4.4 SOFTWARE DE MINERÍA DE DATOS

La oferta de software de Minería de Datos se ha incrementado en los últimos años. Numerosas empresas han desarrollado su propio software de minería de datos para responder a las necesidades actuales de las empresas, ya sea para la obtención de información, así como la búsqueda de patrones en sus datos, que les permita tomar decisiones, mejorar sus procesos, etc.

En esta sección incluimos la revisión de algunos de los productos (software de minería de datos), de los cuales examinamos sus principales características tomando como base a los lineamientos proporcionados por los expertos en la Plataforma Tecnológica del SURE.

4.4.1 CONSIDERACIONES PARA LA REVISIÓN DEL SOFTWARE DE MINERÍA DE DATOS

Antes de elegir un software es necesario identificar con exactitud qué es lo que se requiere. De acuerdo a le entrevista con el personal técnico que administra la Plataforma Tecnológica del SURE, se determinaron los siguientes requisitos básicos o mínimos que deberá poseer el software, en base a sus respuestas. Estos son:

- Características Generales:
    - Reconocimiento del Producto (Popularidad de uso)
    - Precio: Deberá ser gratuito

- Características Técnicas:
    - Compatibilidad Java
    - Compatibilidad con el motor de base de datos existente
    - Compatibilidad con Windows

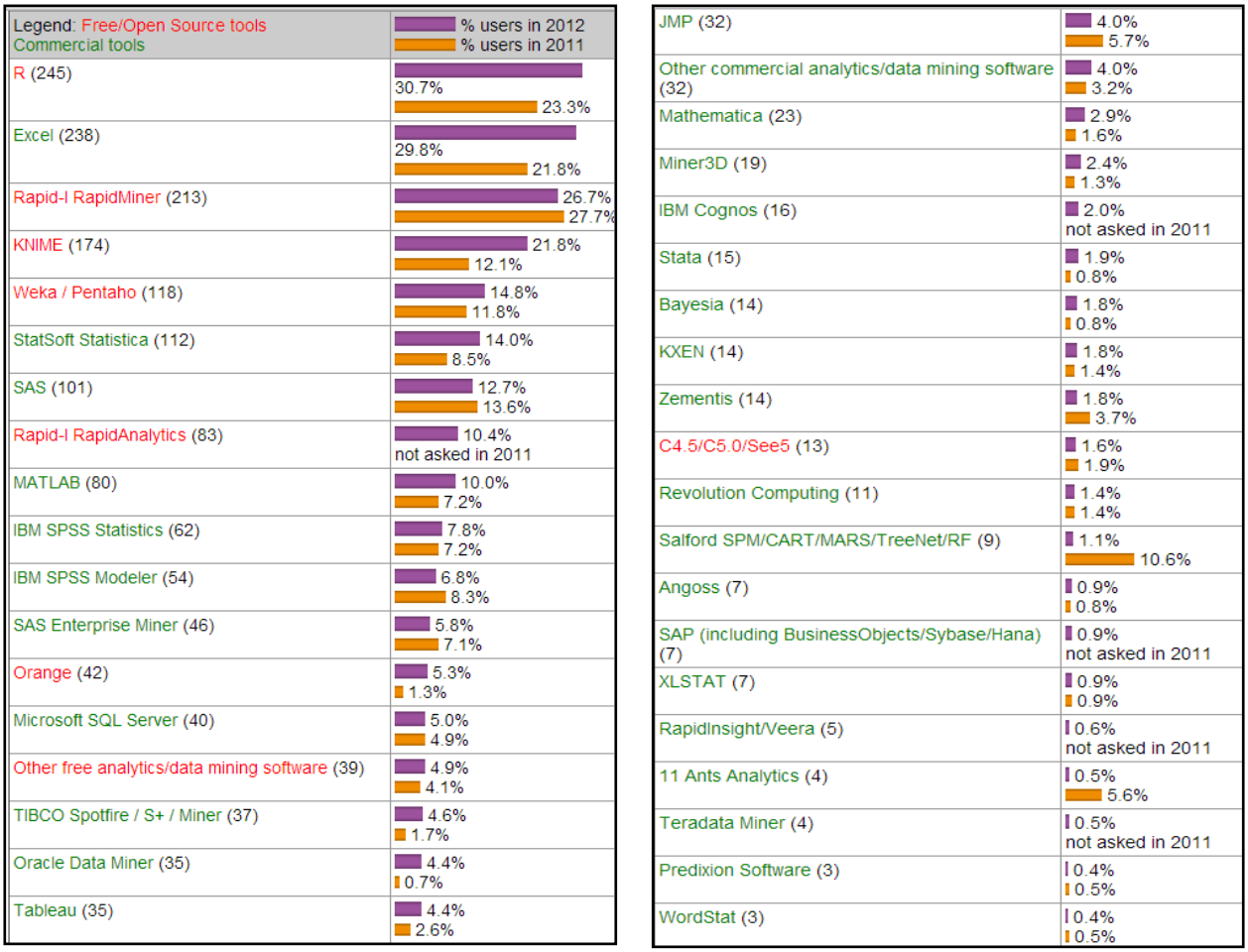
- Características Funcionales:
    - Facilidad de Uso
    - Interfaz Gráfica
    - Facilidad de Instalación

- Características de Soporte
    - Fuentes de consulta y Documentación en línea
    - Disponibilidad de Cursos y capacitaciones

4.4.2 REVISIÓN DE CARACTERÍSTICAS GENERALES DEL SOFTWARE

En nuestros días existen varias opciones de software para aplicar minería de datos, por lo tanto es necesario conocer la el reconocimiento del producto a través de la popularidad de uso del mismo. Para ello recurrimos a la Comunidad de minería de datos denominada KDnuggets.

KDnuggets realizó una encuesta en línea de los productos más conocidos y empleados por la comunidad científica y empresarial para uso en proyectos reales de minería de datos. Concretamente, se estudio 38 aplicativos relacionadas con la minería de datos, ya sea comercial o gratuito. Los resultados de la misma se observan en la figura 14.



Legend: Free/Open Source tools Commercial tools	% users in 2012 % users in 2011
R (245)	30.7% 23.3%
Excel (238)	29.8% 21.8%
Rapid-I RapidMiner (213)	26.7% 27.7%
KNIME (174)	21.8% 12.1%
Weka / Pentaho (118)	14.8% 11.8%
StatSoft Statistica (112)	14.0% 8.5%
SAS (101)	12.7% 13.6%
Rapid-I RapidAnalytics (83)	10.4% not asked in 2011
MATLAB (80)	10.0% 7.2%
IBM SPSS Statistics (62)	7.8% 7.2%
IBM SPSS Modeler (54)	6.8% 8.3%
SAS Enterprise Miner (46)	5.8% 7.1%
Orange (42)	5.3% 1.3%
Microsoft SQL Server (40)	5.0% 4.9%
Other free analytics/data mining software (39)	4.9% 4.1%
TIBCO Spotfire / S+ / Miner (37)	4.6% 1.7%
Oracle Data Miner (35)	4.4% 0.7%
Tableau (35)	4.4% 2.6%

JMP (32)	4.0% 5.7%
Other commercial analytics/data mining software (32)	4.0% 3.2%
Mathematica (23)	2.9% 1.6%
Miner3D (19)	2.4% 1.3%
IBM Cognos (16)	2.0% not asked in 2011
Stata (15)	1.9% 0.8%
Bayesia (14)	1.8% 0.8%
KXEN (14)	1.8% 1.4%
Zementis (14)	1.8% 3.7%
C4.5/C5.0/See5 (13)	1.6% 1.9%
Revolution Computing (11)	1.4% 1.4%
Salford SPM/CART/MARS/TreeNet/RF (9)	1.1% 10.6%
Angoss (7)	0.9% 0.8%
SAP (including BusinessObjects/Sybase/Hana) (7)	0.9% not asked in 2011
XLSTAT (7)	0.9% 0.9%
RapidInsight/Veera (5)	0.6% not asked in 2011
11 Ants Analytics (4)	0.5% 5.6%
Teradata Miner (4)	0.5% not asked in 2011
Predixion Software (3)	0.4% 0.5%
WordStat (3)	0.4% 0.5%

Figura 14. Resultados de la encuesta realizada por KDnuggets - Mayo 2012

Fuente: http://www.kdnuggets.com/2012/05/top-analytics-data-mining-big-data-software.html (s.f.)

Esta figura muestra que el software gratuito para minería de datos está ubicado en los primeros renglones demostrando que en la actualidad el número de usuarios de aplicaciones gratuitas esta aumentado.

Tomando este nivel de popularidad determinamos que los productos gratuitos a revisar con mayor profundamente son los siguientes:

- R DM (Anexo 5)
- Rapid Miner (Anexo 6)
- KNIME (Anexo 7)
- Weka (Anexo 8)

4.4.3 REVISIÓN DE CARACTERÍSTICAS TÉCNICAS DEL SOFTWARE

Para cada una de las aplicaciones que cumplieron con las características generales, se reviso especificaciones técnicas disponibles en sus sitios web, lo cual permitió determinar el cumplimiento de los requisitos expresados por los expertos del Plataforma Tecnológica del SURE. Concretamente, las características que debían cumplir fueron: compatibilidad con Windows en sus diferentes versiones, Java y con el gestor de base de datos existente.

La Tabla 6 presenta los resultados obtenidos en revisión de especificaciones de cada producto. Con una cruz se indica el cumplimiento de la característica asociada.

Tabla 6: Revisión Características Técnicas del Software de Minería de Datos

Software de Minería de Datos	Java	Gestor de Base de Datos existente	Windows
RDataMining	X	X	X
RAPID|MINER	X	X	X
KNIME	X	X	X
WEKA The University of Waikato	X	X	X

De esta inspección encontramos que todos los productos examinados cumplen con las características técnicas requeridas por los expertos SURE.

4.4.4 REVISIÓN DE CARACTERÍSTICAS FUNCIONALES DEL SOFTWARE

Para realizar esta evaluación se procedió a obtener una copia del software e instalarla así como también observas sus funcionalidades.

La tabla 7 presenta los resultados obtenidos en el análisis de cada producto. Con una cruz se indica el cumplimiento de la característica asociada y con un guion se representa el incumplimiento de la misma.

Tabla 7: Revisión Características Funcionales del Software de Minería de Datos

Software de Minería de Datos	Posee interfaz gráfica	Facilidad de Uso	Facilidad de instalación
RDataMining	--	--	X
RAPID MINER	X	X	X
KNIME	X	X	X
WEKA The University of Waikato	X	--	X

Como se puede observar la mayoría de los productos examinados poseen una interfaz grafica. En cuanto al aspecto de facilidad de uso, encontramos que R y Weka son mas orientadas a ser editores de código, mientras que Rapid Miner y KNIME permiten el desarrollo de los procesos de análisis de datos mediante el encadenamiento de operadores a través de su interfaz grafica, lo que permite al usuarios realizar dichas operaciones de manera más intuitiva sin tener que llevar a cabo un aprendizaje de una gran cantidad de comandos.

4.4.5 REVISIÓN DE CARACTERÍSTICAS DE SOPORTE DEL SOFTWARE

Las empresas proveedoras de software generalmente ofrecen servicios de consulta y asesoría, y capacitación, la cual es necesaria para el aprovechamiento del mismo.

Tabla 8: Características de Soporte del Software de Minería de Datos

Software de Minería de Datos	Documentación en Línea	Capacitaciones
RDataMining	X	X
RAPID|MINER	X	X
KNIME	X	X
WEKA The University of Waikato	X	--

De esta inspección encontramos que tanto RapidMiner como Knime cumplen con las características de soporte requeridas por los expertos SURE.

4.4.6 SELECCIÓN DEL SOFTWARE DE MINERÍA DE DATOS

Durante esta revisión comparativa de las características del software de minería de datos se encontró que RapidMiner es un software de código abierto y gratuito, líder a nivel mundial para la minería de datos que permite utilizar una amplia variedad de técnicas descriptivas y predictivas.

Por los cual es apto, compatible y cumple con las características proporcionadas por les técnicos expertos en la Plataforma Tecnológica del SURE.

Tabla 9: Resumen de Evaluación del Software de Minería de Datos

Software de Minería de Datos	Popularidad	Precio	Java	Motor Base de Datos	Windows	Interfaz Grafica	Facilidad de Uso	Facilidad Instalación	Documentación	Entrenamiento	Puntaje
R Data Mining	4	1	1	1	1	0	0	1	1	1	11
RapidMiner	3	1	1	1	1	1	1	1	1	1	12
Knime	2	1	1	1	1	1	1	1	1	1	11
Weka	1	1	1	1	1	1	0	1	1	0	8

4.5 PROTOTIPO DEL MODELO DE MINERÍA DE DATOS

En esta sección se muestra la funcionalidad del software de minería de datos, así como el uso de los algoritmos de detección de anomalías aplicados en la identificación de un tipo de anomalía que se da en el RPI.

4.5.1 PREPARACIÓN DE LOS DATOS

En este caso la detección de anomalías se empleo para identificar aquellas presentaciones de escrituras cuyo tiempo de tramite inmediato. Se selecciono este tipo de anomalía pues se encontró que es uno de los más frecuentes.
El conjunto de datos empleado para este caso contenía presentaciones de escrituras cuyo tiempo de trámite es normal, así como también casos en el que el tiempo es menor al establecido dentro de los procedimientos.

El primer paso fue depurar un conjunto de datos que inicialmente contaba 9 atributos provenientes de 6 tablas diferentes de la base de datos. Posteriormente se reviso y evaluó la relevancia de los mismos para este ejercicio, excluyéndose 4 atributos del conjunto original. Otros de los atributos se emplearon para operaciones donde era necesario obtener información resumida. De acuerdo a estos resultados también se excluyeron registros de presentaciones que solo habían realizado el primer paso del proceso pero nunca se les había dado continuidad a su trámite.

Finalmente el conjunto de datos quedo estructurado de la siguiente forma:

- Identificador único de la Presentación de Escritura
- Fecha en que fue ingresada al sistema
- Fecha en que se finalizo el trámite de dicha presentación
- Total de días tramite de la presentación

Los datos se colocaron en el orden y formato necesarios para alimentar el software base de minería de datos, donde se aplicaron diferentes algoritmos. Toda la información concerniente a la Preparación de los Datos fue registrada en un formato

específicamente diseñado para este propósito, al cual se le denomino “Registro de Esquemas de Anomalías”. El mismo se presenta en el anexo 4 e incluye una descripción del esquema de anomalía y los datos técnicos necesarios para la preparación del conjunto de datos que alimentará el software de minería de datos. Este formato permite replicar este proceso las veces que sea necesario tomándole como guía.

4.5.2 MODELADO: USO k-NN GLOBAL ANOMALY SCORE

Se realizaron varias pruebas con el algoritmo k-NN, en primera instancia se probó con los valores predeterminados (por el software base k=10), luego se fue incrementando los valores (k=15, k=25, k=40) y observando las graficas pudimos identificar que a pesar de incrementar el valor de k siempre obteníamos los mismos resultados; la unica diferencia observada entre las diferentes pruebas eran que los tiempos de ejecucion de se iban incrementando, por lo que al final se configuró el modelo con los valores predeterminados.

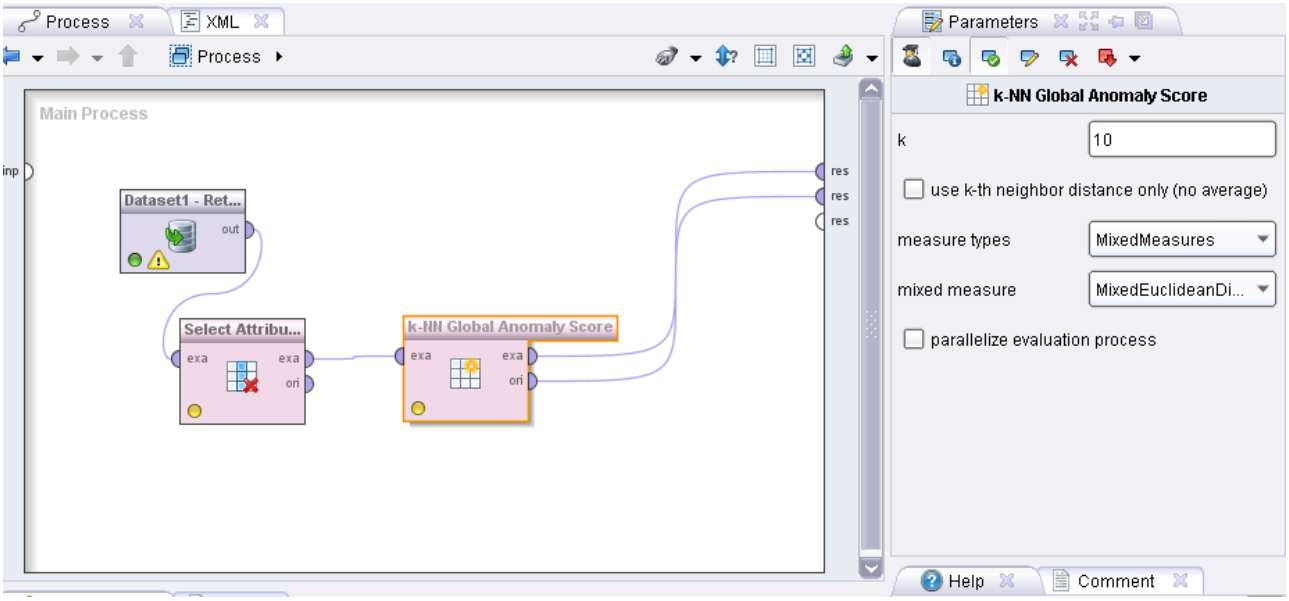


Figura 15. Modelado k-NN (valores por defecto)

De los datos resultantes se observó que el valor promedio del atributo DÍAS (días trámite) es de 61.552 y los valores individuales se encuentran en un rango ente 1 y 1191.910, aunque la mayoría de los datos se encuentran en un rango entre 1 y 415

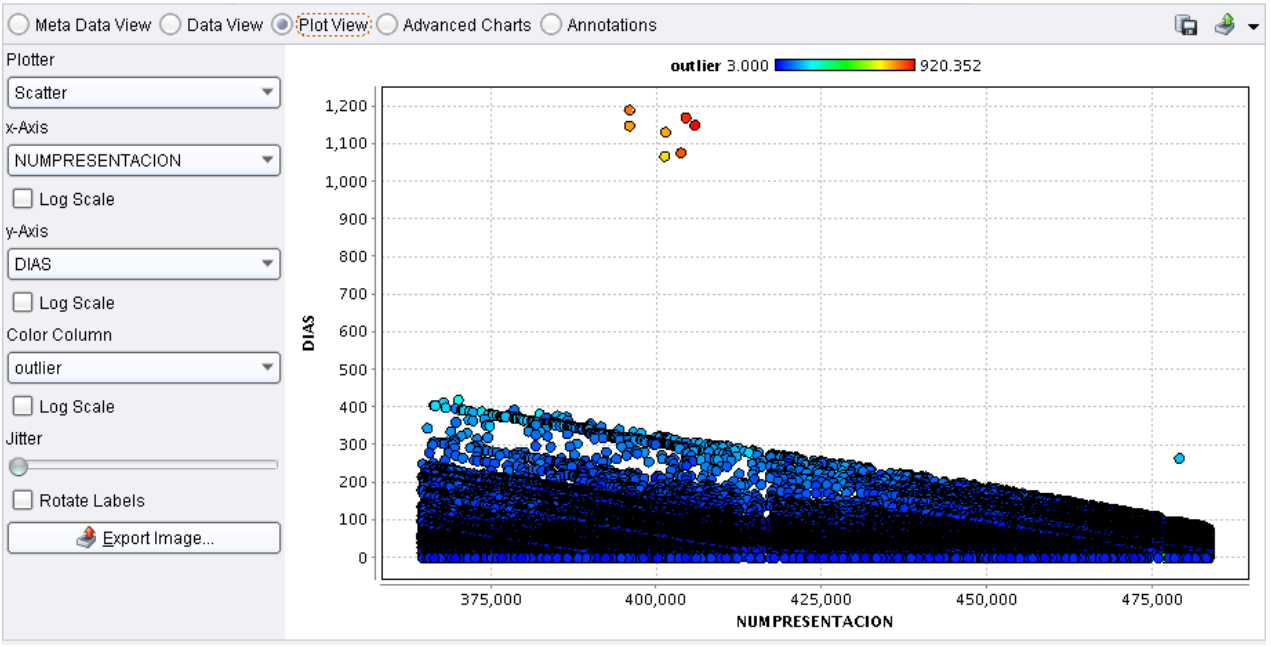
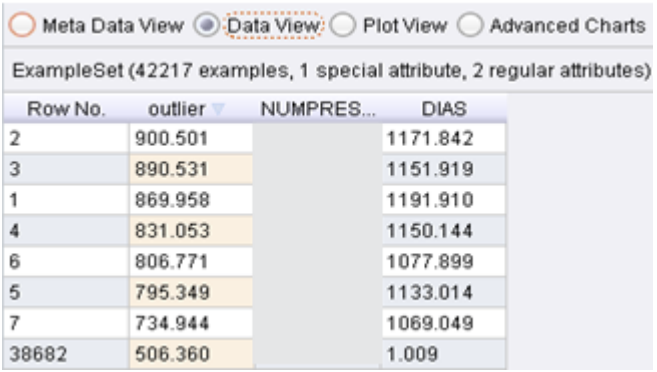


Figura 16. Prueba #1 k-NN (gráfico de dispersión)

En el gráfico se destacan las anomalías que son los elementos que se encuentran alejados del grupo. Los mismos fueron consultados en el sistema y se determinó que no solo presentaban tiempos de resolución inusuales sino que también presentaron irregularidades dentro de los pasos del proceso registral.



Row No.	outlier	NUMPRES...	DIAS
2	900.501		1171.842
3	890.531		1151.919
1	869.958		1191.910
4	831.053		1150.144
6	806.771		1077.899
5	795.349		1133.014
7	734.944		1069.049
38682	506.360		1.009

Figura 17. Extracto de Vista de Datos Prueba k-NN

En la vista de datos (figura 17) se puede observar el listado de los casos con valores atípicos representados en las gráficos y cuyo atributo “outlier” contiene los valores más altos de irregularidad del conjunto de datos procesados.

4.5.3 MODELADO: USO DEL LOF LOCAL OUTLIER FACTOR

Se realizaron varias pruebas con el algoritmo LOF, en primera instancia con los valores predeterminados (por el software base kmin=10 y kmax=20) como se muestra en la figura 18.

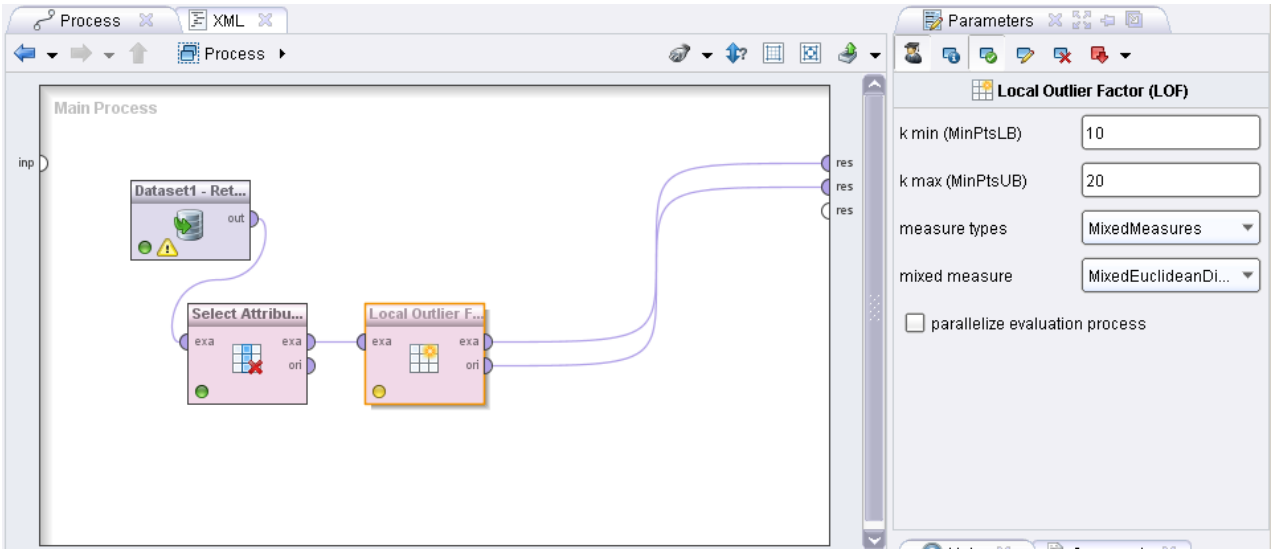


Figura 18. Modelado con LOF (valores por defecto)

Luego se fue incrementando los valores (kmin=10 y kmax=25; kmin=10 y kmax=30; kmin=15 y kmax=30) y observando las graficas pudimos identificar que al de incrementar los valores de kmin y kmax las variaciones en los resultados eran mínimas.

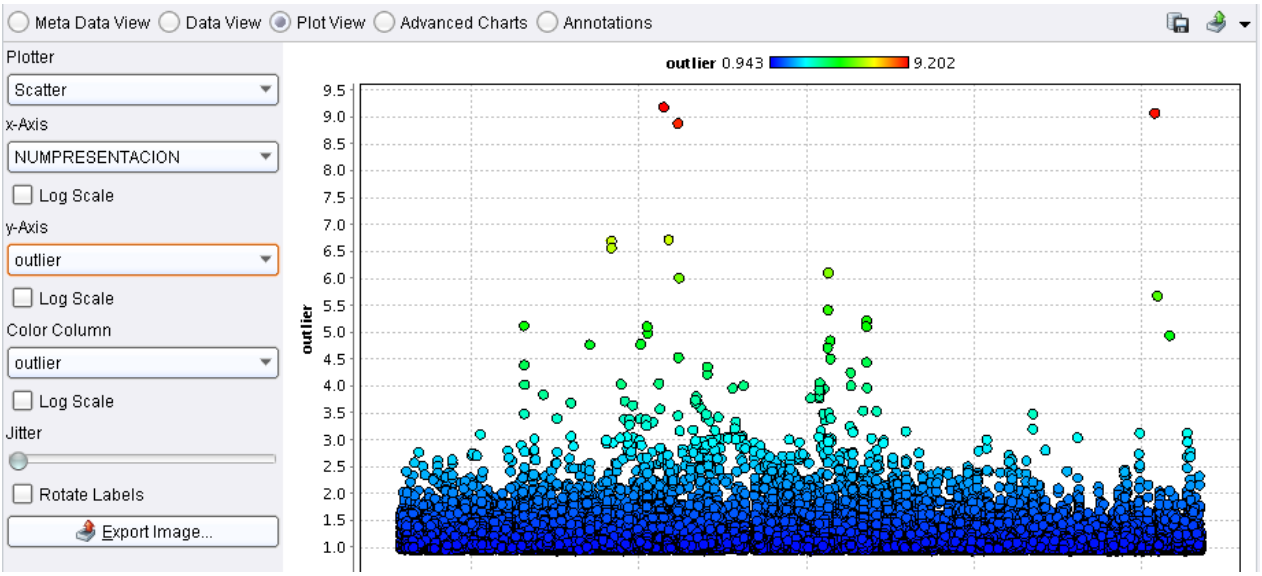


Figura 19. Prueba #1 LOF (gráfico de dispersión)

Se observó que con la con la modificación de los parámetros, la base del gráfico se expandía (figura 19) según se incrementaba la diferencia entre kmin y kmax e igualmente los valores atípicos "medios" aumentaron no así los valores atípicos extremos que seguían siendo los mismos, por lo que se determino utilizar la configuración de parámetros predeterminada.

Meta Data View | Data View | Plot View | Advanced Charts

ExampleSet (42217 examples, 1 special attribute, 2 regular attributes)

Row No.	outlier	NUMPRES...	DIAS
38682	9.498		1.009
1	6.362		1191.910
2	6.292		1171.842
3	6.153		1151.919
4	6.099		1150.144
38290	5.848		3.076
6	5.703		1077.899
526	5.400		265.781

Figura 20. Extracto de Vista de Datos LOF

4.5.4 MODELADO: USO DE CONNECTIVITY-BASED OUTLIER FACTOR (COF)

Igualmente que con los algoritmos anteriores con COF se realizaron varias pruebas iniciando con los valores predeterminados y modificándolos hasta determinar una configuración de parámetros adecuada para el conjunto de datos.

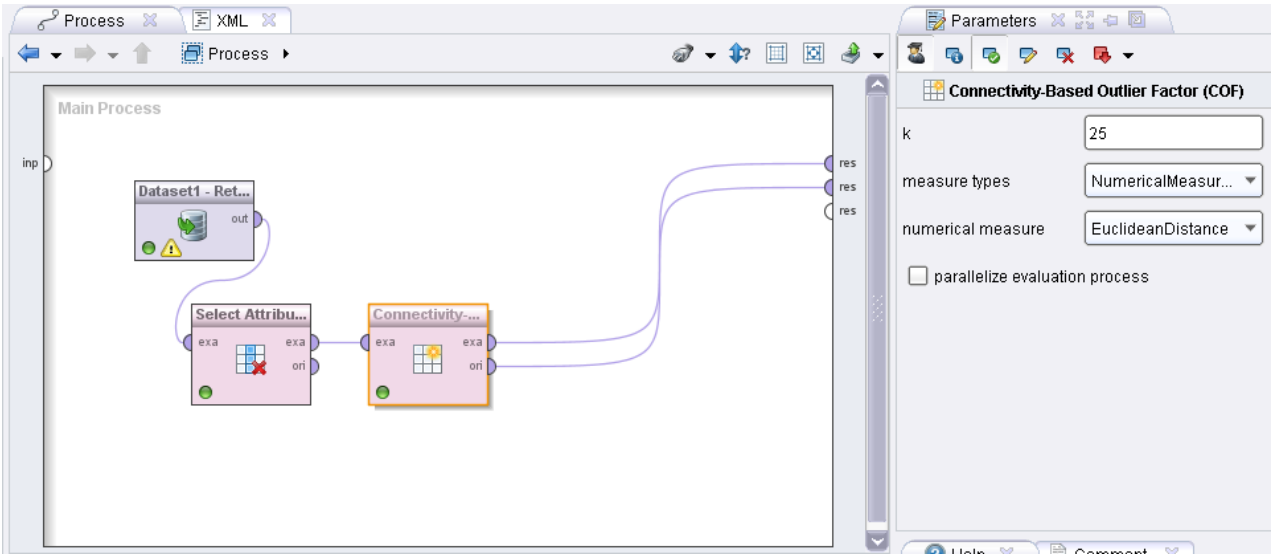


Figura 21. Modelado con COF

En la figura se muestra el modelado con COF, donde para este algoritmo los valores por defecto no eran los idóneos ya que inicialmente se tenían identificados muchos registros como anomalías a medida que aumentamos el valor de K se pudo llegar a un número de anomalías que correspondiera con los resultados arrojados con algoritmos anteriores.

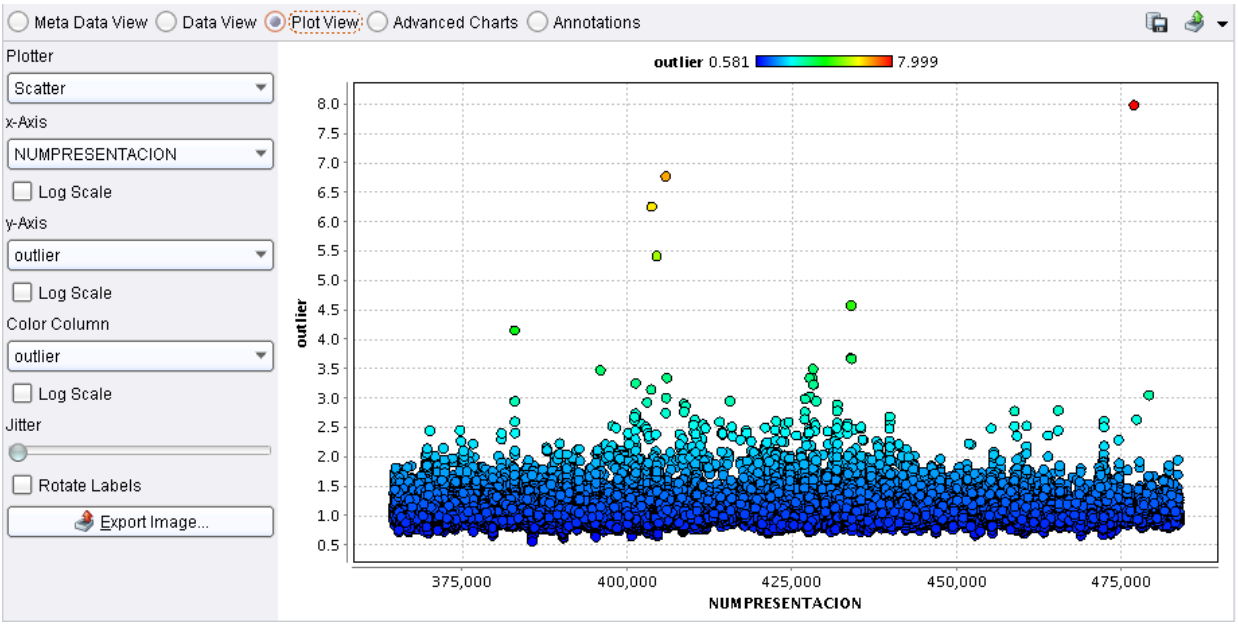


Figura 22. Prueba COF (gráfico de dispersión)

La figura muestra el gráfico resultante de COF el cual es similar a LOF, aunque la cantidad de anomalías disminuye.

Al revisar en el SURE se puedo constatar que en diferentes modelos los mismos casos están siendo catalogados como anomalías.

Meta Data View | Data View | Plot View | Advanced Charts

ExampleSet (42217 examples, 1 special attribute, 2 regular attributes)

Row No.	outlier	NUMPRES...	DIAS
38682	8.489		1.009
2	5.997		1171.842
6	5.821		1077.899
3	5.577		1151.919
1	4.217		1191.910
4	4.217		1150.144
5562	4.004		117.025
526	3.998		265.781

Figura 23. Extracto de Vista de Datos de COF

4.5.5 MODELADO: USO DE k-MEANS

Se realizaron varias pruebas con el algoritmo k-means, en primera instancia se probó el con los valores predeterminados (por el software base k=2; max runs =10). Luego se fue incrementando el número de conglomerados (k=4, k=6, k=10, k=12, k=20).

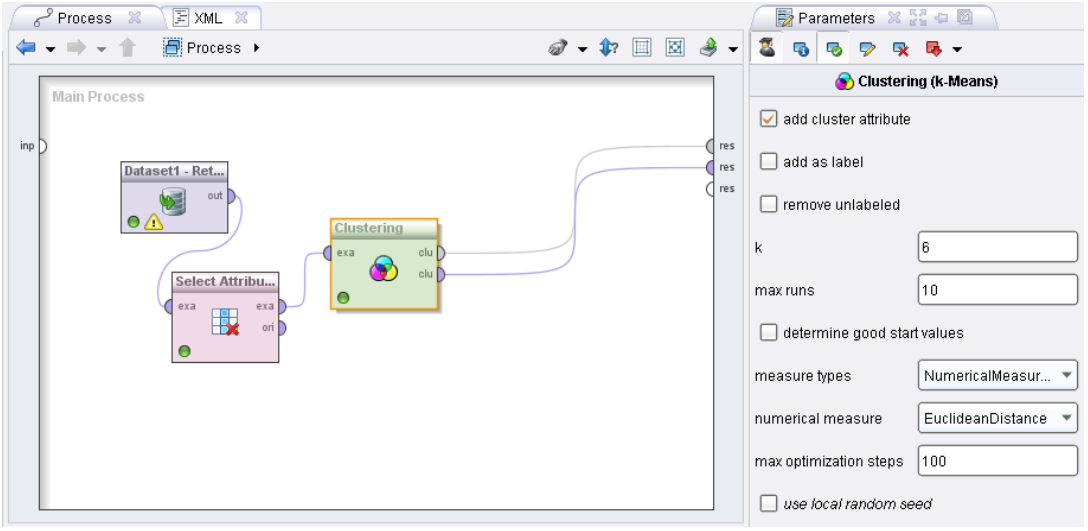


Figura 24. Modelado con k-Means (k=6)

Al modificar los parámetros para este algoritmo como se muestra en la figura 25, pudimos identificar que para nuestro conjunto de datos aumentar el número de conglomerados no tenia impacto alguno en los resultados.

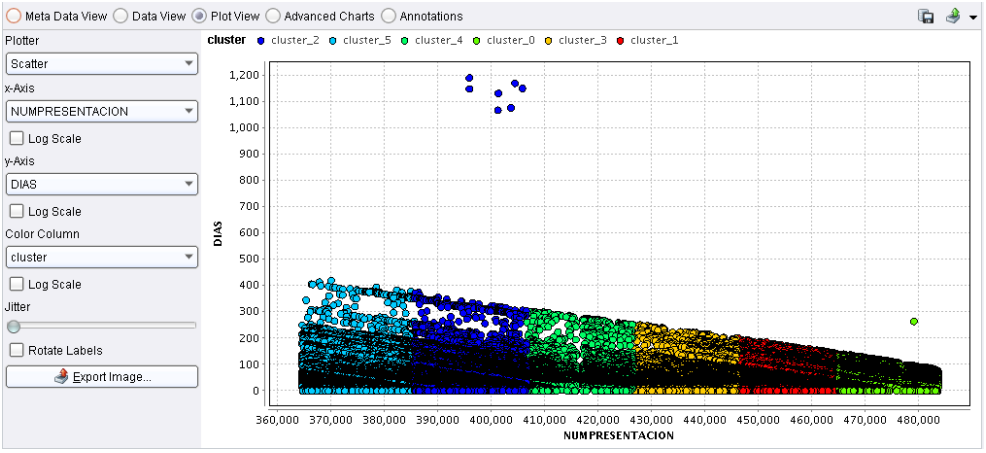


Figura 25. Prueba #1 k-Means (gráfico de dispersión)

La grafica de dispersión muestra el comportamiento que se ha venido dando a lo largo de las pruebas con los diferentes algoritmos, es decir identificar como anomalías un grupo de 7 registros, con la variante que ahora tenemos un atributo de grupo al cual asociar las anomalías.

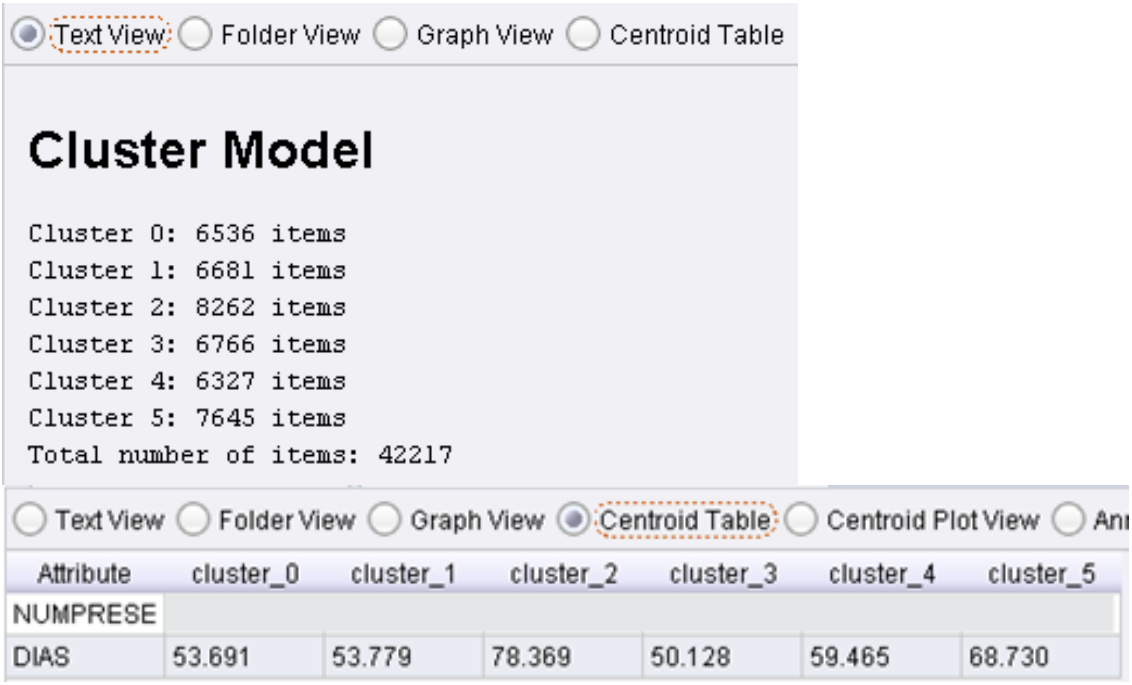


Figura 26. Vista de Datos k-Means (Cluster Model)

El Cluster Model anterior nos indica cuantos datos contiene cada grupo además que nos brinda información acerca del valor que representa el centroide de cada grupo

Nuestro Cluster Model contiene 6 grupos y al revisar cuantos datos contiene cada grupo vemos que ningún de ellos es atípico, ya que la cantidad de registros en cada uno es similar. Este algoritmo nos permite identificar además de anomalías individuales dentro del conjunto de datos a grupos que pueden ser considerados anómalos. Ese no fue el caso de nuestro estudio donde si bien es cierto se identifico un grupo de 7 registros como valores atípicos estos por si solos no constituyeron un conglomerado.

4.5.6 MODELADO: USO DE CLUSTER-BASER LOCAL OUTLIER FACTOR (CBLOF)

Se realizaron varias pruebas con el algoritmo CBLOF, al igual que con los algoritmos anteriores se inició con los valores predeterminados y se fueron modificando para ir evaluando el resultado.

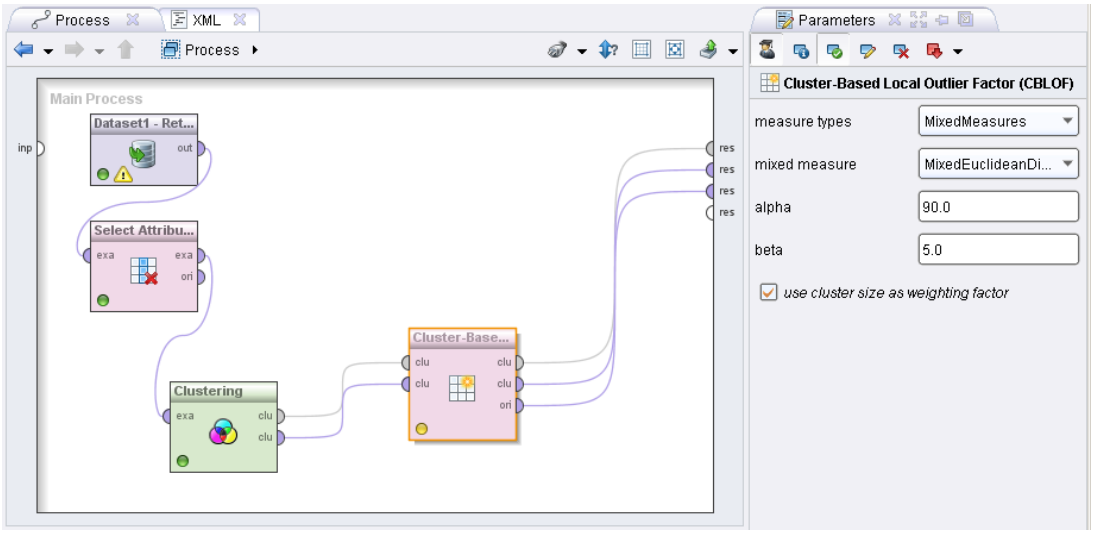


Figura 27. Modelado con CBLOF

El valor de alfa indica el porcentaje de datos que se espera que sean normales, como desconocemos la distribución de nuestro conjunto de datos lo recomendable es dejar el valor predeterminado (alpha=90).

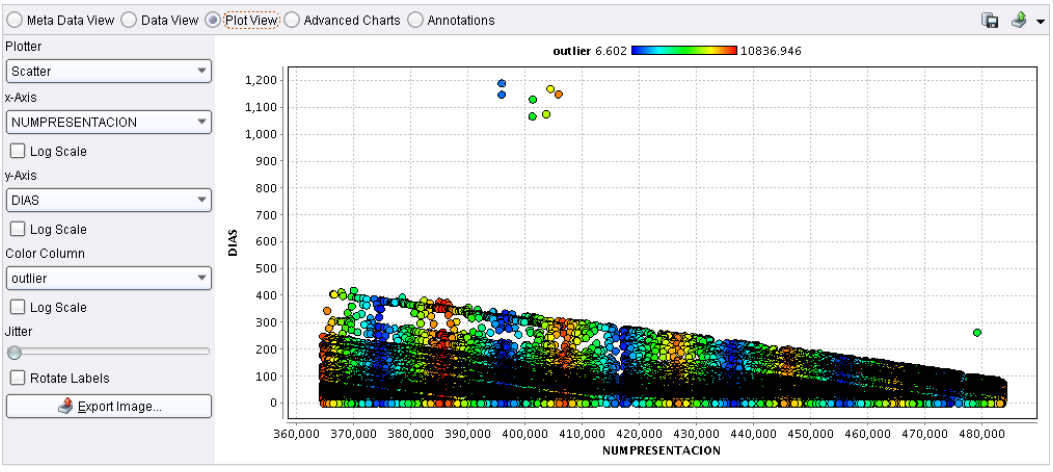


Figura 28. Prueba CBLOF (gráfico dispersión)

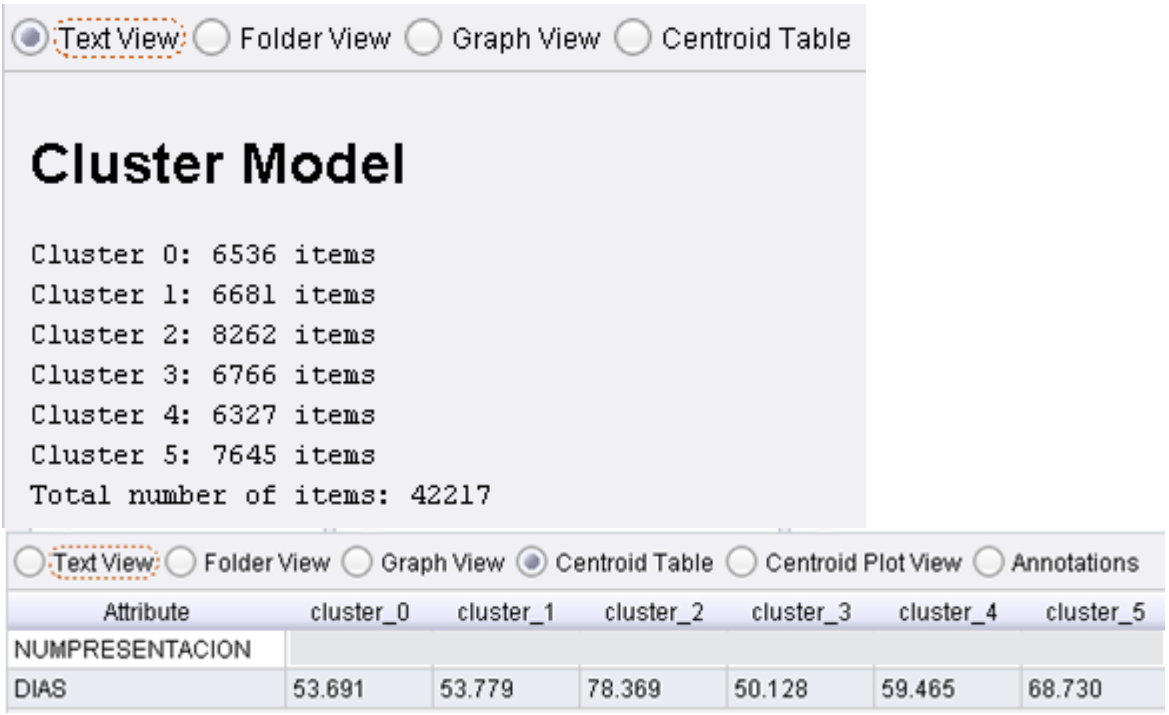


Figura 29. Prueba CBLOF (Cluster Model)

De manera general, para cualquier algoritmo, se debe tener en cuenta que los casos identificados como anomalías no necesariamente se tratan de casos fraudulentos o irregularidades, sino de registros con valores atípicos pero que son totalmente validos o podría significar que se requieren ajustes en la extracción de los datos durante la fase de preparación de datos.

4.5.7 EVALUACIÓN DEL MODELO

Dentro de los resultados obtenidos, hubo varios registros identificados como anomalías. Se debe destacar que los casos detectados como anómalos no necesariamente se tratan de casos de fraudulentos o irregularidades, sino que pueden existir otras razones, por ejemplo que quizás esté dentro del conjunto de datos todavía se encuentra información que necesita ser depurada o transformada.

Tabla 10: Comparación de Algoritmos de Detección de Anomalías

Algoritmo	Cantidad de Anomalías Identificadas	Cantidad de Anomalías Validas	Tiempo Mínimo de Ejecución	Porcentaje de Precisión
k-NN	8	8	2m 39s	100.0%
LOF	14	6	5m 54s	42.9%
COF	6	4	2m 36s	66.7%
k-Means	8	8	0m 51s	100.0%
CBLOF	8	8	0m 57s	100.0%

Como puede verificarse en la tabla los algoritmos que tuvieron mejor porcentaje de precisión fueron: K-NN, k-Means y CBLOF. Sin embargo si tomamos en cuenta el tiempo de ejecución los algoritmos que requieren menor tiempo son k-Means y CBLOF.

CAPÍTULO V. CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

- Después de las entrevistas con los expertos, inspectores generales y la revisión de informes de apoyo de investigación, se identificaron los principales escenarios que han motivado quejas y denuncias, lo que permitió construir un catálogo inicial de esquemas de anomalías que quedaron plasmadas en esta investigación.

- Se logró la evaluación de los diferentes algoritmos así como la determinación de los parámetros o criterios requeridos de detección de anomalías a través de la utilización de la metodología CRISP-DM sobre un conjunto de datos histórico y mediante un método de aprendizaje no supervisado durante la fase de modelado.

- La elección de un algoritmo es una de las tareas más difíciles, ya que dentro de la literatura no existe un consenso en cuanto a la clasificación de los diferentes elementos que componen la minería de datos. Seleccionar el más apropiado depende de los objetivos de minería de datos y el conjunto de datos de que se dispone. Para esta investigación, la cual implica la detección de anomalías, la técnica de agrupamiento mediante la aplicación de los algoritmos K-means y CBLOF resultó ser la más adecuada, haciendo uso del software de minería de datos RapidMiner.

- El desarrollo del prototipo del modelo de minería de datos permitió constatar la generación de conocimiento para enfocar las investigaciones sobre actividades irregulares, ya que se identificaron casos anómalos durante las pruebas. Esto se logró a través de un proceso iterativo basado en los conceptos y técnicas de la minería de datos. Dicho proceso puede ser replicado basado en los elementos aplicados en esta investigación y que se encuentran condensados en la sección de aplicabilidad.

5.2 RECOMENDACIONES

- Para facilitar el proceso de extracción, transformación y carga (ETL, por sus siglas en ingles) durante la fase de preparación de datos es recomendable contar con una estructura intermedia de datos entre la base de datos transaccional y el software base de minería de datos, el cual en esta investigación denominamos estructura de minería de datos, pero que podría corresponder a un almacén de datos (datawarehouse), un subconjunto de un almacén de datos (datamart) o simplemente un conjunto de vistas que sean actualizadas periódicamente.

- Para permitir que el modelo siga creciendo o adaptándose a las nuevas necesidades se sugiere seguir un estándar de registro y documentación de los esquemas de anomalías que sea independiente del software base de minería de datos. En el Anexo 4 se puede ver el formato utilizado en este proyecto de investigación y que se sugiere se convierta en el estándar de la organización.

- Para mantener integridad de la información y proteger así este activo tan importante se sugiere incluir un esquema de seguridad a la estructura de minería de datos basado en roles que incluya un detalle acerca de la propiedad de datos, la definición de niveles apropiados de acceso y controles de protección.

- Para fortalecer la seguridad del SURE se recomienda seguir la guía para la implementación del modelo de detección de anomalías para el RPI, propuesto en la sección de aplicabilidad de esta investigación.

CAPÍTULO VI. APLICABILIDAD

Este capítulo contiene la Guía para la Implementación del Modelo de Detección de Anomalías para el RPI, el cual se basa en los resultados obtenidos durante la fase de análisis y resultados.

ESTRUCTURA DEL CAPÍTULO

6.1 TÍTULO DE LA PROPUESTA

Guía para la Implementación del Modelo de Detección de Anomalías para el RPI

6.2 INTRODUCCIÓN

Con el propósito de implementar el modelo propuesto, se incluye esta guía, que describe las etapas que componen el proceso, los requerimientos técnicos, el cronograma de actividades y el presupuesto. Es importante mencionar que el PATH como brazo técnico del IP sería quien ejecute esta guía; y bajo este supuesto hemos desarrollado la misma.

6.3 MODELO DE DETECCIÓN DE ANOMALÍAS PARA EL RPI

Fundamentados en el concepto de modelo de minería de datos definido en el marco teórico tenemos que este se compone de diferentes elementos que describimos a continuación:

- Origen de Datos: fuente de datos, se trata de la Base de Datos del SURE ya que en ella se almacena las transacciones diarias del RPI. Asimismo se requiere de datos de referencia externos como los catálogos de esquemas de anomalías.

- Estructura de Minería de Datos: es un contenedor intermedio que es resultado de la fase de preparación de datos, el cual puede ser algo tan sencillo como un archivo CSV o tan complejo como un Datawarehouse. Es el insumo para el software de minería de datos.

- Software de Minería de Datos: Permite la aplicación de los algoritmos de minería de datos y enriquece el conjunto de datos originales con información derivada del procesamiento estadístico y permite visualizar estos de manera grafica.

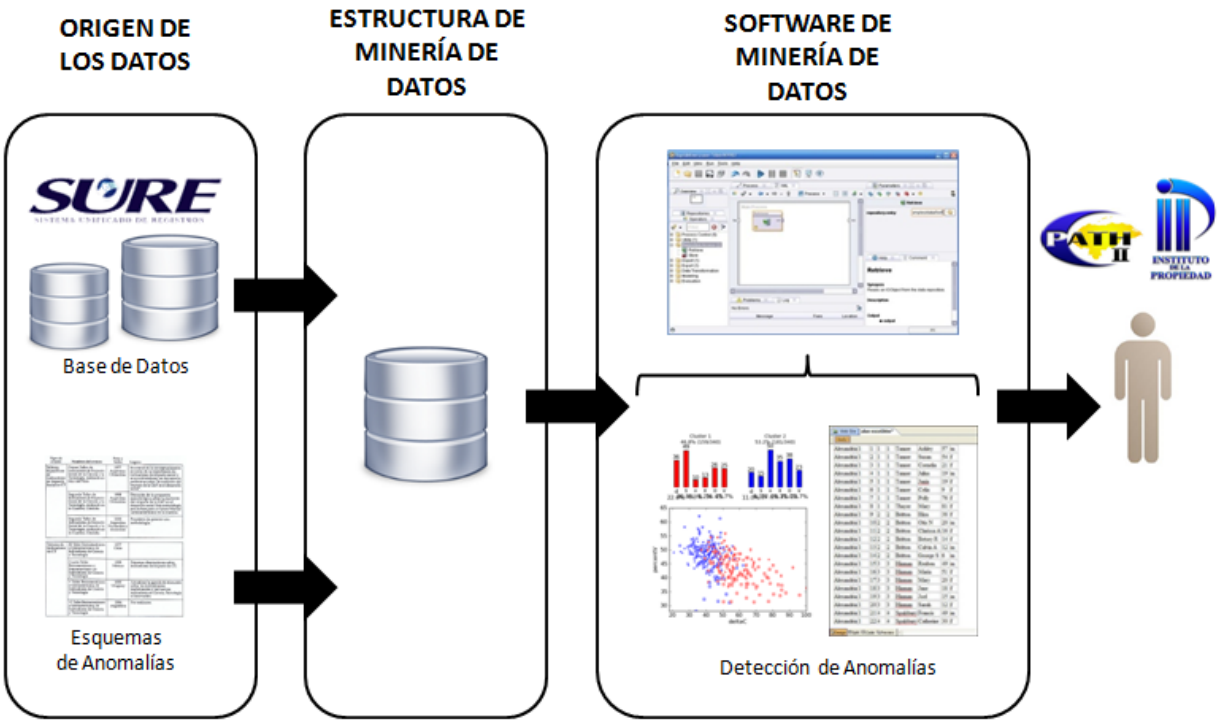


Figura 30. Elementos del Modelo de Detección de Anomalías para el RPI

6.4 PERSONAL INTERESADO (STAKEHOLDERS)

Para la correcta implementación del proyecto es necesario definir claramente al personal interesado, que incluye: patrocinador o patrocinadores, equipo de implementación y usuarios.

6.4.1 PATROCINADOR

Es la persona activamente involucrada con el proyecto. Se encargará de proveer u obtener los recursos financieros para la ejecución del mismo. Es su principal defensor o promotor y se convierte en el vocero frente a las autoridades superiores. En nuestro caso será el Coordinador Nacional del PATH.

6.4.2 EQUIPO DE IMPLEMENTACIÓN (GESTIÓN)

Como en toda actividad que requiere un trabajo ejecutado por un grupo de personas es necesario que ciertos miembros del equipo ejerzan un rol de control para asegurar que los esfuerzos sean orientados a la obtención de los objetivos de la implementación del modelo de minería de datos.

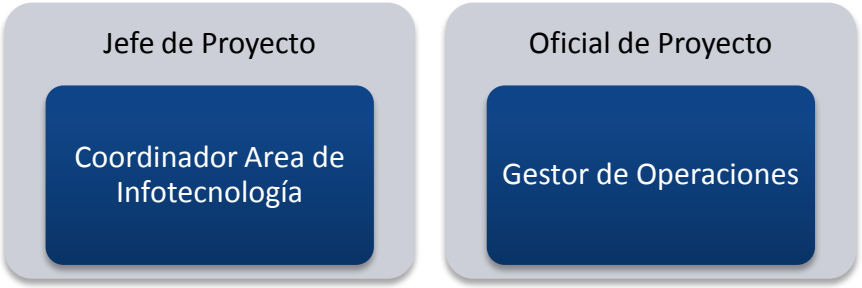


Figura 31. Relación Rol - Puesto Específico

Jefe de Proyecto: tiene la responsabilidad de administrar de la mejor manera los recursos, tecnológicos, humanos y económicos durante la implementación del modelo.

Oficial de Proyecto: fundamental en la supervisión y apoyo al cumplimiento de las tareas que se tienen programadas a diario

6.4.3 EQUIPO DE IMPLEMENTACIÓN (COLABORADORES)

La tarea de implementación se llevará a cabo por un equipo multidisciplinario, a continuación se mencionan los puestos de estos miembros: Administrador de Base de Datos, Analista de Sistemas, Programador, Oficial de Infraestructura, Oficial de Seguridad Informática, Asesor Legal, Personal de Soporte Técnico, Especialista en Minería de Datos e Implementadores del SURE.

6.4.4 USUARIOS

Los tipos de usuarios involucrados en la implementación del modelo de minería de datos son:

- Finales Directos: Que posean conocimientos generales en el uso de las computadoras personales así como de correo electrónico, donde reciban reportes según se estime conveniente.
- Finales Indirectos: Que conozcan de la existencia de este mecanismo, de manera que sientan confianza al realizar sus transacciones en el RPI.
- Usuario Administradores: Suficiente experiencia en el uso de sistemas web, minería de datos y base de datos en este caso el SURE o previamente capacitado. Se encargara del uso y configuración de los algoritmos y se encargara de realizar una actualización periódica del catalogo de anomalías.

6.5 CONDICIONES MÍNIMAS

Para el funcionamiento del modelo existen requerimientos mínimos esenciales, entre ellos: un especialista en minería de datos; un equipo de trabajo, un equipo de cómputo; una política y un proceso.

El especialista en minería de datos sería el encargado de realizar de manera continua la recopilación de esquemas de anomalías y administración del software de minería de datos.

El equipo de implementación es un multidisciplinario y está definido en la sección 6.4.2 y 6.4.3 de esta guía.

El equipo de cómputo requerido debe cumplir con las condiciones mínimas descritas la tabla 11 de esta guía.

La política de uso de minería de datos es parte fundamental debe incluir una definición de un marco de trabajo para el uso de herramientas de minería de datos así como su aporte a los objetivos de negocios.

El proceso (o procesos) además de una descripción clara de las actividades que se requieren para el uso de la minería de datos, debe identificar claramente sus objetivos, quien es el dueño del mismo y definir las métricas para medir su desempeño.

Tabla 11. Requerimientos mínimos

Requerimientos de Software	Sistema Operativo	Windows XP SP3 Windows 7
Requerimientos de Hardware	Procesador	Modelo Intel Core Duo Tipo T6500 @2.1GHz
	Memoria RAM	Mínimo: 1GB Recomendado: 2GB Optimo: 4GB
	Espacio en Disco	Mínimo: 8 GB Recomendado: 40GB

6.6 ASPECTOS FINANCIEROS

El presupuesto incluye el uso de software gratuito de minería de datos (RapidMiner), Se contempla las capacitaciones relacionadas al software así como la adquisición de un computador para la instalación y funcionamiento del mismo. (Los precios están expresados en dólares).

Tabla 12. Presupuesto Recurso Humano

Rubro de Inversión / Detalle	Valor $
Recurso Humano	
Jefe de Proyecto	900.00
Oficial de Proyecto	750.00
Administrador de Base de Datos	1,500.00
Analista de Sistemas	1,125.00
Programador	1,200.00
Oficial de Infraestructura	600.00
Oficial de Seguridad Informática	450.00
Asesor Legal	360.00
Personal de Soporte Técnico	300.00
Especialista en Minería de Datos	6,000.00
Implementadores del SURE	450.00
Total Recurso Humano	**13, 635**

Tabla 13. Presupuesto Rubros Adicionales

Rubro de Inversión / Detalle	Valor $
Hadware	
Servidor con las especificaciones Mínimas requeridas	1,000.00
Software	
Software Base de Minería de Datos	0.00
Capacitación	
Curso Nivel 1 para 3 personas	4,678.20
Curso Nivel 2 para 3 personas	4,678.20
Total	**10,356.40**

Las tablas anteriores describen los montos que se han de invertir para la implementación del Modelo de Minería de Datos. La suma de todos los rubros asciende a un total de $23,991.00.

Se adjunta análisis de retorno de la inversión en seguridad (ROSI, por sus siglas en ingles) ver Tabla 14 y Tabla 15. Este análisis se realizó con valores estimados ya que no se posee una estadística de las perdidas relacionadas a las irregularidades o anomalías, de manera que sirva de ejemplo para el cálculo del mismo. (Valores expresados en lempiras)

Tabla 14. Cálculos ROSI

Calculo ROSI					
Años		0	1	2	3
Ahorro Bruto Anual			350,000.00	350,000.00	350,000.00
Ahorro Bruto Anual a valor actual con factor de descuento al 19%			294,117.65	247,157.69	207,695.53
Valor Contramedida	748,970.87				
Costos Contramedida		23,991.00	6,000.00	6,000.00	6,000.00
Costos contramedida a valor actual con factor de descuento al 19%		23,991.00	5,042.02	4,236.99	3,560.49
Costo de contramedida	36,830.50				
Retorno (valor - costo)	712,140.37		284,873.95	239,389.87	201,167.96
ROSI (Retorno / Costo)	19.34				

Tabla 15. Sumario de Valores para Analisis ROSI

Sumario de Valores para Análisis ROSI	
Pérdidas anuales por Anomalías - sin tratar	450,000.00
Pérdidas anuales por Anomalías - residual luego de ser mitigados	100,000.00
Ahorro anual bruto por contra medida	350,000.00
Costo Inicial por contramedida	23,991.00
Costos anules recurrentes contramedida	6,000.00

Por lo tanto, se determina que una inversión en seguridad es rentable si el valor de retorno (valor - costo) es positivo.

6.7 PRODUCTOS ESPERADOS

- Entendimiento del Negocio
    - Antecedentes: contiene un diagnostico situacional de la organización previo a la implementación
    - Objetivos del Negocio y Criterio de Éxito: es un registro desde la perspectiva del negocios sobre los objetivos a ser alcanzados y cuáles serán los resultados esperados para considerar como exitosa la implementación
    - Inventario de Recursos: documento que contiene una lista de recursos disponibles. Incluye: personal, datos disponibles, hardware, software.
    - Objetivos de Minería de Datos y Criterio de Éxito: es un registro desde la perspectiva técnica de los objetivos a ser alcanzados y cuáles serán los resultados esperados para considerar como exitosa la implementación

- Entendimiento de los Datos
    - Reporte Inicial de Recolección de Datos: Listado de conjunto de datos explorados. Métodos de acceso utilizados. Problemas encontrados y solución aplicada a los mismos.
    - Reporte de Descripción de Datos: Informe descriptivo del conjunto de datos explorados, incluyendo formatos, tipos de datos. De preferencia se

pueden incluir estadísticas de atributos claves. Mínimos, Máximos, Medias o Modas, porcentaje de atributos nulos.

- Reporte de Exploración de Datos: Informe descriptivo del conjunto de datos explorados donde se incluyen los hallazgos resultantes de la observación de relaciones entre atributos, entidades, subconjuntos específicos de datos u otro descubrimiento interesante que se considere debe ser registrado.
- Reporte de Calidad de Datos: Informe donde se reporten problemas encontrados en los datos, alternativas de solución y el posible impacto en el proceso de minería de datos en caso de no resolverse.

- Preparación de los Datos
  - Reporte de Selección de Datos: Listado de de datos o conjunto de datos a ser utilizados. Su principal objetivo es registrar la razón por la que fueron seleccionados.
  - Reporte de Limpieza de Datos: Informe sobre las acciones tomadas para solventar los problemas reportados en el reporte de calidad de datos.
  - Reporte de Transformación de Datos: Informe en el cual debe quedar registrado las acciones realizadas que implicaron la generación de registros, inclusión de atributos derivados, combinación de datos y/o aplicación de formatos

- Modelado
  - Descripción de Técnicas de Modelado Seleccionadas: Documento que incluye la lista y descripción de las técnicas a utilizar
  - Diseño de Pruebas: Documento que describe el procedimiento a seguir para la realización de las pruebas
  - Modelo: descripción narrativa del modelo que incluye los elementos incluidos en el mismo: filtro de registros, filtro de atributos, muestreo, limpieza de datos, reemplazo de datos nulos, algoritmos de

agrupamiento, algoritmos de detección de anomalías, etc. Igualmente se incluyen los parámetros utilizados en cada uno de los elementos.

- Evaluación del Modelo: este documento debe describir los resultados de probar el modelo de acuerdo a lo especificado en el diseño de las pruebas

- Evaluación
  - Reporte de Evaluación de Resultados de Minería de Datos: se comparan los resultados de la minería de datos contra los objetivos de negocios y el criterio de éxito desde el punto de vista de negocios. Debe concluir sobre el grado de éxito alcanzado y sobre la conveniencia del proceso de minera de datos. Definir si existen nuevos objetivos de negocio que se tratarán más adelante en el proyecto o en nuevos proyectos de minerías de datos, así como una lista de recomendaciones sobre los próximos pasos a seguir.

- Implementación
  - Plan de Despliegue: debe especificar la estrategia de implementación y los pasos necesarios para ejecutarla
  - Plan de Monitoreo y Mantenimiento: debe especificar la estrategia de monitoreo y mantenimiento y los pasos necesarios para ejecutarla.
  - Reporte Final: Documento final con un resumen de la experiencia adquirida y las lecciones aprendidas. Puede incluirse una presentación final con los resultados de la minería de datos.

6.8 CRONOGRAMA DE EJECUCIÓN

Para la ejecución de la Implementación se ha definido el siguiente cronograma basado en las etapas del ciclo de vida de un proyecto de minería de datos según la metodología CRIPS-DM.

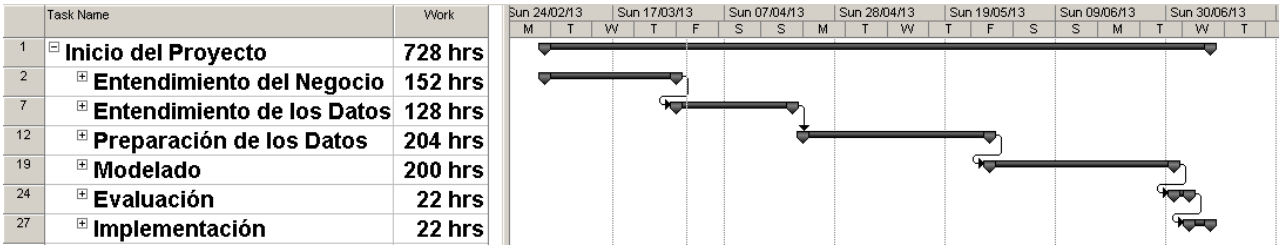


Figura 32. Cronograma de Ejecución Resumido

La figura anterior muestra un resumen de las etapas con un detalle de las horas requeridas para cada una de ellas. A continuación se detalla cada una de las etapas con sus respectivas tareas y su duración en horas.

Entendimiento del Negocio	**152 hrs**
Determinar Objetivos de Negocios (Institucionales)	32 hrs
Diagnostico Situacional	48 hrs
Determinar Objetivos de Minería de Datos	32 hrs
Evaluación Inicial de Técnicas y Herramientas de Minería de Datos	40 hrs
Entendimiento de los Datos	**128 hrs**
Recolección Inicial de Datos	16 hrs
Descripción de Datos	16 hrs
Exploración de Datos	48 hrs
Verificar Calidad de los Datos	48 hrs
Preparación de los Datos	**204 hrs**
Selección de Datos	36 hrs
Limpieza de Datos	48 hrs
Construcción de Datos (de ser necesario)	36 hrs
Integración de Datos	36 hrs
Formateo de Datos	36 hrs
Producir Conjunto de Datos para Modelado	12 hrs
Modelado	**200 hrs**
Selección de Técnicas de Modelado	64 hrs
Generar Diseño de Pruebas	24 hrs
Construcción del Modelo	64 hrs
Evaluar Modelo	48 hrs
Evaluación	**22 hrs**
Evaluación de Resultados	16 hrs
Revisión del Proceso en General	6 hrs
Implementación	**22 hrs**
Realizar un Plan de Despliegue	3 hrs
Realizar un Plan de Monitoreo y Mantenimiento	3 hrs
Generar Reporte Final	4 hrs
Ejecución Plan de Despliegue	12 hrs

Figura 33. Etapas de la Ejecución con sus respectivos tiempos

BIBLIOGRAFÍA

1. Álvarez Caperochipi, J. (2006). *Derecho Inmobiliario Registral* (2.ª ed.). Comares.
2. Auditor General of Alberta. (2010). Report of the Auditor General of Alberta
3. Banco Mundial. (2011). Documento Evaluación de Proyecto.
4. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0.
5. Congreso Nacional. (2004, mayo 28). Ley de la Propiedad.
6. Congreso Nacional. (2011). Decreto No. 164-2011 Convenio Financiero No. 4641-HN
7. Cornejo, A. A. (1994). *Derecho Registral*. Astrea.
8. Diccionario de la lengua española - Vigésima segunda edición. (s. f.). Recuperado 25 de abril de 2013, a partir de http://lema.rae.es/drae/?val=bien%20mueble
9. Diccionario Jurídico. (s. f.). Recuperado 25 de abril de 2013, a partir de http://www.diccionariojuridico.mx/
10. Diccionario jurídico: Iuris tantum. (s. f.). Recuperado 25 de abril de 2013, a partir de http://www.ic-abogados.com/diccionario-juridico/iuris-tantum/34
11. Elmasri, R., & Navathe, S. (2007). Fundamento de Sistemas de Base de Datos (5.ª ed.). Pearson.
12. Faudos Pons, P., Gómez Valle, M., González García, I., Martín Alías, J. I., & Santos Lloro, M. (2008). *Lecciones de Derecho Registral*. ATELIER LIBROS, S.A.
13. Guevara Manrique, Rubén (1988). Derecho Registral. Ojeda.

14. Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2.ª ed.). Morgan Kaufmann.
15. Hernández Gil, F. (1983). *Introducción al Derecho Hipotecario* (3.ª ed.). Derecho Reunidas S.A.
16. Hernández, R., Fernández, C., & Baptista, M. del P. (2010). *Metodología de la Investigación* (5.ª ed.). McGrawHill.
17. IBM Corporation. (2011). IBM SPSS Modeler CRISP-DM Guide.
18. Instituto de la Propiedad (s. f.). Recuperado a partir de http://www.ip.gob.hn
19. Instituto de la Propiedad. (2010, enero 11). Reglamento de La Ley de Propiedad.
20. Instituto de la Propiedad. (2012). Manual de Unificación de Criterios Registrales
21. LaCruz Berdejo, J. L. (2011). *Derecho Inmobiliario Registral*. Civitas Ediciones.
22. Larose, D. T. (2005). *Discovering Knowledge in data An Introduction to Data Mining*. Wiley.
23. Linoff, G. S., & Berry, M. J. A. (2011). *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management* (3.ª ed.). Wiley.
24. Liu, B. (2007). *Web Data Mining Exploring Hyperlinks, Contents, and Usage Data*. Springer.
25. López de Zavalía, F. (1983). *Curso Introductorio al Derecho Registral*. Astrea.
26. Maimon, O., & Rokach, L. (2010). *Data mining and Knowledge Discovery Handbook* (2.ª ed.). Springer.
27. Matignon, R. (2005). *Neural Network Modeling Using Sas Enterprise Miner*. AuthorHouse.
28. Modelos de minería de datos. (s. f.). Recuperado a partir de http://msdn.microsoft.com/es-es/library/cc645779.aspx

29. North, D. M. (2012). *Data Mining for the Masses*.
30. Olson, D., & Delen, D. (2008). *Advanced Data Mining Techniques*. Springer.
31. Palma, C., Palma, W., & Pérez, R. (2009). Data Mining El Arte de Anticipar (Primera.). Santiago de Chile: RIL Editores.
32. PATH (2005). Concepto del Negocio Inmueble
33. PATH. (2006). Documento de Análisis y Diseño Sistema de Seguridad y Auditoria SINAP
34. PATH. (2012). Manual de Inducción - Conociendo el PATH
35. PATH. (s. f.). SINAP - Estructura de la Información. *Estructura de la informacion del SINAP*. Recuperado 22 de enero de 2013, a partir de http://www.sinap.hn/portal/page?_pageid=52,125520&_dad=portal&_schema=PORTAL
36. Prabhu, S., & Venatesan, N. (2007). *Data Mining and Warehousing*. New Age International.
37. Principios Registrales. (s. f.). Recuperado 17 de febrero de 2013, a partir de http://www.ilustrados.com/tema/10034/Principios-Registrales.html
38. Proenza, F. J. (2007). Esquema Estratégico y Operativo de Institucionalización del SINAP.
39. Refaat, M. (2007). *Data Preparation for Data Mining using SAS*. Morgan Kaufmann.
40. Revista del Colegio de Notarios de Jalisco. (s. f.). Recuperado 5 de marzo de 2013, a partir de http://www.revistanotarios.com/?q=node/415
41. Rokach, L., & Maimon, O. (2008). *Data Mining with Decision Trees Theory and Applications* (Vol. 69). World Scientific.

42. SAS Enterprise Miner - SEMMA. (s. f.). Recuperado a partir de http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html
43. Tufféry, S. (2011). *Datamining and statistics for decision making*. Wiley.
44. Turban, E., Sharda, R., & Delen, D. (2009). *Decision Support and Business Intelligence Systems* (Ninth.). Prentice Hall.
45. Vercelli, C. (2009). *Business Intelligence - Data Mining and Optimization for Decision Making*. Wiley.
46. Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques* (3ed ed.). Morgan Kaufmann.
47. Zhang, J. (2013). Advancements of Outlier Detection: A Survey, 26.

ANEXOS

ANEXO 1. PREGUNTAS GUÍA A EXPERTOS PATH - PROCESO REGISTRAL

ENTREVISTA

1. ¿Existe alguna categorización o clasificación de irregularidades que se dan dentro del sistema?

2. ¿Con qué casos de irregularidades se ha encontrado Ud. en el ejercicio de sus funciones diarias?

3. ¿Cuáles de estas irregularidades mencionadas, se dan con mayor frecuencia?

4. ¿Cuáles de estas se dan con menor frecuencia?

ANEXO 2. PREGUNTAS GUÍA A EXPERTOS PATH - HARDWARE Y SOFTWARE

ENTREVISTA PLATAFORMA TECNOLÓGICA SURE

PREGUNTAS GUÍA

1. Características de Gestor de Base de Datos:

_____

2. Sobre el Lenguaje de Programación utilizado:

_____

3. Ambiente de operación del sistema:

_____

ENTREVISTA ASPECTOS SOFTWARE MINERÍA DE DATOS

1. Mencione características generales para la selección del software de minería de datos:

   _____

2. Mencione características funcionales para la selección del software de minería de datos:

   _____

3. Mencione características técnicas (compatibilidad) para la selección del software de minería de datos:

   _____

4. Mencione características de soporte (compatibilidad) para la selección del software de minería de datos:

   _____

ANEXO 3. PREGUNTAS CUESTIONARIO IRREGULARIDADES – INSPECTORES

1. ¿Existen algún procedimiento documentado y establecido para detectar irregularidades en el sistema?

   _____

   _____

2. Mencione los tipos de denuncias relacionadas con irregularidades dentro del sistema (Ej.: Rompimiento del orden de prelación, alteración de derechos del inmueble, etc.)

   _____

   _____

   _____

3. Describa en qué consisten las irregularidades que menciono

   _____

   _____

   _____

Toda la información contenida en este documento será manejada con confidencialidad y privacidad.

ANEXO 4. REGISTRO DE ESQUEMAS DE ANOMALÍAS

I. Datos Generales de la Anomalía

Nombre		No.
Descripción:		
Fuente:		

II. Datos Técnicos

a. Fuente de Datos

Esquema de Base de Datos:					
Nombre de las Tablas y Campos requeridos:					
Detallar Query (Consulta) a la Base de Datos:					

b. Preparación de los Datos

Actividades de construcción y formateo de los datos (marque con una X en la casilla correspondiente):					
Agregaciones ☐	Sumarización ☐	Campos Dummy ☐	Valores Calculados ☐	Combinaciones ☐	Registros Generados ☐
Registros que deberán ser excluidos:					
Nombre de los atributos del conjunto de datos final:					

c. Datos de Control

Elaborado por: _____

Fecha:_____ _____

ANEXO 5. SOFTWARE DE MINERÍA DE DATOS - R DM

RDataMining

What is R

R is a free software environment for statistical computing and graphics. It runs on Windows, Linux and MacOS.

R is widely used in academia and research, as well as industrial applications.

R Documents

If you are new to R, *an Introduction to R* and *R for Beginners* are good references to start with.

More details on R Language and data access are documented respectively by *the R Language Definition* and *R Data Import/Export*.

Other R manuals and many contributed documentations are available at CRAN.

This website also collects links to some free online documents for R.

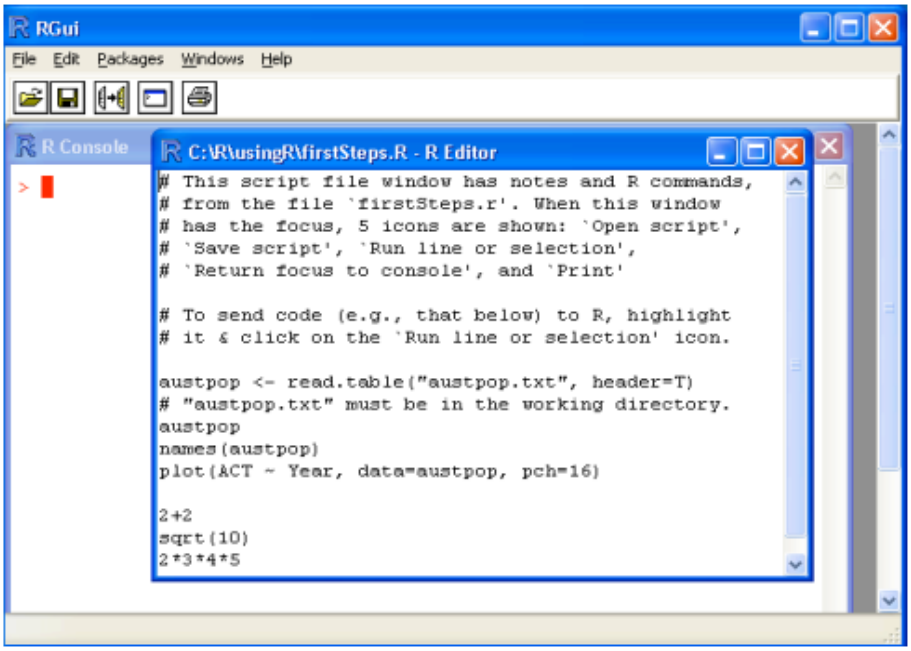
R Packages

R can be extended easily via packages. There are more than 3000 packages available in the CRAN package repository (as on 25 May 2011).

Especially, package RWeka provides an interface to Weka, enabling to use most Weka functions in R.

CRAN task views provide collections of packages for different tasks. Some task views related to data mining are

- CRAN Task View: Machine Learning & Statistical Learning,
- CRAN Task View: Cluster Analysis & Finite Mixture Models,
- CRAN Task View: Time Series Analysis,
- CRAN Task View: Multivariate Statistics, and
- CRAN Task View: Analysis of Spatial Data.

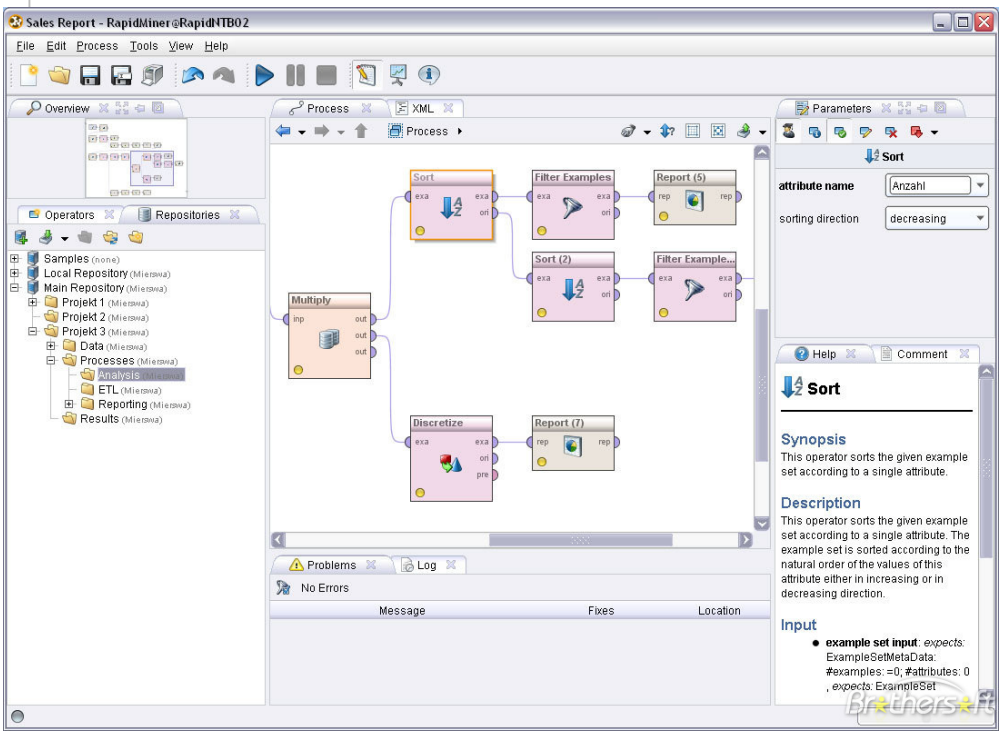


ANEXO 6. SOFTWARE DE MINERÍA DE DATOS – RAPID MINER

RAPID|MINER

Overview | **Features** | Screenshots | Testimonials | Editions and I

- Freely available **open-source** data mining and analysis system
- Runs **on every major platform** and operating system
- Most **intuitive** process design
- Multi-layered data view concept ensures **efficient data handling**
- GUI mode, server mode (command line), or access via Java API
- Simple **extension mechanism**
- Powerful high-dimensional **plotting** facilities
- Most comprehensive solution available: **more than 500 operators** for data integration and transformation, data mining, evaluation, and visualization
- Automatic meta optimization schemes
- Definition of **re-usable building blocks**
- **Standardized XML** interchange format for processes
- Graphical process design for standard tasks, **scripting language** for arbitrary operations
- Machine learning library WEKA fully integrated
- Access to **data sources** like Excel, Access, Oracle, IBM DB2, Microsoft SQL, Sybase, Ingres, MySQL, Postgres, SPSS, dBase, Text files and more
- **Most comprehensive data mining solution** with respect to data integration, transformation, and modeling methods
- **Winner** of several user and jury awards



ANEXO 7. SOFTWARE DE MINERÍA DE DATOS - KNIME

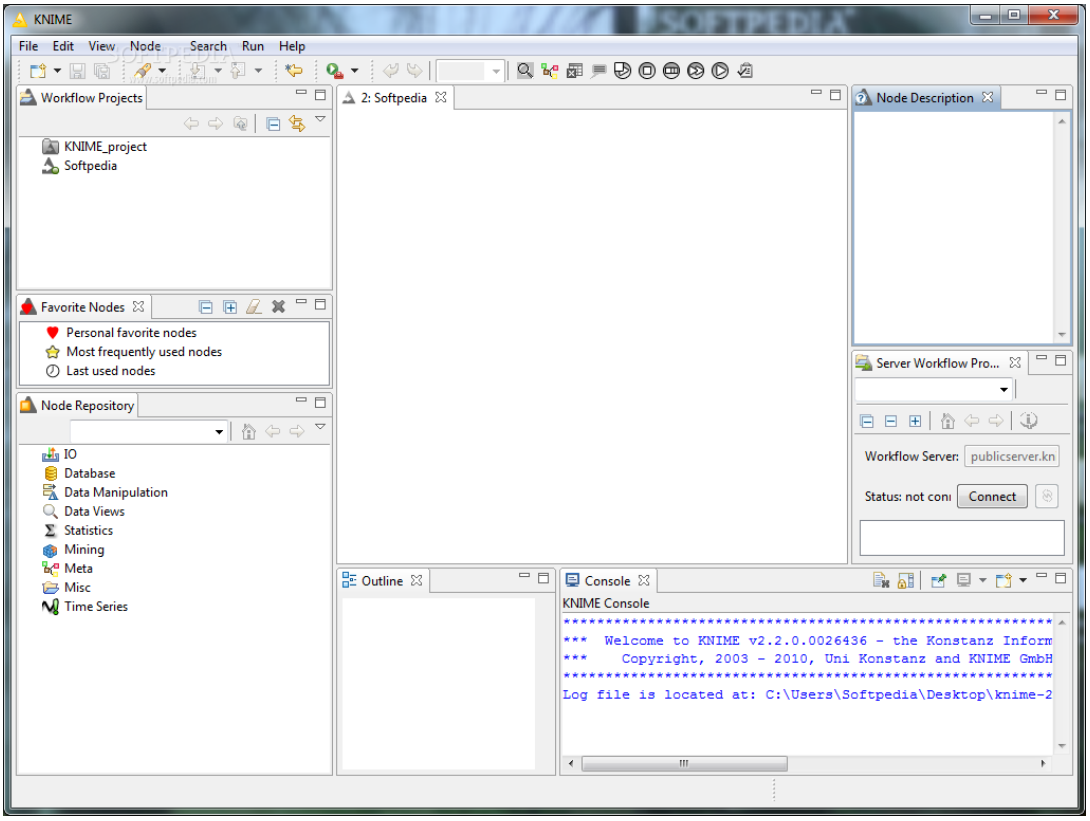
KNIME - Professional Open-Source Software

KNIME [naim] is a user-friendly graphical workbench for the entire analysis process: data access, data transformation, initial investigation, powerful predictive analytics, visualisation and reporting. The open integration platform provides over 1000 modules (nodes), including those of the KNIME community and its extensive partner network.

KNIME can be downloaded onto the desktop and used free of charge. KNIME products include additional functionalities such as shared repositories, authentication, remote execution, scheduling, SOA integration and a web user interface as well as world-class support. Robust big data extensions are available for distributed frameworks such as Hadoop. KNIME is used by over 3000 organizations in more than 60 countries.

/ More information about KNIME.

Identified by Gartner as a Cool Vendor, KNIME is now the open source data analytics platform ranked No. 1 in customer satisfaction for open source analytics and leads in 70% of all categories.



ANEXO 8. SOFTWARE DE MINERÍA DE DATOS - WEKA

Weka 3: Data Mining Software in Java

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like **this**, and the bird sounds like **this**.

Weka is open source software issued under the **GNU General Public License**.