UNIVERSIDAD CENFOTEC

Maestría en Bases de Datos

Proyecto de investigación aplicada 2:

Diseño e implementación de un modelo de Minería de Datos para la prevención de casos de agresión y abandono infantil en Costa Rica

Elaborado por:

Ing. José A. Cabezas Jaikel

Ing. Esteban Oviedo Blanco, MBA-MPM

Junio, 2013

Tabla de contenido

Introducción

Antecedentes

El Patronato Nacional de la Infancia (PANI) cumple su rol social de facilitación y guía, búsqueda de soluciones a los problemas sociales, interfaz entre la misión institucional y su aplicación práctica a la sociedad costarricense; todas circunscritas en sus cuatro funciones básicas: Atención, Protección, Defensa y Garantía, para lo cual debe contar con una plataforma de apoyo adecuada y que brinde todas las herramientas necesarias para facilitar el cumplimiento de dicha tarea.

La implementación de una plataforma tecnológica de software confiable y robusta, soportada en las Tecnologías de Información y Telecomunicaciones, con garantía de funcionamiento adecuado a las necesidades de obtención, procesamiento y producción de soluciones, necesaria para la realización de sus actividades básicas dentro de la sociedad, permite al PANI dar un mejor y más continuado servicio a la ciudadanía en general. Tecnologías de Información, se constituye como eje de la mayoría de los servicios que requieren acceso a información, especialmente aquellos referidos a las bases de datos y sistemas organizacionales.

Después de utilizar varios sistemas de información transaccionales (OLTP), los cuales presentaban deficiencias para atender los requerimientos institucionales actuales y futuros, y a raíz del cambio de paradigma que se ha dado en el quehacer del PANI, al pasar de la doctrina de situación irregular con un enfoque de bienestar, a una doctrina de protección integral con un enfoque centrado en los derechos de los niños, niñas y adolescentes, la Institución se vio en la necesidad de tomar la decisión de desarrollar un nuevo sistema transaccional integral, denominado INFOPANI, tomando como partida el enfoque de la Doctrina de Protección Integral, donde se pueda digitalizar información y automatizar procesos relacionados con los casos que se reportan de **abandono o agresión en menores de edad**.

El desarrollo e implementación del *Sistema de Expediente Electrónico del Patronato Nacional de la Infancia* (INFOPANI), permite llevar un registro de expedientes de las personas menores de edad y de los procesos y actividades que surgen a partir de la atención de cada caso que se reporta como una posible agresión o abandono infantil o juvenil.

El proyecto INFOPANI surgió de la necesidad organizacional, de modernizar el accionar institucional y de dotar a la institución de herramientas de software para el cumplimiento de su misión. Al ser un sistema estructurado permite llevar un control interno del cumplimiento de los procesos y actividades por parte de los profesionales técnicos encargados, cuyo horizonte será cumplir con la Doctrina de Protección Integral

con un enfoque en los derechos de los niños, niñas y adolescentes. Por consiguiente el control de los procesos y actividades incidirán en forma sustancial en la calidad de los servicios brindados, así como generar información oportuna, eficiente y eficaz que coadyuve a la toma de decisiones.

El desarrollo del INFOPANI, tuvo como único fin simplificar y agilizar los servicios al ciudadano, tales como denuncias, atención integral y adopciones, logrando un orden de funcionalidad que permitiera tener toda la información automatizada, siempre actualizada y accesible con los niveles óptimos de seguridad, garantizando en todo el sistema la confidencialidad de la información colocada en las bases de datos.

La primera fase del proyecto ya fue desarrollada y liberada en Producción y consistió en la reorganización de los procesos atencionales en las Oficinas Locales, así como la información suministrada por el Departamento de Atención Integral y el Centro de Orientación e Información.

Adicionalmente, a solicitud del PANI, como parte de la primera fase del proyecto, se construyó un Data Warehouse para seguidamente implementar una solución de cubos multidimensionales, los cuales facilitarían el análisis y el cruce de la información que se va registrando a lo largo del flujo de atención de casos de agresión y abandono infantil y juvenil. El PANI nunca antes había desarrollado algún proyecto de Inteligencia de Negocios, pero la idea de desarrollar estos Data Marts era precisamente incursionar en el mundo del análisis de la información y determinar cómo podría la tecnología apoyar la toma de decisiones estratégicas de la organización.

Si bien el flujo completo de la administración del ciclo de vida de un caso de agresión o abandono inicia con la creación del expediente y pasa posteriormente por varios procesos como la adopción y la acreditación, el INFOPANI Fase I se limita a almacenar información sobre los casos reportados y a generar un expediente que luego es analizado por técnicos especialistas, quienes al final del día lo categorizan como un caso de **riesgo inminente** o no; lo que quiere decir que un caso puede tener altas o bajas probabilidades de que se finalice en un proceso judicial.

La apertura de un expediente es el primer paso en el flujo de atención a reportes de abandono o maltrato en niños y jóvenes. La información que se registra está relacionada con los datos personales del menor de edad involucrado, así como alguna información de sus padres o responsables legales y también del medio ambiente en que se desenvuelve el caso.

El presente Proyecto de Investigación Aplicada consistió en el análisis de los procesos que cubre el sistema transaccional desarrollado en la primera fase de INFOPANI, así como el análisis de las estructuras del sistema transaccional como tal y del cubo de información existente. Todo esto con la finalidad de

diseñar y construir un modelo de Minería de Datos que utilice la información almacenada en el transaccional y el Data Warehouse existentes para poder aplicar algoritmos predictivos y de clasificación que sirvan como herramienta de prevención de más casos de agresión y de abandono infantil y juvenil.

A lo largo de varios años, se desarrollaron un gran número de métodos de análisis de datos basados en la estadística. Sin embargo, en la medida en que se incrementaba la cantidad de información almacenada en las bases de datos, estos métodos empezaron a enfrentar problemas de eficiencia y escalabilidad y es aquí donde aparece el concepto de minería de datos.

Una vez diseñadas las estructuras del modelo de minería, se utilizó toda la información que generó el sistema transaccional desde su puesta en producción a mediados del año 2012 hasta el mes de mayo del año 2013 y que fue consolidada en el Data Warehouse, para cargarlas. Estas transformaciones y cargas permitieron convertir todos estos datos en conocimiento útil para detectar diversas situaciones de riesgo que hasta el momento con las herramientas con que contaba el PANI eran imposible o muy difícil de detectar.

Con las técnicas de Minería de Datos utilizadas en este proyecto, se obtienen elementos de información que permiten la prevención de futuros casos de agresión infantil, utilizando algoritmos estadísticos que identifican y agrupan diferentes ambientes o que predicen probabilidades de ocurrencia dadas ciertas características socio-ambientales.

Justificación

Día con día el PANI genera información y esto le lleva a tener grandes cantidades de esta, la cual les puede ayudar a controlar, optimizar, administrar, examinar, investigar, planificar, predecir, someter, negociar o tomar decisiones en relación con la problemática social de abandono y agresión en menores de edad.

Conforme se va analizando la información que se recolecta en el sistema transaccional INFOPANI, ésta va ganando importancia y dejando de ser simples datos para ser una fuente importante de conocimiento para la institución, y es en este punto donde el presente proyecto encuentra sentido, pues mediante la utilización de algoritmos y herramientas de Minería de Datos, se pretende dotar al PANI de una plataforma que pueda satisfacer sus necesidades de identificar patrones de conductas asociadas a agresiones o abandonos, antes de que estos ocurran.

Dentro de los beneficios que aporta la Inteligencia de Negocios, mediante el análisis a través de Cubos de Información, como el que ya está en capacidad de realizar el PANI desde que se finalizó la fase I del proyecto INFOPANI, está la definición, construcción y monitoreo de diversos indicadores como por ejemplo los rangos de fechas con mayor incidencia de casos de agresión o abandono, la cantidad promedio de casos por mes que se atienden en cada oficina regional, la frecuencia de reporte de casos por rangos de edades de los menores involucrados, entre otros. Todo esto de una manera automatizada y evaluando grandes cantidades de datos. Por su parte, el modelo de Minería de Datos propuesto viene a aportar un complemento idóneo al aplicar algoritmos estadísticos enfocados en prevención y predicción, mediante la identificación de patrones de conductas y perfiles de las personas que se ven involucradas en los casos de agresión o abandono que han sido reportados.

La motivación fundamental de este Proyecto de Investigación Aplicada es poder brindar al PANI una plataforma de análisis que sus funcionarios tomadores de decisiones puedan utilizar para optimizar la inversión de su tiempo, su esfuerzo y sus recursos, de manera que se logren mejores resultados en la efectividad de sus campañas preventivas y en sus procesos internos de atención; sin necesidad de aumentar dicha inversión.

Las técnicas de Minería de Datos le permitirán al PANI analizar patrones de datos donde no se podían detectar a simple vista y donde con otras técnicas de manipulación de datos es imposible su extracción.

Actualmente el PANI cuenta con cubos de información, los cuales presentan información histórica, es decir, hechos que ya ocurrieron. Lamentablemente, debido a que es la primera incursión del PANI en temas de Inteligencia de Negocios, el diseño del Data Mart desarrollado es algo básico y excluye información que hubiera servido para generar análisis muy valiosos. Por ejemplo, debido a que no se dio importancia a la situación económica del padre en el sistema transaccional, no se pudo llevar esa información hacia el Data Warehouse, privando al modelo de análisis de cruces de variables que relacionaran el tema socio-económico con la agresión o abandono infantil y juvenil.

El uso eficaz de la Inteligencia de Negocio es todo un reto para una organización. También representa un beneficio potencialmente grande, que no puede ser demostrado fácilmente. De ahí que el PANI tendrá un papel elemental en el uso que se le dé a la tecnología que se está implementando.

Planteamiento del Problema

Problema General

¿Cómo proveer una herramienta de Minería de Datos que facilite el proceso de análisis ante las situaciones de amenaza o violación de los derechos de los niños, niñas y adolescentes?

Sub-problemas

1. ¿Cómo seleccionar algoritmos de Minería de Datos que se adapten a las necesidades del PANI y que además arrojen datos oportunos?
2. ¿Cómo desarrollar un modelo de Minería de Datos lo suficientemente robusto que permita satisfacer las necesidades de detección de amenazas?
3. ¿Cómo cargar la información del Data Warehouse en el modelo de Minería de Datos?
4. ¿Cómo realizar un análisis de los datos arrojados por el modelo de Minería de Datos implementado y validar que la información obtenida sea coherente?

Objetivo General

Proveer una herramienta de Minería de Datos que facilite el proceso de análisis ante las situaciones de amenaza o violación de los derechos de los niños, niñas y adolescentes, en el PANI.

Objetivos Específicos

1. Seleccionar los algoritmos de Minería de Datos que se adapten a las necesidades del PANI y que además arrojen datos oportunos.
2. Desarrollar un modelo de Minería de Datos lo suficientemente robusto que permita satisfacer las necesidades de detección de amenazas.
3. Desarrollar paquetes de carga y transformación para cargar la información del Data Warehouse en las estructuras de minería.
4. Realizar un análisis de los datos arrojados por el modelo de Minería de Datos implementado y validar que la información obtenida sea coherente.

Alcances

La herramienta brindará la información que el modelo de Minería de Datos pueda arrojar, de acuerdo a los algoritmos seleccionados para la investigación.

La información podrá ser consultada utilizando Microsoft Excel como la herramienta de despliegue y de interfaz al usuario.

Se proveerá de mecanismos de extracción, transformación y carga, para mantener las estructuras de Minería de Datos actualizadas.

Limitaciones

La información que alimente los modelos de Minería de Datos, será solamente aquella que brinde el PANI, y a esta información no se le considerarán problemas de sintaxis, ortografía o concordancia, por lo que se descarga la responsabilidad de los resultados que los análisis puedan arrojar.

La posibilidad de validar la información estadística, estará limitada al conocimiento que puedan tener los usuarios del PANI acerca de la información que se almacene en cada una de las estructuras desarrolladas para los diferentes modelos de Minería de Datos.

Como fuente de datos, se utilizará el Data Warehouse actual con el que cuenta el PANI para trabajar con sus cubos de información, por lo que cualquier omisión o error que éste tenga, no será corregido durante los procesos de extracción a las estructuras de Minería de Datos. Además dado que ya se ha hecho un proceso de limpieza de datos durante los procesos de carga del Data Warehouse, no se implementarán validaciones adicionales durante las extracciones a las nuevas estructuras de minería.

Metodología

Tipo de investigación

De acuerdo con las características de la presente investigación, se establece el enfoque cuantitativo ya que se sustenta en aspectos observables y susceptibles de cuantificar. Además, utiliza la estadística para analizar los datos recopilados. En cuanto al alcance, se define como correlacional/explicativo, pues pretende predecir el comportamiento de una variable al conocer el comportamiento de otras variables relacionadas; con lo que se estarían estableciendo las causas de casos de agresión o abandono.

Área de estudio

El estudio abarcará todo el territorio nacional, ya que los datos que se analizarán durante el proyecto corresponden a registros almacenados en la Base de Datos del sistema InfoPANI, por personal de las diferentes Sedes Regionales del PANI, a lo largo y ancho de Costa Rica.

Objeto y sujeto de estudio

El objeto de estudio es el medio ambiente que rodea a los niños que se ven envueltos en casos de agresión o abandono por parte de sus familiares. Se buscarán patrones que reflejen semejanzas en las características asociadas a todos los casos registrados (ubicación geográfica por ejemplo), para proyectar comportamientos basados en estadísticas.

Por su parte, el sujeto de estudio serán los niños que son agredidos o abandonados, así como sus familiares más cercanos, ya que se pretende analizar su situación social, económica y cultural, de manera que se puedan identificar patrones por medio de la agrupación de características similares de las personas, obtenidas de los datos registrados en la Base de Datos de InfoPANI.

Población y muestra

La población para esta investigación se compone de todos los niños que han sido sujetos de agresión o de abandono por parte de sus padres o familiares, y cuyos casos hayan sido registrados a través del sistema InfoPANI.

El sistema InfoPANI es relativamente reciente y no existe ninguna iniciativa para cargar la información que se haya generado en papel o en otros sistemas, por lo que la cantidad de registros actualmente está limitada al tiempo de funcionamiento del sistema.

Además se debe tener en cuenta que no se estará teniendo acceso a información de otras entidades que pudieran tener datos importantes, lo cual crea una limitante inicialmente, pues no se contará con históricos para el entrenamiento de los conjuntos de aprendizaje.

Fuentes de información

La fuente primaria de la información será el Data Warehouse del sistema InfoPANI, el cual cuenta con todos los registros de casos de agresión y abandono infantil reportados en las diferentes sedes del PANI.

Diseño de técnicas e instrumentos

El desarrollo de la investigación se apoyará en la utilización de métodos correlacionales que pretenderán analizar diferentes variables entre sí, las cuales representan características de los casos de agresión y abandono registrados en el sistema InfoPANI, para determinar patrones o semejanzas entre casos que permitan al PANI tomar acciones preventivas y no solo reactivas como lo ha venido realizando hasta el momento.

Para llevar a cabo estos análisis de datos, se hará uso de técnicas de Minería de Datos. Específicamente del Algoritmo de Árboles de Decisión y del Algoritmo de Clústeres. Así mismo, la herramienta de software que se utilizará para el desarrollo del proyecto será Microsoft SQL Server 2012, la cual provee de funcionalidades técnicas para realizar dichos modelos de minería.

Estado del Arte

A continuación se incluyen una serie de conceptos importantes que dan una visión mucho más amplia de lo que abarca el presente proyecto. Al definir cada una de estas palabras o áreas relacionadas con la Minería de Datos se pretende sentar las bases de la investigación y a la vez, crear un lenguaje común entre los desarrolladores del proyecto y todas aquellas personas que se encuentren interesados en la lectura del mismo.

Diferencia entre Conocimiento e Información

En la economía del conocimiento, debemos distinguir el concepto de conocimiento del de información: "*Poseer conocimiento, sea en la esfera que sea, es ser capaz de realizar actividades intelectuales o manuales. El conocimiento es por tanto fundamentalmente una capacidad cognoscitiva. La información, en cambio, es un conjunto de datos, estructurados y formateados pero inertes e inactivos hasta que no sean utilizados por los que tienen el conocimiento suficiente para interpretarlos y manipularlos*".

A pesar de que el conocimiento se basa en la información, ésta por sí sola no genera conocimiento.

Conocer y pensar no es simplemente almacenar, tratar y comunicar datos. Serán procesos de generalización de distinto tipo y sus resultados, los que determinarán el saber cómo actuar sobre algo en una situación dada. El desarrollar procesos de pensamiento alternativos, creativos e idiosincrásicos. La información no es en sí conocimiento. El acceso a ella no garantiza en absoluto desarrollar procesos originales de pensamiento.

Gestión del Conocimiento

La Gestión del Conocimiento es la obtención del conocimiento necesario por las personas adecuadas, en el tiempo, forma y lugar adecuados (Ackermans, Speel & Ratcliffe). Es un proceso sistemático e intencionado de creación, compartición y aplicación de conocimiento crítico para el desarrollo de la estrategia de negocio, las decisiones u operaciones que conlleva.

Son procesos pre-acordados que permiten mejorar la utilización del conocimiento y de la información que manejan las personas y los grupos. No es un proceso aleatorio, sino intencionado, que permite que las organizaciones que desean alcanzar mayores niveles de logro en sus resultados, lo hagan mediante una inversión consciente en la gestión del conocimiento que involucra a personas, nuevas instancias de trabajo colaborativo, recursos materiales y técnicos, etc.

El objetivo que persigue es lograr primeramente mentalizar a la organización del valor que efectivamente tiene para la empresa el desarrollo del conocimiento, transformándolo así en un nuevo y óptimo ACTIVO, un patrimonio, un capital efectivo de la organización.

En la medida que las personas viven procesos de formación permanente, ligados a sus tareas organizacionales y actualizan sus conocimientos y sus prácticas laborales, la empresa podrá obtener mejores resultados, sean estos productivos, afectivos, de inserción social, de bien común, etc.

Así y no perdiendo esta perspectiva se puede afrontar y enfrentar a la evolución y el progreso de las nuevas tecnologías de tal forma que en un futuro se cree una sociedad más humana y justa donde lo tecnológico y lo humano se integren.

La Minería de Datos pretende convertir la información que se obtiene a partir de los datos, en conocimiento que incluso va más allá del simple hecho de conocer lo que está sucediendo o ha sucedido hasta ahora, y busca proyectar a futuro lo que podría suceder, generando un conocimiento predictivo. En el presente proyecto, al aplicar la minería a un tema crítico como lo es el abandono y maltrato en niños y jóvenes, el objetivo que se busca es proyectar un futuro basado en datos reales del pasado y presente, para poder tener el conocimiento a buen tiempo para tomar acciones preventivas que disminuyan los problemas sociales con una inversión más acertada de la misma cantidad de recursos incluso, que la que se invierte actualmente.

OLAP vs Minería de Datos

Las herramientas OLAP proporcionan facilidades para “manejar” y “transformar” los datos. Producen otros “datos” (más agregados, combinados) y ayudan a analizar los datos porque producen diferentes vistas de los mismos. Las herramientas de Minería de Datos son muy variadas: permiten “extraer” patrones, modelos, descubrir relaciones, regularidades, tendencias, etc. Y producen “reglas” o “patrones” (“conocimiento”).

El análisis que realizan las herramientas OLAP es dirigido por el usuario, son consultas lanzadas sobre cubos OLAP que tienen la información precalculada y almacenada. Por el contrario, la Minería de Datos permite razonar de forma inductiva a partir de lo que se llaman vistas "minables" de datos para llegar a una hipótesis general que modele el problema.

Un ejemplo clarificará la diferencia entre ambas técnicas:

Una pregunta típica de un sistema OLAP sería: “El año pasado, ¿se reportaron más casos de agresión o abandono en el Valle Central o en Guanacaste?”. La respuesta del sistema a través de una consulta OLAP sería del tipo “En el Valle Central se reportaron 3.000 casos de agresión o abandono, mientras que, durante el mismo intervalo, en Guanacaste se reportaron 873”. Obviamente es una información interesante y útil, pero restringida por las hipótesis realizadas a priori.

En cambio, un problema típico para resolver utilizando Minería de Datos sería, por ejemplo: “Hallar un modelo que determine las características más relevantes de los padres de los niños o adolescentes que sufren de agresión o abandono”. Habría que construir una vista "minable" a partir de los datos del pasado que nos interese valorar como parámetros de entrada, y el sistema de Minería de Datos proporcionaría una respuesta del tipo: “Depende de la situación geográfica, rango de edad y situación económica familiar. Los habitantes del Valle Central que pertenecen a un cierto grupo de edad y nivel de ingresos probablemente reportarán más casos de agresión o abandono que gente de las mismas características en Guanacaste”.

Como puede verse, se trata de problemas distintos, de modo que según los objetivos perseguidos deberá utilizarse una técnica u otra. Además, puesto que sus conclusiones son complementarias, en general será conveniente combinar ambas para obtener los mejores resultados.

Interrelación entre OLAP y Minería de Datos

Como se aprecia en la Figura 1, tanto para un análisis OLAP, como de Minería de Datos, las fuentes de datos son las mismas Bases de Datos o archivos que utiliza una organización en su día a día. De igual manera, el proceso de extracción y transformación que da forma al Data Warehouse es el mismo para ambos análisis.

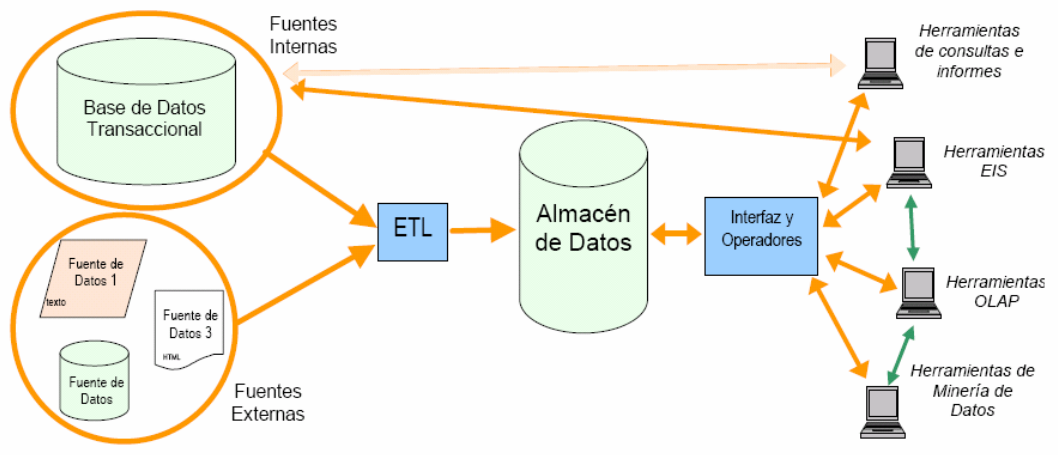


Figura 1. Fuentes de datos para Minería de Datos y Sistemas OLAP

Situación Actual

Capacidad actual en el análisis de la información

Actualmente el PANI cuenta con análisis limitado a través de reportes empotrados en su sistema InfoPANI, así como con un cubo de información OLAP, el cual se alimenta de la misma Base de Datos transaccional. A continuación se presenta una breve descripción de las dos fuentes de información con que dispone el PANI a través de los sistemas mencionados.

Reporteo

El sistema InfoPANI presenta reportes estáticos, los cuales están contenidos dentro de la aplicación, desde donde se conectan al servidor de la Base de Datos transaccional directamente, y tienen una interfaz estandarizada como se puede ver en la Figura 2.

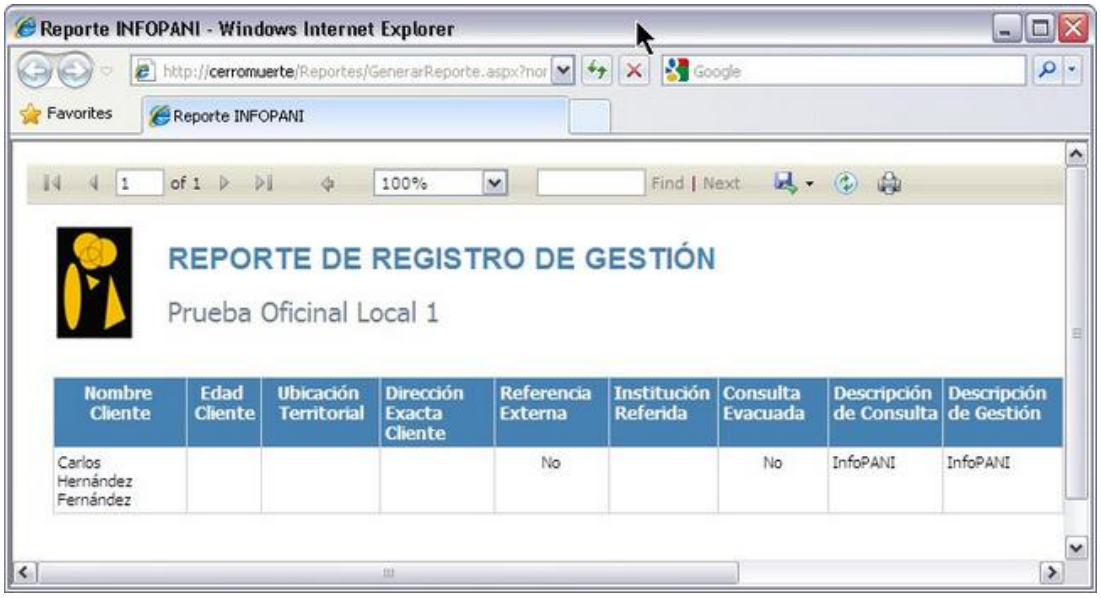


Figura 2. Reportes estáticos de InfoPANI.

Nótese cómo la información al ser estática, limita el conocimiento a la estructura única que fue diseñada por un usuario de la organización durante la etapa de análisis de la aplicación InfoPANI. Posiblemente esta vista de la información satisface una necesidad específica, por lo que se convierte en conocimiento pero sobre un tema muy puntual.

Si se quisiera utilizar el mismo método de reporteo para analizar otros conceptos que pudieran relacionarse con la información que se despliega en este reporte u otro, habría que hacer un esfuerzo muy significativo, para ampliar el alcance de la consulta que se realiza, o desarrollar un nuevo reporte. En ambos escenarios se requiere de un tiempo considerable de un desarrollador de software y acceso al código fuente de la aplicación. Habría que hacer una cirugía al sistema transaccional.

La principal ventaja de este tipo de análisis a través de reporteo, es que la información generalmente se encuentra en línea, por lo que si se debe tomar una decisión en el momento, es una excelente y confiable fuente de información, mientras que por lo general los análisis a través de OLAP tienen un "delay" en su procesamiento, precisamente para no saturar los sistemas transaccionales con extracciones pesadas de datos, y por otro lado los análisis de Minería de Datos tienen otros propósitos más estratégicos muy diferentes a la simple consulta de datos.

Cubos de Información OLAP

Cuando se construyó la primera fase del sistema InfoPANI se diseñaron las estructuras del Data Warehouse y los DataMarts necesarios para realizar una serie de análisis muy importantes para la organización. A continuación se presenta un análisis realizado sobre el modelo multidimensional de SQL Server Analysis Services que se desarrolló.

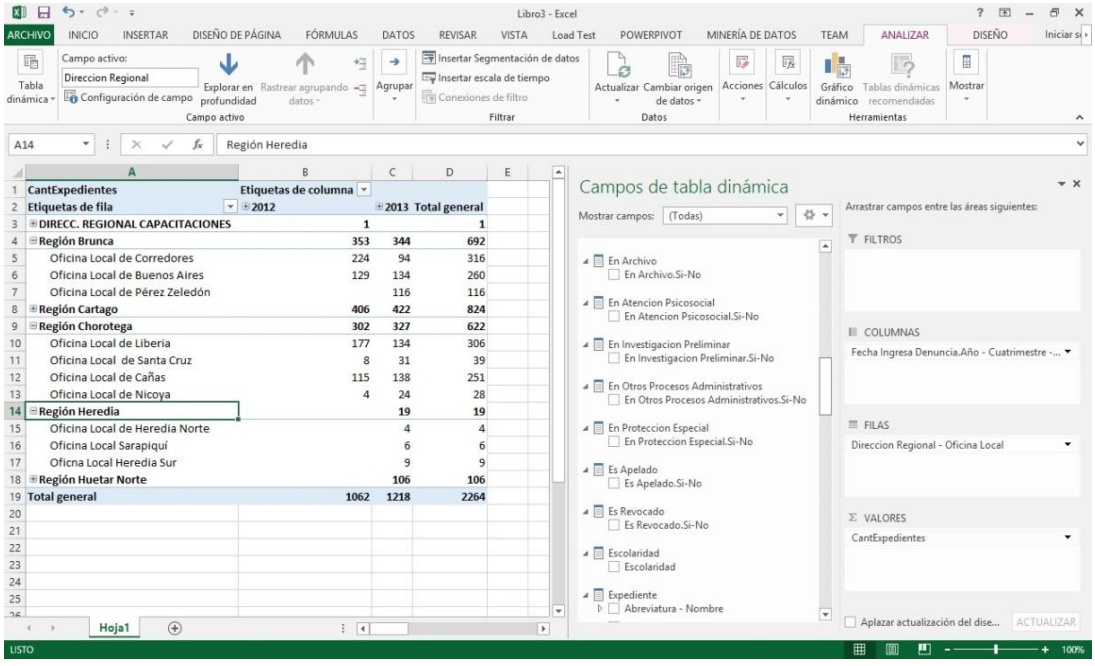


Figura 3. Cubo de información de InfoPANI.

En la Figura 3, se puede observar, el cubo de información puede ser "pivoteado" desde Excel para analizar los datos que han sido registrados en el sistema transaccional, es decir, estamos realizando un análisis post-mortem de los eventos que han acontecido en relación a los casos de agresión y abandono infantil y juvenil.

Capacidad deseada en el análisis de la información

La utilización de Minería de Datos pretende llevar el análisis y la toma de decisiones a un nivel predictivo que permita actuar proactivamente y no de manera reactiva como se ha venido haciendo hasta el momento. El escenario que se desea lograr es aquél en el que se pueda proyectar mediante un sistema de información, lo que podría suceder en la sociedad en relación con los casos de agresión y abandono a niños y jóvenes, si se presentan diversos factores ambientales que se han identificado como relevantes y que son almacenados en el sistema transaccional, día a día, en las diferentes oficinas localizadas a lo largo del país.

Adicionalmente, al aplicar la minería sobre los datos que actualmente se almacenan en la Base de Datos de InfoPANI, se desea llegar a obtener la capacidad de analizar distintos grupos o clústeres de casos que muestren características similares, para ir determinando patrones que marquen un comportamiento similar y que terminen en casos de alto riesgo. La intención de adquirir este conocimiento, es enfocar esfuerzos en aquellos grupos que cumplan con esas características en común que conforman los clústeres o grupos y así realizar un esfuerzo preventivo más enfocado y acertado, que las campañas que actualmente se publicitan por medios de comunicación masivos, que van dirigidos a toda la población en general y que posiblemente no estén llegando a los grupos más propensos a verse involucrados en casos de agresión y abandono infantil y juvenil.

Además, la Minería de Datos se utilizará para poder identificar las probabilidades de que un caso específico sea de alto riesgo o no, de acuerdo con la historia almacenada y los diferentes pesos que el algoritmo predictivo le dé a las variables de entrada, o características que rodean el caso. Así, se pueden realizar análisis "what if" en donde se simulen casos que cumplan con características específicas, para obtener un porcentaje de probabilidad de que dicho caso sea de alto riesgo o no. Este análisis a su vez debería indicar qué tipo de casos son los más propensos a convertirse en alto riesgo, sin necesidad de realizar escenarios ficticios.

Método What if? Para la determinación de riesgo inminente en los eventos reportados y registrados

El método what if es un método inductivo que utiliza información específica de un proceso, en este caso el flujo de atención de casos de agresión o abandono de menores de edad, para generar una serie de preguntas que son pertinentes durante el tiempo. Consiste en definir tendencias, formular preguntas, desarrollar respuestas y evaluarlas, incluyendo las posibles consecuencias.

Al contar con los modelos de Árboles de decisión y de Clustering se puede utilizar el método what if para analizar un caso específico una vez que es reportado y de acuerdo con las características del entorno que rodea el evento que está siendo reportado, se pueden responder las interrogantes mediante la información que ambos modelos proporcionan.

A manera de ejemplo, si se reporta un posible maltrato a un niño, con edad entre los 6 y los 10 años, que no asiste a la escuela, de padres divorciados, que vive con un familiar, cuya madre tiene cierta nacionalidad y el padre otra, además de que se ha determinado que su situación económica no es alentadora, entre otras características, podríamos realizar un análisis what if utilizando los modelos construidos, donde en el caso del árbol de decisión será el modelo quien nos irá guiando e indicando las principales variables a considerar y que debemos revisar para determinar si hay riesgo inminente o no. En el caso del modelo de clustering la evaluación se realiza diferente pues se debe revisar si las características reportadas en el caso coinciden con algún clúster identificado como un clúster de alta probabilidad de riesgo inminente.

Con el presente proyecto, se pretende llegar a crear un conjunto de aprendizaje lo suficientemente robusto, que cada vez que se ingrese un nuevo caso en la estructura de minería, se pueda clasificar según sus variables en sí es un caso de riesgo inminente o si no lo es, y así poder tomar acciones ya sea preventivas o reactivas para lograr minimizar el impacto de la denuncia. Además, los resultados de las ejecuciones de los algoritmos arrojaran indicios de cuáles son las variables que tienen más peso dentro de un caso de abuso o abandono infantil para poder tomar medidas preventivas con los grupos sociales de alto riesgo y así prevenir futuras denuncias.

Actualmente la base de datos del Data Warehouse, incluye los históricos de casos que han sido detectados como de alto riesgo y estos podrán ser considerados el conjunto de aprendiza con el cual, los nuevos casos tendrán base para ser clasificados como de alto riesgo o no.

Análisis de la brecha de estados: actual vs. deseado

Como se mencionó anteriormente, el PANI cuenta con un Data Warehouse construido a partir del sistema transaccional InfoPANI, por lo que ya posee mucho camino avanzado en la consecución de un modelo de Minería de Datos.

Una vez realizado el análisis del negocio, del sistema y de las estructuras de datos existentes, se deberá proceder a realizar el diseño de la estructura de minería, que estará conformada por las variables que se utilizarán para definir el medio ambiente que rodea un caso. Esta estructura deberá cargarse posteriormente mediante una consulta que se realizará para leer directamente del Data Warehouse existente.

Finalmente, se deberán aplicar los diferentes algoritmos de minería a la estructura creada, para obtener los resultados y tomar las mejores decisiones.

Desarrollo de la Solución

Análisis del Negocio

En esta primera etapa se establecieron reuniones con funcionarios del PANI para entender la necesidad que satisface el sistema transaccional InfoPANI en su primera fase y para comprender el ciclo completo de un caso, desde que se reporta un posible abandono o agresión, hasta que se descarta o se toman las medidas legales correspondientes.

Se determinó que el proyecto InfoPANI forma parte de un plan a largo plazo, por lo que no abarca el ciclo completo de la administración de los casos y por ahora lo más lejos que se puede llegar en su seguimiento, es hasta identificar si representa alto riesgo o no, y este determina el alcance del sistema de soporte de toma de decisiones del proyecto.

El interés del PANI al desarrollar esta primera fase, era digitalizar los casos que ingresan por diferentes medios, para poder llevar un expediente y dar trazabilidad a cada uno, en cualquier momento.

Se pudo determinar que hay un alto grado de interés en la información de la persona agredida, como por ejemplo su información personal, su nivel de escolaridad, su edad, su nacionalidad, entre otros.

Con la implementación del sistema InfoPANI y el cubo de información que se desarrolló en la primera fase de este proyecto estratégico a largo plazo, el PANI está incursionando en la utilización de Inteligencia de Negocios (B.I.) para analizar el medio ambiente en que se generan los casos de agresión y abandono que debe atender diariamente. Esta nueva tecnología le permitirá realizar análisis más flexibles y extensos que los que permiten los reportes tradicionales que vienen incluidos dentro de la aplicación.

Se determinó también que hay un comité encargado del proceso que automatiza el InfoPANI, pero no involucra personal de planificación, por lo que hay una gran capacidad de análisis que podría utilizarse para planificar la estrategia de la organización, pero no se está aprovechando.

Análisis del Sistema transaccional InfoPANI

Una vez que se contó con la descripción del negocio sobre la necesidad que cubre el sistema transaccional y el tipo de análisis que se desea realizar, se procedió a revisar el sistema transaccional, pantalla por

pantalla, para determinar la información que se va registrando en cada paso del flujo del proceso. Para esto se solicitó acceso al sistema actual, el cual se puede observar en la Figura 4.

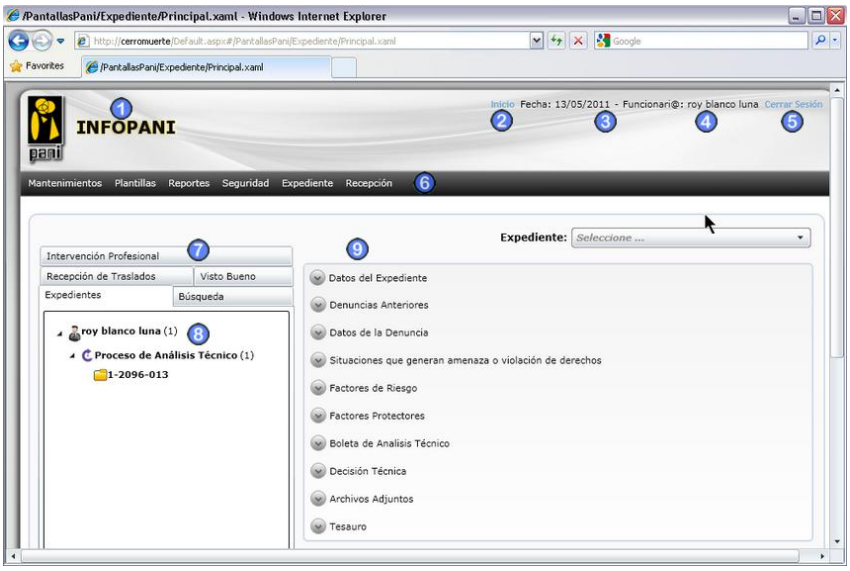


Figura 4. Sistema InfoPANI.

Se determinó que existe una cantidad importante de catálogos y que el alcance del sistema inicia con la apertura de un caso, pasando por el ingreso de información relacionada, y finalizando cuando se determina si es de alto riesgo o no. El sistema permite la trazabilidad de un caso de manera que si pasa por varios procesos de revisión, se pueda seguir el rastro y el resultado en cada revisión.

El sistema cuenta con los siguientes catálogos:

- Países
- Plantillas por proceso
- Direcciones Regionales
- Oficinas Locales
- Disciplinas
- Instituciones
- Grupo Étnico
- Forma de Notificación
- Tipo de Discapacidad
- Ubicación Geográfica
- Factores
- Forma de Ingreso

- Situaciones que generan amenaza

Además, el sistema automatiza los siguientes procesos:

- Creación de un nuevo expediente
- Análisis Técnico
- Investigación Preliminar
- Atención Psicosocial
- Protección en Sede Administrativa
- Procesos Administrativos y Judiciales

Adicionalmente, el sistema InfoPANI cuenta con reportes y búsquedas, así como con un módulo de seguridad para autorizar el ingreso y establecer los permisos de los usuarios.

Análisis técnico de las estructuras fuentes

Tal como se ha mencionado anteriormente, el sistema InfoPANI posee su propia Base de Datos transaccional sobre SQL Server, así como un Data Warehouse que hospeda un Data Mart. Se logró determinar que el Data Warehouse posee información limitada para poder desarrollar el modelo de minería, pero que la Base de Datos transaccional posee información adicional que no está siendo trasladada al Data Warehouse y que es de importancia para el análisis de los casos.

Se estableció una lista de consultas preliminares que podrían ser resueltas por el modelo de minería, para proceder a buscar la información en el Data Warehouse o el transaccional, sin embargo poco a poco se fueron descartando algunas de ellas, ya que tanto el OLTP como el Data Warehouse no almacenan cierta información que podría ser importante, como por ejemplo información personal de los padres o la situación socio-económica de la familia de la persona agredida.

En el Anexo A se muestra el diseño entidad-relación de la Base de Datos del OLTP, mientras que en el Anexo B se muestran los diseños entidad-relación de las Bases de Datos Staging y Data Warehouse. Se puede apreciar como hay información como el estado civil de la madre o el padre que se encuentra en el OLTP pero no en el Data Warehouse, así como también hay carencia de información en ambas Bases de Datos.

Se comprendieron las estructuras de estrella diseñadas en el Data Warehouse y se revisaron las dimensiones y las métricas existentes para asegurar que los datos se estaban cargando correctamente a través de los ETL's.

Análisis del problema

El problema en la prevención de casos de abandono o agresión a niños y jóvenes se centraliza en la determinación de cuándo un evento reportado se puede clasificar rápidamente como un potencial caso con riesgo inminente o no, sin necesidad de que avance en el flujo de atención administrativo definido.

Para ello se deberán medir otras variables y calcular a partir de ellas mediante un modelo adecuado la predicción deseada. No se dispone de dicho modelo porque se conocen sólo las principales variables de entrada, pero no las relaciones existentes entre ellas. Para este tipo de problemas, la minería de datos ayuda a encontrar un modelo que represente una aproximación de las relaciones con un grado de probabilidad.

Selección del modelo

Para el modelado del problema se han seleccionado dos modelos distintos para analizar los resultados desde diferentes perspectivas y ver si con ambos se pueden determinar conclusiones similares o si existe discordancia en sus resultados.

Árboles de decisión

El algoritmo de árboles de decisión es un algoritmo de clasificación y regresión para el modelado de predicción de atributos discretos y continuos.

Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos. Utiliza los valores, conocidos como estados, de estas columnas para predecir los estados de una columna que se designa como elemento de predicción. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción. El árbol de decisión realiza predicciones basándose en la tendencia hacia un resultado concreto.

Para los atributos continuos, el algoritmo usa la regresión lineal para determinar dónde se divide un árbol de decisión.

Clustering

El algoritmo de clústeres es un algoritmo de segmentación. El algoritmo utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones.

Los modelos de agrupación en clústeres identifican las relaciones en un conjunto de datos que no se podrían derivar lógicamente a través de la observación casual.

El algoritmo de agrupación en clústeres se diferencia de otros algoritmos de minería de datos, como el algoritmo de árboles de decisión, en que no se tiene que designar una columna de predicción para generar un modelo de agrupación en clústeres. El algoritmo de agrupación en clústeres entrena el modelo de forma estricta a partir de las relaciones que existen en los datos y de los clústeres que identifica el algoritmo.

Diseño de la nueva estructura para minería

La solución de Minería de Datos planteada, se basa en un origen de datos relacional, lo cual trae como ventaja el poder reunir datos, entrenar y actualizar el modelo sin la complejidad de crear un cubo.

Una vez que se revisaron las variables con que se cuenta en el sistema InfoPANI para definir los ambientes que rodean los casos de agresión o abandono de niños y adolescentes, se diseñó la estructura de minería a utilizar, a partir de la información relevante que se pudo extraer de las tablas que componen el Data Warehouse y que el mismo PANI indicó es importante tomar en cuenta en el momento de determinar si un caso específico es un riesgo inminente de agresión o abandono.

Se determinó que existen variables dentro del Data Warehouse que se pueden utilizar para dar mayor peso a las hipótesis que se puedan extraer con los algoritmos de minería, sin embargo hay algunas de ellas que no fueron cargadas correctamente al Data Warehouse por lo que se tuvo que recurrir a su extracción directamente de los sistemas transaccionales. Al ser variables importantes como lo son la escolaridad y la edad de los padres, se decidió hacer su extracción y posterior unión a los datos del Data Warehouse.

Después de analizar todas las variables con las que se cuenta en los sistemas, se procedió a discriminar bajo la supervisión del PANI cuales son relevantes y cuáles no, y además se discriminó cual sería la variable objetivo, determinando que se utilizaría la variable riesgo inminente, como predictivo de los modelos.

A continuación se detallan las variables que se obtendrán del Data Warehouse:

- **FechaIngresaDenuncia**: fecha en que se registró la denuncia
- **Expediente**: código del caso dentro del sistema
- **GrupoEtnico**: grupo étnico del menor
- **Nacionalidad**: nacionalidad del menor
- **Escolaridad**: grado de escolaridad del menor
- **CondicionLegal**: condición legal del menor
- **Género**: sexo del menor agredido o abandonado
- **FormaIngreso**: forma en que ingresó el caso al PANI
- **GrupoEdad**: rango de edad en que se ubica el menor
- **Provincia**: provincia en que vive el menor
- **Canton**: cantón en que vive el menor
- **Distrito**: distrito en que vive el menor

Las variables que se obtendrán directamente del OLTP son las siguientes:

- **EstadoCivilPadre**: estado civil del papá del niño o joven afectado
- **NacionalidadPadre**: nacionalidad del papá del niño o joven afectado
- **GeneroPadre**: sexo del papá del niño o joven afectado
- **GrupoEdadPadre**: rango de edad del papá del niño o joven afectado
- **EstadoCivilMadre**: estado civil de la mamá del niño o joven afectado
- **NacionalidadMadre**: nacionalidad de la mamá del niño o joven afectado
- **GeneroMadre**: sexo de la mamá del niño o joven afectado
- **GrupoEdadMadre**: rango de edad de la mamá del niño o joven afectado
- **RiesgoInminente**: indicador de que el caso ha sido identificado como alto riesgo

Una vez seleccionadas las variables a utilizar, sus tipos de dato y definir las fuentes de donde se va a proceder a hacer la extracción de los datos, se procedió a crear una base de datos en SQL Server 2012 que va a contener la estructura de Minería de Datos a utilizar. Luego, dentro de esta base de datos, se hizo la creación de una tabla en la cual se incluyeron todos los campos con los tipos de datos requeridos para poder almacenar la información a utilizar. Esta estructura se puede observar en la Figura 5. Se debe

mencionar que durante la creación de la estructura se tuvo que tomar importantes decisiones de diseño, entre las cuales se puede mencionar:

- Las columnas dentro de la tabla utilizada para albergar la estructura de Minería de Datos no contendrá códigos transaccionales, sino que contendrá las descripciones de los atributos requeridos. Por ejemplo en lugar de guardar el valor 1 para nacionalidad costarricense, dentro de la estructura se guarda el texto costarricense; esto lleva a que la estructura sea totalmente desnormalizada.

- En los sistemas OLTP no se encuentra definido el grupo de edad ni del menor, ni de los padres, lo que se encuentra es la edad exacta, por lo que se debió proceder a realizar una clasificación dentro de grupos preestablecidos que fueron de interés para los funcionarios del PANI.

- Los tipos de datos de cada uno de los atributos de la tabla se hicieron coincidir con los tipos de datos de los sistemas transaccionales para así no tener que hondar en transformaciones en el momento de las extracciones y así hacer los ETL's lo más simple posible.

Figura 5. Estructura de minería a utilizar.

Después de crear la estructura de la tabla, se creó un proyecto de Integration Services, el cual se encargará de mantener actualizada la estructura de Minería de Datos mediante la carga diaria de la nueva información que se vaya incluyendo a los sistemas transaccionales de los cuales se alimenta esta solución.

Se decidió crear una base de datos y una tabla independiente en lugar de crear simplemente una vista en el Data Warehouse, ya que esto permite poder independizar el servidor de Minería de Datos del servidor de sistemas transaccionales y además del servidor de cubos, pudiendo así no interferir un proceso con otro y a la vez mejorar el rendimiento de los mismos. Esta arquitectura propuesta, además se utiliza con el fin de poder crear una independencia total de los sistemas OLTP durante los tiempos de consulta y solamente consultarlos durante los tiempos de carga para así no interferir en su rendimiento.

Para la implantación del presente proyecto, se requirió contar con un servidor con Microsoft SQL Server 2012 instalado y configurado adecuadamente para contar con los servicios de base de datos, Integration Services y Analysis Services en modalidad multidimensional, además se requirió de contar con un usuario con los suficientes permisos para la creación de bases de datos y estructuras de tablas. Además se requirió un usuario con los suficientes permisos para poder ejecutar paquetes de Integration Services y además que su contraseña contara con la política de no expiración de contraseña, esto para poder satisfacer la necesidad de automatización y simplicidad en los procesos de carga de los datos.

Es importante mencionar que la solución de Minería de Datos planteada es totalmente escalable, pues permite la incorporación de nuevas variables de entrada, modificación del predictivo y adición de variables informativas. Además los procesos de carga creados son también escalables y no deben ser recreados por cada modificación que se requiera, lo cual hace que sea una solución muy flexible que se pueda ir adaptando a las necesidades del PANI.

Carga de la estructura para minería

Para realizar la carga y mantenimiento de los datos a la estructura de Minería de Datos, se utilizó un paquete de Integration Services el cual se instaló en el servidor para automatizar esta tarea.

En el anexo C se puede ver la consulta que se realiza al Data Warehouse y al OLTP, donde se obtiene del primero toda la información del niño o adolescente, que ya fue cargada por los ETL’s existentes y que se encargan de poblar el Data Warehouse. La información de los padres o representantes legales, así como la identificación de si el caso es un riesgo inminente o no, se obtiene del OLTP y se liga al query por medio de JOINS entre tablas de las diferentes Bases de Datos, lo cual es posible de una forma transparente debido a que la extracción de la información del Data Warehouse proviene de este mismo sistema OLTP.

En la Figura 6, se muestra los 2 tiempos de carga que se utilizaron para poblar la estructura de Minería de Datos creada. Se puede mencionar que se hará una única carga inicial o completa de toda la información que existe en la base de datos del sistema InfoPANI, y luego, diariamente se cargará únicamente la información que se vaya ingresando al sistema para así mantener la estructura de minería actualizada.

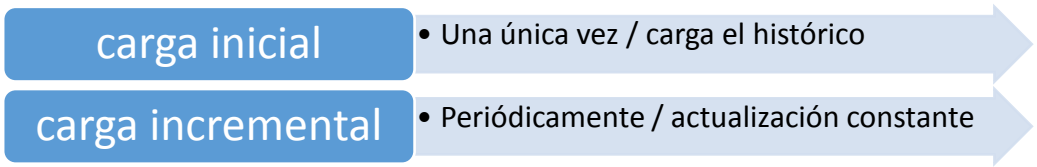


Figura 6. Tiempos de carga de la estructura de minería.

Esta consulta se almacenó en un Procedimiento Almacenado, el cual a su vez se incrustó en un paquete de Integration Service de carga inicial y otro de carga incremental, de manera que se haga un primer llenado con el histórico de todos los datos registrados y posteriormente se haga una actualización de la estructura periódicamente para mantener el modelo actualizado.

Procesamiento del modelo de minería

Mediante la plataforma de Microsoft Visual Studio Shell 2010 se logró procesar el modelo de minería a partir de la estructura de datos descrita anteriormente, ingresándola como fuente de datos.

Es importante mencionar que los modelos de Minería de Datos en la plataforma SQL Server 2012 de Microsoft, son parte del servicio Analysis Services, por lo tanto este es un prerrequisito indispensable para el presente proyecto. Además, este servicio a partir de la versión 2012 de SQL Server, incluye 2 modalidades de configuración, el modo tabular y el modo multidimensional, siendo este último el único que tiene la capacidad de manejar modelos de Minería de Datos.

Para la creación del modelo de Minería de Datos en Microsoft Visual Studio Shell 2010, el cual es el IDE incluido en Microsoft SQL Server 2012, se siguieron los siguientes pasos:

1. Se seleccionó un proyecto multidimensional y de Minería de Datos de Analysis Services, pues como ya se mencionó es el único que actualmente soporta estructuras de Minería de Datos.

2. Se creó un origen de datos, el cual apunta al servidor de bases de datos donde se encuentra la instancia de la base de datos que contiene la estructura de minería.

3. Luego, se realizó la creación de un componente de vista de origen de datos, el cual tiene como función filtrar todas aquellas tablas que puedan existir en la base de datos de origen y solamente mostrar las que realmente vayan a ser utilizadas en el modelo.

4. Después de cumplir con los pasos anteriores, se procedió a crear la estructura de Minería de Datos requerida. Durante la creación de esta estructura de minería, se pide se seleccione cual es el algoritmo a utilizar, cuáles serán las variables de entrada, cuál será el predictivo y si existen algunas otras variables que serán utilizadas meramente como información adicional y que no se incluirán como variables de entrada.

5. Para el presente proyecto, una vez que se creó el modelo y se le aplicó el primer algoritmo, se procedió a aplicar otros algoritmos sobre la misma estructura ya creada, lo cual es fácilmente hecho una vez que la estructura está definida y el primer algoritmo aplicado.

6. Se procede a hacer la implementación en el servidor y a procesar la estructura para incluir los datos que se encuentran en la tabla de Minería de Datos creada anteriormente.

El análisis predictivo pretende determinar la incidencia de un conjunto de variables de entrada, sobre una variable de salida, de manera que podamos determinar cuáles variables tienen mayor influencia sobre el comportamiento que pueda presentar esta última. En la Figura 7 se puede observar el proyecto de Minería de Datos desarrollado, en el cual se aprecian las variables de entrada (Input) seleccionadas para los algoritmos involucrados en la solución.

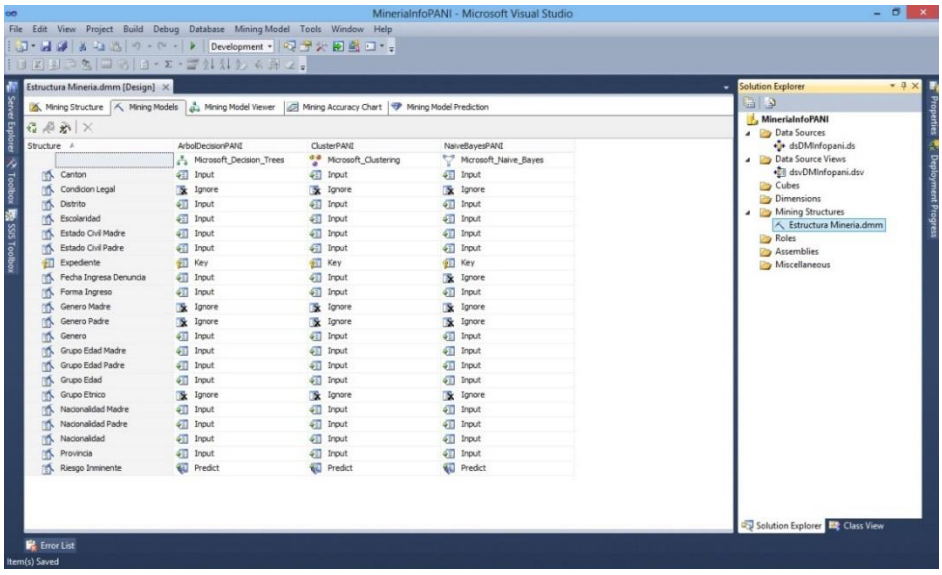


Figura 7. Proyecto de Minería de Datos en Visual Studio.

De igual manera, se puede apreciar que se ha determinado indicar como variable de salida (Predict) la variable llamada “Riesgo Inminente”, ya que es precisamente el objetivo del análisis.

El posterior procesamiento de la solución será automatizado mediante un paquete de Integration Services, el cual se encargará de procesar la estructura de Minería de Datos una vez que se haya extraído de las fuentes la nueva información y cargado a la estructura de Minería de Datos. Este procesamiento se hace la recomendación que sea ejecutado diariamente para así mantener la información lo más fresca posible, teniendo en cuenta que el sistema para la toma de decisiones tendrá un atraso de un día con los sistemas transaccionales.

Sin embargo, el procesamiento de la información también podrá hacerse manual en caso de ser requerido con solo ejecutar la tarea programada que orquestará los diferentes paquetes de Integration Services requeridos para la carga de los datos.

El proceso de carga de la información se puede apreciar en la Figura 8. En este proceso, la extracción y consolidación del sistema InfoPANI y el Data Warehouse actual del PANI se hace mediante ETL’s de Integration Services hacia la base de datos de minería que se encuentra en el servidor de bases de datos y de ahí mediante otro ETL de Integration Services, se hace el procesamiento de la base de datos de Analysis Services donde está el modelo de minería y los algoritmos disponibles para los usuarios.

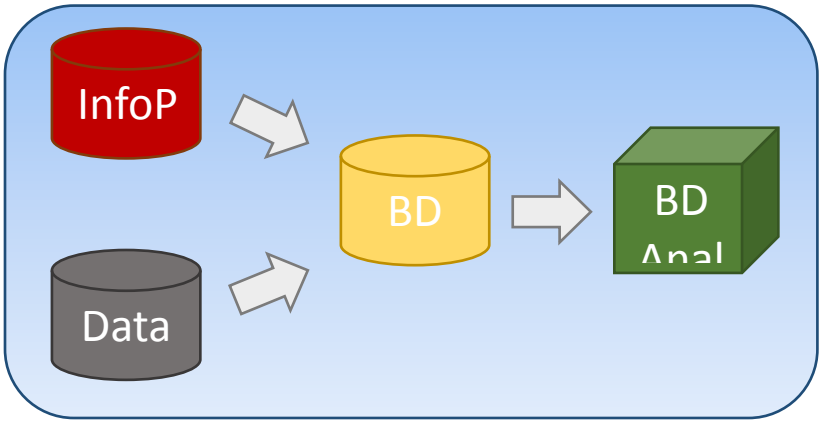


Figura 8. Procesamiento de modelo de Minería de Datos.

En el proceso descrito en la Figura 8, es importante rescatar que no se hace limpieza de información para los datos provenientes del Data Warehouse pues se nos indicó que los datos ya pasaron por un proceso de QA antes de entrar a él; sin embargo para la información proveniente del sistema InfoPANI, si se aplican técnicas simples para asegurar que los valores que se ingresan no contengan información nula o inconsistente con los datos esperados por la estructura de minería.

Riesgo Inminente

Debido a que el proyecto InfoPANI abarca un flujo de procesos que actualmente se encuentra en su primera fase: "Expediente Electrónico", lo más avanzado en el proceso que se puede analizar es la determinación de riesgo inminente, sin poder afirmar o descartar si realmente se trata de un caso de abuso o abandono, ya que para esto se requiere información que se obtendrán en próximas etapas.

De acuerdo con el análisis realizado al inicio del proyecto, el riesgo inminente significa que el caso en estudio tiene altas probabilidades de ser dictaminado como agresión o abandono al menor de edad por parte de sus padres o familiares.

El modelo se procesó mediante la aplicación de diferentes algoritmos para analizar posteriormente los resultados de cada uno y determinar cuál algoritmo servirá más para lograr cubrir las expectativas del PANI, en torno a esta herramienta.

Visualización y administración de la solución

Para la visualización de los algoritmos y el análisis por parte de los usuarios finales, se utiliza la herramienta Microsoft Excel 2010 o 2013, a la cual se debe instalar un complemento llamado "Microsoft SQL Server 2012 Data Mining Add-ins for Microsoft Office", el cual realiza la conexión a la base de datos de Analysis Services y muestra los algoritmos existentes.

Tal como ya se ha mencionado, los objetos de Minería de Datos se almacenan dentro de una instancia de Analysis Services y este permite tanto la modificación de sus objetos en modo conectado como desconectado, lo que da una mayor flexibilidad a los usuarios avanzados del sistema, los cuales podrían implementar un cambio a los objetos en producción, o mediante una solución desconectada la cual deberá ser instalada en el servidor posterior a su desarrollo.

Al utilizarse una herramienta conocida como los es Excel, la resistencia que se pudiera presentar a utilizar la aplicación disminuye, viéndose más como una extensión a la aplicación de cubos que como algo nuevo.

Resultados

Presentación, análisis e interpretación de los resultados

Análisis mediante el método de Árboles de Decisión

El primer algoritmo aplicado fue el algoritmo de Árboles de decisión, obteniendo los siguientes resultados:

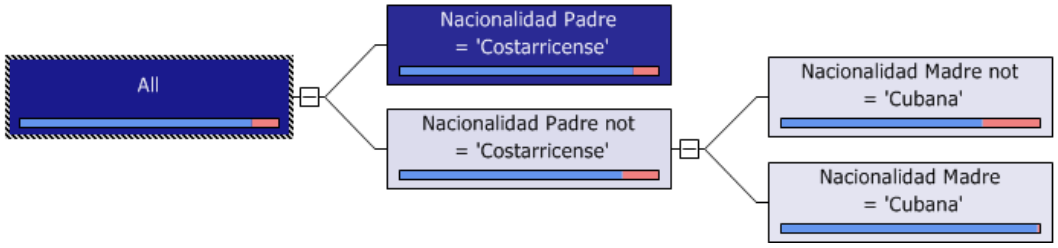


Figura 9. Resultado de algoritmo de Árboles de Decisión.

En la Figura 9 se muestra el árbol resultante y se puede observar cómo la Nacionalidad del Padre se convierte en la variable de mayor peso al determinar si un caso será o no de riesgo inminente. Al ser este un algoritmo de clasificación y al estar aplicado sobre atributos discretos, la predicción se hace sobre la correlación que tienen los atributos con el predictivo. En este caso se puede ver como la mayor correlación existente entre los datos y el predictivo (riesgo inminente) se da en el atributo nacionalidad del padre, el cual se determinó por el algoritmo como un buen elemento de predicción para un riesgo inminente. Este algoritmo realiza sus predicciones basándose en la tendencia de búsqueda de un riesgo inminente.

En la ejecución realizada para este proyecto se puede observar en la Figura 9, que la división en el árbol o nodos sólo se realiza por el atributo nacionalidad del padre, pues fue el único atributo que mostró una correlación lo suficientemente significativa con el predictivo para ser representado. Dado el resultado de la ecuación que utiliza el algoritmo para la obtención de la información, se puede notar que el atributo antes mencionado fue el que obtuvo el mejor puntaje para poder dividir los casos en subconjuntos, y estos a su vez ser analizados por independiente.

De este segundo análisis independiente es de donde se puede observar que para los casos en los cuales la nacionalidad del padre es diferente de “costarricense”, el siguiente atributo que obtuvo el mejor puntaje en la ecuación y que por consiguiente es el que muestra mayor correlación con el predictivo fue la nacionalidad de la madre en especial cuando su valor es “cubana”.

Para la ejecución del algoritmo que se realizó para el presente proyecto, se obtuvo que la nacionalidad de los padres de los niños o adolescentes, tiene una alta correlación con el predictivo, es decir, la presencia de un riesgo inminente para los menores, lo cual es un factor a tomar en cuenta durante los análisis de la información.

Esta información no debe servir para estereotipar a ninguna persona o grupo social, pero sí para prestar mayor atención a los casos que puedan provenir de aquellos que indique el sistema que están teniendo mayor porcentaje de incidencia en riesgos inminentes.

Con este resultado arrojado por el algoritmo de árboles de decisión, podemos comenzar a ver patrones importantes dentro de los casos que se registran diariamente en las oficinas del PANI a lo largo del país, donde en este caso el algoritmo está indicando que se debe prestar mucha atención a aquellos casos que ingresan donde el padre es extranjero y la madre también, pero específicamente esta segunda es de nacionalidad cubana. A estos casos el PANI debería brindarles especial atención desde el primer momento en que la información es registrada en el sistema y no esperar a que pase por todo el flujo de procesos para ser clasificado como riesgo inminente.

Igualmente, ya con este resultado el PANI podría enfocar sus esfuerzos en conjunto con embajadas de países extranjeros por ejemplo, para hacer campañas de prevención a la agresión y el abandono.

Al analizar la red de dependencias del mismo algoritmo, se obtiene el grafo que se muestra en la Figura 10.

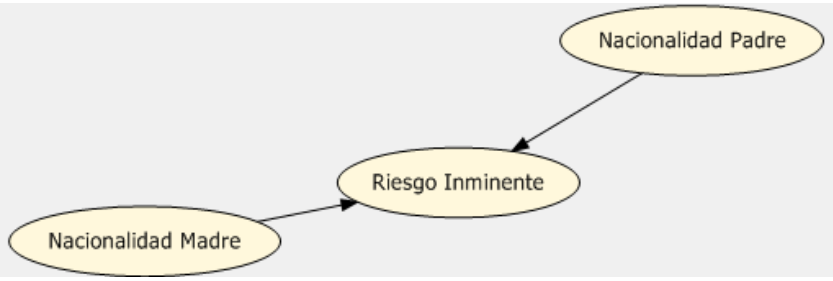


Figura 10. Red de dependencias del algoritmo de árboles de decisión.

Se puede apreciar cómo de todas las variables introducidas, el algoritmo detectó solamente dos como las más influyentes en el resultado final, las cuales corresponden a la Nacionalidad del Padre y la Nacionalidad de la Madre y descartó el resto de variables, pues no influyen en el resultado buscado que es la indicación de si un caso será o no Riesgo Inminente.

Análisis mediante el método de Clustering

El segundo algoritmo utilizado se conoce como Clustering y pretende identificar grupos o clústeres que presenten características similares, que se relacionen fuertemente entre sí y que se diferencien de otros clústeres que posean características diferentes.

A pesar de la separación de los clústeres, lo más importante es identificar en cuáles el riesgo inminente tiene alta probabilidad para perfilar esos grupos y tratar de ubicarlos en la sociedad de una manera más específica y atender rápidamente los casos que presenten las mismas características.

El resultado arrojado por el algoritmo de Clustering es el siguiente:

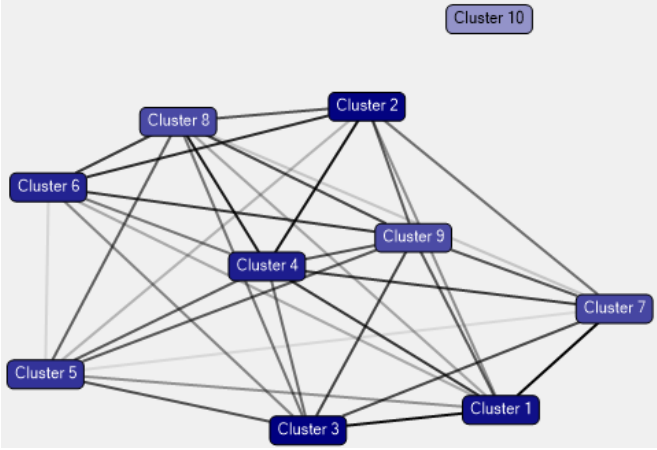


Figura 11. Resultado de algoritmo de Clústeres.

Nótese en la Figura 11, como el algoritmo realizó 1 agrupaciones o clústeres. Cada grupo o clúster generado por el algoritmo corresponde a la identificación de patrones similares en los casos registrados y el color azul se intensifica conforme más grande sea el conjunto y mayor similitud haya entre sus componentes. Las líneas que unen el grafo también se intensifican de acuerdo con la semejanza en las características de un clúster y otro.

Se puede notar que el clúster 10 fue separado del resto automáticamente por el algoritmo, ya que los casos que posean dichas características tienen una probabilidad muy baja de convertirse en riesgo inminente, debido a que sus valores están por debajo del umbral especificado para tener correlaciones fuertes con otros clústeres, lo cual implica que dado un caso la probabilidad de que se clasifique en un grupo con las características asignadas al clúster 10 sea muy baja.

Si se analizan los clústeres, se podrá observar que nuevamente las variables de nacionalidad de los padres juega un papel fundamental, al igual que en la aplicación del algoritmo de árboles de decisión.

Sin embargo, en esta ocasión podemos observar en la tabla de los perfiles de los clústeres, cómo la variable Nacionalidad de la Madre indica que dentro de este clúster un 84% de los casos corresponden a madres de nacionalidad cubana. Este dato es de suma importancia, ya que con el algoritmo anterior se pudieron adelantar criterios y generar prejuicios o estereotipos sobre las madres cubanas, cuando en este otro análisis podemos ver como si bien hay muchos casos de madres cubanas registrados, la historia nos ha dicho que con una combinación con otras variables ambientales, como por ejemplo donde la Nacionalidad del Padre es Cubana también (78%), tenemos resultados muy diferentes a los que se podrían suponer, luego de ver el algoritmo de Árboles de Decisión.

Continuando con el análisis de clústeres, se puede analizar toda la población en términos de probabilidades, de acuerdo a una distribución en todas las combinaciones de valores de las variables que la determinan. Así por ejemplo, en la Figura 12 se puede observar cómo la probabilidad de que el riesgo no sea inminente para toda la población de casos registrados equivale al 90% de probabilidad, lo cual es un dato positivo.

Variables	Valores	Probabilidad
Riesgo Inminente	False	90 %
Nacionalidad	Costarricense	90 %
Nacionalidad Padre	Costarricense	88 %
Nacionalidad Madre	Costarricense	87 %
Genero	Femenino	53 %
Genero	Masculino	46 %
Distrito	ausente	44 %
Estado Civil Madre	Soltero(a)	39 %
Estado Civil Padre	Soltero(a)	37 %
Escolaridad	SIN DEFINIR	36 %
Estado Civil Padre	Casado(a)	34 %
Grupo Edad Padre	60+	32 %
Estado Civil Madre	Casado(a)	31 %
Grupo Edad Madre	60+	30 %

Figura 12. Análisis de clústeres.

En este análisis también se puede observar igualmente cómo la nacionalidad se convierte en una variable que genera relaciones fuertes entre los clústeres, lo que significa que es determinante para predecir el riesgo inminente de un caso.

Finalmente, cada clúster puede analizarse individualmente desde la perspectiva de su complemento, obteniendo un gráfico tal como se muestra en la Figura 13.

Variables	Valores	Favorece Cluster 1	Favorece Complemento de Cluster 1
Grupo Edad	Infancia		
Escolaridad	SIN DEFINIR		
Estado Civil Padre	Soltero(a)		
Grupo Edad Madre	15 - 20		
Estado Civil Madre	Soltero(a)		
Grupo Edad Padre	21 - 25		
Escolaridad	III CICLO (7,8 Y 9 ANO DE SECUNDARIA)		
Grupo Edad	Adolescencia Primera Etapa		
Escolaridad	II CICLO (4,5 Y 6 GRADO DE PRIMARIA)		
Grupo Edad Madre	41 - 45		
Nacionalidad Madre	Costarricense		
Grupo Edad Padre	41 - 45		
Grupo Edad Madre	36 - 40		
Provincia	Guanacaste		
Grupo Edad Madre	21 - 25		
Grupo Edad Padre	46 - 50		
Grupo Edad	Adolescencia Segunda Etapa		
Nacionalidad Madre	Nicaraguense		
Estado Civil Padre	Divorciado(a)		
Estado Civil Madre	Divorciado(a)		

Figura 13. Análisis de clúster y su complemento.

En la Figura 13 se pueden apreciar las variables y los valores de las variables que más favorecen un clúster en específico, así como su complemento.

Conclusiones y Recomendaciones

La Minería de Datos se convierte en una herramienta muy potente para el análisis del negocio, basado en su propia información. Lejos de desplazar el BI tradicional, la minería pone al descubierto algunos patrones de comportamiento, difícilmente perceptibles en un análisis tradicional realizado mediante reporteo o consultas OLAP.

El PANI al ser una institución encargada de la prevención y detección de casos de abuso y agresión infantil, y al ser una institución que tiene alto impacto sobre la sociedad costarricense, al incursionar en el ámbito de Minería de Datos, está logrando agilizar sus procesos de detección de riesgos en los menores y logrando actuar de una manera más rápida para lograr evitar que los casos presentados lleguen a extremos lamentables. Es aquí donde el proyecto actual ha cobrado importancia, pues colabora directamente en lograr este objetivo, permitiendo alcanzar un nivel de mayor agilidad y disminuyendo los costos al realizar esfuerzos más focalizados a la necesidad de los diferentes segmentos de la sociedad.

Al finalizar el desarrollo del modelo de Minería de Datos y utilizar el mismo para realizar algunos análisis sobre la información que se procesó, podemos obtener las siguientes conclusiones y brindar algunas recomendaciones:

Conclusiones

- El contar con una herramienta de Minería de Datos, colabora al PANI en la prevención de casos de agresión y abandono de menores, pues cada vez que se ingrese un nuevo caso, este podrá ser comparado contra los que ya han sido clasificados como riesgos inminentes y así poder tomar medias rápidamente para evitar llegar a casos extremos.

- La Minería de Datos se convierte en el PANI en una herramienta muy potente para el análisis de su propia información. Lejos de desplazar el BI tradicional, la minería pone al descubierto algunos patrones de comportamiento, difícilmente perceptibles en un análisis tradicional realizado mediante reporteo o consultas OLAP, lo cual colabora sensiblemente con el PANI colaborando a prevenir de una forma más directa los posibles casos de abandono o agresión infantil.

- Dado que las bases de datos del PANI tienen varias inconsistencias en cuanto a campos que deberían contener datos y simplemente están con valores nulos, esto causa que variables que serían de interés para las predicciones, no puedan ser utilizadas y tengan que ser descartadas del

modelo hasta que los sistemas transaccionales sean corregidos, pues ingresar estas variables con valores default podrían causar un sesgo en los resultados y no reflejar la realidad.

- La Minería de Datos aplicada en un entorno social, como lo es el PANI, puede ayudar a prevenir la ocurrencia de problemas, como casos de abandono o agresión, mediante el enfoque de campañas dirigidas a los grupos de mayor riesgo, de acuerdo con las características ambientales que los definen; lo que repercute a nivel interno en ahorro económico al no lanzar actividades o publicidad a todos los sectores, sino poder hacer segmentación de mercados y aplicar campañas o publicidad de acuerdo a los problemas que presenta cada grupo de conflicto y así maximizar su aprovechamiento por los sectores a los que va dirigida.

- En la Base de Datos transaccional de InfoPANI no se está almacenando información que podría contribuir sustancialmente a un análisis más específico, como por ejemplo el estatus socio-económico de los padres del niño o a adolescente, la escolaridad de los padres o el grupo étnico de los padres. Esto causa que se dé una limitante a los análisis y proyecciones que se pueden hacer por medio de la herramienta de minería.

- Al momento de aplicar cualquiera de los dos algoritmos (árboles de decisión y clústeres), no se obtiene una segregación muy amplia debido a que el sistema recientemente se puso en Producción y no cuenta con mucha información registrada.

- El contar con una solución de Minería de Datos, lleva al PANI a poder realizar un análisis de los casos ingresados de una manera más eficiente y con resultados más certeros, pues anteriormente los análisis de los casos se realizaban únicamente con el criterio de la persona que está analizando el caso. Esto lleva a un mayor aprovechamiento del personal con el que se cuenta, que pueden dedicarse a las labores propias de su puesto en lugar de estar revisando cuales casos podrían clasificarse como riesgos inminentes.

Recomendaciones

- Debido a que este es el primer proyecto de Minería de Datos en el cual se incursiona en el PANI, se deben formar varios equipos que puedan dar mantenimiento a la solución, o recargar funciones a algunos empleados ya existentes. Entre estas funciones se podrían citar:

- Data Warehouse: debe estar compuesto por analistas y por desarrolladores. Entre las responsabilidades que debe cumplir este equipo se encuentra el mantenimiento a las estructuras de Data Warehouse y de Minería de Datos, así como a los ETL's encargados de todos los movimientos de información. Otra función que deberán cumplir los miembros de este equipo es la implementación de todos aquellos requerimientos planteados por el equipo de calidad de datos o el equipo de Minería de Datos.

- Calidad de datos: este equipo debe estar compuesto por personas que se encarguen de validar la información tanto en los sistemas transaccionales, en el Data Warehouse y en las estructuras de Minería de Datos, para asegurar la completitud y exactitud de la información y que así la información que se requiera sea lo más consistente posible y que cumpla con todos los requerimientos de las diferentes herramientas que la utilicen.

- Minería de Datos: Este será el equipo encargado de revisar los resultados arrojados por las soluciones de Minería de Datos, además de solicitar a los desarrolladores cualquier cambio o implementación de un nuevo algoritmo para ir ajustando las soluciones cada vez más a la realidad del negocio, o si se requiere la incorporación de alguna nueva variable al sistema. Este equipo debe estar conformado por especialistas del negocio y estadistas que tengan conocimiento en Minería de Datos.

- Se recomienda que se hagan revisiones y ajustes a los diferentes procesos del sistema transaccional y a las extracciones del Data Warehouse, esto con el fin de evitar que se esté almacenando información que no es valiosa. Además es importante que en el sistema transaccional se incluyan validaciones para no permitir que información sensible de cada uno de los casos pueda ser dejada en blanco. Y, se debe concientizar a los usuarios de los sistemas de la importancia de incluir todos los datos durante el ingreso de los nuevos casos para lograr obtener resultados más certeros en los diferentes sistemas de toma de decisiones.

- Para la solución de este proyecto, se recomienda como visualizador de datos Microsoft Excel, por la facilidad de adaptación que tendrán los usuarios y también por el ahorro significativo en licencias, ya que el PANI ya cuenta con las licencias de este producto. Sin embargo se hace la recomendación que se pueda a futuro utilizar herramientas como Microstrategy Suite, la cual permite la conexión a proyectos de Analysis Services y de Minería de Datos, permitiendo mayor cantidad de opciones de visualización y además incorporando funcionalidades interesantes como lo es la visualización y consulta de los modelos por medio de dispositivos portátiles como lo son celulares y tabletas.

- Se recomienda a las unidades de detección de casos en riesgo inminente que puedan cuantificar la cantidad de tiempo ahorrado y dinero antes y después de implementar las soluciones de Minería de Datos, y poder presentar informes a otras unidades para que puedan adoptar modelos similares orientados a cada área de interés dentro del PANI, lo cual podría llevar a una reestructuración de procesos internos y a una eficacia mayor de toda la organización.

- Se podría a nivel transaccional hacer conexiones con los sistemas del Tribunal Supremo de Elecciones y con los sistemas de Migración y Extranjería para obtener todos los datos faltantes de los casos que tiene el PANI para así lograr tanto completar datos sensibles como poder hacer comparativas de si la información suministrada es verdadera, lo cual llevaría a poder tratar cada caso con mayor fiabilidad de los resultados y de las variables involucradas.

- Dado que este proyecto está completamente basado en la primer etapa del sistema InfoPANI, se hace la recomendación que por cada etapa subsiguiente, se haga una revisión del modelo actual, e incluir algunas variables que se puedan considerar importantes o considerar crear un modelo de minería independiente para cada una de las etapas dependiendo de su nexo con la etapa anterior, esto con el fin de continuar con el crecimiento de los sistemas de Minería de Datos en paralelo al crecimiento de las soluciones transaccionales.

- Involucrar a áreas o departamentos de planificación dentro del proceso de análisis de la información, ya que los presupuestos y proyectos que se desarrollan en la institución podrían tener mejores resultados si están alineados con los resultados arrojados por la Minería de Datos, y así crear una cultura organizacional en la cual la Minería de Datos sea importante para todos y pueda también extenderse a otras áreas de la organización o en el mejor de los casos a otras organizaciones del sector público.

Referencias Bibliográficas

Espinoza, Roberto. (2009). *Herramientas ETL. ¿Que son, para que valen? Productos más conocidos. ETL´s Open Source*. Recuperado de http://www.dataprix.com/blogs/respinosamilla/herramientas-etl-que-son-para-que-valen-productos-mas-conocidos-etl-s-open-sour

Microsoft Developer Network. (2012). Algoritmos de Minería de Datos (Analysis Services: Minería de Datos). Recuperado de http://msdn.microsoft.com/es-es/library/ms175595.aspx

Microsoft Developer Network. (2012). *Algoritmos de Minería de Datos (Analysis Services: Minería de Datos).* Recuperado de http://technet.microsoft.com/es-mx/library/ms175595.aspx

Microsoft Developer Network. *Conceptos de Minería de Datos*. Recuperado de http://msdn.microsoft.com/es-es/library/ms174949.aspx

Microsoft Developer Network. (2012). *Tutorial de SSIS: Crear un paquete ETL simple*. Recuperado de: http://msdn.microsoft.com/es-cr/library/ms169917(v=sql.105).aspx

Patronato Nacional de la Infancia (2012). *Sobre el PANI*. Recuperado de http://www.pani.go.cr/index.php?option=com_content&view=article&id=50&Itemid=60

Sivakumar Harinathand, Stephen R. Quinn. (2006). *Professional SQL Server Analysis Services 2005 with MDX*. Wiley Publishing, Inc, Canada.

Xindong Wu, et al. (2007). *Top 10 Alorithms in Data Mining*. Springer-Verlag, London.

ICONS Soluciones Informáticas, S. L. (2010). OLAP vs Minería de Datos. Recuperado de http://www.icons.es/business-intelligence/61-OLAP-datamining

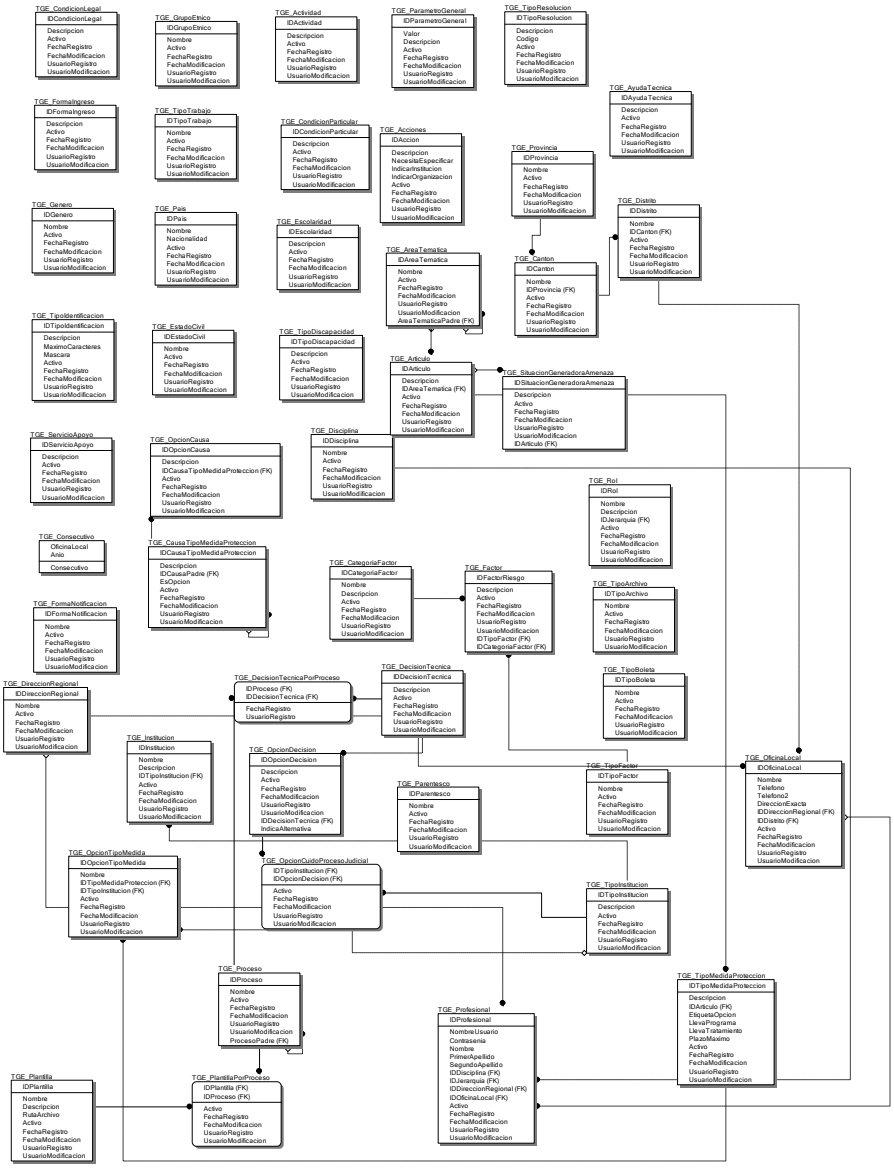
Maneiro, Mariela Yanina (2008). Monografía de Adscripción: "Minería de Datos". Recuperado de http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosYany2008.pdf

Anexo A

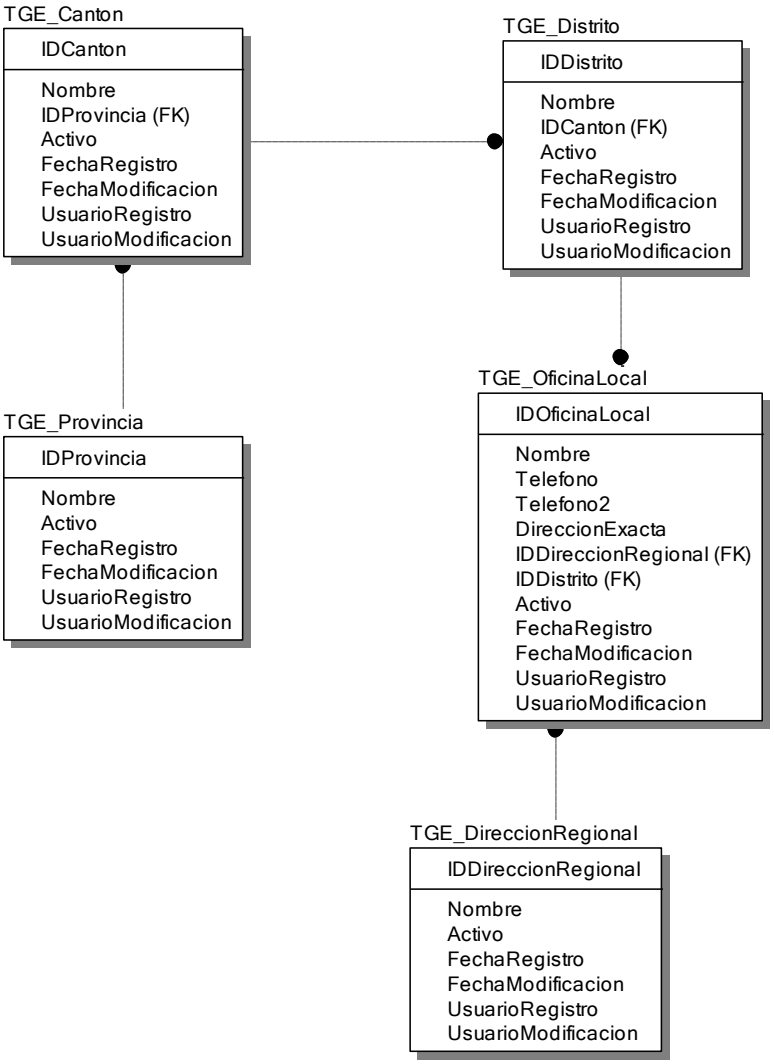
Modelo de Entidad-Relación de la Base de Datos transaccional actual:

A continuación se muestra el diagrama de la Base de Datos transaccional donde se almacenan todos los datos relacionados con los casos de agresión o abandono diariamente.

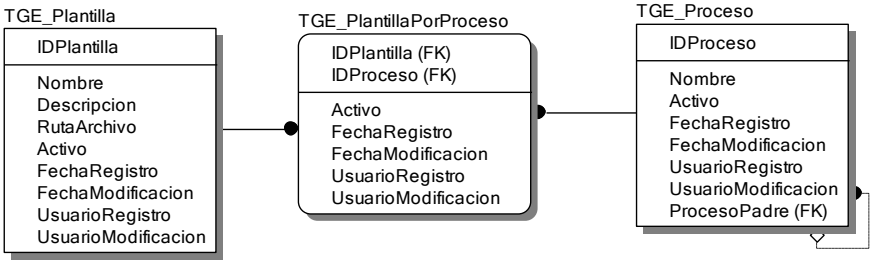
Catálogos



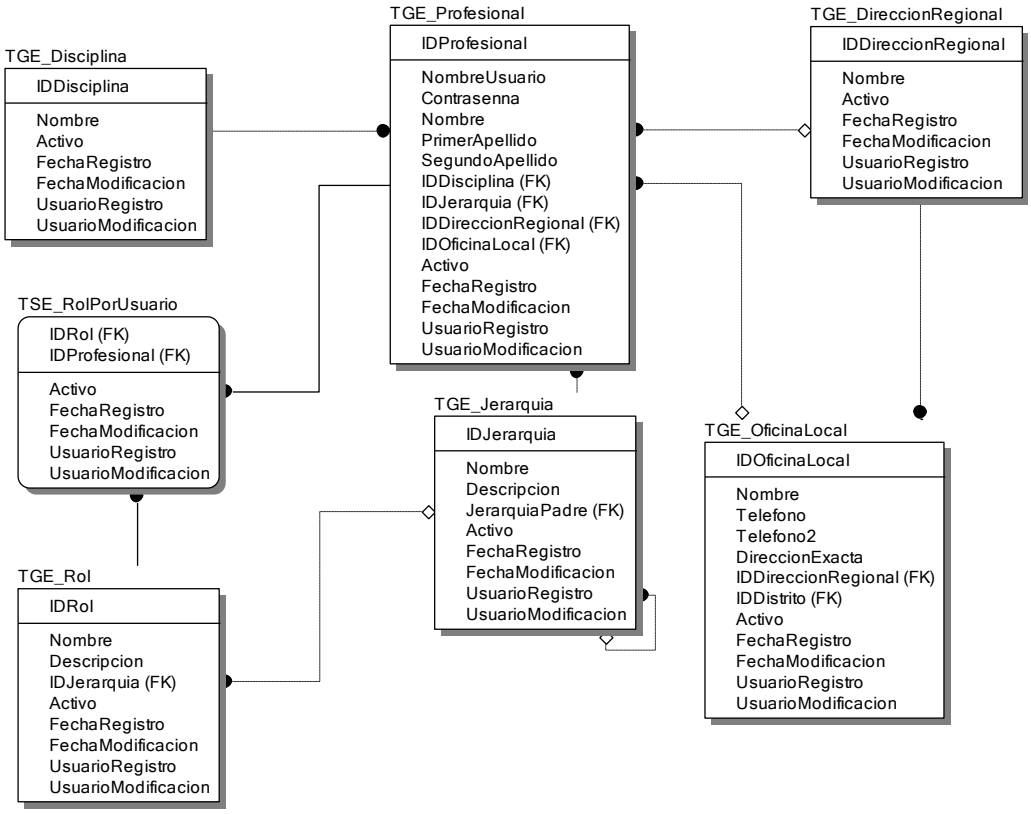
Ubicación Geográfica



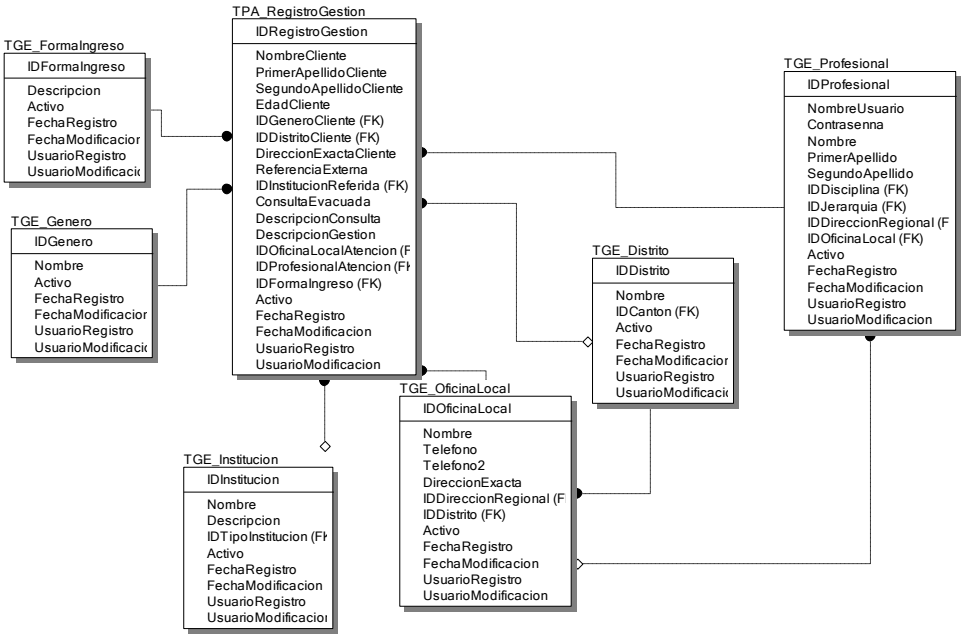
Plantillas



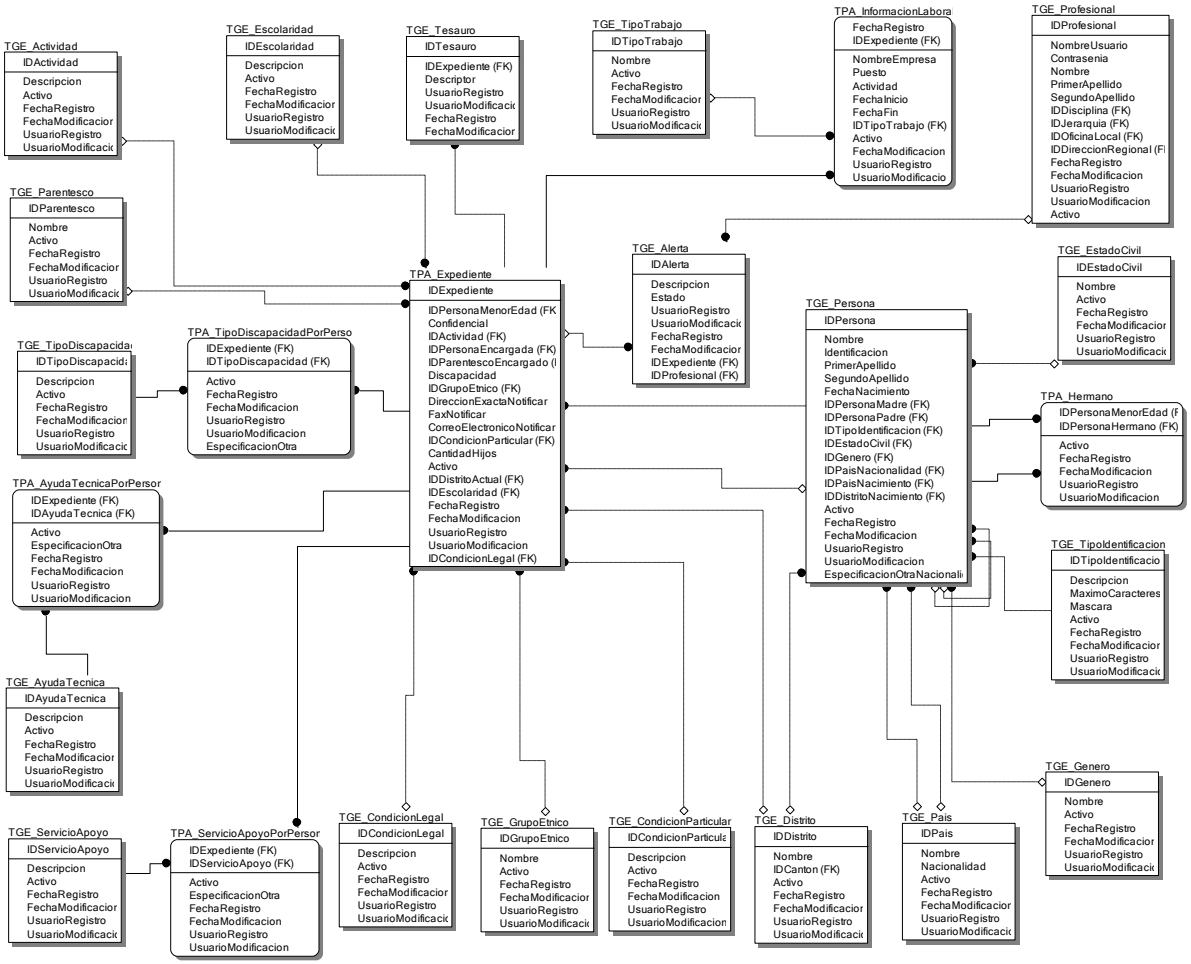
Funcionarios



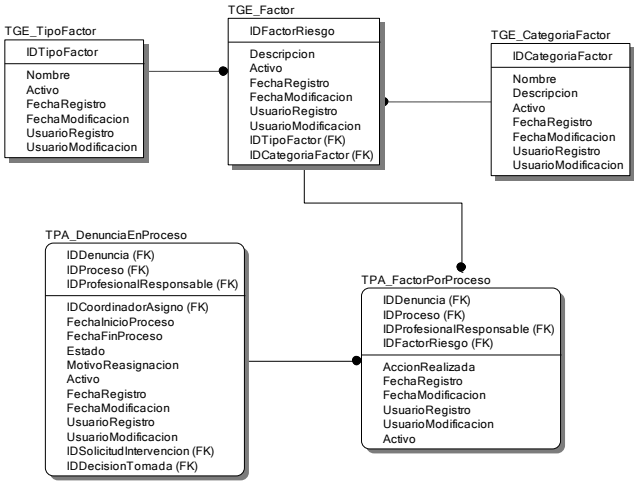
Registro de Gestión



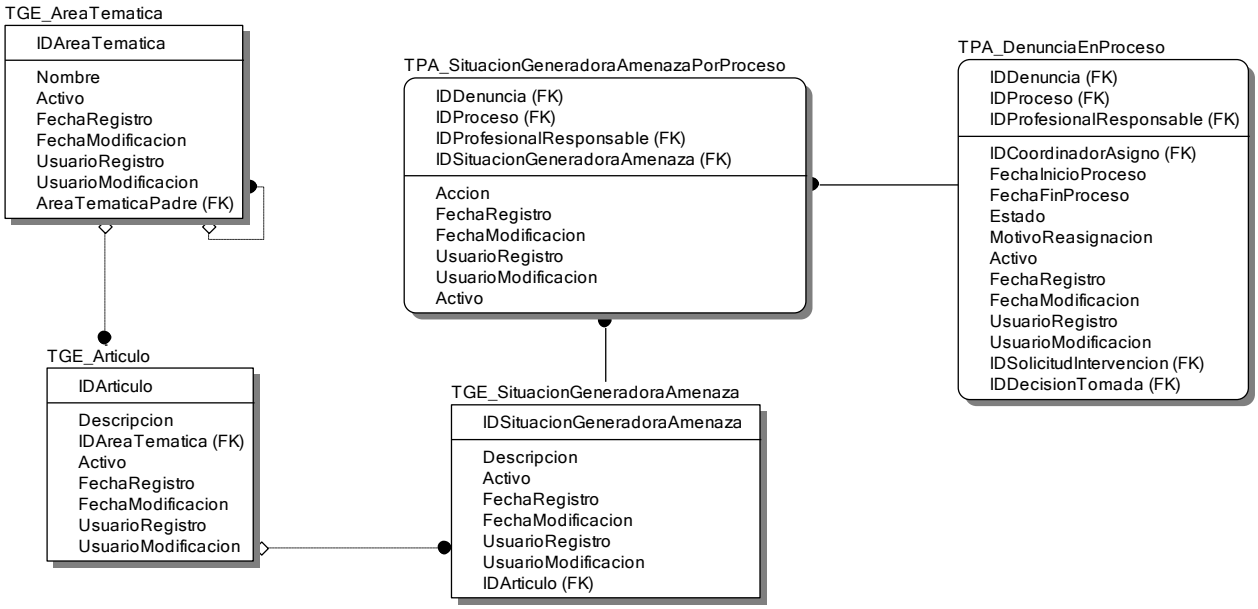
Expedientes



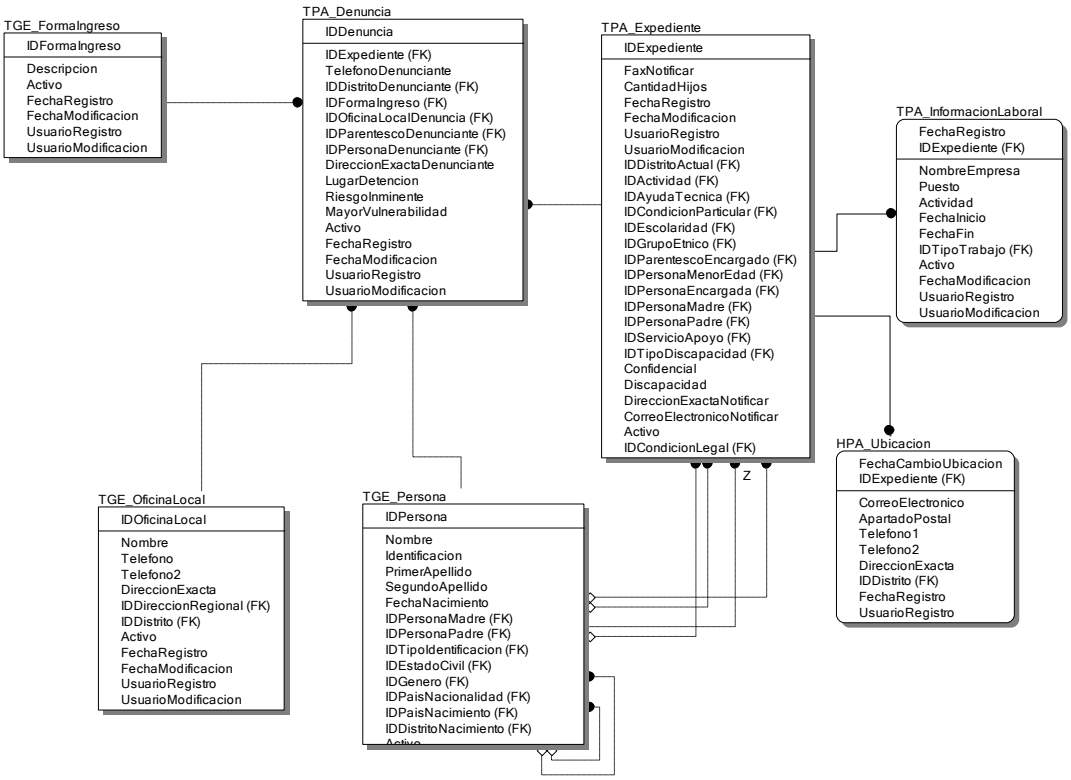
Factores de Riesgo y Factores Protectores



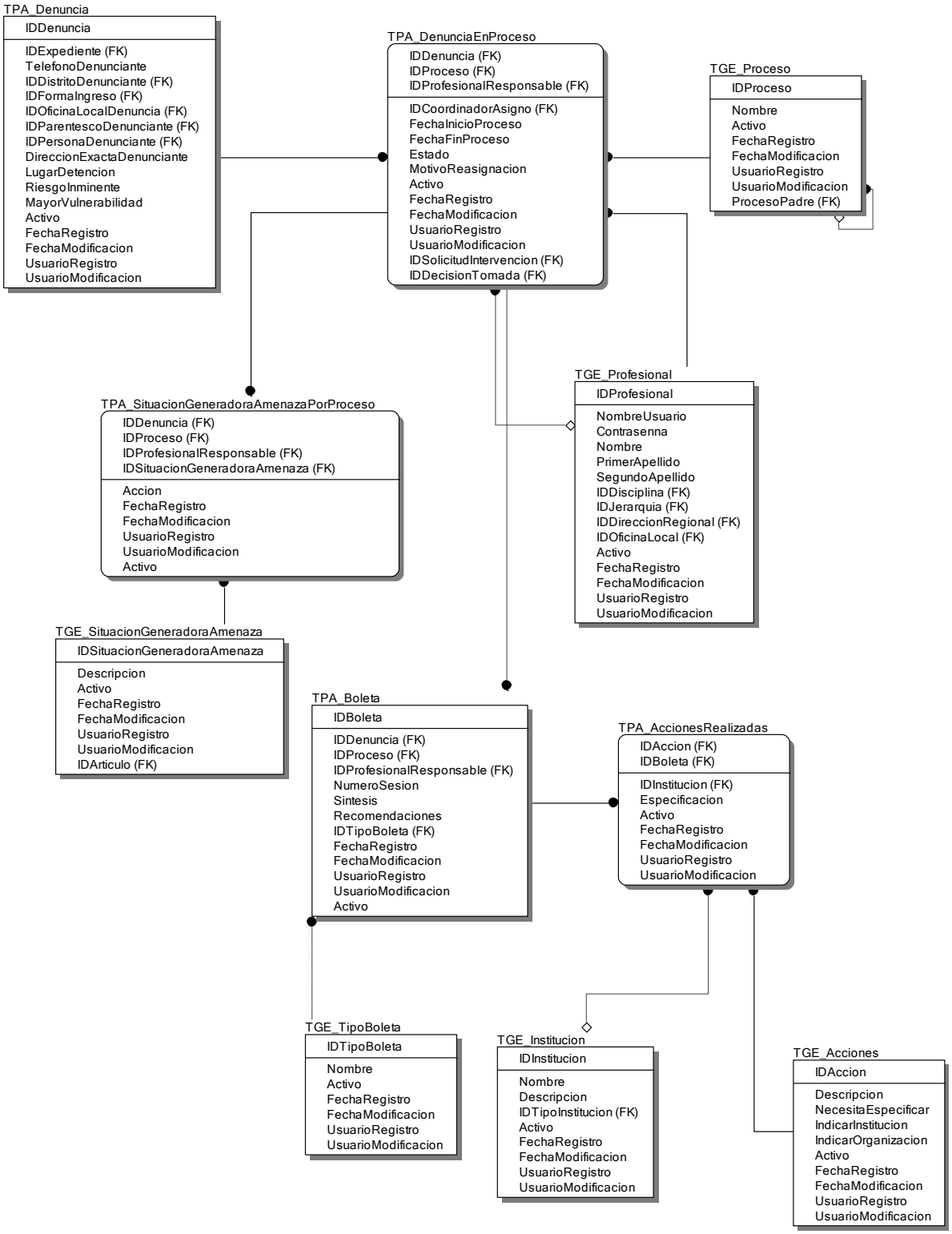
Situaciones que generan amenaza o violación de derechos



Denuncias

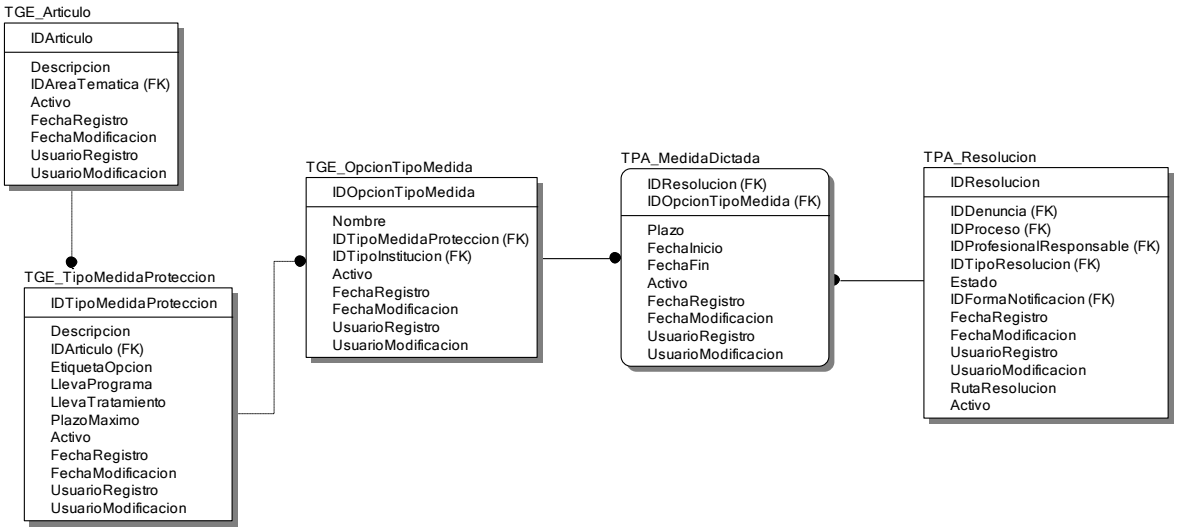


Procesos y servicios

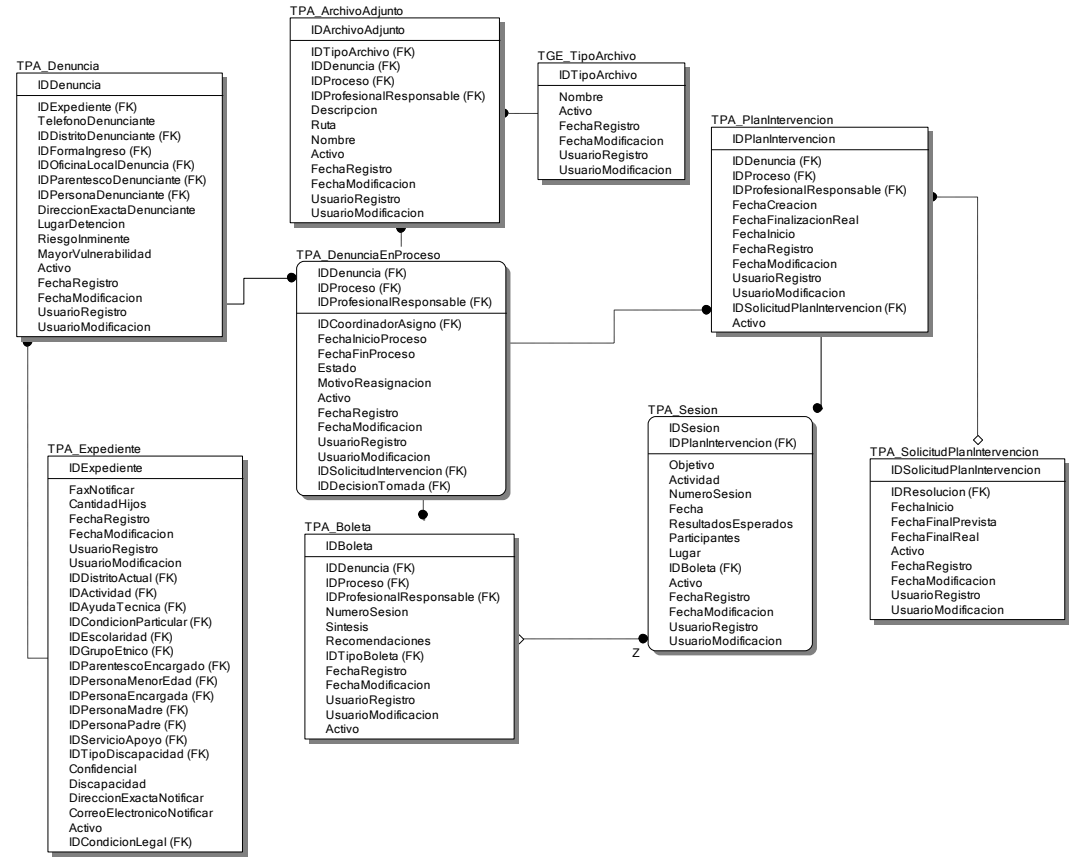


Proceso Especial de Protección en Sede Administrativa

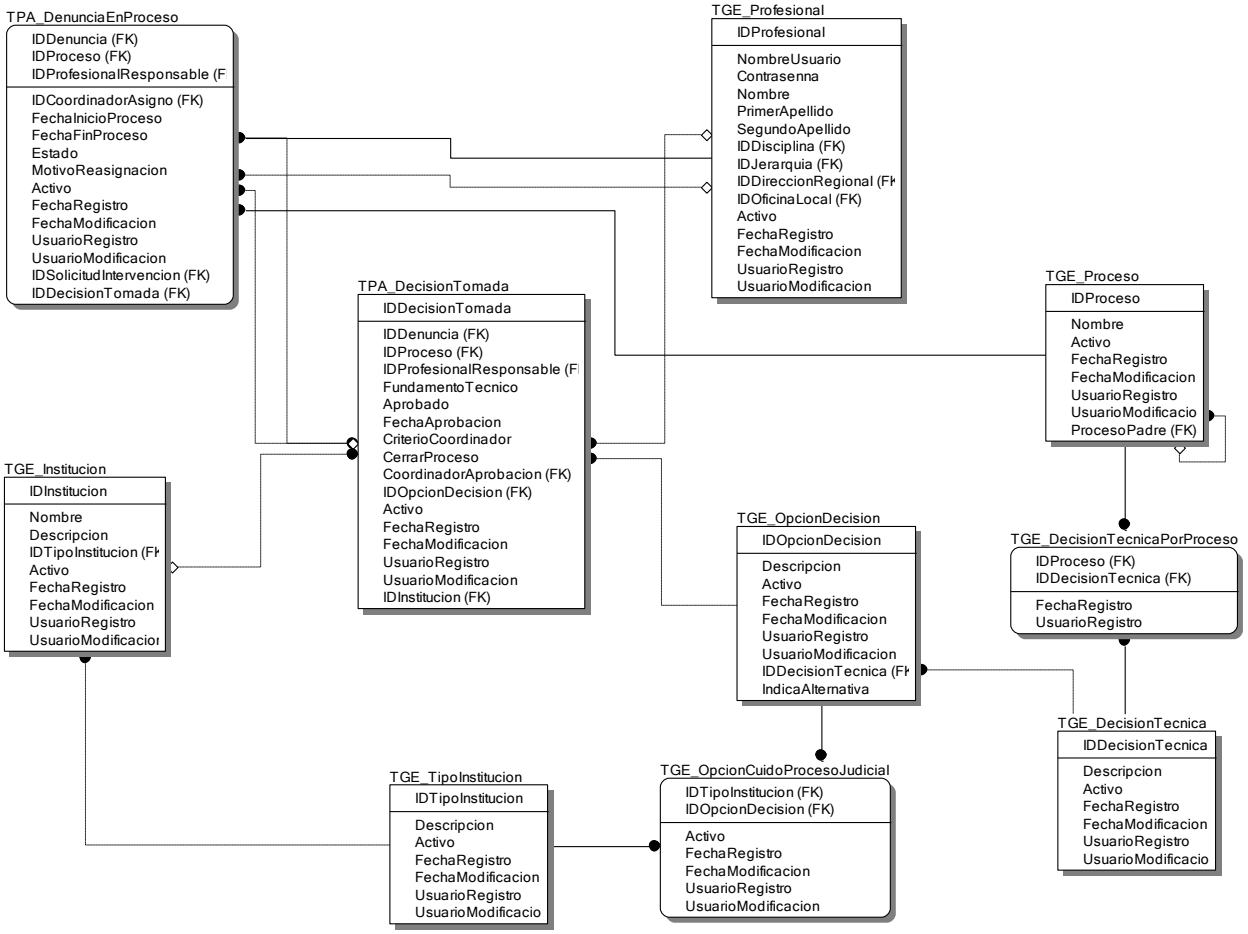
Tipos de medidas de protección



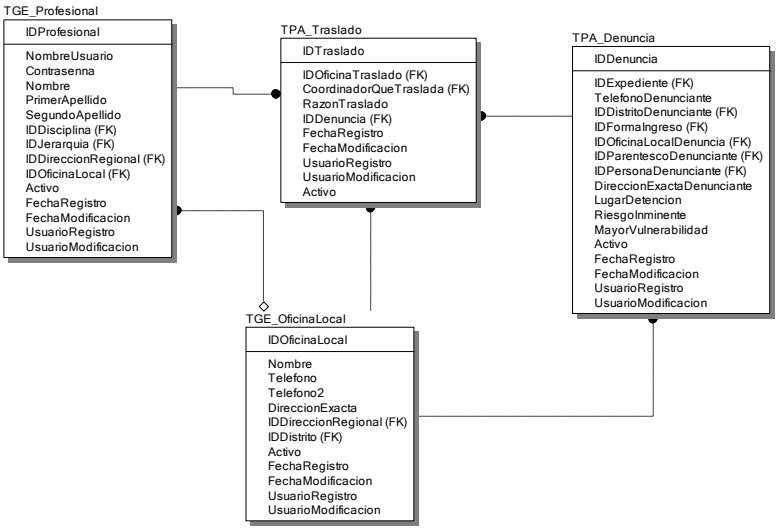
Probatorios y archivos adjuntos



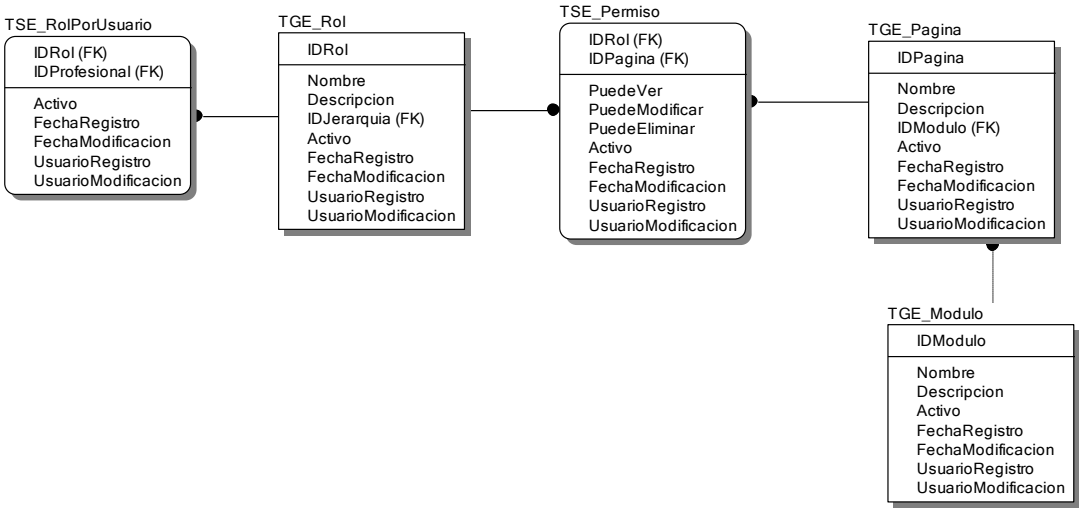
Decisión Técnica



Traslados de expedientes entre oficinas



Esquema de Seguridad



Anexo B

Modelo de Entidad-Relación del Data Warehouse actual:

En esta sección se muestran los diagramas de entidad relación tanto del Staging Area como del Data Warehouse actual con el que cuenta el PANI.

Modelo Entidad Relación del Staging Area

A continuación se muestra el diagrama de la Base de Datos "Staging Area" donde se realizan las extracciones de la Base de Datos transaccional del InfoPANI.

Diagrama SA.oxps

Modelo Entidad Relación del Data Warehouse

A continuación se muestra el diagrama de la Base de Datos "Data Warehouse" donde se realizan las transformaciones de la Base de Datos "Staging Area".

Diagrama DW.oxps

Anexo C

Esta es la consulta que se realiza para la extracción de todos los valores del Data Warehouse y del sistema OLTP para cargar la estructura de Minería de Datos creada para el presente proyecto.

```
USE DM_Infopani
INSERT INTO dbo.EstructuraMineria
SELECT DISTINCT
        FechaIngresaDenuncia = CONVERT(VARCHAR(10), CONVERT(DATETIME, DPR.ID_FechaIngresaDenuncia), 121),
        Expediente = EXE.Ref_Expediente,
        GrupoEtnico = GEI.Descripcion,
        Nacionalidad = NAC.Descripcion,
        Escolaridad = ESC.DescripcionEscolaridad,
        CondicionLegal = COL.Descripcion,
        Genero = GEN.Descripcion,
        FormaIngreso = FIN.Descripcion,
        GrupoEdad = GED.Descripcion,
        Provincia = UBG.DescripcionProvincia,
        Canton = UBG.DescripcionCanton,
        Distrito = UBG.DescripcionDistrito,
        EstadoCivilPadre = TESCPA.Nombre,
        NacionalidadPadre = TNAPA.Descripcion,
        GeneroPadre = TGENPA.Descripcion,
        GrupoEdadPadre =
        CASE
                WHEN DATEDIFF(MM, TPERPA.FechaNacimiento, GETDATE()) / 12 BETWEEN 15 AND 20 THEN '15 - 20'
                WHEN DATEDIFF(MM, TPERPA.FechaNacimiento, GETDATE()) / 12 BETWEEN 21 AND 25 THEN '21 - 25'
                WHEN DATEDIFF(MM, TPERPA.FechaNacimiento, GETDATE()) / 12 BETWEEN 26 AND 30 THEN '26 - 30'
                WHEN DATEDIFF(MM, TPERPA.FechaNacimiento, GETDATE()) / 12 BETWEEN 31 AND 30 THEN '31 - 35'
                WHEN DATEDIFF(MM, TPERPA.FechaNacimiento, GETDATE()) / 12 BETWEEN 36 AND 40 THEN '36 - 40'
                WHEN DATEDIFF(MM, TPERPA.FechaNacimiento, GETDATE()) / 12 BETWEEN 41 AND 45 THEN '41 - 45'
                WHEN DATEDIFF(MM, TPERPA.FechaNacimiento, GETDATE()) / 12 BETWEEN 46 AND 50 THEN '46 - 50'
                WHEN DATEDIFF(MM, TPERPA.FechaNacimiento, GETDATE()) / 12 BETWEEN 51 AND 55 THEN '51 - 55'
                WHEN DATEDIFF(MM, TPERPA.FechaNacimiento, GETDATE()) / 12 BETWEEN 56 AND 60 THEN '56 - 60'
                ELSE '60+'
        END,
        EstadoCivilMadre = TESCMA.Nombre,
        NacionalidadMadre = TNAMA.Descripcion,
        GeneroMadre = TGENMA.Descripcion,
        GrupoEdadMadre =
        CASE
                WHEN DATEDIFF(MM, TPERMA.FechaNacimiento, GETDATE()) / 12 BETWEEN 15 AND 20 THEN '15 - 20'
                WHEN DATEDIFF(MM, TPERMA.FechaNacimiento, GETDATE()) / 12 BETWEEN 21 AND 25 THEN '21 - 25'
                WHEN DATEDIFF(MM, TPERMA.FechaNacimiento, GETDATE()) / 12 BETWEEN 26 AND 30 THEN '26 - 30'
                WHEN DATEDIFF(MM, TPERMA.FechaNacimiento, GETDATE()) / 12 BETWEEN 31 AND 30 THEN '31 - 35'
                WHEN DATEDIFF(MM, TPERMA.FechaNacimiento, GETDATE()) / 12 BETWEEN 36 AND 40 THEN '36 - 40'
                WHEN DATEDIFF(MM, TPERMA.FechaNacimiento, GETDATE()) / 12 BETWEEN 41 AND 45 THEN '41 - 45'
                WHEN DATEDIFF(MM, TPERMA.FechaNacimiento, GETDATE()) / 12 BETWEEN 46 AND 50 THEN '46 - 50'
                WHEN DATEDIFF(MM, TPERMA.FechaNacimiento, GETDATE()) / 12 BETWEEN 51 AND 55 THEN '51 - 55'
                WHEN DATEDIFF(MM, TPERMA.FechaNacimiento, GETDATE()) / 12 BETWEEN 56 AND 60 THEN '56 - 60'
                ELSE '60+'
        END,
        RiesgoInminente = TDEN.RiesgoInminente
FROM DW_Infopani.dbo.TH_DenunciaProceso DPR
INNER JOIN DW_Infopani.dbo.TD_GrupoEtnico GEI
        ON GEI.ID_GrupoEtnico = DPR.ID_GrupoEtnico
INNER JOIN DW_Infopani.dbo.TD_Nacionalidad NAC
        ON NAC.ID_Nacionalidad = DPR.ID_PaisNacionalidad
INNER JOIN DW_Infopani.dbo.TD_Escolaridad ESC
        ON ESC.ID_Escolaridad = DPR.ID_Escolaridad
INNER JOIN DW_Infopani.dbo.TD_CondicionLegal COL
        ON COL.ID_CondicionLegal = DPR.ID_CondicionLegal
INNER JOIN DW_Infopani.dbo.TD_Genero GEN
        ON GEN.ID_Genero = DPR.ID_Genero
```

```
INNER JOIN DW_Infopani.dbo.TD_FormaIngreso FIN
        ON FIN.ID_FormaIngreso = DPR.ID_FormaIngreso
INNER JOIN DW_Infopani.dbo.TD_GrupoEdad GED
        ON GED.ID_GrupoEdad = DPR.ID_GrupoEdad
INNER JOIN DW_Infopani.dbo.TD_Expediente EXE
        ON EXE.ID_Expediente = DPR.ID_Expediente
INNER JOIN DW_Infopani.dbo.TD_UbicacionGeografica UBG
        ON UBG.Ref_Distrito = DPR.ID_DistritoActual
INNER JOIN Infopani_Test.dbo.TGE_Persona TPER
        ON TPER.IDPersona = EXE.Ref_PersonaMenor
INNER JOIN Infopani_Test.dbo.TGE_Persona TPERPA
        ON TPERPA.IDPersona = TPER.IDPersonaPadre
INNER JOIN Infopani_Test.dbo.TGE_EstadoCivil TESCPA
        ON TESCPA.IDEstadoCivil = TPERPA.IDEstadoCivil
INNER JOIN DW_Infopani.dbo.TD_Nacionalidad TNAPA
        ON TNAPA.Ref_Nacionalidad = TPERPA.IDPaisNacionalidad
INNER JOIN DW_Infopani.dbo.TD_Genero TGENPA
        ON TGENPA.Ref_Genero = 'h'
INNER JOIN Infopani_Test.dbo.TGE_Persona TPERMA
        ON TPERMA.IDPersona = TPER.IDPersonaMadre
INNER JOIN Infopani_Test.dbo.TGE_EstadoCivil TESCMA
        ON TESCMA.IDEstadoCivil = TPERMA.IDEstadoCivil
INNER JOIN DW_Infopani.dbo.TD_Nacionalidad TNAMA
        ON TNAMA.Ref_Nacionalidad = TPERMA.IDPaisNacionalidad
INNER JOIN DW_Infopani.dbo.TD_Genero TGENMA
        ON TGENMA.Ref_Genero = 'm'
INNER JOIN Infopani_Test.dbo.TPA_Denuncia TDEN
        ON TDEN.IDExpediente = EXE.Ref_Expediente
```

Anexo D

A continuación se muestra el cuadro que contiene toda la información obtenida de la ejecución del algoritmo de clústeres.

Variables	Estados	Población (Todo)	Cluster 3	Cluster 2	Cluster 1	Cluster 4	Cluster 6	Cluster 5	Cluster 7	Cluster 8	Cluster 9	Cluster 10
Tamaño		1107	138	136	129	122	119	110	100	98	97	58
Canton	Puntarenas	38	9 %	2 %	1 %	4 %	0 %	0 %	3 %	12 %	1 %	0 %
Canton	Turrialba	37	1 %	0 %	6 %	6 %	0 %	2 %	11 %	4 %	0 %	3 %
Canton	San Carlos	34	2 %	1 %	3 %	2 %	5 %	2 %	6 %	2 %	4 %	8 %
Canton	Cartago	34	0 %	6 %	6 %	7 %	0 %	0 %	4 %	4 %	0 %	2 %
Canton	San José	28	9 %	0 %	2 %	0 %	4 %	4 %	0 %	2 %	0 %	3 %
Canton	Desamparados	28	0 %	1 %	3 %	0 %	6 %	2 %	4 %	0 %	9 %	5 %
Canton	La Unión	27	0 %	7 %	5 %	4 %	0 %	0 %	1 %	4 %	0 %	0 %
Canton	San Ramón	25	0 %	6 %	2 %	2 %	0 %	0 %	6 %	1 %	0 %	5 %
Canton	...	...	...	...	...	...	...	...	...	...	...	...
Distrito	ausente	492	40 %	46 %	52 %	53 %	41 %	26 %	38 %	56 %	45 %	46 %
Distrito	San Rafael	26	5 %	0 %	2 %	3 %	3 %	3 %	4 %	0 %	1 %	3 %
Distrito	Concepción	20	1 %	4 %	4 %	0 %	1 %	0 %	3 %	3 %	0 %	2 %
Distrito	San Isidro	20	1 %	2 %	1 %	0 %	7 %	5 %	0 %	1 %	0 %	2 %
Distrito	San Juan	19	1 %	4 %	3 %	1 %	1 %	1 %	3 %	0 %	0 %	2 %
Distrito	San Pedro	17	1 %	2 %	0 %	1 %	2 %	2 %	1 %	2 %	4 %	3 %
Distrito	Santa Rosa	16	0 %	1 %	1 %	6 %	0 %	4 %	1 %	0 %	0 %	0 %
Distrito	San Antonio	15	4 %	0 %	1 %	0 %	1 %	0 %	0 %	4 %	3 %	3 %
Distrito	...	...	...	...	...	...	...	...	...	...	...	...
Escolaridad	SIN DEFINIR	397	54 %	26 %	62 %	25 %	8 %	31 %	59 %	1 %	9 %	81 %
Escolaridad	I CICLO (1, 2 Y 3 GRADO DE PRIMARIA)	216	18 %	8 %	20 %	35 %	29 %	24 %	5 %	28 %	20 %	5 %

Escolaridad	II CICLO (4,5 Y 6 GRADO DE PRIMARIA)	206	8 %	23 %	5 %	25 %	26 %	25 %	3 %	35 %	42 %	3 %
Escolaridad	III CICLO (7,8 Y 9 ANO DE SECUNDARIA)	157	1 %	36 %	1 %	9 %	27 %	14 %	5 %	32 %	12 %	3 %
Escolaridad	3. KINDER	45	8 %	2 %	2 %	1 %	5 %	0 %	13 %	0 %	12 %	0 %
Escolaridad	1. MATERNAL	35	9 %	2 %	5 %	1 %	0 %	0 %	9 %	0 %	0 %	0 %
Escolaridad	OTRO	19	0 %	1 %	3 %	1 %	4 %	3 %	4 %	1 %	0 %	0 %
Escolaridad	2. PREKINDER	13	3 %	1 %	1 %	0 %	0 %	0 %	3 %	0 %	2 %	1 %
Escolaridad	...	...	...	...	...	...	...	...	...	...	...	...
Estado Civil Madre	Soltero(a)	437	60 %	20 %	65 %	56 %	11 %	18 %	45 %	35 %	39 %	26 %
Estado Civil Madre	Casado(a)	343	10 %	56 %	24 %	6 %	72 %	10 %	35 %	42 %	29 %	33 %
Estado Civil Madre	Unión Libre	212	25 %	12 %	9 %	14 %	6 %	65 %	13 %	10 %	12 %	33 %
Estado Civil Madre	Divorciado(a)	103	5 %	11 %	2 %	21 %	9 %	7 %	5 %	12 %	17 %	8 %
Estado Civil Madre	Viudo(a)	12	1 %	1 %	0 %	3 %	1 %	0 %	1 %	2 %	2 %	0 %
Estado Civil Padre	Soltero(a)	413	59 %	13 %	64 %	55 %	15 %	20 %	41 %	30 %	38 %	19 %
Estado Civil Padre	Casado(a)	375	15 %	63 %	26 %	8 %	73 %	12 %	37 %	49 %	22 %	38 %
Estado Civil Padre	Unión Libre	213	20 %	13 %	9 %	13 %	4 %	62 %	19 %	9 %	20 %	35 %
Estado Civil Padre	Divorciado(a)	95	5 %	10 %	1 %	21 %	7 %	6 %	3 %	10 %	19 %	8 %
Estado Civil Padre	Viudo(a)	11	0 %	1 %	1 %	4 %	2 %	0 %	0 %	1 %	1 %	0 %
Fecha Ingresa Denuncia	12/18/2012 12	18/12/2012 00:00	24/01/2013 00:00	18/11/2012 00:00	11/12/2012 00:00	13/12/2012 00:00	25/12/2012 00:00	10/11/2012 00:00	22/12/2012 00:00	19/02/2013 00:00	19/12/2012 00:00	12/11/2012 00:00
Fecha Ingresa Denuncia		83.18:23:36.4530000	65.00:01:18.4050000	87.10:48:32.9590000	89.09:13:35.6240000	81.17:34:26.2560000	74.19:17:23.6730000	88.02:55:23.7060000	75.09:23:52.7220000	52.07:34:49.1600000	81.21:26:06.6990000	77.22:25:44.1250000
Forma Ingreso	Referencia de instancia Pública	294	10 %	23 %	25 %	23 %	30 %	32 %	16 %	24 %	28 %	90 %
Forma Ingreso	Denuncia vía telefónica	167	33 %	12 %	17 %	17 %	16 %	12 %	12 %	11 %	9 %	1 %
Forma Ingreso	Denuncia Personal	160	18 %	12 %	18 %	12 %	3 %	16 %	23 %	15 %	20 %	5 %
Forma Ingreso	Denuncia por el 9-1-1	131	10 %	18 %	17 %	6 %	17 %	12 %	13 %	4 %	16 %	0 %

Forma Ingreso	Progenitor(a)	107	9 %	10 %	7 %	5 %	10 %	8 %	17 %	15 %	14 %	0 %
Forma Ingreso	Referencia de Autoridad Judicial	70	1 %	11 %	5 %	11 %	9 %	7 %	2 %	9 %	4 %	0 %
Forma Ingreso	Adulto Responsable de la PME	55	9 %	5 %	3 %	7 %	4 %	5 %	4 %	7 %	4 %	0 %
Forma Ingreso	COI	33	4 %	2 %	3 %	3 %	3 %	0 %	4 %	8 %	1 %	0 %
Forma Ingreso	...	...	...	...	...	...	...	...	...	...	...	...
Genero	Femenino	586	54 %	56 %	49 %	39 %	58 %	47 %	48 %	82 %	62 %	32 %
Genero	Masculino	513	46 %	44 %	51 %	61 %	42 %	53 %	44 %	18 %	38 %	68 %
Genero	SIN DEFINIR	8	0 %	0 %	0 %	0 %	0 %	0 %	8 %	0 %	0 %	0 %
Grupo Edad Madre	60+	330	20 %	29 %	18 %	49 %	14 %	59 %	17 %	42 %	30 %	25 %
Grupo Edad Madre	26 - 30	238	35 %	2 %	31 %	11 %	9 %	20 %	21 %	22 %	38 %	37 %
Grupo Edad Madre	36 - 40	163	6 %	23 %	3 %	19 %	31 %	9 %	10 %	27 %	7 %	17 %
Grupo Edad Madre	21 - 25	131	25 %	1 %	24 %	5 %	1 %	5 %	42 %	1 %	0 %	5 %
Grupo Edad Madre	41 - 45	101	3 %	15 %	0 %	12 %	32 %	0 %	2 %	5 %	16 %	11 %
Grupo Edad Madre	15 - 20	60	8 %	2 %	20 %	1 %	0 %	4 %	8 %	1 %	1 %	0 %
Grupo Edad Madre	46 - 50	55	3 %	18 %	2 %	2 %	10 %	0 %	1 %	2 %	7 %	3 %
Grupo Edad Madre	51 - 55	23	0 %	9 %	0 %	2 %	4 %	2 %	0 %	0 %	0 %	2 %
Grupo Edad Madre	...	...	...	...	...	...	...	...	...	...	...	...
Grupo Edad Padre	60+	349	37 %	17 %	36 %	24 %	10 %	62 %	15 %	34 %	50 %	46 %
Grupo Edad Padre	36 - 40	175	7 %	17 %	7 %	24 %	14 %	12 %	15 %	37 %	12 %	21 %
Grupo Edad Padre	26 - 30	150	28 %	2 %	21 %	8 %	3 %	3 %	37 %	6 %	11 %	13 %
Grupo Edad Padre	41 - 45	141	9 %	20 %	2 %	24 %	26 %	7 %	6 %	17 %	8 %	6 %
Grupo Edad Padre	46 - 50	105	2 %	23 %	1 %	11 %	31 %	6 %	2 %	6 %	7 %	5 %
Grupo Edad Padre	21 - 25	75	12 %	0 %	21 %	0 %	0 %	3 %	20 %	0 %	0 %	3 %
Grupo Edad Padre	51 - 55	61	2 %	12 %	4 %	6 %	13 %	4 %	3 %	0 %	6 %	4 %

Grupo Edad Padre	56 - 60	37	3 %	9 %	3 %	2 %	3 %	2 %	2 %	1 %	5 %	2 %
Grupo Edad Padre	...	...	...	...	...	...	...	...	...	...	...	...
Grupo Edad	Niñez Media	312	32 %	11 %	22 %	50 %	37 %	35 %	6 %	37 %	42 %	13 %
Grupo Edad	Niñez Temprana	243	48 %	17 %	32 %	11 %	1 %	5 %	66 %	0 %	12 %	11 %
Grupo Edad	Adolescencia Primera Etapa	222	5 %	30 %	6 %	18 %	30 %	32 %	6 %	43 %	31 %	5 %
Grupo Edad	Adolescencia Segunda Etapa	167	2 %	38 %	5 %	17 %	28 %	14 %	3 %	17 %	13 %	14 %
Grupo Edad	Infancia	129	13 %	1 %	34 %	3 %	1 %	11 %	10 %	0 %	1 %	53 %
Grupo Edad	Mayor de Edad	25	0 %	5 %	2 %	1 %	2 %	4 %	1 %	3 %	0 %	5 %
Grupo Edad	SIN DEFINIR	9	0 %	0 %	0 %	0 %	1 %	0 %	8 %	0 %	0 %	0 %
Nacionalidad Madre	Costarricense	965	92 %	93 %	98 %	92 %	99 %	65 %	98 %	89 %	97 %	12 %
Nacionalidad Madre	Nicaraguense	77	9 %	2 %	0 %	8 %	1 %	34 %	0 %	11 %	1 %	3 %
Nacionalidad Madre	Cubana	55	0 %	0 %	1 %	0 %	0 %	0 %	0 %	0 %	1 %	84 %
Nacionalidad Madre	Panameña	3	0 %	1 %	0 %	0 %	0 %	0 %	1 %	0 %	0 %	0 %
Nacionalidad Madre	China	2	0 %	1 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
Nacionalidad Madre	Guatemalteca	2	0 %	1 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
Nacionalidad Madre	Colombiana	1	0 %	0 %	1 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
Nacionalidad Madre	Norteamericana	1	0 %	1 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
Nacionalidad Madre	...	...	...	...	...	...	...	...	...	...	...	...
Nacionalidad Padre	Costarricense	979	94 %	95 %	95 %	97 %	100 %	74 %	97 %	94 %	90 %	9 %
Nacionalidad Padre	Nicaraguense	54	5 %	1 %	3 %	3 %	0 %	26 %	1 %	6 %	4 %	3 %
Nacionalidad Padre	Cubana	51	0 %	0 %	1 %	0 %	0 %	0 %	1 %	0 %	0 %	78 %
Nacionalidad Padre	Norteamericana	5	0 %	1 %	1 %	0 %	0 %	0 %	2 %	0 %	0 %	2 %
Nacionalidad Padre	Panameña	3	1 %	1 %	0 %	0 %	0 %	0 %	0 %	0 %	1 %	0 %
Nacionalidad Padre	Ecuatoriana	3	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	5 %

Nacionalidad Padre	Colombiana	3	0 %	0 %	1 %	0 %	0 %	0 %	0 %	0 %	1 %	2 %
Nacionalidad Padre	Guatemalteca	2	0 %	1 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
Nacionalidad Padre	...	...	...	...	...	...	...	...	...	...	...	...
Nacionalidad	Costarricense	992	98 %	96 %	97 %	98 %	98 %	78 %	90 %	95 %	97 %	10 %
Nacionalidad	Ecuatoriana	28	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	45 %
Nacionalidad	Nicaraguense	26	1 %	0 %	0 %	1 %	0 %	17 %	0 %	3 %	0 %	3 %
Nacionalidad	Cubana	24	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	38 %
Nacionalidad	SIN DEFINIR	23	1 %	1 %	2 %	1 %	2 %	5 %	8 %	2 %	1 %	0 %
Nacionalidad	Norteamericana	4	0 %	2 %	0 %	0 %	0 %	0 %	1 %	0 %	0 %	1 %
Nacionalidad	Colombiana	3	0 %	0 %	1 %	0 %	0 %	0 %	0 %	0 %	1 %	2 %
Nacionalidad	Guatemalteca	2	0 %	1 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
Nacionalidad	...	...	...	...	...	...	...	...	...	...	...	...
Provincia	San José	283	48 %	4 %	22 %	2 %	44 %	27 %	11 %	29 %	56 %	31 %
Provincia	Alajuela	243	2 %	41 %	24 %	26 %	18 %	11 %	37 %	7 %	21 %	36 %
Provincia	Cartago	145	3 %	18 %	20 %	29 %	0 %	5 %	23 %	17 %	0 %	6 %
Provincia	Puntarenas	141	20 %	5 %	6 %	16 %	15 %	12 %	10 %	25 %	10 %	10 %
Provincia	Guanacaste	126	4 %	19 %	24 %	11 %	10 %	14 %	2 %	10 %	1 %	10 %
Provincia	Heredia	112	20 %	8 %	4 %	5 %	13 %	28 %	8 %	6 %	3 %	3 %
Provincia	Limón	57	3 %	5 %	1 %	12 %	0 %	4 %	9 %	6 %	9 %	4 %
Riesgo Inminente	False	995	90 %	91 %	84 %	93 %	89 %	85 %	95 %	86 %	93 %	100 %
Riesgo Inminente	True	112	10 %	9 %	16 %	7 %	11 %	15 %	6 %	14 %	7 %	0 %