UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

TÍTULO: MARCO DE TRABAJO PARA LA CAPTURA Y ANÁLISIS DE INFORMACIÓN GENEALÓGICA.

TESIS PRESENTADA

EN OPCIÓN AL TÍTULO DE

MÁSTER EN INFORMÁTICA APLICADA

Autor: Ing. Reynaldo Rosado Roselló

Tutores: Dra.C. Beatriz Marcheco Teruel

MSc. Maykel Yelandi Leyva Vázquez

MSc. Yeleny Zuelueta Véliz

Consultante: Dr. C. Rafael Arturo Trujillo Rasúa

La Habana, diciembre de 2012

"AÑO 54 DE LA REVOLUCIÓN"

DECLARACIÓN JURADA DE AUTORÍA

Yo, Reynaldo Rosado Roselló con carné de identidad 84081718181, declaro que soy el autor principal del resultado que expongo en la presente tesis titulada Marco de trabajo para captura, integración y análisis de información genealógica, para optar por el título académico de Máster en Informática Aplicada. Este trabajo fue desarrollado durante el período comprendido entre 2009 y 2012. En especial, agradecer a la Dra.C Beatriz Marcheco Teruel, MSc. Maikel Yelandi Leyva Vázquez y MSc. Yeleny Zulueta Veliz quienes fungieron como tutores, y al Dr.C. Rafael Arturo Trujillo Rasúa por su confianza y apoyo. A todos ellos, así como a otros colegas y amigos que no he mencionado por razones de espacio, les doy las más sinceras gracias. Finalmente, declaro que todo lo anteriormente expuesto se ajusta a la verdad, y asumo la responsabilidad moral y jurídica que se derive de este juramento profesional. Y para que así conste, firmo la presente declaración jurada de autoría en La Habana, a los __ días del mes de diciembre del año 2012.

_____

Ing. Reynaldo Rosado Roselló

Firma del autor

Resumen

El análisis de la información genealógica encierra un elevado valor debido al conocimiento que permite descubrir sobre las causas biológicas y ambientales de las distintas discapacidades que afectan la calidad de vida de las personas. Sin embargo persisten dificultades que limitan el análisis de datos genealógicos: la falta de integración entre las herramientas de representación de arboles genealógico y las herramientas de análisis de datos; ausencia de indicaciones metodológicas que permitan tal integración; y consumo elevado de tiempo y esfuerzo pues los especialistas tienen que extraer los datos de las herramientas genealógicas para luego preparar la información en los ficheros y formatos que interpreta alguna herramienta de análisis de datos.

En el presente trabajo se propone un marco de trabajo flexible e integrado para el análisis de información genealógica. Además se propone un método multicriterio de toma de decisión en grupo basado que utiliza el Grupo Focal, Proceso de Análisis Jerárquico y Operadores de Agregación, para la identificación y priorización de los criterios para la selección de herramientas de representación genealógicas. Otro aporte obtenido de este trabajo es la integración de técnicas de análisis de datos en una aplicación informática para la representación de árboles genealógicos.

La validación del marco de trabajo se realizó a través de un estudio de caso y evaluaciones de expertos. La triangulación de estos métodos científicos permitió comprobar que los altos niveles flexibilidad e integración del marco de trabajo posibilitaron disminuir el esfuerzo y tiempo empleados en el proceso de análisis de datos.

Palabras clave: Análisis de datos, Árboles genealógicos, Integración, Método multicriterio.

Abstract

The analysis of genealogical information encloses a great value due to the knowledge about environmental and biological causes of human handicaps that is obtained. However there are some difficulties restricting the genealogical analysis: the lack of integration between genealogical representation tools and data analysis tools; lack of methodological indication for accomplish such integration; and the high time and effort implication because the specialist have to extract manually the information from genealogical tools to develop files as input for some of the data analysis tools.

A flexible and integrated framework for genealogical information analysis is proposed in this work. Also is proposed a group multicriteria decision making method based on focus group, AHP and aggregator operators, for the identification and order of criteria related with the selection of genealogical representation tools. The integration of data analysis technique in a genealogical representation tool is another contribution of this work.

The validation of the framework was made trough a study case and expert's evaluations. The triangulation of these scientific methods allowed to confirm that the high levels of flexibility and integration of the framework aided the decreasing the effort and time employed in the data analysis process.

Keywords: Data analysis, Genealogical trees, Integration, Multicriteria method.

Índice

Introducción

La genética médica es la ciencia de la variación biológica humana dada su relación con la salud y con las enfermedades. La genética clínica es la parte de la genética médica que se ocupa de la salud individual del ser humano y su familia. Es la ciencia que se ocupa del diagnóstico, la prevención y el manejo de las enfermedades genéticas. En las últimas cinco décadas los genetistas clínicos han encontrado su órgano específico de atención –el genoma humano– justo como es el corazón para los cardiólogos y el sistema nervioso para los especialistas en neurología. Se trata de una de las especialidades médicas que ha realizado grandes avances en breve tiempo, con un notable impacto sobre las demás ramas de la medicina, la biología y las ciencias en general [1].

Ya en Cuba se cuenta con 9 años de experiencia en el trabajo de la red nacional de genética médica dirigida por su Centro Nacional (CNGM) y más de 30 años en el desarrollo del programa de diagnóstico, manejo y prevención de enfermedades genéticas y defectos congénitos. Estos avances, dirigidos a ofrecer una mayor seguridad de salud y felicidad a la familia cubana, tuvieron su punto de partida luego del estudio nacional realizado a personas con retraso mental y otras discapacidades, que por orientaciones de Fidel, se realizó entre julio del 2001 y abril del 2003, insertado en la Batalla de Ideas [2].

Con el transcurso de los años este centro ha acumulado un número importante de datos, derivados de disímiles estudios e investigaciones realizados por lo que la gestión y análisis de esta información es de vital importancia para realizar investigaciones. Representa de un valor incalculable poder gestionar y analizar esta información teniendo en cuenta el conocimiento que se almacena en la misma. Partiendo de un grupo de variables resulta interesante caracterizarlas, construir modelos y hasta predecir comportamientos. Conocer probabilidades de ocurrencia puede ser útil a la hora de avaluar si es pertinente realizar un estudio o someter algún individuo a pruebas, que en la mayoría de los casos se realizan con reactivos que son muy costosos en el mercado internacional [3].

El análisis de las relaciones entre variables, las tendencias de una población con respecto a una enfermedad, la formación de comunidades dentro de una población estudiada que comparten un grupo de características, resulta de mucho interés científico para ayudar a tomar decisiones. Con los grandes volúmenes de datos es imposible evidenciar estos resultados para

poder convertirlo en conocimiento y a su vez utilizarlo en la prevención de enfermedades. El proceso de descubrimiento de conocimiento a partir de un conjunto de datos se le conoce por sus siglas en inglés (KDD). Este proceso es soportado por un conjunto de técnicas y algoritmo de Minería de Datos (MD).

Cuba dispone hoy de una red de profesionales entrenados, servicios y centros de genética médica cuya fortaleza principal radica en su integración a la atención primaria de salud. La presencia en la comunidad de especialistas con conocimientos de genética, preparados para trabajar en la medicina comunitaria con un enfoque multidisciplinario e integrador, permite colocar a la genética médica y sus avances tecnológicos en el primer punto de contacto entre los individuos, la familia y la comunidad en general. El sistema nacional de salud garantiza de manera permanente, universal y equitativa la atención sanitaria en este campo, algo que solo es posible con una voluntad política que prioriza la atención primaria de salud como el eslabón fundamental para mejorar cada vez más el estado de salud de la población a través de un enfoque preventivo [4, 5].

Lo propios genetistas como profesionales de la salud, coinciden en la importancia de la representación del árbol genealógico, como herramienta básica para poder analizar la historia familiar de los pacientes. Un grupo importante de investigadores coinciden en el carácter interdisciplinario de las investigaciones teniendo en cuenta a la hora de analizar la tradición familiar, no solo los elementos genéticos sino los factores sociales y ambientales que pueden influir en el paciente [6].

La directora nacional de genética en Cuba planteó: “Los árboles genealógicos familiares, son el recurso más utilizado por los profesionales de la genética médica desde hace un siglo, se realizan a un paciente referido a la consulta ante un posible diagnóstico genético, para evaluar la transmisión del rasgo o enfermedad en la familia y asesorar sobre el riesgo a los miembros que lo soliciten. La identificación de familias con varios miembros afectados por estas enfermedades crónicas, permite realizar una mejor caracterización clínica de la misma, identificar tempranamente a otros miembros de la familia enfermos y abordar con un propósito preventivo a los demás familiares en riesgo”[7].

Esta forma de representación familiar se ha convertido en un mecanismo que no solo utilizan los profesionales de la salud. Especialistas de diversas ramas lo emplean para estudios e investigaciones de la historia familiar. El desarrollo de las tecnologías de la información ha sabido responder a esta demanda de especialistas médicos y no médicos, se han desarrollado por la comunidad internacional numerosas herramientas informáticas para la representación de

árboles genealógicos. Son disímiles las funcionalidades que brindan. Con el tiempo han ido evolucionando, sobre todo en la mejorar de la visualización. En ocasiones no se tienen todos los elementos de la gama de herramientas disponibles y se seleccionan sin hacer un análisis previo de las mismas.

Propio de la evolución en el área, el número de personas que acuden diariamente a las consultas de genética es cada vez mayor. Esto implica un crecimiento de los niveles de información que almacenan las bases de datos (BD) de las herramientas genealógicas. En las mismas se obtienen diferentes árboles con varias generaciones de pacientes y enfermedades. Algunas han incorporado componentes de análisis de riesgo, aplicando algunas reglas mendelianas básicas Un ejemplo de esto lo constituye el paquete WGAF de R [8], que se puede afirmar que es de las que más ha avanzado en estos temas de análisis.

Las limitaciones para hacer análisis a partir de la información gestionada constituyen de forma general una debilidad importante de las herramientas para la representación del árbol genealógico. se han alcanzado algunos resultados discretos en el tema, pero en la literatura consultada no se identifican herramientas que permitan hacer análisis a nivel poblacional. Esta es una deuda pendiente de la informática con la genética, dada la necesidad de poder cerrar el ciclo de la información con el análisis de la misma.

Los datos son observaciones y medidas científicas que, una vez que han sido analizados e interpretados, pueden ser desarrollados como evidencia para tratar una cuestión. Los datos ocupan el centro de las investigaciones científicas y todos los científicos recogen datos de una u otra manera [9]. En la actualidad los especialistas tienen que extraer los datos de las herramientas genealógicas para luego preparar la información en los ficheros y formatos que interpreta alguna herramienta de análisis de datos. Esto tiene muchas limitantes en el proceso ya que no está sincronizado, imposibilitando actualizar los análisis a partir de incorporación de nueva información. Los hace dependientes de varias herramientas y de personal calificado en las mismas.

El estudio realizado al proceso de análisis de la información genealógica evidencia que existen restricciones que influyen en todo el proceso, afectando considerablemente la evolución de los estudios clínico-genéticos. Esto provoca demoras que en este tipo de investigaciones resulta crítico, teniendo en cuenta que los resultados son aplicados directamente a elevar la calidad de vida de las personas. Muchas veces se trata de personas con discapacidad o limitantes que constituyen el sector más vulnerable de la sociedad.

Considerando la importancia y el impacto directo en el bienestar social que puede tener el proceso de análisis de información genealógica para trazar políticas y estrategias del sector de la salud a nivel poblacional se plantea como ***problema de la investigación*****:** ¿Cómo disminuir el esfuerzo en el proceso de análisis de información genealógica?

A partir del problema planteado en la investigación, se identifica *como* ***objeto de estudio*** Proceso de análisis de información genealógica, y dentro de este se precisa como ***campo de acción*** Integración de técnicas de análisis de datos a los sistemas para la representación de árboles genealógicos. Para dar solución al problema de la investigación, se propone como ***objetivo general*****:** Desarrollar un marco de trabajo flexible e integrado para el análisis de información genealógica. Del mismo se derivan los siguientes ***objetivos específicos:***

- Analizar las tendencias actuales en el análisis de la información genealógica soportado por herramientas informáticas.
- Desarrollar un marco de trabajo para el análisis de información genealógica.
- Validar el marco de trabajo mediante el método de experto y el estudio de caso.

La investigación se desarrolla teniendo como ***hipótesis*****:** Si se desarrolla un marco de trabajo flexible que permita realizar de manera integrada la captura y análisis de la información genealógica, entonces disminuirá el esfuerzo en el procesamiento de la información genealógica.

Para el desarrollo de la presente investigación se propone seguir una Estrategia Explicativa pues los conocimientos precedentes acerca del problema han sido suficientes para plantear una hipótesis explicativa y la representación del problema es clara en lo referente a la caracterización del fenómeno en sus aspectos externos. Esta estrategia podrá llevarse a cabo con la utilización de métodos científicos:

Métodos teóricos:

- Análisis Histórico-Lógico: para profundizar en los antecedentes y las tendencias actuales del análisis de la información genealógica.
- Analítico-Sintético: para el estudio y el establecimiento del estado del arte del análisis de la información genealógica.
- Modelación: para la conceptualización del proceso de análisis de la información genealógica.

Métodos empíricos:

- Entrevista: para recopilar información de tendencias, comportamientos o estadística en la investigación, en el estudio del objeto y el campo de acción de la investigación.
- Observación: para la recopilación de la información in situ de las características y comportamientos de las unidades de estudio. Se utilizará en este caso una observación participativa.

Métodos matemáticos:

- Estadística Descriptiva: para determinar los valores promedio o más frecuentes obtenidos por la aplicación de los métodos empíricos utilizados.
- Análisis Porcentual: para determinar los valores porcentuales significativos en los métodos empíricos utilizados.

Novedad y aporte de la investigación:

- Marco de trabajo para el análisis de la información genealógica.
- Conjunto de criterios para la selección de herramientas para la representación genealógicas.
- Integración de técnicas de análisis de datos en una aplicación informática para la representación de árboles genealógicos.

La tesis está estructurada de la siguiente manera:

- **Capítulo 1.** Se muestra el marco teórico de la investigación. Muestra el estado del arte de la representación genealógica y las técnicas de análisis de datos. Se revisan los principales niveles de integración de las herramientas informáticas. Por la importancia de los métodos de decisión multicriterio se hace un estudio de los principales.

- **Capítulo 2.** Se presenta el marco de trabajo propuesto, sus etapas y principales actividades. Se aplican métodos de decisión multicriterio obteniendo un conjunto de criterios para la selección de herramientas de representación genealógica y otros para las de análisis de datos. En el caso de los criterios de selección de las herramientas de

representación genealógica mediante la aplicación de AHP con operadores OWA se les asigna una ponderación.

- **Capítulo 3.** Se evalúa la flexibilidad de la propuesta mediante métodos de experto utilizando etiquetas lingüísticas. Luego se emplea un estudio de caso que permite evaluar la aplicabilidad del marco de trabajo, terminando con una comparación de escenarios antes de la propuesta y con la propuesta. Como parte del estudio de caso se integran las herramientas alasARBOGEN y Weka a nivel de aplicación.

Finalmente se presentan las **Conclusiones** y **Recomendaciones** derivadas de la investigación, las **Referencias bibliográficas**, así como los **Anexos** que dan información adicional sobre el trabajo realizado.

Capítulo 1 Fundamentos teóricos del análisis de la información genealógica.

Introducción

En el presente capítulo se muestran los principales conceptos que sustentan teóricamente la solución propuesta, resultado de la investigación. Se realiza un estudio de las tendencias actuales de la representación genealógica y de las técnicas de análisis de datos, así como las principales herramientas informáticas que lo sustentan respectivamente. Se desarrolla un análisis de los niveles de integración de herramientas informáticas y sus tendencias actuales. Se estudian los principales métodos de decisión multicriterio.

1.1. Los árboles genealógicos y la genética.

Tanto los historiadores familiares como los genealogistas, sustentan su trabajo sobre la importancia de preservar el pasado con el fin de mejorar el futuro. El legado de los antepasados va desde los genes mismos hasta los valores morales, la cultura y los bienes materiales. El árbol genealógico es una representación gráfica que expone los datos genealógicos de un individuo en una forma organizada y sistemática, sea en forma de árbol o tabla. Puede ser ascendiente, es decir que expone los antepasados o ancestros de un individuo o puede ser descendiente, es decir que expone todos los descendientes del individuo [10].

Un genograma es un formato para dibujar un árbol genealógico que registra información sobre los miembros de una familia y sus relaciones sobre por lo menos tres generaciones [11]. Mediante estos se realiza la representación gráfica del árbol genealógico, a través de ciertos símbolos y otros datos que representan a las personas y sus relaciones e integran el árbol. Los mismos permiten visualizar una fuente rica de hipótesis acerca de cómo un problema clínico puede tener relación con el contexto familiar.

Dado el crecimiento acelerado en la utilización del árbol genealógico surge la necesidad de estandarizar su representación. Los especialistas de la rama comenzaron la definición de estándares para la representación gráfica de árboles genealógicos, siendo la más aceptada la “Pedigree Standardization Task Force" (PSTF) [12] de un grupo de investigadores de la Sociedad Nacional de Asesores Genéticos de los Estados Unidos, planteada en 1995. Este grupo en la actualidad ha adoptado el nombre de Pedigree Standardization Work Group (PSWG). El estándar planteado reduce las posibilidades de realizar interpretaciones incorrectas de la información médica y genética de los pacientes y sus familiares, además de impulsar la

calidad de la atención a los pacientes por profesionales de la genética y facilitar la comunicación entre investigadores involucrados en el estudio de la familia [13].

Según datos de OMS el 6 % de la población mundial padecen de alguna enfermedad determinada parcial o totalmente por factores genéticos [14]. Mientras que otras enfermedades más comunes, aquellas en las que se combinan factores genéticos y no genéticos en su origen, son de alta prevalencia y generalmente ocupan las primeras causas de muerte, como las cardiovasculares (cardiopatía isquémica), hipertensión, diabetes, demencias, esquizofrenia y depresión. Existen factores de riesgo no genéticos para estos padecimientos que la persona debe desechar, como son el hábito de fumar, consumir bebidas alcohólicas, una dieta poco saludable y el sedentarismo. Un individuo que tiene factores de riesgo genéticos para algunos de estos padecimientos, si mantiene hábitos y un estilo de vida saludables, puede que no desarrolle la enfermedad. De aquí la importancia en el trabajo preventivo de las mismas.

Como ha planteado Marcheco [15], el árbol genealógico es la principal herramienta de trabajo para la gestión de la información de los pacientes. Una vez que llega un paciente a la consulta se le construye su árbol, recopilando un grupo de informaciones necesarias en función del padecimiento del mismo. Con la evolución de la genética como disciplina emergente la medicina y su carácter interdisciplinario, cada vez son mayores las evidencias de las consecuencias de factores sociales, ambientales y del modo de vida de las personas en el padecer de las enfermedades. Es por esto que en las investigaciones actuales se crean grupos multidisciplinarios y cada vez son más diversos los cuestionarios que se aplican a la hora de realizar los estudios.

Toda esta panorámica trae como consecuencia que se levanten grandes volúmenes de datos en los estudios que se realizan. Se plantea que la genética es el centro de balance de las investigaciones médicas y como tal, construye grandes árboles genealógicos con información muy diversa en el afán de abarcar todos los factores que puedan influir en la evolución de una enfermedad o tendencia poblacional.

Las TIC han tenido y tienen un impacto importante en todos los ámbitos de la sociedad. Han permitido acelerar el desarrollo de la ciencia y la obtención de conocimiento. En el campo de la genética, no se ha dejado de contribuir a su desarrollo desde la informática. A nivel mundial se han desarrollado un número importante de aplicaciones informáticas para la representación del árbol genealógico. A continuación se evidencia un estudio del estado actual y las tendencias en este sentido de las principales herramientas.

1.2. Aplicaciones para la representación de los árboles genealógicos.

El estudio y análisis de los datos genealógicos constituye actualmente una actividad científica de gran importancia; y su intercambio facilita el trabajo de científicos, médicos e investigadores. Con el desarrollo de la tecnología y los sistemas de cómputo en el campo de la medicina, internacionalmente se han desarrollado diversos sistemas para la representación de árboles genealógicos, con el objetivo fundamental de agilizar los procesos de representación gráfica de árboles genealógicos. En su mayoría han tenido en cuenta al menos parcialmente lo establecido por Bennet, para la estandarización de los árboles genealógicos [12].

A continuación se muestran las principales herramientas para la representación genealógica que fueron estudiadas. Es importante destacar que existen disímiles herramientas con este fin [16]. En el presente estudio se hace referencia a las de más prestigio en cuanto a funcionalidades, aceptación y uso en círculos académicos y científicos.

BitGen v2.0

BitGen v2.0 se encuentra disponible en el mercado para la confección de árboles genealógicos de una familia, partiendo de datos previamente almacenados relativos a esta. Es una aplicación en español, que tiene como objetivo principal, facilitar la elaboración de árboles genealógicos en una computadora [17].

La interfaz de BitGen II para la confección de un árbol genealógico presenta cuatro pergaminos de tamaño DIN-A4, bordeados por una pareja de reglas para medir adecuadamente la posición de cada elemento gráfico. Además, contiene una barra de herramientas y los controles de movimientos, desde donde se pueden indicar todas las funciones que se quieren realizar. BitGen II es la segunda versión de un software al que aún le falta dinamismo, rapidez, intuición y un funcionamiento gráfico más cercano al de aplicaciones destinadas a trabajar en entornos operativos Windows.

Cyrillic

Posee todas las características necesarias para dibujar líneas genéticas. Es el más completo, funcional y fácil de usar. Es un software diseñado para el manejo de datos genéticos en el sistema operativo Windows. Las herramientas que contienen se disponen de forma tal que se facilite el manejo de los datos disponibles [18].

Como características importantes presenta [19]:

- Actualización automática de los datos mientras se está dibujando.
- Perfeccionamiento del manejo para casos de gemelos y embarazos múltiples.
- Cálculo de riesgos para enfermedades familiares, permitiendo su realización a nivel de individuo dentro de una familia donde exista una determinada enfermedad.
- Utilización de BD como Access o Corel Paradox.

Este sistema en su versión 3.0 incluye como funcionalidad importante el análisis de riesgo de padecimiento de cáncer de origen genético, como el cáncer de mama y próstata. Tiene una excelente organización del espacio de trabajo y de los individuos según su generación, su mayor desventaja es su diseño para Windows únicamente, unido a los altos precios de la licencia [20].

Madeline 2.0 PDE

Es un sistema de Licencia GPL de código abierto, desarrollado en C++ que a través de la línea de comandos genera el árbol genealógico basado en las relaciones entre los individuos. Permite la representación de consanguinidad en las relaciones y esto hace que a veces se representen ciclos en el árbol genealógico.

Actualmente está en desarrollo un servicio web para el acceso desde otros sistemas que permitirá la visualización en el navegador de un árbol genealógico. Usa una simbología estándar, implementando la mayoría de los estándares establecidos por PSTF. Maneja su información a través de formatos como XHTML, OASIS, Office Open XML, Madeline XML y, opcionalmente, archivos en formato MySQL desde la computadora local o de una ubicación externa usando el protocolo de red HTTP/S[21]

Haplopainter

Es un sistema libre para Windows que posibilita un fuerte trabajo con los haplotipos, permitiendo el análisis de múltiples marcadores en el árbol genealógico. Tiene la deficiencia de que el sistema es muy ligero, y carece de facilidades para el trabajo gráfico. Además no presenta un manejo de símbolos u otros elementos estándares. Se centra fundamentalmente en cargar árboles desde diversos formatos tales como CSV y Linkage, o desde sistemas de BD en PostgreSQL, MySQL u Oracle. Importa además los haplotipos desde base de datos reconocidos tales como Simwalk, GeneHunter, Merlin y Allegro. Exporta los árboles en varios formatos, entre ellos CSV y PNG [22].

Progeny clínica

Es una herramienta ideal para la gestión de árboles genealógicos y datos clínicos. Es una solución de software para la familia basada en estudios e investigaciones que desde 1996 ha estado proporcionando a las instituciones de investigación y los servicios clínicos genéticos en todo el mundo la posibilidad de establecer genealogías y seguimiento de los datos de pacientes y su historia. Puede configurar la base de datos para incluir campos ilimitados, pantallas de entrada de datos de diseño, permiten funciones de seguridad para restringir el acceso a usuarios específicos, ejecutar consultas sobre los datos entre otros. Estas funcionalidades se integran a Progeny Lab, software de gestión de laboratorio que permite la gestión de todos los datos genéticos [23] y Progeny LIM, software que permite dar seguimiento a cualquier tipo de muestra y los datos asociados a su inventario [24], dos herramientas que le dan un gran valor agregado.

Esta herramienta es una de las más completas en la actualidad pero tiene la limitante de excesivos precios por el pago de licencia. A continuación en la Figura 1.1 una interface de la misma.

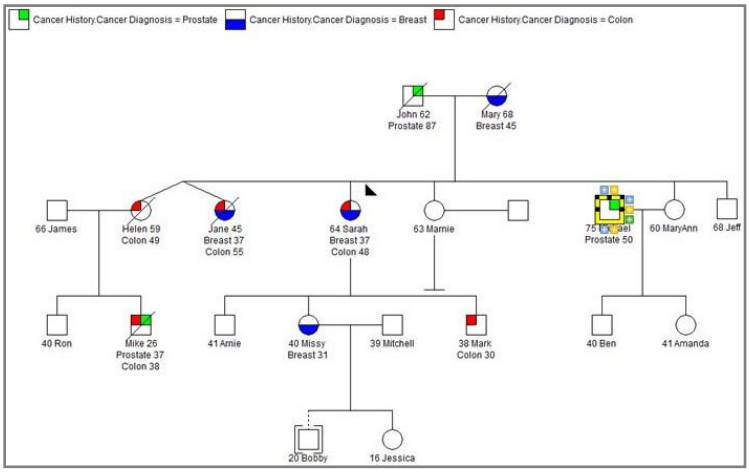


Figura 1.1 Interface de usuario Progeny Clinical.

alasARBOGEN

Primera herramienta de este tipo desarrollada en Cuba. Se desplegó desde el año 2007 en la Red Nacional de Genética Médica en su primera versión. Esta versión no tuvo mucho éxito debido a la falta de un mecanismo de soporte técnico y actualización de versiones. Tenía un

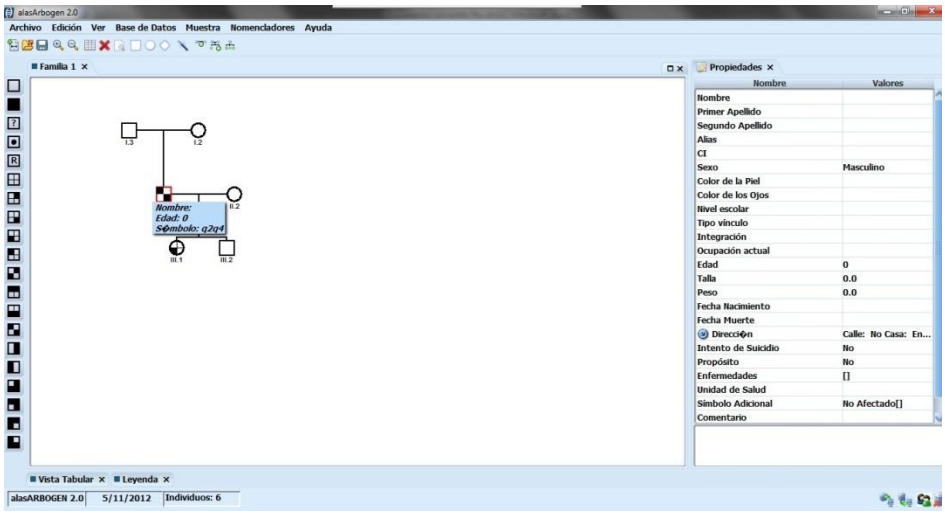
grupo de limitantes relacionada con las facilidades visuales y que fue desarrollada en .net, tecnología privativa. Luego se desarrolla su versión 2.0 en el 2010, con un cambio de tecnología y dando un salto importante en cuanto a interfaces de usuarios amigables con un ambiente parecido al Cyrilic. Se desarrolló una versión de escritorio y otra web, con una BD en PostgreSQL que garantiza los temas de seguridad y fiabilidad de la información. Tiene una arquitectura cliente-servidor que es lo suficientemente flexible para permitir trabajar conectado o desconectado y luego actualizar [25].

Se desarrolló en conjunto con el CNGM y cuenta con las funcionalidades necesarias para el trabajo de los genetistas. Una de sus principales novedades es el cálculo de riesgo a partir de la incorporación de un módulo que además permite definir reglas y luego emplearlas en la predicción. Se presentan a continuación sus principales funcionalidades:

- Gestión avanzada de las muestras de laboratorio.
- Mejora de las relaciones entre los individuos, sobre todo las gemelares (Monocigóticos y Disigóticos).
- Algoritmos que permitan la organización automática de los árboles.
- Métodos computacionales para la representación y análisis de los arboles.
- Estandarización de ficheros de salvas que permitan intercambio con otros sistemas líderes en el mundo.

Su distribución es totalmente gratuita. A continuación en la Figura 1.2 una imagen de la misma.

Figura 1.2 Interface de usuario alasARBOGEN 2.0.



1.2.1. Consideraciones sobre la representación de los árboles genealógicos.

Como se evidencia de la revisión anterior, se pueden clasificar las herramientas en dos grandes grupos, los que tienen que pagar licencia y los que no. Este elemento es de suma importancia pues el investigador puede tener la intensión de utilizar alguna en específico pero si es privada y no cuenta con la posibilidad de adquirirla debe renunciar a esa opción.

Si se analizan los elementos descritos, puede resumirse que las herramientas más completas y con mejores características lo constituyen las privativas, entre estas las más reconocidas son Cyrilic y Progeny Cliniical. Hay que señalar que las herramientas que no implican pago de licencia, presentan deficiencias que limitan su uso en los círculos académicos y científicos. A pesar de esto en el estudio se evidencia que de las herramientas libres, alasARBOGEN y Madeline 2.0 PDE compiten por sus características con muchas de las privadas.

1.3. Técnicas de análisis de datos.

Debido al desarrollo de los microprocesadores y los sistemas de almacenamiento, en los últimos años ha existido un crecimiento acelerado en las capacidades de generar y colectar datos. Contar con muchos datos no significa la meta, puesto que la cantidad no implica por si sola obtener mejores resultados si no se cuenta con la posibilidad de explotarlos. Sin embargo, dentro de estas enormes masas de datos existe una gran cantidad de información oculta de un valor incalculable, a las que no se puede acceder por las técnicas clásicas de recuperación de la información.

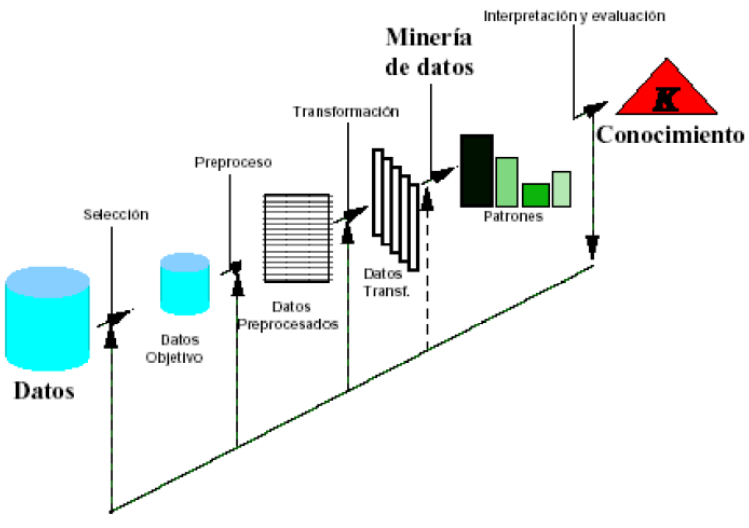
Para dar solución a esta problemática planteada anteriormente muchos expertos investigan hoy en el KDD [26]. El hallazgo de esta información oculta es posible gracias a la MD, que entre las técnicas que aplica se encuentra la inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad.

Conceptualmente se define el KDD es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos [27].

El objetivo fundamental del KDD es encontrar conocimiento útil, válido, relevante y nuevo sobre un fenómeno o actividad mediante algoritmos eficientes, dadas las crecientes órdenes de magnitud en los datos. Al mismo tiempo hay un profundo interés por presentar los resultados de manera visual o al menos de manera que su interpretación sea muy clara. Otro aspecto es que la interacción humano-máquina deberá ser flexible, dinámica y colaboradora. El resultado de la exploración deberá ser interesante y su calidad no debe ser afectada por mayores volúmenes de datos o por ruido en los mismos. En este sentido, los algoritmos de descubrimiento de información deben ser altamente robustos [28].

El proceso de KDD consiste en usar métodos de minería de datos (algoritmos) para extraer (identificar) lo que se considera como conocimiento de acuerdo a la especificación de ciertos parámetros usando una base de datos junto con pre-procesamientos y post-procesamientos [29]. En la Figura 1.3 se muestra el proceso de KDD.

Figura 1.3 El proceso de KDD



La MD se fortalece por su carácter integrador de varias áreas del conocimiento, como son la Estadística, la Inteligencia Artificial, la Computación Gráfica, las BD y el Procesamiento Masivo. Su punto de partida y soporte, puesto que es donde se encuentra la materia prima, son las BD.

La Minería de Datos se define como:

"… un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos." [30]

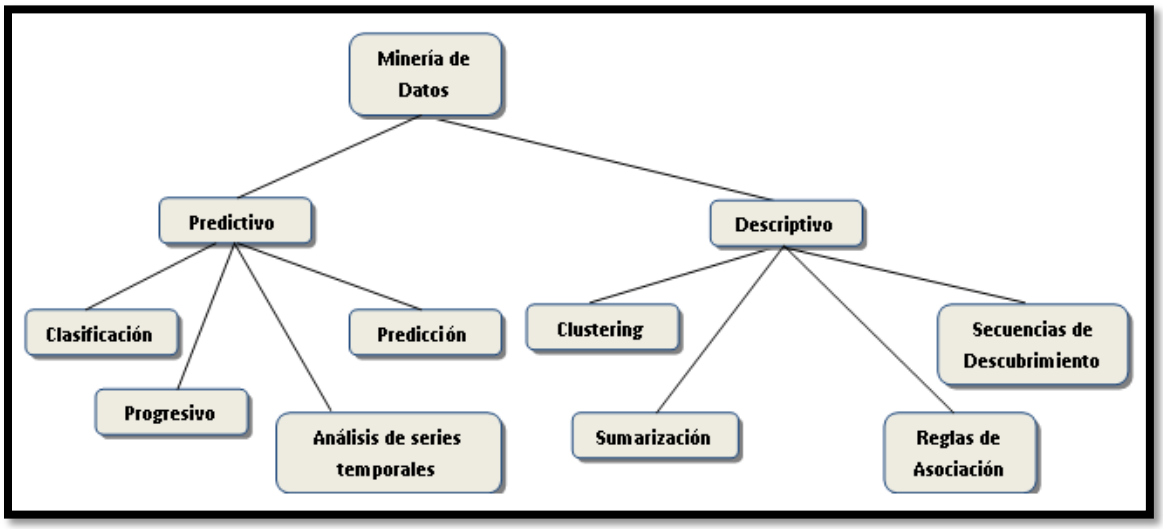
Aunque se suele utilizar indistintamente los términos de KDD y DM es importante aclarar que no son lo mismo. Según lo define Fayyad en 1996 el DM es el paso dentro del proceso de KDD, consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos pre-procesados.

A pesar de que se desconoce cómo lograr que las computadoras aprendan tan bien como las personas, ciertos algoritmos propuestos en el campo han resultado efectivos en varias tareas de aprendizaje. Los algoritmos de la DM los se clasifican en dos grandes grupos [31] como se muestra en la Figura 1.4.

Supervisados o predictivos: predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos. A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción de datos cuya etiqueta es desconocida.

No supervisados o descriptivos: con estos algoritmos se descubren patrones y tendencias en los datos actuales. El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio de ellas.

Figura 1.4 Clasificación de los algoritmos de MD.



1.4. Aplicaciones para el análisis de datos.

Las herramientas de MD permiten extraer patrones, tendencias y regularidades para describir y comprender mejor los datos y predecir comportamientos futuros. Facilitan el acceso a la información para que el análisis sea más efectivo, es decir son instrumentos de apoyo.

Tanto la tecnología informática actual, la madurez de las técnicas de aprendizaje automático y las nuevas herramientas de MD de sencillo manejo, permiten a una pequeña o mediana organización (o incluso individuos) tratar grandes volúmenes de datos almacenados en BD (propias de la organización, externas o en la web).

La aparición de numerosas herramientas y paquetes de "minería de datos", como: Clementine de SPSS, Intelligent Miner de IBM, Enterprise Miner de SAS, DM Suite (Darwin) de Oracle, WEKA (de libre distribución), Knowledge Seeker y Rproject (herramienta gratuita de análisis estadístico), posibilitan el uso de técnicas de minería de datos a personas no especializadas en el tema [32].

Clementine / SPSS

Se trata de un software especializado de MD, creado por el grupo SPSS. Permite una visualización de los datos ofrecidos por la base de datos que se maneja, de forma que el proceso de negocio quede patente mediante el sistema analítico que construye. Algunas características de su funcionamiento se pueden resumir en las siguientes:

- Importa los datos directamente de las bases más importantes (Oracle, Ingres, Sybase o Informix) aunque permite importar igualmente formatos de datos de cualquier otra base o fichero.
- Facilita el trabajo con series de tiempo, permitiendo el filtrado simple o personalizado en los registros e incluye funciones para procesar secuencias.
- Permite visualización intuitiva de los datos al trabajar con diversidad de formatos, diagramas de puntos, gráficos, histogramas, etc., señalando áreas de especial interés.
- Utiliza técnicas de redes neuronales e inducción para aprender de salidas previas en la realización de predicciones de forma tal que el conocimiento de la toma de decisiones se gestiona automáticamente.
- Aprovecha la infraestructura existente con minería de datos reduciendo costos y maximizando la tecnología [33].

A pesar de ser una de las herramientas más utilizadas actualmente, posee licencia comercial, trayendo como consecuencia que empresas con restricciones financieras no la puedan utilizar.

Intelligent Miner

Se trata de un software privativo creado por IBM provisto de una serie de herramientas de Data Mining que permiten el procesamiento de datos, visualización de resultados y el uso de una amplia variedad de algoritmos, de forma individual o en combinación para la resolución de problemas empresariales: análisis de mercados, segmentación de clientes. Permite la posibilidad de soporte para múltiples plataformas: AIX, Windows NT, Sun Solaris, y OS/390, y la facultad de manejar grandes volúmenes de datos.

Se presenta como un software dirigido a clientes tanto expertos como no expertos en técnicas de MD, ya que ofrece una visualización de datos y resultados basada en gráficos que presentan los resultados de las asociaciones y modelos detectados.

Las operaciones de modelización se fundamentan en el descubrimiento de tendencias, clúster demográfico (comportamiento de compra, segmentación del mercado) y clasificaciones de árbol (tendencia a la compra, gasto proyectado) [34].

Enterprise Miner

Este software, desarrollado por el SAS Institute, constituye una herramienta privativa, cuya principal característica es el empleo de un entorno integrado en el que los estadísticos, los

gerentes de la empresa y el departamento de investigación, pueden trabajar conjuntamente consiguiendo una modelización de datos más rápida.

Presenta flexibilidad en su configuración y resulta eficiente tanto para el usuario individual como para los requerimientos de grandes empresas. Incluye algoritmos avanzados, árboles de decisión, redes neuronales, razonamiento basado en el aprendizaje, regresión lineal y logística, análisis clúster, asociaciones y series temporales [35].

DM Suite (Darwin)

Es un sistema de multiestrategia para el desarrollo del conocimiento en BD de gran dimensión. Es distribuido bajo licencia comercial. Está provisto de un sistema integrado con una gran variedad de algoritmos para la construcción y aplicación de modelos de aprendizaje, y problemas de predicción y clasificación. Está ideado para la resolución de conflictos de estrategia y operaciones a los que pueden enfrentarse los decisores de las empresas en la realidad. Creado por Oracle, presenta las siguientes características:

- Escalabilidad y paralelismo.- Cuanto mayor sea el número de procesadores utilizados simultáneamente, más rápidamente se construirán los modelos y mayor será la amplitud de los datos utilizados para el aprendizaje.
- Multi-estrategia.- Presenta flexibilidad a la hora de utilizar distintos algoritmos, eligiendo el mejor para cada problema. Igualmente, existe la posibilidad de combinarlos para obtener mayor poder en la modelización.
- Modularidad.- Permite añadir el aprendizaje y el análisis de algoritmos en el futuro, simplificando la integración y el soporte de aplicaciones [32].

R Project

Realiza análisis estadísticos destinados a la investigación biomédica. Puede asociarse a varias BD y librerías que utilicen lenguajes como Perl, Python, C o Fortran. Soporta múltiples plataformas como: GNU/Linux, Windows XP, Windows 2000, Windows XP, Windows Vista, Windows 7, Windows 8.

Es de distribución libre, su implementación es sencilla y no requiere de mayores recursos informáticos pero tiene como desventajas que no ofrece una interfaz de usuario amigable y las llamadas a R se realizan en línea de comando. Además sus paquetes no siempre se utilizan de

la misma forma, al provenir de desarrolladores diferentes, lo que dificulta la realización de muchas tareas [36].

Weka

Este software está programado en Java. Es independiente de la arquitectura, ya que funciona en cualquier plataforma sobre la que haya una máquina virtual Java disponible. Weka se denomina a sí mismo como un conjunto de Librerías para tareas de MD. Está disponible libremente bajo la licencia pública general de GNU. Contiene una extensa colección de técnicas para preprocesamiento de datos y modelado. Es fácil de utilizar por un principiante gracias a su interfaz gráfica de usuario.

Soporta varias tareas estándar de MD, especialmente, preprocesamiento de datos, clustering, clasificación, regresión, visualización y selección. Todas las técnicas de Weka se basan en que los datos están disponibles en un fichero plano (flat file) o una relación, en la que cada registro de datos está descrito por un número fijo de atributos (generalmente numéricos o nominales, aunque soporta otros tipos). Además proporciona acceso a BD vía SQL gracias a la conexión JDBC (Java Database Connectivity) y puede procesar el resultado devuelto por una consulta hecha a la base de datos [37].

1.4.1. Consideraciones sobre las aplicaciones para el análisis de datos

El estudio de varias herramientas que apoyan la minería de datos arrojó que existen algunas privativas y de código abierto. Las privativas son muy completas pero no es posible reutilizar el código fuente ni sus componentes. Sus valores monetarios son altos, por lo que representan un gasto financiero considerable para el país. En cambio las de código abierto como Rproject y Weka fueron analizadas y cumplen con las necesidades de investigadores de la materia. De las privativas la de mayor referencia es Clementine/SPSS. De manera general, estas son las 3 herramientas punteras en los temas de análisis de dato en la actualidad.

1.5. Tendencias actuales de la integración de herramientas informáticas.

En los últimos años se han logrado avances significativos en los temas relacionados con la integración de aplicaciones informáticas, tratando de dar solución a las islas tecnológicas desarrolladas en las empresas que con el tiempo han ido surgiendo como solución a procesos independientes de las organizaciones.

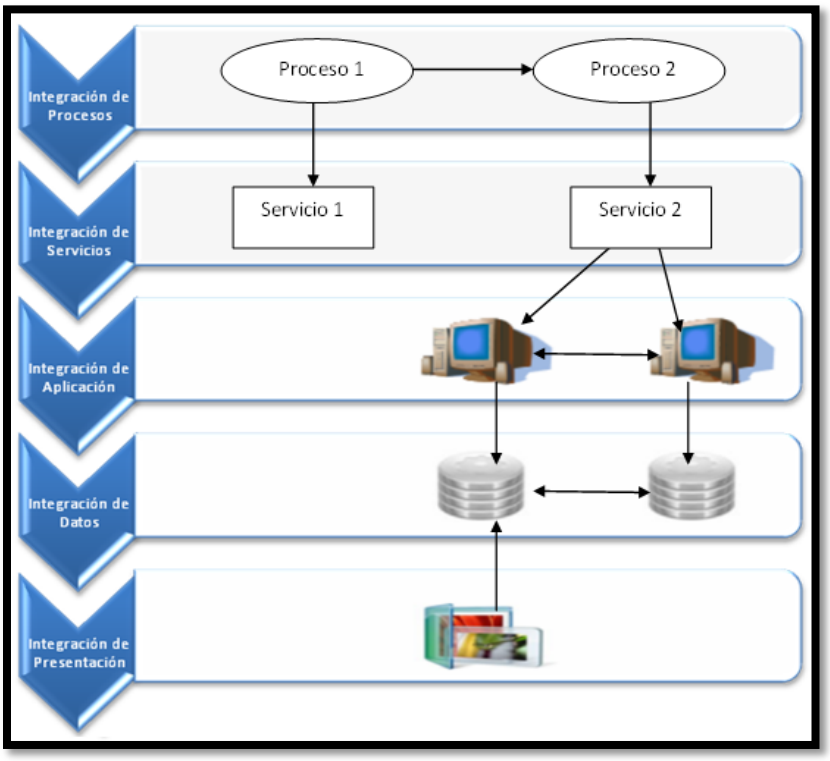
Los métodos tradicionales de integración de sistemas han demostrado no ser escalables, resolviendo sus problemas en su momento pero ocasionando otros a largo plazo. Las tendencias actuales giran sobre soluciones tecnológicas robustas que pueden abordar la problemática en toda su magnitud y complejidad. Sin embargo, no es sólo en la integración de la tecnología donde está el problema, también se trata de los propios procesos y flujo del negocio, por lo que la temática puede implicar una re-conceptualización de cómo funcionan las organizaciones [38].

En la última década la tendencia y los enfoques para atacar el problema de la integración han cambiado. La integración de la tecnología y las técnicas relacionadas con la información no estructurada en los informes y los medios digitales, y los datos estructurados en bases de datos, se están convirtiendo en una parte importante en el ámbito de integración. Esto se debe a una variedad de factores, incluida la aparición de XML como un formato de datos estándar. La información y los datos están en el corazón de cada proyecto de integración. Últimamente, la integración se realiza sobre distintos tipos de intercambio de datos en formatos diferentes. El problema central de los proyectos de integración es la manera de permitir la interoperabilidad entre sistemas con los datos en diferentes formatos y estructuras [39].

La complejidad de los procesos que cubren las aplicaciones de software en la actualidad provoca que las funcionalidades no siempre sean cubiertas por un componente o incluso que sobrepasen los límites de la aplicación. Este problema encuentra solución con la integración a diferentes niveles (Figura 1.5):

- Integración de presentación: permite la agregación de datos provenientes de diferentes sistemas dentro de una única vista.
- Integración de datos: permite la sincronización de los datos ubicados en BD diferentes. Si esta se realiza solo después de un periodo de tiempo, cierta cantidad de datos se perderán.
- Integración de aplicaciones: permite que las funcionalidades de una aplicación puedan ser accedidas directamente por otras.
- Integración de servicios: permite poner a disposición de varias aplicaciones un conjunto común de servicios reutilizables.
- Integración de procesos: permite la definición de modelos de procesos de negocio o flujos de trabajo, a través de la llamada de servicios reutilizables.

Figura 1.5. Niveles de integración.



A continuación se muestran un grupo de ventajas y desventajas de cada uno de los niveles mencionados anteriormente[40].

Integración de presentación:

Ventajas:

- No requiere cambios en los sistemas a integrar.
- Su riesgo en bajo puesto que no modifica ningún sistema.
- El costo en relativamente bajo puesto no implica muchos desarrollos.

Desventajas:

- El rendimiento es muy bajo.
- Es muy susceptible a cambios y actualizaciones en los sistemas.
- En muchas instancias solo prolonga el problema de la integración, se considera una solución intermedia.

Integración de datos:

Ventajas:

- Es económica
- Disponibilidad de tecnologías para su construcción.
- Es rápida a la hora de su desarrollo.

Desventajas:

- No resuelve el problema de la integración de métodos, es más que todo una medida temporal.
- No escala adecuadamente para integración de aplicaciones OLTP.
- Puede esconder asuntos importantes de los Sistemas de Información Empresarial

Integración de aplicación:

Ventajas:

- Proveen mecanismo para compartir tanto métodos como datos.
- Mueven la información fuera de los sistemas que eran difíciles de acceder.
- Proveen la infraestructura para compartir procesos de negocios comunes.
- Soportan problemas de negocios comunes, tales como uniones y adquisiciones.

Desventajas:

- Las interfaces que proveen los vendedores de aplicaciones empaquetadas varían desde buenas hasta inexistentes.
- Se necesita acceso a los datos.

Integración de servicios:

Ventajas:

- Proveen de una infraestructura para la reutilización de código para muchas aplicaciones empresariales.
- Disponibilidad de tecnologías y experticia.
- Es la solución más adecuada para muchas empresas.

Desventajas:

- Mucho más compleja y costosa que los otros enfoques.
- Necesita de mucho tiempo, arquitectura y planificación.
- Las tecnologías existentes pueden no escalar a aplicaciones empresariales o tener otras deficiencias.

Integración de procesos:

Ventajas:

- Permite integración a nivel de flujo de los procesos de negocio de una entidad.
- Permite incluir flujos de aprobación.
- Permite gestionar integralmente un proceso.

Desventajas:

- Es una solución altamente costosa y compleja de desarrollar.
- Se alcanza por un proceso de madurez del propio desarrollo de software de la organización.

Es importante destacar que no existe una solución ideal para integrar aplicaciones. La selección de la misma depende en gran medida del problema a resolver y el escenario en que se encuentre. Es por eso que en ocasiones conviene un nivel u otro. Todos los niveles han sido probados y cada uno puede mostrar ejemplos en los que han sido empleados con éxitos.

La integración a nivel de procesos es de las más novedosas, su implantación requiere de un proceso de madurez en los procesos de la organización. Se centra en el logro de la interrelación necesaria entre las distintas actividades y etapas del proceso más allá de las

barreras técnicas u organizacionales para el logro de un propósito determinado. Se apoya en los distintos niveles de integración pero está orientado al logro de la interoperabilidad de los componentes funcionales[41]. Un nivel de integración intermedio entre actualidad y infraestructura lo constituyen a nivel de aplicación, los niveles más novedosos no siempre es factible utilizarlo, sobre todo si se trata de soluciones que no sean a nivel empresariales con una complejidad mayor.

1.6. Métodos multicriterio

La búsqueda de la eficiencia y la productividad contribuyen a la exploración de metodologías de apoyo para la toma de decisiones en escenarios donde intervienen múltiples variables o criterios de selección.

El análisis multicriterio se define como el mundo de conceptos, aproximaciones, modelos y métodos, usados para auxiliar a los decisores a describir, evaluar, ordenar, jerarquizar, seleccionar o rechazar objetos, con base en una evaluación (expresada por puntuaciones, valores o intensidades de preferencia) de acuerdo con varios criterios. Los métodos de decisión multicriterio son una base, sustentada en elementos científicos, que aporta mejoras distintivas para asumir una decisión [42].

Existen métodos multicriterio para fortalecer y ampliar los resultados técnicos, obtenidos a través de la revelación de las preferencias de actores involucrados en la toma de decisiones.

Proceso Analítico Jerárquico (AHP: The Analytic Hierarchy Process)

El Proceso Analítico Jerárquico (Analytic Hierarchy Process, AHP), propuesto por Saaty en 1980 [43], se basa en la idea de que la complejidad inherente a un problema de toma de decisión con criterios múltiples, se puede resolver mediante la jerarquización de los problemas planteados.

El método AHP es un modelo de decisión que interpreta los datos y la información directamente mediante la realización de juicios y medidas en una escala de razón dentro de una estructura jerárquica establecida. Es un método de selección de alternativas (estrategias, inversiones, etc.) en función de una serie de criterios o variables, las cuales suelen estar en conflicto. Es un método matemático creado para evaluar alternativas cuando se tienen en consideración varios criterios y está basado en el principio que la experiencia y el conocimiento de los actores son tan importantes como los datos utilizados en el proceso [44, 45].

Además optimiza la toma de decisiones complejas cuando existen múltiples criterios o atributos, mediante la descomposición del problema en una estructura jerárquica, desagregado en sus elementos más pequeños. En este sentido, es clara la importancia de cada elemento (criterio). Sobre la base de una secuencia de comparaciones entre pares, las prioridades relativas (pesos) se determinan mediante el método Eigenvector.

Método de Relaciones de Superación (PROMETHEE: Preference Ranking Organization Method for Enrichment Evaluation)

Trata de establecer, mediante la evaluación en función de criterios, una ordenación jerarquizada en un conjunto de alternativas. El método busca establecer un orden jerarquizado de las alternativas, determinado por un flujo neto, este se compone del flujo entrante y saliente de cada alternativa, estos flujos reflejan el nivel de dominar o ser dominado de unas alternativas con otras. Existen 6 tipos de funciones de preferencia mediante las cuales se puede establecer ese orden jerarquizado, cada una de ellas representa diferentes soluciones de decisión [46, 47].

Este método no posee una guía específica para determinar los pesos. Además, los criterios generalizados necesitan ser definidos, lo que puede ser difícil de lograr por un usuario inexperto.

Técnica para ordenar las preferencias mediante la similitud a la solución ideal (TOPSIS)

El método TOPSIS es un modelo de decisión que ordena preferencias por similitud, desarrollado por [48]. TOPSIS es un método de decisión multicriterio de ordenación para identificar las soluciones de un conjunto finito de alternativas. El principio básico es que la alternativa elegida debe tener la menor distancia a la solución ideal positiva y la mayor distancia a la solución ideal negativa. Una solución ideal se define como una colección de puntuaciones o valores en todos los atributos considerados en la decisión, pudiendo suceder que tal solución sea inalcanzable. El vector compuesto por los mejores valores del j-ésimo atributo respecto de todas las alternativas posibles es quien recibe el nombre de “solución ideal positiva” (SIP); recíprocamente, la “solución ideal negativa” (SIN) será aquella cuyo vector contenga los peores valores en todos los atributos.

El concepto intuitivo de alternativa ideal es que sería aquella que, sin dudarlo, siempre elegiría el decisor. De igual modo, la alternativa anti-ideal sería aquella que, sin dudarlo, nunca elegiría el decisor. De este modo, puede suceder que una alternativa seleccionada desde el punto de

vista de su “distancia” más corta respecto de la solución ideal positiva deba competir con otra alternativa seleccionada como la más lejana de la solución ideal negativa. Por ello, y a fin de definir la solución ideal, el método TOPSIS define un índice de similitud que se construye combinando la proximidad al ideal positivo y la lejanía respecto al ideal negativo.

1.6.1. Consideraciones sobre los métodos multicriterio

PROMETHEE en el caso de muchos criterios, se dificulta la toma de decisiones para obtener una visión clara del problema y evaluar los resultados. Este método no posee una guía específica para determinar los pesos. Además, los criterios generalizados necesitan ser definidos, lo que puede ser difícil de lograr por un usuario inexperto. TOPSIS implica necesariamente la definición de un ideal negativo y un ideal positivo.

Sin embargo una de las ventajas más relevantes del modelo AHP consiste en la estructuración de la jerarquía del problema de forma visual. Para aplicar este método no es necesario contar con información cuantitativa sobre el resultado que alcanza cada alternativa en cada uno de los criterios considerados, sino tan sólo los juicios de valor de la persona que tome las decisiones.

Conclusiones

Es importante obtener mejores resultados en el proceso de captura y análisis de la información genealógica para contribuir a mejorar los resultados en las investigaciones clínico genéticas. Las herramientas de representación genealógica no permiten realizar actividades minería de datos. A pesar de que también múltiples herramientas que soportan la minería de datos, no hay integración entre unas y otras. Estos elementos evidencian la necesidad de un marco de trabajo flexible e integrado que permita disminuir el esfuerzo en la captura y análisis de la información genealógica.

Capítulo 2. Marco de trabajo para el análisis de la información genealógica.

Introducción

En el presente capítulo se realiza una descripción del marco de trabajo propuesto para hacer más flexible el proceso de análisis de la información gestionada en los sistemas para la representación del árbol genealógico. Se especifica, en cada una de las tres etapas que lo componen, todas las actividades que se proponen desarrollar en estas. La presente propuesta, a pesar de estar soportada por un grupo de herramientas informáticas da la posibilidad al usuario de escoger las mismas. Se propone la utilización de técnicas multicriterio para la selección de herramientas.

1.1. Descripción del marco de trabajo propuesto.

La constante evolución de las tecnologías para el diagnóstico en los hospitales y centros de investigaciones para la salud hace que cada día aumenten los volúmenes de datos a analizar. La información almacenada por sí sola, aunque sea a grandes escalas, no constituye conocimiento y por ende no permite hacer análisis para tomar decisiones. Proveer a los especialistas, de mecanismos, técnicas y herramientas, que permitan acelerar estos análisis, constituiría un salto cualitativo en el avance de los servicios de salud.

Un marco de trabajo actúa como un mapa que provee un orden coherente a determinadas actividades que permiten solucionar un determinado problema [49]. Un marco de trabajo establece la base para un proceso de análisis de datos completo, al identificar actividades aplicables a cualquier escenario sin importar su tamaño o complejidad [50].

Teniendo en cuenta lo antes planteado, el marco de trabajo propuesto en la presente investigación presenta como principios y características la integración de distintas actividades relacionadas con la captura, integración y análisis de la información genealógica del paciente en un marco de trabajo único. Se orienta a disminuir el esfuerzo en la obtención de conocimiento a partir de los datos analizados. Facilita el trabajo de los especialistas, permitiendo mejorar los servicios sanitarios de la sociedad. Con esta propuesta se logra de una manera integrada cerrar el ciclo de la gestión y el análisis de la información que hoy se gestiona en los estudios clínicos genéticos así como en las consultas del sector. Brinda información para que las autoridades competentes puedan tomar estrategias y medidas en

función de prevenir enfermedades, conductas, así como mejorar el bienestar social y elevar la calidad de vida de las personas.

El marco de trabajo toma comprende un conjunto de 9 actividades dividas en tres etapas de ejecución: Captura, Integración y Análisis, las cuales se relacionan a continuación:

Etapa 1. Captura de la información genealógica del paciente

1. Seleccionar la herramienta para la representación genealógica.
2. Registrar la información genealógica del paciente.

Etapa 2. Integración de las herramientas

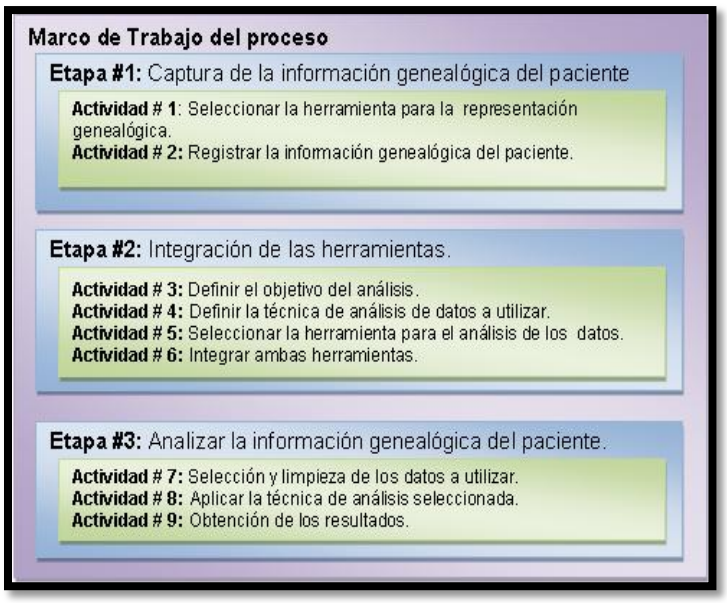
3. Definir el objetivo del análisis.
4. Definir la técnica de análisis de datos a utilizar.
5. Seleccionar la herramienta para el análisis de los datos.
6. Integrar ambas herramientas.

Etapa 3. Analizar la información genealógica del paciente

7. Selección y limpieza de los datos a utilizar.
8. Aplicar la técnica de análisis seleccionada.
9. Obtención de los resultados.

Una vez ejecutadas las actividades se obtiene como salida un informe con una caracterización de los datos analizados, los gráficos resultantes que permiten una mejor interpretación y descripciones del análisis de la información procesada. En la figura 2.2 se muestra una representación gráfica del marco para una mayor comprensión.

Figura 2.2 Marco de trabajo propuesto



A continuación se muestra una descripción detallada de cada una de las etapas con sus actividades que conforman el marco de trabajo.

1.2. Etapa 1. Captura de la información genealógica del paciente

Objetivo: Registrar la información genealógica del paciente en una herramienta de representación genealógica.

Técnicas utilizadas: Grupo Focal, AHP.

Entrada: Información genealógica del paciente.

Salida: Información genealógica del paciente almacenada en una herramienta de representación genealógica.

Actividades:

1. **Seleccionar la herramienta para la representación genealógica.**

Para ganar mayor flexibilidad e independencia no se predefine una herramienta. El especialista que va a aplicar el marco de trabajo tiene la posibilidad de seleccionarla. Para esto se debe conocer una serie de herramientas disponibles para la representación genealógica. Como parte

de la investigación, se aplica primeramente la técnica multicriterio Grupo Focal, para obtener los criterios de selección, epígrafe 2.2.1. Luego se le aplica a estos criterios el método AHP con el operador de agregación OWA para ponderarlos, epígrafe 2.2.2. A continuación se muestra en la tabla 2.1 los criterios con sus respectivas ponderaciones.

Tabla 2.1 Ponderación de los criterios de selección.

Identificador	Criterio	Ponderación
C_1	Pago de licencia	0.450
C_2	Persistencia de los datos mediante la utilización de BD	0.090
C_3	Cubrimiento de las funcionalidades necesarias	0.014
C_4	Garantía de la seguridad y fiabilidad de los datos	0.270
C_5	Estructura de datos para la representación.	0.050

Finalmente se decide la herramienta a utilizar y se instala para su posterior explotación.

2. Registrar la información del paciente

Para registrar la información del paciente se pueden presentar dos escenarios:

Escenario 1: Consulta del especialista en genética. Esta es una forma pasiva de ingresar la información de las personas que asisten a las consultas clínico genéticas.

Escenario 2: Censos poblacionales. Esta es una forma activa de obtener la información de un grupo de personas con características, sospechas o rasgos en común con el objetivo de comprobar una hipótesis.

En ambos casos, una vez introducida la información, se deben revisar las preguntas abiertas a comentarios, que luego dificultan la interpretación por un sistema automatizado, así como evaluar detenidamente el rango de las variables para acotar sus valores. Particularmente en el escenario 2 es fundamental que especialistas en la materia revisen los instrumentos aplicados antes de digitalizarlos, para corregir la mayor cantidad de errores.

Es importante dedicar todo el tiempo necesario a garantizar la calidad del dato para maximizar su precisión.

1.3. Etapa 2. Integración de las herramientas

Objetivo: Integrar las herramientas seleccionadas para la representación del árbol genealógico y para el análisis de los datos.

Técnicas utilizadas: AHP, Grupo Focal, Integración de herramientas a nivel de aplicaciones.

Entrada: Información genealógica del paciente almacenada en la herramienta de representación genealógica.

Salida: Herramientas integradas.

Actividades:

3. **Definir el objetivo del análisis.**

El especialista debe tener claridad de los datos que tiene almacenados para a partir de estos plantearse su propósito con los mismos, o sea, definir qué objetivo persigue como usuario que interpretará los resultados finales. De esta manera permitirá guiar la búsqueda de los resultados, filtros y vistas en función de los objetivos. No resulta aconsejable aplicar técnicas a ciegas para luego formular un resultado.

4. **Definir la técnica de análisis de datos a utilizar.**

El especialista según las necesidades planteadas valora si se trata de un análisis predictivo, cuando ese pretende evaluar cómo será el futuro, o descriptivo, cuando proporciona información sobre las relaciones entre los datos y sus características que a simple vista no es posible determinar.

Luego, de acuerdo al tipo de análisis seleccionado, se decide la técnica de análisis de datos. Para decidir cuál es la más apropiada para aplicar es necesario tener claridad del objetivo que se propone con los datos. De este análisis depende no solamente la técnica sino el algoritmo a aplicar en los próximos pasos.

5. **Seleccionar la herramienta para el análisis.**

El especialista debe conocer una serie de herramientas disponibles para el análisis de datos. Mediante la aplicación de la técnica multicriterio AHP se evalúan las herramientas a partir de los siguientes criterios de selección, obtenidos como resultado de la aplicación de la técnica Grupo Focal. La misma se describe en el epígrafe 2.2:

- Pago de licencias.
- Implementación de los algoritmos necesarios para el análisis.
- Capacidad de integración con la herramienta de representación.

Finalmente con el resultado de la técnica AHP se decide la herramienta a utilizar.

6. Integrar ambas herramientas.

Una vez seleccionadas las herramientas para la representación del árbol genealógico y la de análisis de los datos se procede a integrarlas. Se propone realizar la integración a nivel de aplicaciones teniendo en cuenta los siguientes aspectos [51] :

- Flexibilidad para el acceso a la lógica de negocio.
- Acceso a datos a través de funcionalidades existentes.
- Implementación de un proceso de negocio que se pueda utilizar desde ambas aplicaciones.
- Interoperabilidad mediante interfaces de programación.
- Nuevos desarrollos basados en los códigos originales.

La propuesta de integración no establece dependencia entre las herramientas, permitiendo mantenerlas desacopladas pero sí debe seguir los siguientes principios:

En cuanto a la infraestructura para la integración se debe:

- Establecer y mantener la estrategia de integración del producto.
- Establecer y mantener el entorno necesario para dar soporte a la integración de los componentes del producto.
- Establecer y mantener los procedimientos y los criterios para integración de los componentes del producto.

En cuanto a la gestión de las interfaces entre componentes se debe:

- Revisar las descripciones de la interfaz en cuanto a cobertura y completitud.
- Gestionar las definiciones, diseños y cambios de las interfaces internas y externas para los productos y los componentes de producto.

Y en cuanto al ensamblaje de los componentes se debe:

- Confirmar, antes de ensamblar, que cada componente de producto requerido para ensamblar el producto ha sido identificado correctamente, funciona de acuerdo con su descripción y que las interfaces de componente de producto cumplen con las descripciones de la interfaz.
- Ensamblar los componentes de producto de acuerdo a los procedimientos y estrategia de integración del producto.
- Evaluar los componentes de producto ensamblados para compatibilidad de la interfaz.
- Empaquetar el producto o componente de producto ensamblado y entregarlo al cliente.

1.4. Etapa 3. Analizar los datos procesados

Objetivo: Realizar el análisis de la información genealógica del paciente.

Técnicas utilizadas: Técnica de MD seleccionada.

Entrada: Información genealógica del paciente almacenada en una base de datos.

Salida: Informe resumen de los resultados del análisis.

Actividades:

7. **Selección y limpieza de los datos a utilizar.**

La calidad del conocimiento descubierto no solo depende del algoritmo de minería utilizado, sino también de la calidad de los datos minados [52].

Se deben seleccionar y preparar el subconjunto de datos que se va a minar, los cuales constituyen lo que se conoce como vista minable. Este paso es preciso ya que algunos datos recopilados son irrelevantes o innecesarios.

Es importante elegir las variables más significativas en correspondencia con el resultado que se percibe. Se trata de suministrarle a los algoritmos de minería el grupo de datos necesarios para el análisis en cuestión.

A continuación se proponen los principales elementos a tener en cuenta en esta actividad:

- Seleccionar variables que no se corresponden con el objetivo propuesto, enmascara y altera el resultado final.
- Analizar los valores, que pueden estar originados por errores o ser simplemente extremos de determinadas variables y no se ajustan al comportamiento general de los rangos definidos.
- Revisar la ausencia de datos, teniendo en cuenta las causas antes de tomar cualquier medida.

8. Aplicar la técnica de análisis seleccionada.

De acuerdo a la técnica de análisis de datos seleccionada se escoge el algoritmo que se va a aplicar. Posteriormente se determinan los parámetros de configuración en los que se basará la ejecución del algoritmo.

Se recomienda que dentro de cada técnica se apliquen varios algoritmos debido a que permite comparar diferentes visualizaciones de la información, proporcionándole al especialista un mejor análisis de los datos.

9. Obtención de los resultados.

Después de aplicar la técnica seleccionada se debe realizar una caracterización de la información obtenida a través de su representación gráfica con el objetivo de facilitar su interpretación.

Finalmente se genera un informe para que sea utilizado por los especialistas teniendo en cuenta los siguientes aspectos:

- Objetivo del análisis
- Población estudiada
- Muestra seleccionada

- Variables de análisis
- Caracterización de la representación gráfica de la información.

1.5. Técnicas que soportan el marco de trabajo.

El objetivo principal de la propuesta es dotar a los especialistas de un marco de trabajo a aplicar para analizar grandes volúmenes de información genealógica. Durante su desarrollo el especialista debe que enfrentarse a un grupo de decisiones, consultas y demás procesos. Como parte del marco de trabajo se proponen un conjunto de métodos y técnicas que permiten dar un mayor nivel de flexibilidad a la propuesta.

En las secciones siguientes se realizará la descripción de las técnicas propuestas a utilizar en las distintas etapas del marco de trabajo.

1.5.1. Aplicación de la técnica Grupo Focal para determinar los criterios.

El Grupo Focal [53] es una técnica cualitativa que consiste en 90 a 120 minutos de discusión con un grupo limitado de personas que reúnen ciertas características comunes para su selección y son guiados por un moderador quien conduce la sesión sobre la base de una guía de moderación. Esta técnica permite a través de las discusiones y opiniones conocer cómo piensan los participantes respecto a un asunto o tema determinado [54]. El tamaño de los grupos focales puede variar desde ocho hasta 12 personas.

En la presente investigación se siguen los pasos propuestos para la aplicación de la técnica en la selección de las herramientas de representación genealógica y para el análisis de los datos [55]:

- Diseño de la guía de moderación que permitirá recoger la información de interés para la investigación.
- Definición de la muestra y reclutamiento: selección de la composición correcta de cada grupo e identificación de las personas adecuadas para participar en las sesiones.
- Moderación de las sesiones: uno de los elementos esenciales de la metodología del Grupo Focal es el rol que tiene que jugar el moderador. Esta persona debe que ser un profesional que posea experiencia para moderar sesiones de trabajo en grupo. Al guiar la discusión debe hacer que cada persona participe e interactúe con los demás sin que un participante de manera individual domine la discusión.

- Reporte: se ofrece un resumen de la sesión de trabajo del grupo, el cual incluye los comentarios de los participantes, los resultados, las conclusiones y recomendaciones.

Esta técnica contribuye a la dinámica del grupo, permitiendo que la interacción entre estos obtenga la consideración de interesantes aspectos adicionales o identifique problemas comunes experimentados por muchas personas.

Según lo planteado se decidió aplicar dicha técnica para elegir los elementos a tener en cuenta en la selección de las herramientas para la representación genealógica y el análisis de los datos. A partir de un análisis que se realizó con un grupo de personas que se encuentran en el universo de posibles usuarios del marco de trabajo, un total de 11 especialistas funcionales del dominio, 6 genetistas y 5 informáticos. Luego se tomaron los criterios donde la mayor parte de los especialistas coincidían, los cuales se especifican a continuación. Los criterios para la selección de herramientas para la representación genealógica, se muestran en la tabla 2.2.

Tabla 2.2 Descripción de los criterios para la selección de herramientas

Criterios	**Nombre**	**Descripción**
c1	Pago de licencia.	Pago requerido para la adquisición de la licencia.
c2	Persistencia de los datos.	Persistencia de los datos mediante la utilización de BD.
c3	Cubrimiento de las funcionalidades.	Cubrimiento de las funcionalidades necesarias.
c_4	Fiabilidad.	Garantía de la seguridad y fiabilidad de los datos.
c_5	Estructura de datos para la representación.	Utilización de las estructura de datos adecuadas.

Los criterios para seleccionar la herramienta de análisis de datos se muestran en la Tabla 2.3.

Tabla 2.3 Descripción de los criterios para la selección de herramientas.

Criterios	**Nombre**	**Descripción**

r_1	Pago de licencia.	Pago requerido para la adquisición de la licencia
r_2	Implementación de los algoritmos.	Implementación de los algoritmos necesarios para el análisis.
r_3	Capacidad de integración.	Capacidad de integración con la herramienta de representación.

Para la selección de las herramientas de representación genealógica se decidió como parte de la investigación y posible aporte que facilite la aplicabilidad del marco de trabajo, determinar los pesos que debe tener cada criterio. Para poder ponderar los criterios se aplicó el operador OWA al método AHP. A continuación se describe en el epígrafe 2.2.3 este paso.

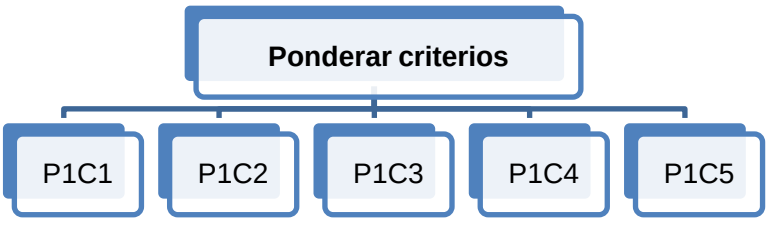
1.5.2. Un método basado en AHP y el Operador OWA para determinar los pesos de los criterios en la selección de herramientas de análisis genealógico.

El método propuesto para la determinación de los pesos de los criterios para la selección de las herramientas está basado en los puntos de vista de diferentes especialistas en la materia. El método está basado en el empleo de AHP para dar pesos a criterios [56] y el empleo del operador OWA (ordered weighted averaging o traducido al español media ponderada ordenada) [57]. El mismo se describe en detalles a continuación.

Paso 1. Desarrollo de la jerarquía

Se desarrolla la jerarquía de elementos para la selección de herramientas genealógicas. En la cima de la jerarquía se encuentra el nodo que representa el objetivo. En este caso es la ponderación de los criterios. Este nodo es descompuesto en un conjunto de criterios $C = \{c_i = 1,2,\dots,5\}$ (Figura 2.3)

Figura 2.3. Jerarquía con el objetivo y los criterios.



Paso 2: Obtención de las matrices individuales

En este paso se obtienen las prioridades a partir del conjunto de especialistas en genética descritos en la tabla 2.2, $\mathrm{E} = \{\mathrm{e}^{\mathrm{k}}, \mathrm{k} = 1,2,\dots,6\}$. Partimos de las matrices de comparación por pares entre los criterios dadas por cada experto.

$$M_C^k = \begin{array}{c} \\ C_1 \\ \vdots \\ C_t \end{array} \begin{array}{c} \begin{array}{ccc} C_1 & \dots & C_5 \end{array} \\ \begin{pmatrix} a_{11}^k & \dots & a_{15}^k \\ \vdots & \dots & \vdots \\ a_{51}^k & \dots & a_{55}^k \end{pmatrix} \end{array} \qquad (2.1)$$

Donde M_C^k es la matriz de comparación entre los criterios, $C = \{C_i, i = 1,2,\dots,5\}$, dada por el experto, e^k.

Se propone el empleo de la escala nominal propuesta por Saaty [58] (Tabla 2.3), siendo $\mathrm{a_{xy}}$ la valoración del criterio Y con respecto a X, el recíproco de la comparación es dado por $1/\mathrm{a_{xy}}$

Tabla 2.3. Escalas empleadas en la valoración.

Definición	Valor relativo	Valor recíproco
X tiene el mismo grado de preferencia que Y	1	1.00
X tiene un grado intermedio entre igual y moderada preferencia sobre Y	2	0.50
X tiene un grado moderado de preferencia sobre Y	3	0.33
X tiene un grado intermedio entre moderado y fuerte preferencia sobre Y	4	0.25
X tiene un grado fuerte de preferencia sobre Y	5	0.20
X tiene un grado intermedio entre fuerte y muy fuerte preferencia sobre Y	6	0.16
X tiene un grado muy fuerte de preferencia sobre Y	7	0.14
X tiene un grado intermedio entre muy fuerte y extremadamente fuerte preferencia sobre Y	8	0.12
X tiene un grado extremadamente fuerte de preferencia sobre Y	9	0.11

Para ilustrar este paso la matriz obtenida por el experto 1se muestra en la Figura 2.4:

Figura 2.4 Matriz de comparación por pares del experto.

$$M_C^1 = \begin{array}{c} \\ c_{11} \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{array} \begin{array}{c} \begin{array}{ccccc} c_1 & c_2 & c_3 & c_4 & c_5 \end{array} \\ \begin{pmatrix} 1 & 5 & 5 & 3 & 9 \\ 0.2 & 1 & 0.33 & 0.14 & 5 \\ 0.2 & 3 & 1 & 0.33 & 5 \\ 0.33 & 7 & 3 & 1 & 7 \\ 0.11 & 0.2 & 0.2 & 0.14 & 1 \end{pmatrix} \end{array}$$

Paso 3: Obtención de la matriz colectiva

Una vez que se han obtenido las matrices anteriores la matriz colectiva es obtenida por medio del operador OWA. El reordenamiento de los valores previo a la agregación de la matriz de peso empleada permite flexibilidad en el proceso. Este operador unifica los criterios clásicos de

decisión en incertidumbre en un solo marco de trabajo. Este operador puede ser definido de la forma siguiente:

Definición 2.1 [57] . Un operador OWA) es una función $\mathrm{F_{OWA}} : \mathbb{R}^{\mathrm{n}} \rightarrow \mathbb{R}$ de dimensión n si tiene un vector asociado W de dimensión n con $\mathrm{w_{ij}} \in [0,1]$ y $\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{w_j} = 1$, de forma tal que:

$$F_{OWA}(a_1, a_2, \dots, a_n) = \sum_{j=1}^{n} w_j b_j \qquad (2.2)$$

donde $\mathrm{b_j}$ es el j-ésimo más grande de los $\mathrm{a_j}$.

De este modo se obtiene la matriz colectiva $\bar{\mathrm{M}}_{\mathrm{c}}$ como,

$$\bar{\mathrm{M}}_{\mathrm{C}} = F_{OWA}\left(\mathrm{M}_{\mathrm{C}}^{1}, \dots, \mathrm{M}_{\mathrm{C}}^{\mathrm{k}}\right) \text{ con } \{\mathrm{k} = 1,2, \dots ,6\} \qquad (2.3)$$

En este caso se escogió un vector de pesos W que ponderará con un mayor peso los valores centrales W= [0.1, 0.2, 0.4, 0.2, 0.1] para disminuir el efecto de las valoraciones extremas de los expertos en el resultado final. El resultado de la agregación (2.3) se muestra en la Figura 2.5.

Figura 2.5. Matriz colectiva.

$$\bar{M}_c = \begin{array}{c} \\ c_{11} \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{array} \begin{array}{c} \begin{array}{ccccc} c_1 & c_2 & c_3 & c_4 & c_5 \end{array} \\ \begin{pmatrix} 1 & 5 & 5 & 3 & 9 \\ 0.2 & 1 & 0.33 & 0.14 & 5 \\ 0.2 & 3 & 1 & 0.33 & 5 \\ 0.33 & 7 & 3 & 1 & 7 \\ 0.11 & 0.2 & 0.2 & 0.14 & 1 \end{pmatrix} \end{array}$$

Paso 4: Calcular la relevancia

En este paso se calcula la relevancia de cada criterio $\mathrm{p_{C_i}}$. Inicialmente se calcula la suma de de cada columna de la matriz. Este cálculo permite posteriormente la normalización de la matriz.

$$\bar{M}_C = \begin{array}{c} \\ C_1 \\ \vdots \\ C_5 \end{array} \begin{array}{c} \begin{array}{ccc} C_1 & \dots & C_5 \end{array} \\ \begin{pmatrix} a_{11} & \dots & a_{15} \\ \vdots & \dots & \vdots \\ a_{51} & \dots & a_{55} \end{pmatrix} \end{array} \qquad v_{C_i} = \sum_{j=1}^{5} a_{ji} \qquad (2.4)$$

Se normaliza cada comparación por pares de la siguiente forma:

$$\bar{M}_{C-Norm} = \begin{matrix} \\ C_1 \\ \vdots \\ C_5 \end{matrix} \begin{matrix} \begin{matrix} C_1 & \dots & C_5 \end{matrix} \\ \begin{pmatrix} \frac{a_{11}}{v_{c_1}} & \dots & \frac{a_{15}}{v_{c_5}} \\ \vdots & \dots & \vdots \\ \frac{a_{51}}{v_{c_1}} & \dots & \frac{a_{55}}{v_{c_5}} \end{pmatrix} \end{matrix} \qquad (2.5)$$

donde v_{C_i} es obtenido mediante la ecuación 2.4

Calcular la relevancia (peso) de cada criterio p_{C_i} de la siguiente forma:

$$p_c = \begin{pmatrix} \frac{1}{5}\sum_{i=1}^{i=5} a_{1i} \\ \vdots \\ \vdots \\ \frac{1}{5}\sum_{i=1}^{i=5} a_{5i} \end{pmatrix}; \; p = \begin{pmatrix} p_{C_1} \\ \vdots \\ \vdots \\ p_5 \end{pmatrix} \qquad (2.6)$$

A continuación se muestran los resultados del proceso descrito anteriormente.

$$p_c = \begin{pmatrix} 0.45_{c_1} \\ 0.09_{c_2} \\ 0.014_{c_3} \\ 0.27_{c_4} \\ 0.05_{c_5} \end{pmatrix}$$

Figura 2.6 Pesos asignados a los criterios.

Con este resultado se aporta una ponderación diferente para los criterios de selección que permitirá realizar una mejor selección. El resultado muestra que el criterio de pago de licencia es catalogado como el de mayor importancia a la hora de la selección. El resultado muestra que los expertos le asignan una mayor importancia al criterio pago de licencia. Estos criterios con sus respectivos pesos fortalecen y facilitan las actividades implicadas.

Conclusiones

En este capítulo fue descrito el marco de trabajo mediante sus 9 actividades, plasmando una breve descripción de cada una. Se definieron un conjunto de criterios con sus pesos respetivos para la selección de herramientas de representación genealógica. Se definieron un conjunto de criterios para la selección de la herramienta de análisis de datos.

Capítulo 3 Validación del marco de trabajo para el análisis de la información genealógica.

Introducción

La validación de toda investigación resulta de gran interés. Mediante la misma se demuestra si es válida o no la propuesta. En el presente capítulo se realiza una primera evaluación del marco del trabajo mediante un grupo de expertos. Luego se realiza un estudio de caso que permite evaluar por un grupo de especialistas futuros usuarios de la propuesta de manera integral.

3.1 Evaluación de la flexibilidad y la integración en el marco de trabajo.

La hipótesis planteada en la investigación: Si se desarrolla un marco de trabajo flexible que permita realizar de manera integrada la captura y análisis de la información genealógica, entonces disminuirá el esfuerzo en el procesamiento de la información genealógica; es del tipo causal multivariada, estableciendo la relación que se muestra en la figura 3.1.

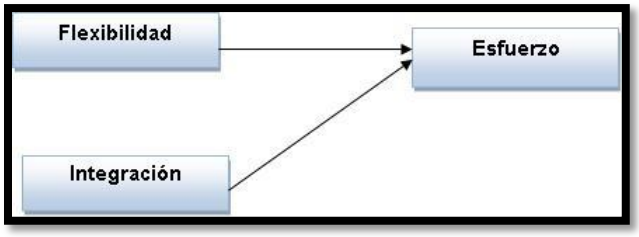


Figura 3.1 Relación entre variables.

La definición conceptual del “nivel de flexibilidad” tiene en cuenta los siguientes aspectos recogidos en la bibliografía [59, 60]:

- Adaptabilidad a las características particulares del entorno de aplicación.
- Facilidad para su aplicación en distintos escenarios.
- Independencia de dominios o sistemas específicos.
- Capacidad para incluir adecuadamente las preferencias del decisor.
- Posibilidad de incluir múltiples criterios de diferente naturaleza.

La definición conceptual del “nivel de integración” se basa en los siguientes principios según plantea:

- Capacidad de proveer un mecanismo para compartir tanto métodos como datos.
- Permitir mover la información fuera de los sistemas que eran difíciles de acceder previo a la integración.
- Tener acceso a datos a través de funcionalidades existentes.
- Implementación de un proceso de negocio común.
- Mantener independencia y desacoplamiento entre las herramientas a integrar.

Se parte de un método para la evaluación basado en [61-63] y que plantea las siguientes actividades:

1. Establecer marco evaluación: Se seleccionan los expertos y criterios a evaluar. El marco de trabajo queda definido de la siguiente forma:
   1.1. $C=\{c_1, c_2, c_3, c_4\}$ los criterios a ser evaluados y que se definen en la Tabla 3.1

Tabla 3.1 Criterios seleccionado para al evaluación.

Variable	Criterios	Nombre	Descripción
Flexibilidad	**C_1**	Adaptabilidad	Adaptabilidad a las características particulares del entorno de aplicación.
	C_2	Generalidad	Facilidad para su aplicación en distintos escenarios.
	C_3	Independencia	Independencia de dominios o sistemas específicos.
	C_4	Capacidad para representar las preferencias de los decisores	Capacidad para incluir adecuadamente las preferencias del decisor. Capacidad para modificar los criterios de selección y sus pesos. Posibilidad de incluir múltiples criterios de diferente naturaleza.
Integración	**C_5**	Autonomía	Desacoplamiento entre las herramientas a integrar. Garantizar la independencia de las aplicaciones una vez desacopladas.
	C_6	Capacidad para gestionar la integración	Compatibilidad de las interfaces. Claridad en el flujo de la integración.
	C_7	Facilidad en el acceso a los datos	Mueven la información fuera de los sistemas que eran difíciles de acceder. Acceso a la lógica de negocio.

			Acceso a datos a través de funcionalidades existentes.

1.2. Seleccionar una pequeña cantidad de expertos exagera el papel de cada uno de ellos, y cuando es muy grande, resulta un tanto difícil lograr una opinión concordante. Son diferente las formas de encontrar el número de expertos deseado, como por ejemplo [64]:

- Gráfico de Dalkay (Anexo 5).
- Ley de probabilidad binomial utilizando la expresión:

$$m = \frac{p * (1 - p) * k}{i^2}$$

Donde:

m: Número de expertos.

p: Proporción estimada de errores de los expertos.

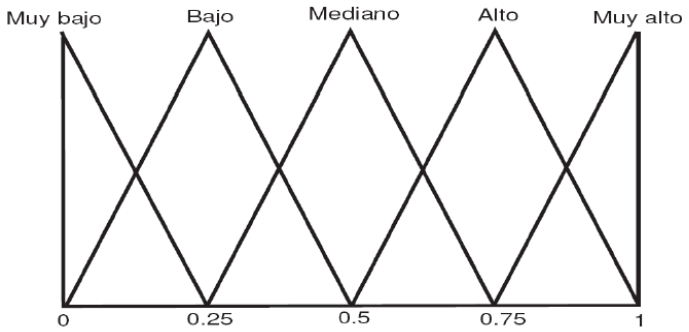
i: Nivel de precisión deseado

k: Constante asociada al nivel de confianza seleccionado.

En la investigación se adoptó el criterio de Dalkay, que con un margen de error propuesto de 0,1, recomendado para investigaciones de este tipo arrojó que se debían y utilizar 10 expertos. De esta manera obtenemos que E= $\{e_1, e_2, \ldots, e_{10}\}$ es el conjunto de 10 expertos que evalúan la flexibilidad, mientras que E= $\{e_{11}, e_{12}, \ldots, e_{20}\}$ son lo que evalúan la integración.

2. Recoger la información: Se obtiene la valoración lingüística de los expertos con respecto a los criterios utilizando un conjunto de cinco términos lingüísticos, con representación y semántica como la que se muestra en la figura 3.2.

Figura 3.2 Conjunto de términos lingüísticos



La semántica de lo términos es dada por funciones de pertenencia triangulares:

- Muy Alto (s_4)= (0.75, 1,1)
- Alto (s_3)= (0.5, 0.75,1)
- Mediano (s_2)= (0.25, 0.5, 0.75)
- Bajo (s_1)= (0, 0.25, 0.5)
- Muy bajo (s_0)= (0,0,0.25)

Las valoraciones de cada experto para cada criterio de muestran en las tablas 3.2 y 3.3

Tabla 3.2 Valoraciones de los criterios de “flexibilidad” dadas por los expertos.

Variable	**Expertos**										
Flexibilidad		1	2	3	4	5	6	7	8	9	10
	1	MA	M	A	MA	MA	B	A	MA	A	MA
	2	M	MA	B	A	A	MA	MA	MA	MA	MA
	3	A	B	M	MA	MA	A	MA	M	A	A
	4	A	MA	A	M	MA	MA	MA	A	MA	MA

Tabla 3.3 Valoraciones de los criterios de “integración” dadas por los expertos.

Variable	**Expertos**										
Integración		11	12	13	14	15	16	17	18	19	20
	5	A	MB	A	MB	A	MB	A	MB	MB	MB
	6	M	A	M	A	M	MB	A	MB	A	MB
	7	A	MB	A	MB	A	A	MB	MB	MB	A

3. Agregar la información. La agregación de las valoraciones se realiza en el siguiente orden:
   3.1. Se calcula el valor final de cada criterio mediante la agregación de la valoración dada por cada experto.
   3.2. Se agrega y se obtiene la evaluación final de la flexibilidad a partir de la agregación del valor obtenido anteriormente para cada criterio.
   3.3. Para la agregación de la información se propone la utilización del operador Media Aritmética sobre 2-tuplas (2-TMA) [65].

Definición 3.1. Siendo $A = \{(r_1, \alpha_1) \dots, (r_m, \alpha_m)\}$ un conjunto de 2-tuplas lingüísticas, la 2-tupla que simboliza la media aritmética, (2-TMA, se calcula de la siguiente forma

$$MA\big((r_1,\alpha_1)\ldots,(r_m,\alpha_m)\big) = \Delta\big(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\beta_i\big)$$

donde β_i es $\Delta^{-1}\big((r_i,\alpha_i)\big)$

En las tablas 3.4 y 3.5 se aprecian las valoraciones colectivas de cada uno de los criterios.

Tabla 3.4. Valoración colectiva para cada criterio de flexibilidad.

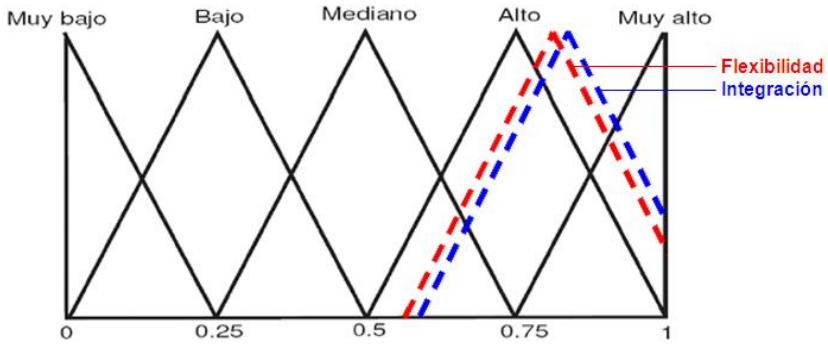
Criterio	Valor
1	(Alto,0.2)
2	(Alto,0.3)
3	(Alto,0)
4	(Muy Alto,-0.5)

Tabla 3.5. Valoración colectiva para cada criterio de integración.

Criterio	Valor
5	(Alto, 0.4)
6	(Alto, 0)
7	(Muy Alto, -0.4)

Como se puede apreciar todos los criterios fueron evaluados de alto a excepción de los criterios 4 y 7 evaluados de muy alto. La evaluación final del "nivel de flexibilidad" es (Alta, 0.25) y del "nivel de integración" es (Alta, 0.33) por lo que se puede afirmar que la evaluación dada por los expertos en ambos casos es satisfactoria. La representación de los términos lingüísticos obtenidos en la escala original, se muestra en la figura 3.3.

Figura 3.3. Términos lingüísticos para "Flexibilidad" e "Integración".



3.2 Resultados del Estudio de Caso.

El estudio de caso [66] es un método de investigación empírico ampliamente utilizado en la informática para la valoración de los resultados de la investigación [67-70]. La estrategia de

investigación de estudio de caso es la más adecuada cuando se plantea en el estudio una pregunta de investigación donde su forma es del tipo “¿cómo?” y “¿porqué?” y/o se tiene poco o nulo grado de control sobre el comportamiento real de los objetos/eventos en estudio [71].

El estudio de caso representa una herramienta muy útil de hacer investigación, ya que permite tener como resultado un enfoque holístico de una situación o evento en estudio, lo cual concede al investigador un abanico muy amplio de posibilidades para abordar un problema de investigación [71]. El propósito de los estudios de casos es comprender la interacción entre las distintas partes de un sistema y de las características importantes del mismo, para que este análisis pueda ser aplicado de manera genérica, incluso a partir de un único caso [72].

A continuación se describe un estudio de caso para evaluar el funcionamiento de la propuesta y obtener información tanto cuantitativa como cualitativa que contribuya a la validación de la propuesta, desarrollando todas las actividades que el mismo propone. Seguidamente una descripción del desarrollo de cada una de las etapas con sus actividades:

Primera etapa:

1. Seleccionar la herramienta para la representación genealógica:

En esta actividad se recomiendan cinco criterios con sus respectivos pesos para seleccionar la herramienta aplicando el método AHP. La misma dio como resultado para este estudio de caso que alasARBOGEN es la herramienta con mayores condiciones para el desarrollo del marco de trabajo. Seguidamente una muestra de la aplicación del método AHP.

Los criterios en este caso los constituyen las principales herramientas para la representación genealógica estudiadas en el capítulo 1 y los criterios, los propuestos en el marco de trabajo.

Criterios.

- Pago de licencia.(P1C1)
- Persistencia de los datos mediante la utilización de BD.(P1C2)
- Cubrimiento de las funcionalidades necesarias.(P1C3)
- Garantía de la seguridad y fiabilidad de los datos.(P1C4)
- Estructura de datos para la representación.(P1C5)

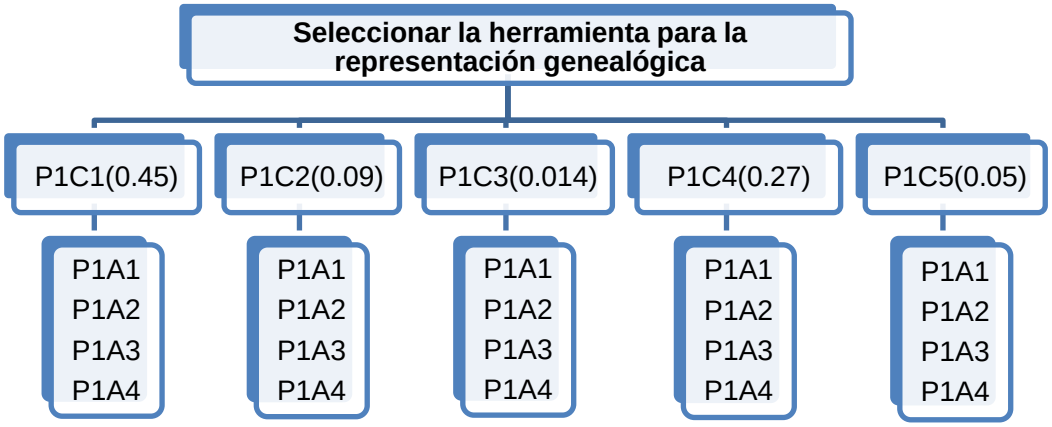
Alternativas

- Progeny Clinical (P1A1)

- Cyrillic. (P1A2)
- Madeline 2.0 PDE. (P1A3)
- alasARBOGEN. (P1A4)

A continuación se muestra la Figura 3.4 que muestra la estructura jerárquica del proceso de selección de la herramienta para la representación genealógica.

Figura 3.4. Estructura jerárquica para la selección de la herramienta para la representación genealógica.



El próximo paso es obtener las matrices de comparación por pares de alternativas para cada uno de los criterios. Se muestra en la tabla 3.6 la del primer criterio y el resto se encuentran en el Anexo 1.

Tabla 1.6. Matriz de comparación por pares de alternativas respecto al criterio Pago de licencia.

	P1A1	P1A2	P1A3	P1A4	Ponderación
P1A1	1	3	0,33	0,33	0,153
P1A2	0,33	1	0,2	0,2	0,069
P1A3	3	5	1	1	0,389
P1A4	3	5	1	1	0,389
	7,33	14	2,53	2,53	1,000

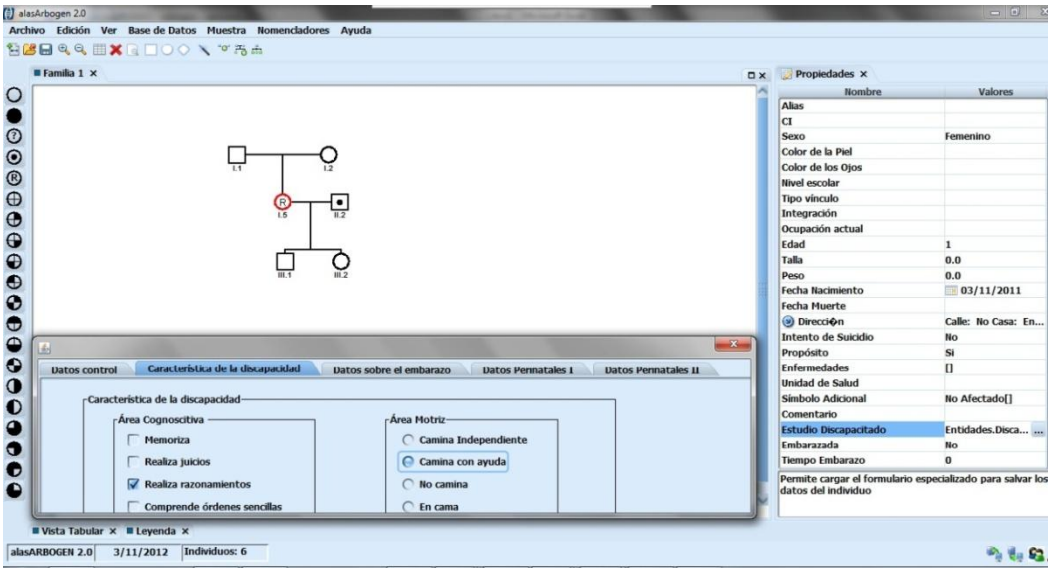
Finalmente se obtuvo la matriz (Anexo 1) resultado de la aplicación de AHP. Se evidencia que la mejor alternativa a utilizar como herramienta para la representación genealógica es alasARBOGEN, se muestra en los anexos un resumen de los principales resultados obtenidos. Con esto concluye la primera actividad del marco de trabajo.

2. Registrar la información genealógica del paciente:

La tarea de analizar información se ha venido desarrollando indistintamente sin contar con el resultado de la presente investigación. Esto permite tener referencia de algunos criterios cuantificables que permiten, a partir de este estudio de caso, comparar con resultados anteriores para poder evaluar si se mejora el proceso respecto a estos criterios. Es por esto que se decide utilizar una sesión del instrumento que se empleó en un estudio realizado entre los años 2009 y 2010 sobre las Personas con Discapacidad en la República del Ecuador. Se selecciona este, puesto que los especialistas cuentan con una referencia de un análisis similar que se realizó con los resultados obtenidos del mismo estudio en Venezuela con un instrumento similar. Se toma la misma cantidad de personas y número de variables a analizar, para lograr un problema similar y hacer más fiel su comparación.

En este estudio de Ecuador se pesquisaron alrededor de 394 mil personas. Fueron seleccionadas 87 variables del instrumento original. Estas abarcan un resumen que caracteriza a una persona, priorizando la parte clínico genética. Para una mayor comprensión siguiendo la lógica del estudio se agrupan en 4 secciones: Datos de Control, Características de la discapacidad, Datos sobre el embarazo de la madre y Datos perinatales. A continuación se muestra (Figura ·3.5) una pantalla de la aplicación alasARBOGEN registrando los datos.

Figura 3.5 Herramienta alasARBOGEN



Teniendo en cuenta los rangos de valores y algunos parámetros estadísticos publicados como resultados de este estudio se generó una población de cien mil personas que representa el 25

% del total captados en el estudio originalmente. Para la generación de los datos se empleó la herramienta Data Generator [73] que permite poblar una BD en poco tiempo. Esta herramienta permite configurar los valores que pueden tomar los atributos.

De esta manera concluye la primera etapa del marco de trabajo con los datos registrados en la BD mediante la herramienta seleccionada, para la representación genealógica.

Segunda etapa:

3. Definir el objetivo del análisis.

El objetivo que se propone es caracterizar las personas con discapacidad. Se quieren obtener patrones de comportamiento, similitudes, grupos que compartan características entre otras informaciones de interés que describan la población captada.

4. Definir la técnica de análisis de datos a utilizar.

Partiendo del objetivo de análisis que se quiere con los datos que se tienen se evidencia la necesidad de un análisis descriptivo. Dentro de las técnicas más utilizadas para este tipo de análisis se encuentran los clúster que satisfacen los objetivos propuestos permitiendo además obtener las variables dominantes en la población.

5. Seleccionar la herramienta para el análisis de los datos

En esta actividad se proponen 3 criterios para seleccionar la herramienta aplicando AHP. La misma dio como resultado que Weka es la herramienta que cumple con mayores condiciones para la ejecución del marco de trabajo. Seguidamente una muestra de la aplicación del método AHP para su selección.

Las alternativas en este caso los constituyen las principales herramientas para la análisis de datos estudiadas en el capítulo 1 y los criterios, los propuestos en el marco de trabajo.

Criterios.

- Pago de licencias.(P5C1)
- Implementación de los algoritmos necesarios para el análisis.(P5C2)
- Capacidad de integración con la herramienta de representación.(P5C3)

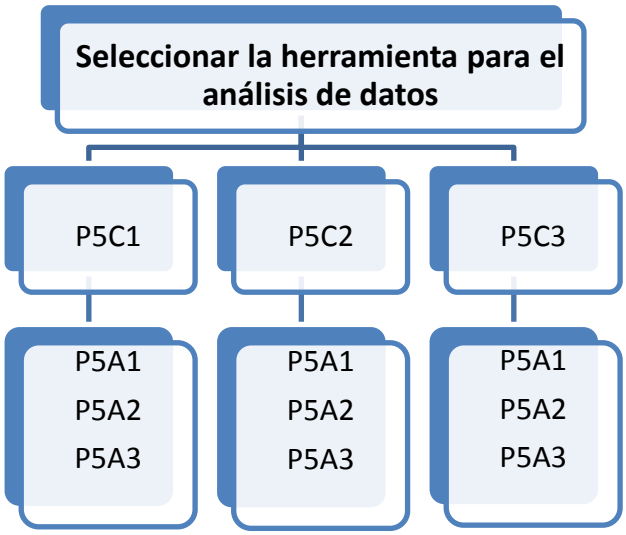
Alternativas

- Clementine / SPSS.(P5A1)
- R Project.(P5A2)

- Weka.(P5A3)

A continuación se muestra la Figura 3.6 que muestra la estructura jerárquica del proceso de selección de la herramienta para el análisis de datos.

Figura 3.6 Herramienta alasARBOGEN



Teniendo en cuenta los criterios de selección se elabora la matriz de comparación por pares y se determina la ponderación que toma cada criterio como se muestra en la Tabla 3.7. Como resultado se evidencia que los criterios de mayor ponderación son el pago de licencia y la seguridad y fiabilidad de los datos respectivamente.

Tabla 3.7 Matriz de comparación por pares de criterios

	P5C1	P5C2	P5C3	Ponderación
P5C1	1	3	3	0,575
P5C2	0,33	1	0,33	0,139
P5C3	0,33	3	1	0,286

El próximo paso es realizar las matrices de comparación por pares de alternativas para cada uno de los criterios. Se muestra en la tabla 3.8 respecto al tercer criterio que fue donde se evidenciaron las mayores diferencias en las ponderaciones por alternativas. Es importante destacar que el segundo criterio no dio prioridad a ninguna de las herramientas. El resto de las matrices se encuentran en el anexo 2.

Tabla 3.8 Matriz de comparación por pares de alternativas respecto al criterio Capacidad de integración con la herramienta de representación.

	P5A1	P5A2	P5A3	Ponderación
P5A1	1	0,33	0,14	0,087
P5A2	3	1	0,33	0,243
P5A3	7	3	1	0,670

Resultado de la aplicación de AHP se evidencia que la mejor alternativa a utilizar como herramienta para el análisis de la información genealógica es Weka, terminando de esta manera la quinta actividad del marco de trabajo.

6. Integrar ambas herramientas.

La integración se realiza tal y como plantea el marco de trabajo a nivel de aplicación. Estar ambas herramientas desarrolladas sobre una misma tecnología Java, facilitó su integración. Otro elemento que facilitó la integración fue poder contar con el código de las dos herramientas. La integración se realizó recompilando el código necesario de Weka desde alasARBOGEN. Con esto se logró utilizar las funcionalidades de Weka. La integración permitió que desde el código de Weka se pudiera acceder a la BD donde estaba almacenada toda la información de los pacientes, almacenada en una BD en PostgreSQL mediante una conexión JDBC. Concluye esta actividad con la integración exitosa de ambas herramientas.

Tercera etapa:

7. Selección y limpieza de los datos a utilizar.

Para la selección de los datos, se realizó un análisis de las variables recogidas de la persona, quedando solamente 29 variables. Entre las principales causas por las que se desecharon las variables se encuentran:

- Campos abiertos a comentarios.
- Fechas.
- Algunos datos de control como dirección.
- Datos de los especialistas que realizaron el estudio.

No se desarrollaron tareas de limpieza de datos teniendo en cuenta que los mismos fueron generados automáticamente con sus respectivos rangos de valores y sin permitir espacios vacios.

8. Aplicar la técnica de análisis seleccionada.

La aplicación para el análisis fue ejecutada en una PC con microprocesador Core i3 y 4 GB de RAM, sobre una la distribución 10.10 de Ubunto. Se aplicaron 2 algoritmos, SimpleKmeans y EM.

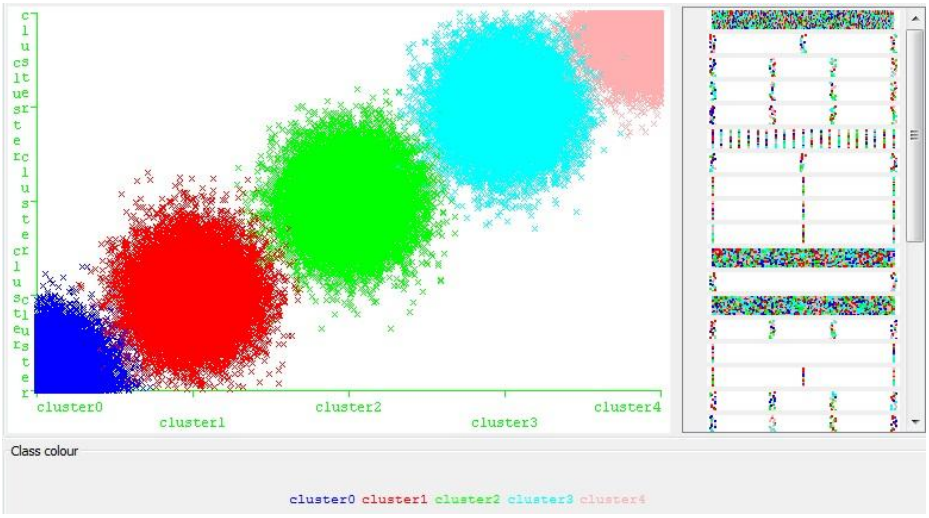
SimpleKmeans: Este algoritmo se basa en agrupar por vecindad. Entre los métodos de partición es de los más populares. Se basa en situar los prototipos en el espacio y calculando la distancia. Generalmente se utiliza la euclidiana y ubicando al resto de los elementos según sea menor el resultado de la distancia con respecto a los prototipos. Se configuró con K=5 para obtener esta cantidad de agrupamientos, a continuación una imagen en la Figura 3.7 que visualiza algunos de los resultados del mismo: no se referencia la tabla

Tabla 3.9 Cantidad de personas agrupadas.

Clúster	Cantidad de Personas
1	15332
2	21276
3	20215
4	23012
5	20165

Un elemento a tener en cuenta es el valor de k. No se aconseja que sea un número muy alto ya que de esto dependen los agrupamientos y si se fraccionan inadecuadamente puede dividir grupos de forma artificial, fraccionando grupos reales. Este método no es suficientemente eficiente cuando los puntos de un grupo están muy cercanos del centroide de otro.

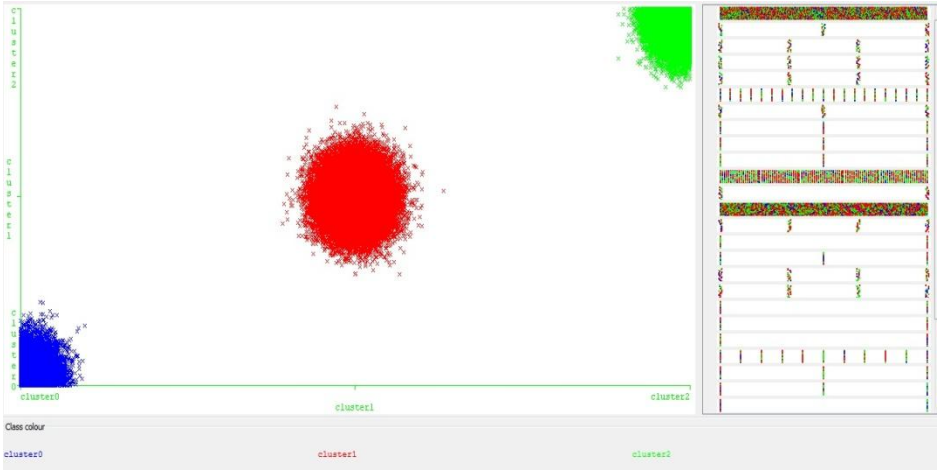
Figura 3.7 Vista del clúster con K=5.



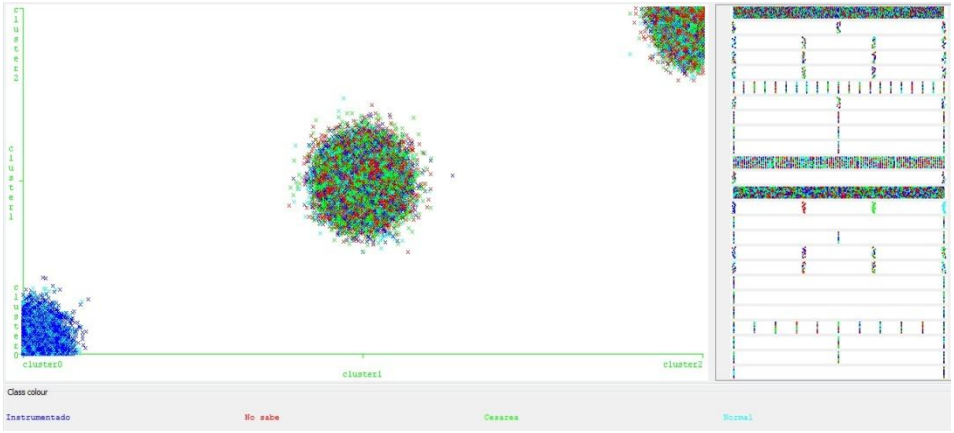
EM: En este caso el algoritmo funciona semejante al anterior pero él se encarga de determinar cuál es la cantidad óptima de clúster a partir de las iteraciones definidas. En este caso se configuró para que trabajara con 100 iteraciones con K=-1. El algoritmo arrojó como resultado tres clúster. A continuación algunas imágenes que muestran los resultados obtenidos. Figura 3.8.

Figura 3.8 Resultado de la aplicación del clúster con el algoritmo EM.

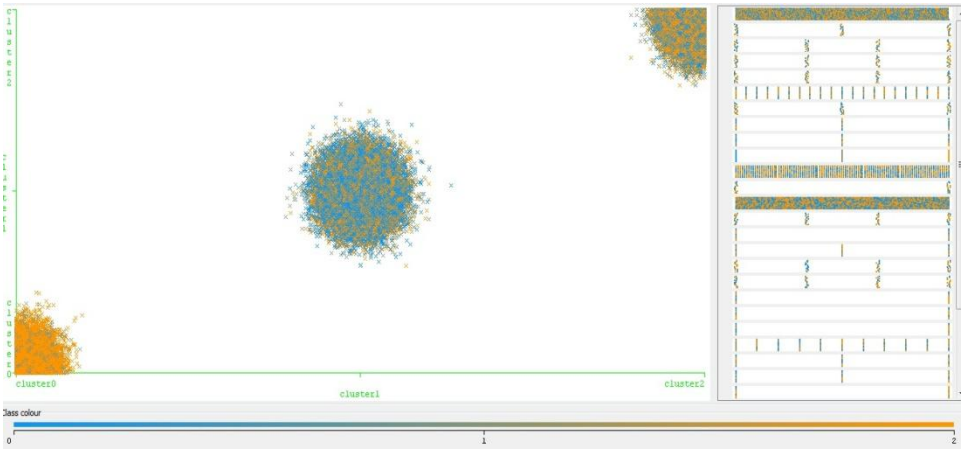
A



B



C



Obtención de los resultados.

Este es el último paso del marco de trabajo y es el encargado de elaborar el informe con los resultados del análisis. Es muy importante hacer una buena descripción del proceso para que los especialistas posean todos los elementos a la hora de interpretar los resultados. A continuación se muestra un resumen de los elementos que debe mostrar el informe:

- Objetivo del análisis: Encontrar similitudes entre las personas con discapacidad.
- Población estudiada: Personas con discapacidad.
- Muestra seleccionada: 100 mil personas para un 25%.
- Variables de análisis: Ver anexo 16.

- Caracterización de la representación gráfica de la información. Aquí se realiza una descripción de los principales clúster obtenidos para ayudar a un mayor entendimiento por parte de los especialistas.

Conclusiones del estudio de caso:
Para evaluar si la aplicación de esta propuesta contribuye a reducir el esfuerzo en el proceso, se realiza un análisis del nivel de integración y el esfuerzo empleado para este estudio de caso. Con este resultado se realiza un análisis respecto a como se hacía sin este marco de trabajo comparando con un estudio similar desarrollado anteriormente en Venezuela.

Se define *Nivel de Integración* (N_i) como:

$$N_i = \frac{H_i}{H_t} \qquad (3.1)$$

donde H_i significa la cantidad de herramientas informáticas que soportan el marco de trabajo y H_t el total de aplicaciones que intervienen en el maro.

Por otra parte se define Esfuerzo (E) como:

$$E = \frac{t}{P} \qquad (3.2)$$

donde t significa el tiempo total expresado en horas dedicadas al proceso de análisis y P la cantidad de personas que intervinieron.

Según datos aportados por especialistas en un análisis anterior que se organizó con una muestra de los resultados obtenidos en el mismo estudio en Venezuela los resultados fueron:

Total de herramientas empleadas: 2

Total de herramientas integradas: 0

Tiempo total desde la concepción hasta obtener el resultado: 425 horas.

Cantidad de personas empleadas: 5

En el estudio de caso resultado de la presente investigación los resultados fueron:

Total de herramientas empleadas: 2

Total de herramientas integradas: 2

Tiempo total desde la concepción hasta obtener el resultado: 167 horas.

Cantidad de personas empleadas: 4

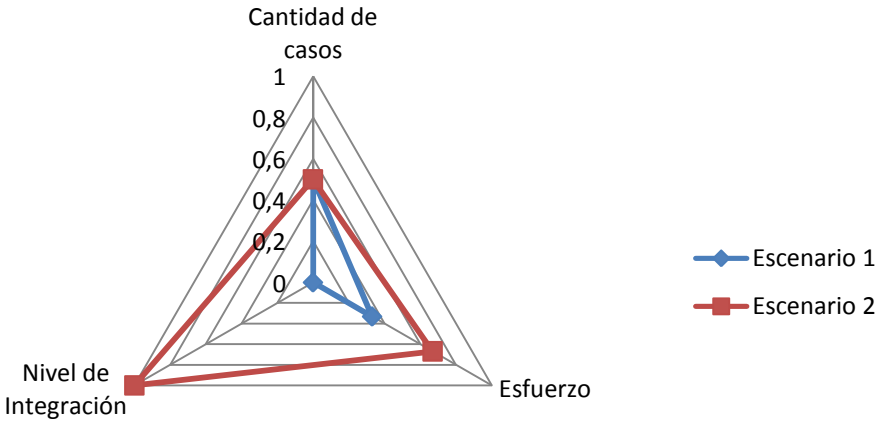
Aplicado las definiciones 3.1 y 3.2 obtenemos: referencia tabla y poner número y nombre a la tabla

Tabla 3.10 Análisis de los escenarios.

Criterio	Escenarios		% del resultado en el segundo escenario respecto al primero.
	1-Sin el marco de trabajo	2-Con el marco de trabajo	
N_i	0	1	100
E	85	42	49

Aunque se evidencia la mejora del segundo escenario respecto al primero para una mayor comprensión se decide realizar un gráfico radial que visualiza los resultados, para esto se normalizan los valores [74]. El resultado se muestra en la Figura 3.9:

Figura 3.9. Análisis de los escenarios.



Conclusiones

La aplicación del método de experto permitió evaluar a partir de la opinión de estos la flexibilidad del marco de trabajo. El marco de trabajo propuesto funciona y cumple sus objetivos, evidenciándose en el estudios de casos que se desarrolló. La comparación a partir del nivel de integración y el esfuerzo de cómo de realizaban el proceso de análisis de información genealógica antes de la propuesta y después de la misma demostró que la propuesta es factible. Además se integraron las herramientas alasARBOGEN y Weka para validar el marco de trabajo.

Conclusiones Generales

- La evaluación de las tendencias actuales en el análisis de la información genealógica demostró la necesidad de un marco de trabajo que permite disminuir los esfuerzos en el proceso.
- Se desarrolló un marco de trabajo que permite la captura y análisis de la información genealógica a partir de la definición de los criterios para la selección de herramientas de representación genealógica y análisis de datos.
- Se validó la propuesta mediante el método de expertos y estudio de caso, pudiendo evidenciar una mejora en la flexibilidad y el nivel de integración, lo cual permitió comprobar la aplicabilidad del marco de trabajo desarrollado y sus efectos en la disminución del esfuerzo empleado en el análisis de la información genealógica.

Recomendaciones

- Aplicar el marco de trabajo propuesto a nuevos casos que permitan obtener retroalimentación a partir de los resultados y opiniones de los usuarios.
- Incluir en la etapa de integración el análisis de sensibilidad de manera que permita una selección más robusta de las herramientas de análisis de datos.

Referencias bibliográficas

[1] G. P. Copenhaver*, et al.*, "Genetic definition and sequence analysis of Arabidopsis centromeres," *Science,* vol. 286, pp. 2468-2474, 1999.

[2] B. M. Teruel. (2009) El Programa Nacional de Diagnóstico, Manejo y Prevención de Enfermedades Genéticas y Defectos Congénitos de Cuba, 1981-2009. *Revista Cubana de Genética Cumunitaria* Available: http://bvs.sld.cu/revistas/rcgc/v3n2_3/rcgc1623010%20esp.htm

[3] B. M. Teruel. (2008) Genética Médica y Enfermedades Crónicas: el camino de la Prevención. *Revista Cubana de Genética Comunitaria*.

[4] B. M. Teruel, "El Programa Nacional de Diagnóstico, Manejo y Prevención de Enfermedades Genéticas y Defectos Congénitos de Cuba: 1981-2009."

[5] A. G. Álvarez Pérez*, et al.*, "Voluntad política y acción intersectorial: Premisas clave para la determinación social de la salud en Cuba," *Revista cubana de higiene y epidemiología,* vol. 45, pp. 0-0, 2007.

[6] M. G. Guzmán*, et al.*, "Enfermedades virales emergentes," *Rev Cubana Med Trop,* vol. 53, pp. 5-15, 2001.

[7] B. M. TERUEL, "Genética médica y enfermedades crónicas: el camino de la prevención," *Rev Cubana Genet Comunit,* vol. 2, pp. 3-4, 2008.

[8] M. H. Chen and Q. Yang, "GWAF: an R package for genome-wide association analyses with family data," *Bioinformatics,* vol. 26, pp. 580-581, 2010.

[9] A. C. Anne E. Egger. (2008) Datos: Análisis e interpretación. *Visionlearning*. Available: http://www.visionlearning.com/library/module_viewer.php?mid=154&l=s

[10] "Babylon. Arbol Genealógico," ed.

[11] "Atención Primaria," 2007.

[12] R. L. Bennett*, et al.*, "Standardized human pedigree nomenclature: update and assessment of the recommendations of the National Society of Genetic Counselors," *Journal of Genetic Counseling,* vol. 17, pp. 424-433, 2008.

[13] R. L. Bennett*, et al.*, "Recommendations for standardized human pedigree nomenclature," *Journal of Genetic Counseling,* vol. 4, pp. 267-279, 1995.

[14] O. M. d. l. Salud, *Estadisticas Sanitarias Mundiales 2005*: World Health Organization, 2005.

[15] B. M. TERUEL, "Genética comunitaria: la principal prioridad para la genética médica en Cuba," *Rev Cubana Genet Comunit,* vol. 2, pp. 3-4, 2008.

[16] L. M. Solís and J. J. S. Baena, "Herramientas de análisis de información genealógica: Estudio y evaluación," *Biblios,* 2009.

[17] O. M. Llorente, "Componente visual para la representación de árboles genealógicos," UCI, La Habana, 2011.

[18] (2000). *CyrillicSoftware*. Available: http://www.cyrillicsoftware.com/

[19] Ò. Coltell and D. Corella, "ESPECIAL BIOINFORMÁTICA."

[20] R. L. Álvarez, "Métodos computacionales para la representación y análisis de árboles genealógicos," ed, 2011.

[21] C. Reilly, *Statistics in human genetics and molecular biology*, 2009.

[22] H. T. a. P. Nürnberg. (2005) Haplopainter: a tool for drawing pedigrees with complex haplotypes. *Bioinformatics*.

[23] *Progeny Genetics-lab*. Available: http://www.progenygenetics.com/lab/&usg=ALkJrhhgBFav1PhbEVZc6tVwT

[24] *Progeny Genetisc- Lims*. Available: http://www.progenygenetics.com/lims/&usg=ALkJrhgKLOVDXGOlMTgmLW6P

[25] A. L. Reynaldo, "SLD031-ARQUITECTURA PARA EL SISTEMA DE REPRESENTACIÓN DE ÁRBOLES GENEALÓGICOS ALASARBOGEN EN SU VERSIÓN 2.0," in *VIII Congreso Internacional de Informática en la Salud. II Congreso Moodle Salud*, 2010.

[26] J. C. Riquelme*, et al.*, "MINERÍA DE DATOS: CONCEPTOS Y TENDENCIAS Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, primavera, año/vol. 10, número 029 Asociación Española para la Inteligencia Artificial Valencia, España," *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial,* pp. 11-18, 2006.

[27] I. n. I. y. P. L. Abdelmalik Moujahid. (1996). *Introducción a la Miner´ıa de Datos*.

[28] Vallejos, "Conceptos del KDD," ed, 1996.

[29] L. M. E. A. Aguilera, "MINERÍA DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO."

[30] F. y. otros, 1996.

[31] A. T. Valero, "Extracción de Información con Algoritmos de Clasificación," 2005.

[32] M. J. R. Q. José Hernández Orallo, Cesar Ferri Ramírez, "Introducción a la Minería de datos."

[33] D. S. G. Cesar Pérez López, *Minería de datos. Técnicas y herramientas*, 2007.

[34] H. P. CABENA Peter, STADLER Rolf, VERHEES Jaap, ZANASI Alessandro, *Discovering data mining from concept to implementation*, 1998.

[35] P. Randall Matignon, "Data Mining Using SAS Enterprise Miner," ed, 2007.

[36] B. V. y. D. Smith, "Introducción a R," 2000.

[37] E. F. Mark Hall, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann. (2009). *The WEKA data mining software: an update*. *11*. Available: http://dl.acm.org/citation.cfm?id=1656278

[38] D. Pérez and M. Dressler, "Tecnologías de la información para la gestión del conocimiento," *Pérez, Daniel; Dressler, Matthias." Tecnologías de la información para la gestión del conocimiento". Intangible Capital, enero-marzo de 2007, vol. 3, núm. 15, p. 31-59.,* 2007.

[39] V. S. Wing Lam. (2007). *Enterprise Architecture and Integration*. Available: http://www.info-sci-ref.com

[40] D. W. Rodríguez, "Integración de Aplicaciones," presented at the Postgrado en Computación, 2009.

[41] R. Klischewski, "Information integration or process integration? How to achieve interoperability in administration," *Electronic Government,* pp. 57-65, 2004.

[42] E. E. Julieta Martínez, Enrique E. Tarifa, "de Métodos para Análisis Multicriterio: PROMETHEE y AHP. Parte I: Fundamentos Teóricos," 2009.

[43] T. L. Saaty and L. G. Vargas, "The Seven Pillars of Analytic Hierarchy Process Moddels, Methods, Concepts & Application of the Analytic Hierarchy Process," 2001.

[44] M. N. Yngrid Naime Velasquez, Carlos Rodríguez Monroy. (2010) Aplicación de la técnica AHP para evaluar el efecto de los valores organizacionales en la productividad.

[45] F. L. R. Sergio A. Berumen, "La utilidad de los métodos de decisión multicriterio (como el AHP) en un entorno de competitividad creciente.," 2010.

[46] E. E. Julieta Martínez, Enrique E. Tarifa. (2008). *Evaluación de Métodos para Análisis Multicriterio: PROMETHEE y AHP.*

[47] C. CHEN-TUNG*, et al.*, "Applying multiple linguistic PROMETHEE method for personnel evaluation and selection. Industrial Engineering and Engineering Management," 2009.

[48] C. HWANG and K. YOON, "Multiple attribute decision methods an applications.," 1981.

[49] S. A. Bernard, "The Coherency Architect," ed, 2010.

[50] R. Pressman, "Ingeniería de software," *ARQUITECTURA,* vol. 2, p. 2, 2005.

[51] C. P. Team, "CMMI for Development, version 1.2," 2006.

[52] Y. Naranjo*, et al.*, "VI Congreso Internacional de Informática en Salud Temática."

[53] J. Kontio*, et al.*, "The Focus Group Method as an Empirical Tool in Software Engineering " in *Guide to Advanced Empirical Software Engineering*, F. Shull*, et al.*, Eds., ed: Springer London, 2008, pp. 93-116.

[54] F. Atienza*, et al.*, "Encuentros difíciles en atención primaria: una aproximación mediante grupos focales," *Análisis y Modificación de Conducta,* vol. 37, 2012.

[55] E. Morales Lara, "Enfoque alternativo de construccion en trabajo social| Politicas publicas sobre los programas de violencia en la television," UNIVERSITY OF PUERTO RICO, RIO PIEDRAS (PUERTO RICO), 2010.

[56] F. J. Domínguez-Mayo*, et al.*, "Quality evaluation for Model-Driven Web Engineering methodologies," *Information and Software Technology,* vol. 54, pp. 1265-1282, 2012.

[57] R. R. Yager*, et al.*, *Recent Developments in the Ordered Weighted Averaging Operators: Theory and Practice*: Springer, 2011.

[58] T. Saaty, *The analytic hierarchy process.* New York: McGraw Hill, 1980.

[59] R. Sadiq*, et al.*, "Integrating indicators for performance assessment of small water utilities using ordered weighted averaging (OWA) operators," *Expert Systems with Applications,* vol. 37, pp. 4881-4891, 2010.

[60] G. Beliakov*, et al.*, "Aggregation of Preferences in Recommender Systems," in *Recommender Systems Handbook*, F. Ricci*, et al.*, Eds., ed: Springer US, 2011, pp. 705-734.

[61] M. Espinilla*, et al.*, "A 360-degree performance appraisal model dealing with heterogeneous information and dependent criteria," *Information Sciences,* 2012.

[62] M. Espinilla*, et al.*, "A heterogeneous evaluation model for assessing sustainable energy: A Belgian case study," in *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*, 2010, pp. 1-8.

[63] M. Espinilla Estévez, "Nuevos modelos de evaluación sensorial con información lingüística," DEA, Departamento de Informática, Universidad de Jaén, Jaen, 2009.

[64] Y. Z. Véliz, "Modelo de Gestión de Riesgo en Proyectos de Desarrollo de Software," UCI, 2007.

[65] F. Herrera and L. Martínez, "A 2-tuple fuzzy linguistic representation model for computing with words," *Fuzzy Systems, IEEE Transactions on,* vol. 8, pp. 746-752, 2000.

[66] R. K. Yin, *Case study research: Design and methods* vol. 5: Sage Publications, Incorporated, 2008.

[67] C. W. Ping, "A Methodology for Constructing Causal Knowledge Model from Fuzzy Cognitive Map to Bayesian Belief Network," PhD Thesis, Department of Computer Science, Chonnam National University, 2009.

[68] A. Singh, "Architecture value mapping: using fuzzy cognitive maps as a reasoning mechanism for multi-criteria conceptual design evaluation," PhD Thesis, Missouri University of Science and Technology, Missouri, 2011.

[69] J. Merigó, "New extensions to the OWA operators and its application in decision making," PhD Thesis, Department of Business Administration, University of Barcelona, 2008.

[70] M. A. Iqbal*, et al.*, "A New Requirement Prioritization Model for Market Driven Products Using Analytical Hierarchical Process," in *International Conference on Data Storage and Data Engineering*, 2010.

[71] J. E. Macluf*, et al.*, "El estudio de caso como estrategia de investigación en las ciencias sociales," *Ciencia Administrativa,* p. 7.

[72] S. M. B. Hernández., "CONCEPCIÓN METODOLÓGICA DE CÓMO DISEÑAR TAREAS PARA EL APRENDIZAJE DE LOS ESTUDIANTES EN LA UNIVERSIDAD AGRARIA DE LA HABANA Y SUS SEDES MUNICIPALES," En opción al grado científico de Doctor en Ciencias Pedagógicas, 2010.

[73] D. Garcia and M. Millan, "A prototype of synthetic data generator," in *Computing Congress (CCC), 2011 6th Colombian*, 2011, pp. 1-6.

[74] S. Opricovic and G.-H. Tzeng, "Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS," *European Journal of Operational Research,* vol. 156, pp. 445-455, 2004.

Anexos

Anexo 1. Descripción de AHP para la selección de la herramienta de representación genealógica.

Tabla 1. Matriz de comparación por pares de criterios.

	P1C1	P1C2	P1C3	P1C4	P1C5	Ponderación
P1C1	1	5	5	3	9	0,472
P1C2	0,2	1	0,33	0,14	5	0,084
P1C3	0,2	3	1	0,33	5	0,131
P1C4	0,33	7	3	1	7	0,280
P1C5	0,11	0,2	0,2	0,14	1	0,032

Tabla 2. Matriz de comparación por pares de alternativas respecto al criterio Pago de licencia.

	P1A1	P1A2	P1A3	P1A4	Ponderación
P1A1	1	3	0,33	0,33	0,153
P1A2	0,33	1	0,2	0,2	0,069
P1A3	3	5	1	1	0,389
P1A4	3	5	1	1	0,389
	7,33	14	2,53	2,53	1,000

Tabla 3. Matriz de comparación por pares de alternativas respecto al criterio Persistencia de los datos mediante la utilización de BD

	P1A1	P1A2	P1A3	P1A4	Ponderación
P1A1	1	0,33	5	0,33	0,164
P1A2	3	1	7	1	0,394
P1A3	0,22	0,14	1	0,14	0,048
P1A4	3	1	7	1	0,394
	7,22	2,47	20	2,47	1,000

Tabla 4. Matriz de comparación por pares de alternativas respecto al criterio Cubrimiento de las funcionalidades necesarias.

	P1A1	P1A2	P1A3	P1A4	Ponderación
P1A1	1	0,33	3	0,2	0,121
P1A2	3	1	5	0,33	0,262
P1A3	0,33	0,33	1	0,14	0,063
P1A4	5	3	7	1	0,554
	9,33	4,66	16	1,67	1,000

Tabla 5. Matriz de comparación por pares de alternativas respecto al criterio Garantía de la seguridad y fiabilidad de los datos

	P1A1	P1A2	P1A3	P1A4	Ponderación
P1A1	1	0,33	3	0,33	0,153
P1A2	3	1	5	1	0,389
P1A3	0,33	0,2	1	0,2	0,069
P1A4	3	1	5	1	0,389
	7,33	2,53	14	2,53	1,000

Tabla 6. Matriz de comparación por pares de alternativas respecto al criterio Estructura de datos para la representación

	P1A1	P1A2	P1A3	P1A4	Ponderación
P1A1	1	1	3	1	0,300
P1A2	1	1	3	1	0,300
P1A3	0,33	0,33	1	0,33	0,099
P1A4	1	1	3	1	0,300
	3,33	3,33	10	3,33	1,000

Tabla 7. Matriz de comparación por pares de criterios

	P5C1	P5C2	P5C3	Ponderación
P5C1	1	3	3	0,575
P5C2	0,33	1	0,33	0,139
P5C3	0,33	3	1	0,286

Anexo 2. Descripción de AHP para la selección de la herramienta de análisis de datos.

Tabla 1. Matriz de comparación por pares de alternativas respecto a la Capacidad de integración con la herramienta de representación

	P5A1	P5A2	P5A3	Ponderación
P5A1	1	0,33	0,14	0,087
P5A2	3	1	0,33	0,243
P5A3	7	3	1	0,670

Tabla 2. Matriz de comparación por pares de alternativas respecto al Pago de licencias

	P5A1	P5A2	P5A3	Ponderación
P5A1	1	0,2	0,2	0,091
P5A2	5	1	1	0,455
P5A3	5	1	1	0,455

Tabla 3. Matriz de comparación por pares de alternativas respecto a la Implementación de los algoritmos necesarios para el análisis

	P5A1	P5A2	P5A3	Ponderación
P5A1	1	1	1	0,333
P5A2	1	1	1	0,333
P5A3	1	1	1	0,333

Tabla 4. Análisis final de AHP para seleccionar las Alternativas.

	Puntuación de la alternativa por criterio					**Criterios**					
	P1C1	**P1C2**	**P1C3**	**P1C4**	**P1C5**	**P1C1**	**P1C2**	**P1C3**	**P1C4**	**P1C5**	**Resultado**
Alternativa	**a**	**b**	**c**	**d**	**e**	**c1=a*0,472**	**c2=b*0,084**	**c3=c*0,131**	**c4=d*0,280**	**c5=e*0,032**	**R=C1+C2…C5**
P1A1	0,153	0,164	0,121	0,153	0,3	0,072	0,014	0,016	0,043	0,010	0,15425
P1A2	0,069	0,394	0,262	0,389	0,3	0,032	0,033	0,034	0,109	0,010	0,21833
P1A3	0,389	0,048	0,063	0,069	0,099	0,184	0,004	0,008	0,019	0,003	0,21846
P1A4	0,389	0,394	0,554	0,389	0,3	0,184	0,033	0,073	0,109	0,010	0,40796

Tabla 5. Análisis final de AHP para seleccionar las Alternativas.

	Puntuación de la alternativa por criterio			**Criterios**			
	P5C1	**P5C2**	**P5C3**	**P5C1**	**P5C2**	**P5C3**	**Resultado**
Alternativa	**a**	**b**	**c**	**c1=a*0,575**	**c2=b*0,139**	**c3=c*0,286**	**R=C1+C2+C3**

P5A1	0,15289523	0,16392766	0,12131427	0,088	0,023	0,035	0,14540
P5A2	0,06863795	0,39380727	0,26156012	0,039	0,055	0,075	0,16901
P5A3	0,38923341	0,04845781	0,06312939	0,224	0,007	0,018	0,24860

Anexo 3. Variables utilizadas en el análisis del estudio de caso.

Tabla 1. Descripción de las variables utilizadas en el análisis del estudio de caso.

Variable	Descripción	Posibles valores
color_de_piel (Nom)	Describe el color de la piel en el individuo.	Blanco, Negro y Mestizo
color_de_ojo (Nom)	Describe el color de los ojos del individuo.	Negro, Verde , Azul y Marrón
area_autonomia_social (Nom)	Describe la autonomía en el ámbito social del individuo.	Independiente, Semidependiente y Dependiente
area_motriz (Nom)	Describe la motricidad del individuo.	Camina independiente, camina con ayuda, no camina y en cama
edad_madre (Num)	Describe la edad de la madre en el embarazo del individuo	Un valor numérico que representa la edad.
movimientos_fetales (Nom)	Describe los movimientos fetales de la madre durante el embarazo del individuo	Fuertes, débiles y No sabe
consanguinedad_padre (Num)	Describe la existencia de consanguinidad en los padres del individuo.	0 : No, 1:Si y 2: No sabe
dan_fetal_x_alochol (Num)	Describe la existencia de daño fetal por la ingestión de bebidas alcohólicas de la madre durante el embarazo del individuo	0 : No, 1:Si y 2: No sabe
dan_fetal_x_droga (Num)	Describe la existencia de daño fetal por el consumo de droga de la madre durante el embarazo del individuo	0 : No, 1:Si y 2: No sabe
peso (Num)	Describe el peso corporal del individuo.	Un valor numérico que indica el peso en kilogramos.
sex (Nom)	Describe el sexo del individuo.	M y F
talla (Num)	Describe la talla corporal del individuo	Un valor numérico que indica el tamaño en centímetros
tipo_parto (Nom)	Describe el tipo de parto de la madre del individuo.	Normal, Instrumentado, Cesárea y No sabe.
id_nevus (Num)	Describe la existencia de nevus (lunares) en el individuo.	0: No y 1: Si
id_psicosis_discapcidad (Num)	Describe la existencia de alguna psicosis en el individuo.	0: No y 1: Si
color_piel (Nom)	Describe el color de la piel del individuo al nacer.	Cianótico , Normal, Ictero intenso y No sabe
id_signos_dismorficos (Num)	Describe la existencia de signos dismorficos en el individuo	0: No y 1: Si

id_hemangiomas (Num)	Describe la existencia de hemangiomas en el individuo.	0: No y 1: Si
id_malformacion_congenita_ext (Num)	Describe la existencia de algún tipo de malformación congénita externa en el individuo	0: No y 1: Si
apgar (Num)	Describe el tiempo del apgar en el individuo al nacer	Valor numérico que va desde 0 hasta 10, donde el valor 0 indica que no sabe.
paralisis_cerebral (Num)	Describe la existencia de un diagnostico de parálisis cerebral en el individuo	0 : No, 1:Si y 2: No sabe
edad_gestacional (Num)	Describe la edad gestacional del parto de la madre del individuo	0: No sabe, 1: Pretérmino 2: A término
id_malformacion_congenita_int (Num)	Describe la existencia de algún tipo de malformación congénita interna en el individuo	0: No y 1: Si
id_enfermedades_infecciosas (Num)	Describe la existencia de enfermedades infecciosas en la madre del individuo durante el embarazo.	Un valor numérico que indica el tipo de enfermedad
id_exposicion_fuentes_calor (Num)	Describe si la madre durante el embarazo del individuo estuvo expuesta a fuentes de calor	0 : No, 1:Si y 2: No sabe
tratamiento_escaras (Num)	Describe si el individuo recibe tratamiento por las escaras	0: No y 1: Si
escaras (Num)	Describe si el individuo presenta escaras en la piel	0: No y 1: Si
grado_profundida_intelectual (Nom)	Describe el grado de la discapacidad intelectual que presenta el individuo	Profunda, Media y Ligera

Anexo 4. Selección de la cantidad de expertos.