UNIVERSIDAD DE VALPARAÍSO
Facultad de Ingeniería
Escuela de Ingeniería Civil Informática
Ingeniería Civil en Informática

Sistema de Minería de Datos para analizar casos de cánceres y diabetes

Propuesta de Trabajo de Título

José Manuel Arenas Alarcón
jose.arenasa@alumnos.uv.cl

4 de abril de 2014

Profesor Guía: Eliana Providel Godoy

Resumen

La Minería de Datos es un campo de desarrollo e investigación que intenta descubrir patrones interesantes y desconocidos en grandes volúmenes de datos, para ser utilizados en sistemas informáticos que apoyan la toma de decisiones. Es así como para el Departamento de Epidemiología de la Secretaría Regional Ministerial (SEREMI) de Salud (DESS) de Valparaíso está en desarrollo un Sistema de Ayuda en Epidemiología (SADEPI) el cual tiene módulos para el registro y análisis de casos de cánceres y diabetes. Sin embargo el sistema de análisis que posee sólo abarca la generación de estadística descriptiva y no genera nuevo conocimiento útil y no trivial que ayude en la toma de decisiones. Por esto es que surge la necesidad de crear un sistema que permita entregar información útil y no trivial de los datos almacenados en el sistema SADEPI acerca de los casos de cánceres y diabetes, que es el objetivo principal de desarrollo del presente trabajo de título.

Índice

1. Introducción

El Ministerio de Salud (MINSAL), tiene como misión contribuir a elevar el nivel de salud de la población, además de desarrollar armónicamente los sistemas de salud, centrados en las personas [1]. Perteneciente al MINSAL, por región, se encuentra la Secretaría Regional Ministerial (SEREMI) de Salud, que entre sus departamentos se encuentra el Departamento de Epidemiología[1] (DESS).

DESS está encargado de organizar y mantener funcionando el Sistema de Vigilancia en Salud Pública e Investigación Epidemiológica (SVE) para la prevención y control de problemas de salud, así como el procesamiento y análisis de datos para la información epidemiológica en apoyo a la gestión sanitaria. Es así, como SVE tiene el propósito de contribuir a mejorar la calidad de vida y nivel de salud de la población chilena, a través de la entrega de información para la planificación y evaluación de las políticas y programas de prevención y control de enfermedades no transmisibles y sus factores de riesgos. Para apoyar este propósito SVE cuenta con datos asociados a enfermedades no transmisibles y sus factores de riesgos, generando insumos para la toma de decisiones en salud.

Las enfermedades no transmisibles se pueden clasificar en agudas y crónicas. Dentro de las enfermedades crónicas no transmisibles se encuentran: enfermedad isquémica del corazón, accidentes cerebrovasculares, diabetes mellitus (tipo 1 y tipo 2), cánceres (estómago, colon y recto, mama, cervicouterino, tráquea, bronquios y pulmón) y enfermedades crónicas de las vías respiratorias inferiores. Y asociado a las enfermedades agudas no transmisibles se encuentran accidentes del tránsito e intoxicaciones agudas por plaguicidas.

Actualmente el DESS cuenta con un Sistema de Ayuda en Epidemiología (SADEPI)[2], el cual es una herramienta que permite acceder a través de distintos módulos del sistema a información relevante de datos asociados a Diabetes Mellitus, Causas de muerte, Egresos Hospitalarios y Cáncer. Esta información corresponde a tablas y gráficos que se utilizan en informes para así mantener la vigilancia de las enfermedades con información de utilidad.

SADEPI sólo cuenta con el registro e información acerca de dos enfermedades crónicas no transmisibles que son los casos de cánceres (que es una de las principales causas de muerte a nivel mundial, siendo responsable de 7,6 millones de defunciones ocurridas en 2008) y diabetes mellitus (de la cual la diabetes mellitus tipo 2 tiene una prevalencia del 9,4 % en lo mayores de 15 años que viven en Chile). Sin embargo existe un problema y es que SADEPI no cuenta con un sistema que permita entregar

[1] La Epidemiología [2] es la ciencia que estudia cuándo y dónde ocurren las enfermedades y cómo se transmiten a las poblaciones.

[2] `http://ssrv.cl/sadepi`

información útil y no trivial, como puede ser patrones o relaciones interesantes entre los datos, asociado a un análisis para la generación de conocimiento sobre los datos almacenados. Por lo que es de importancia, considerando estas dos enfermedades crónicas no transmisibles (cáncer y diabetes), contar con un sistema que permita la generación de información útil y no trivial sobre los datos que actualmente cuenta SADEPI. Considerando esta falencia es que el presente trabajo de título tiene por objeto principal el desarrollo de un sistema, utilizando técnicas de minería de datos, entregando información que apoye la toma de decisiones.

2. Definición del Problema

2.1. Problema

Actualmente el DESS de Valparaíso cuenta con un sistema llamado SADEPI el cual puede registrar y analizar datos asociados a cáncer y diabetes, utilizando sólo estadística descriptiva. Esto está apoyado por la generación de informes desarrollados bajo demanda por el encargado de epidemiología.

De acuerdo a lo descrito, se detecta un problema y es que el sistema SADEPI no cuenta con un sistema que permita entregar información útil y no trivial, como puede ser detectar o predecir patrones, como también identificar relaciones interesantes entre los datos, que permita un análisis para la generación de conocimiento sobre los datos almacenados de cáncer y diabetes.

2.2. Solución Propuesta

Para dar solución al problema señalado anteriormente, se propone crear un módulo que pertenezca a SADEPI que permita la generación y visualización de información útil y no trivial utilizando técnicas de Minería de Datos, detectando por ejemplo patrones de comportamiento, relaciones entre los datos, asociación y dependencia de datos, como puede ser las relaciones de casos de cánceres o diabetes y ubicación geográfica, como también datos asociados a edad, sexo, antecedentes hereditarios, entre otros, utilizando los datos de cáncer y diabetes. Para que de esta forma apoye al DESS en el procesamiento y análisis de los datos.

Este módulo obtendrá los datos desde la base de datos de SADEPI, y permitirá al usuario generar un reporte con la información obtenida luego de procesar los datos utilizando técnicas de Minería de Datos.

2.3. Importancia del trabajo

Considerando las enfermedades crónicas no transmisibles, cáncer y diabetes, es de importancia contar con un sistema que permita la generación de información útil y no trivial sobre los datos que actualmente cuenta SADEPI, ya que esta información permitirá apoyar la gestión y toma de decisión para que el DESS pueda:

- Establecer nuevos protocolos de acción para cada una de las enfermedades.
- Administrar eficiente y oportunamente los recursos hospitalarios.
- Gestionar la adquisición temprana de bienes/servicios.
- Proyectar y administrar inventarios.
- Determinar patrones de las enfermedades permitiendo el diseño de medidas preventivas y correctivas.
- Debido a conocimiento anterior y nuevos datos generados, determinar grupos de riesgo por criterios geográficos, socioeconómicos entre otros, para la correcta entrega de insumos o servicios como medida de control.

3. Objetivos

3.1. Objetivo General

El objetivo de este trabajo de título es crear un módulo que será integrado en el sistema SADEPI, tal que permita entregar información útil y no trivial utilizando técnicas de Minería de Datos con los datos de cáncer y diabetes.

3.2. Objetivos Específicos

Para cumplir con el objetivo general, se detallan aquí los objetivos específicos:

- Analizar y comparar distintas técnicas de Minería de Datos.
- Establecer patrones de los datos almacenados, utilizando las técnicas de Minería de Datos analizadas.
- Detección y predicción para la toma de decisiones, basado en los datos existentes.
- Implementar las técnicas de Minería de Datos seleccionadas.
- Generar reportes según la técnica de Minería de Datos utilizada.

4. Metodología

Para cumplir con los objetivos propuestos anteriormente, es necesario distinguir dos principales fases:

1. La primera fase donde se analizan y comparan distintas técnicas de Minería de Datos, con el objetivo de escoger la(s) técnica(s) que puedan entregar mejor información útil y no trivial.

2. La segunda fase de desarrollo del módulo que permita ejecutar la(s) técnica(s) de Minería de Datos seleccionadas en la fase anterior, y generar un reporte con la información generada.

La metodología que se utilizará para desarrollar la primera fase es utilizar el Modelo de Transformación de Balzer [3] el cual se utiliza para cambiar la transformación de los datos, seleccionar algoritmos, optimizar y compilar. Este modelo, como se muestra en la Figura 1, permitirá encontrar la técnica de Minería de Datos más adecuada, donde de una especificación formal, pasa por un proceso de transformación sobre los algoritmo.

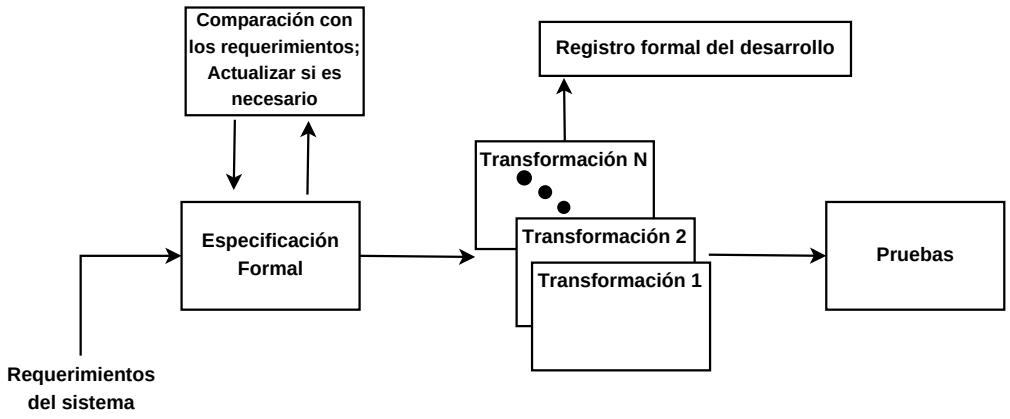

Figura 1: Modelo de transformación.

Para la segunda fase se utilizará la metodología incremental/iterativo. El módulo a desarrollar contará de dos incrementos, un incremento para ejecutar las técnicas de Minería de Datos para los casos de cáncer y otro incremento para los casos de diabetes, esto permite la posibilidad de que en el desarrollo de ambos módulos pueda mejorar. Junto con esto, como se muestra en la Figura 2 esta fase abarca la etapa de análisis de requerimientos, diseño, desarrollo y pruebas, donde se puede iterar desde las pruebas al diseño, para mejorar el desarrollo, para luego finalizar con la etapa de implantación e integración del sistema.

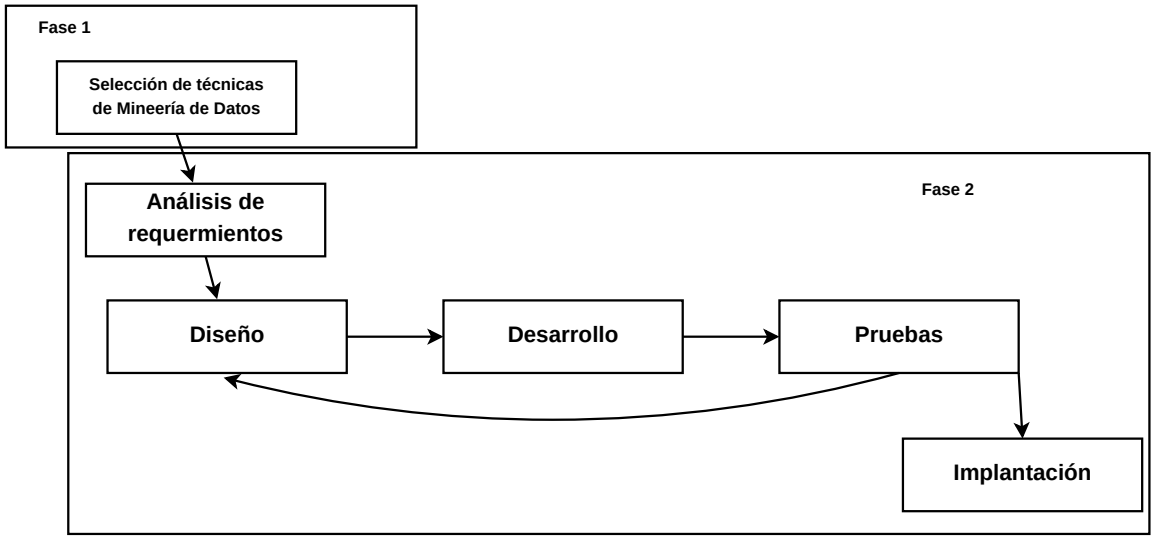

Figura 2: Modelo incremental/iterativo.

5. Planificación

En la Tabla 1 se presentan todas las actividades que se desarrollaran durante el primer semestre, asociado a su fecha de inicio y fin, junto con el producto a entregar. Además de lo presentado en la tabla, existe al menos una reunión de trabajo, por semana, con la profesora guía y continua comunicación a través de correo electrónico. Además todas las etapas tienen asociados su documentación, la cual estará en constante revisión por la profesora guía.

6. Recursos

6.1. Recursos Humanos

Durante el desarrollo de este trabajo de título, se requiere de un desarrollador de software con conocimientos en técnicas de minería de datos. Esta responsabilidad la adquirirá el alumno quien realiza este trabajo de título. Además, para los trabajos de pruebas se requiere la participación de usuarios del sistema y usuarios expertos.

6.2. Recursos Materiales

Como recursos materiales se necesitará acceso al servidor donde se aloja SADEPI, para poder implantar el software, además para tener acceso a la base de datos de SADEPI con los registros de las enfermedades crónicas no transmisibles que gestiona.

Actividad	**Inicio**	**Fin**	**Producto**
Propuesta de trabajo de título	17-03-2014	03-04-2014	Informe y presentación de propuesta de Trabajo de Título
Marco conceptual	07-04-2014	24-04-2014	Informe y presentación Marco conceptual, estado del arte, definición del problema, análisis de solución propuesta
- Estado del arte	07-04-2014	11-04-2014	Definición del estado del arte
- Definición del problema	14-07-2014	18-04-2014	Definición del problema
- Análisis de solución	21-04-2014	24-04-2014	Análisis de la solución
Diseño de la solución	28-04-2014	16-05-2014	Informe y presentación del diseño de la solución
- Análisis y selección de técnicas de Minería de Datos	28-04-2014	09-05-2014	Selección e implementación de las técnicas de Minería de Datos
- Diseño arquitectónico	12-05-2014	13-04-2014	Diseño de la arquitectura
- Diseño de interfaz	14-05-2014	15-05-2014	Diseño de la interfaz
Implementación	19-05-2014	04-07-2014	Informe y presentación de la implementación
- Implementación de los módulos	16-06-2014	04-07-2014	Implementación de los módulos que entregan información útil y no trivial.

Tabla 1: Planificación primer semestre.

6.3. Recursos del Desarrollador

Los recursos que el desarrollador de este trabajo de titulo posee son:

- Computador portátil para el desarrollo e investigación, sus características son:
  - Procesador Intel(R) Core(TM) i7-3630QM CPU @ 2.40GHz (8 CPUs).
  - Memoria 8192MB RAM.
  - Windows 8.1 Pro 64-bit.
  - 250GB de disco duro SSD.
- Sistema de Gestión de Bases de Datos MySQL.
- Sistema Operativo Centos virtualizado configurado como servidor web, con apache, php y MySQL, para realizar pruebas.

Bibliografía

[1] Ministerio de Salud. `http://web.minsal.cl/mision_vision`. Último acceso 03-04-2014.

[2] Gerard J. Tortora, Berdell R. Funke, y Christine L. Case. *Introducción a la Microbiología*. Editorial Médica Panamericana, 2007. Novena Edición.

[3] Shari Lawrence Pfleeger. *Software Engineering: Theory and Practice*. Prentice Hall PTR Upper Saddle River, NJ, USA, 2001. Segunda Edición.