Facultad de Ingeniería
Escuela de Ingeniería Civil Informática

SISTEMA DE MINERÍA DE DATOS PARA ANALIZAR CASOS DE CÁNCERES Y DIABETES

Por

José Manuel Arenas Alarcón

Trabajo realizado para optar al Título de
INGENIERO CIVIL EN INFORMÁTICA
Prof. Guía: Eliana Providel Godoy
Julio 2014

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 como herramientas informáticas 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 general

Índice de tablas

Índice de figuras

Capítulo 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 [12]. 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.

Considerando las enfermedades no transmisibles, es que estos 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.

[1] La Epidemiología [19] es la ciencia que estudia cuándo y dónde ocurren las enfermedades y cómo se transmiten a las poblaciones.

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 los 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 información útil y no trivial, como puede ser patrones o relaciones interesantes entre los datos, así como también datos geográficos, 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 generar información antes desconocida e interesante 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] `http://ssrv.cl/sadepi`

Capítulo 2

Marco Conceptual

2.1. Estado del Arte

2.1.1. Terminología

El **Ministerio de salud** (MINSAL) es un organismo del estado cuya misión es contribuir a elevar el nivel de salud de la población, además de desarrollar armónicamente los sistemas de salud, centrados en las personas [12]. En la región de Valparaíso, el MINSAL tiene a cargo la **Secretaría Regional Ministerial** (SEREMI) de salud el cual tiene por misión contribuir en el mejoramiento sostenido de la salud y la calidad de vida de la población de la Región de Valparaíso [24]. Uno de los departamentos de la SEREMI de salud de Valparaíso es el **Departamento de Epidemiología** (DESS) el que está encargado de organizar y mantener funcionando el **Sistema de Vigilancia** en Salud Pública e Investigación Epidemiológica. La **Epidemiología** [19] es la ciencia que estudia cuándo y dónde ocurren las enfermedades y cómo se transmiten a las poblaciones. La **Vigilancia Epidemiológica** es la recolección sistemática, análisis e interpretación de datos de salud necesarios para la planificación, implementación y evaluación de políticas de salud pública [26]. En Chile se ha incorporado la **Vigilancia de Enfermedades No Transmisibles** (VENT) [25], estas enfermedades 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.

En apoyo a la Vigilancia de Enfermedades No Transmisibles está en desarrollo un **Sistema de Ayuda en Epidemiología** (SADEPI) el cual es una herramienta que permite acceder a información relevante respecto a diabetes mellitus, causas de muerte, egresos

hospitalarios y cáncer. Considerando que la diabetes y el cáncer son las enfermedades con las que trabaja SADEPI, así como también este trabajo de título, es de importancia la descripción de estas enfermedades.

La **diabetes mellitus** [5] (DM) comprende un grupo de trastornos metabólicos frecuentes que comparten el fenotipo de la hiperglucemia. La DM se clasifica con base en el proceso patógeno que culmina en hiperglucemia, las dos categorías amplias de la DM se designan tipo 1 y tipo 2.

La **diabetes tipo 1** es resultado de la deficiencia completa o casi total de insulina, y la **diabetes tipo 2** es un grupo heterogéneo de trastornos que se caracterizan por grados variables de resistencia a la insulina, menor secreción de dicha hormona y una mayor producción de glucosa.

El **cáncer** es un grupo de enfermedades que se produce por el crecimiento anormal y desordenado de las células del cuerpo, esto es causado por alteraciones celulares [8].

Este trabajo de título tiene por objetivo aplicar técnicas de Minería de Datos sobre los datos de diabetes y cáncer que actualmente tiene SADEPI, esto se realiza utilizando el **proceso de descubrimiento de conocimiento en bases de datos** conocido como KDD (*Knowledge Discovery from Databases*), el cual es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos [21], una de las partes esenciales de este proceso es la aplicación de diferentes técnicas de **Minería de Datos** la cual se define como el proceso de descubrimiento de conocimiento y patrones interesantes de grandes cantidades de datos [20].

2.1.2. Estado actual de cáncer en el mundo

El cáncer es una de las principales causas de defunción en el mundo. En 2012 se registraron 8,2 millones de muertes por su causa [9].

Los que más muertes causan cada año son los cánceres de pulmón, hígado, estómago, colon y mama, la Tabla 2.1 muestra la cantidad de muertes en el mundo de los principales tipos de cáncer en 2012 [10], como porcentaje del total de muertes por cáncer[1].

Aproximadamente un 30 % de las muertes por cáncer son debidas a cinco factores de riesgo conductuales y dietéticos: índice de masa corporal elevado, ingesta reducida de frutas y verduras, falta de actividad física, consumo de tabaco y consumo de alcohol.

[1] El total de muertes por cáncer para el 2012 es de 8.201.000 [10]

Cáncer	**Muertes**
Pulmonar	19,4 %
Hepático	9,1 %
Gástrico	8,8 %
Colorrectal	8,5 %
Mamario	6,4 %

Tabla 2.1: Principales tipos de cáncer.

El consumo de tabaco es el factor de riesgo más importante. Es la causa de casi el 20 % de las muertes mundiales por cáncer en general y alrededor del 70 % de las muertes mundiales por cáncer de pulmón.

Los cánceres causados por infecciones víricas, tales como las infecciones por virus de las hepatitis B (VHB) y C (VHC) o por papilomavirus humanos (PVH), son responsables de hasta un 20 % de las muertes por cáncer en los países de ingresos bajos y medios [4].

El consumo de tabaco y alcohol, la dieta malsana y la inactividad física son los principales factores de riesgo de cáncer en todo el mundo. Las infecciones crónicas por VHB, VHC y algunos tipos de PVH son factores de riesgo destacados en los países de ingresos bajos y medianos. El cáncer cervicouterino, causado por PVH, es una de las principales causas de defunción por cáncer en las mujeres de países de ingresos bajos.

Más del 60 % de los nuevos casos anuales totales del mundo se producen en África, Asia, América Central y Sudamérica. Estas regiones representan el 70 % de las muertes por cáncer en el mundo.

2.1.3. Estado actual de cáncer en Chile

El cáncer es la segunda causa de defunciones en Chile con una tasa de 130,2 por 100.000 habitantes, luego de las enfermedades cardiovasculares. Además es responsable del 24,9 % del total de muertes (2011) [14, 16]. Sin embargo se observa que en las regiones de Arica y Parinacota, Iquique, Antofagasta y Aisén el cáncer pasa a ser la primera causa de muerte. Además, en el 2007, nueve de las Regiones sobrepasa la tasa de mortalidad por cáncer del país, estas son: Arica y Parinacota, Tarapacá, Antofagasta, Coquimbo, Valparaíso, Maule, Biobío, Araucanía, Los Ríos y Magallanes [15].

En la primera Encuesta Nacional de Salud de Chile desarrollada en el año 2003 por

el MINSAL, se observa un predominio de estilos de vida poco saludables, como lo son el tabaquismo, el sedentarismo y la obesidad; y estos resultados se repiten en la encuesta 2009-2010 [15]. La Tabla 2.2 muestra los factores de riesgo según la Encuesta Nacional de Salud de Chile 2003 y 2009-2010.

Factor de riesgo	**2003**	**2009-2010**
Tabaco	42 %	40,6 %
Sedentarismo	89 %	88,6 %
Sobrepeso y obesidad	60 %	64,5 %

Tabla 2.2: Principales factores de riesgo según Encuesta Nacional de Salud de Chile 2003 y 2009-2010

2.1.4. Estado actual de diabetes en el mundo

En el mundo hay más de 347 millones de personas con diabetes[2]. Más del 80 % de las muertes por diabetes se registran en países de ingresos bajos y medios [11]. La mayoría de las personas con diabetes tiene entre 40 y 59 años de edad y se calcula que la diabetes causó 5,1 millones de muertes en 2013, además más de 79.000 niños desarrollaron diabetes tipo 1 [18]. La Tabla 2.3 muestra una estimación de la prevalencia y cantidad de personas con diabetes en el año 2035.

Diabetes (20-79 Años)	**2013**	**2035**
Prevalencia global (%)	8,3	10,1
Personas con diabetes (millones)	382	592

Tabla 2.3: Estimación diabetes año 2035.

La diabetes tipo 2 representa entre el 85 % y el 95 % del total de la diabetes en los países de ingresos altos y puede representar un porcentaje aún mayor en los países de ingresos medios y bajos. La diabetes tipo 1, aunque menos común que la diabetes tipo 2, está aumentando cada año. En la mayoría de los países de ingresos altos, la mayor parte de la diabetes en niños y adolescentes es la diabetes tipo 1 [18].

La diabetes impone una gran carga económica para las personas y sus familias, los sistemas nacionales de salud y los países. El gasto sanitario por la diabetes representó el

[2]paciente con glucemia en ayunas $\geq$ 7,0 mmol/l o medicado

10,8 % del gasto sanitario total de todo el mundo en 2013. El gasto sanitario incluye el gasto médico por diabetes de los sistemas de salud, así como las personas que viven con la diabetes y sus familias [18].

2.1.5. Estado actual de diabetes en Chile

La diabetes mellitus tipo 2 es una enfermedad con una alta prevalencia en los mayores de 15 años que viven en Chile, 9,4 % de acuerdo a la Encuesta Nacional de Salud 2009 - 2010 [13]. La tasa de mortalidad por causa de diabetes mellitus es de 19,2 por 100.000 habitantes, lo que corresponde a 3.253 defunciones [17]. La principal causa de muerte en los diabéticos es la enfermedad cardiovascular. Las personas adultas con diabetes tienen un riesgo entre 2 a 4 veces mayor que los adultos no diabéticos de presentar un evento cardiovascular [7].

En Chile, la tasa de amputaciones en diabéticos durante la década pasada aumentó en 28 %, de 3,5 a 4,5 por 1.000 diabéticos, lo que corresponde para el año 2006 a 3.192 amputaciones en personas diabéticas [7].

2.1.6. Proceso de descubrimiento de conocimiento en bases de datos

Frecuentemente se utilizan algunos términos como sinónimos de la Minería de Datos. Uno de ellos es la extracción o "descubrimiento de conocimiento en bases de datos" (Knowledge Discovery in Databases, KDD) y suele utilizarse ambos indistintamente, aunque existen claras diferencias entre los dos. KDD se utiliza como un proceso que consta de una serie de fases [21], mientras que la minería de datos es sólo parte de estas fases.

Knowledge Discovery in Databases (Figura 2.1) (KDD) es un modelo y análisis exploratorio automático de grandes repositorios de bases de datos. KDD es un proceso organizado para identificar patrones válidos, nuevos, útiles y comprensibles desde grandes y complejos conjuntos de datos. La Minería de Datos es el núcleo principal del proceso de KDD, que implica la inferencia de algoritmos que exploran los datos, desarrolla el modelo y descubre patrones previamente desconosidos. El modelo es usado para la comprensión de los fenómenos de los datos, análisis y predicción [23].

El proceso de descubrimiento de datos consiste en nueve etapas y es iterativo, esto significa que se puede mover a cada etapa para volver a ajustar las etapas anteriores si es necesario. Estas etapas son:

1. **Desarrollo de una comprensión del dominio de la aplicación**. Este es el paso preparatorio inicial, donde se establece el escenario para la compresión de los que se

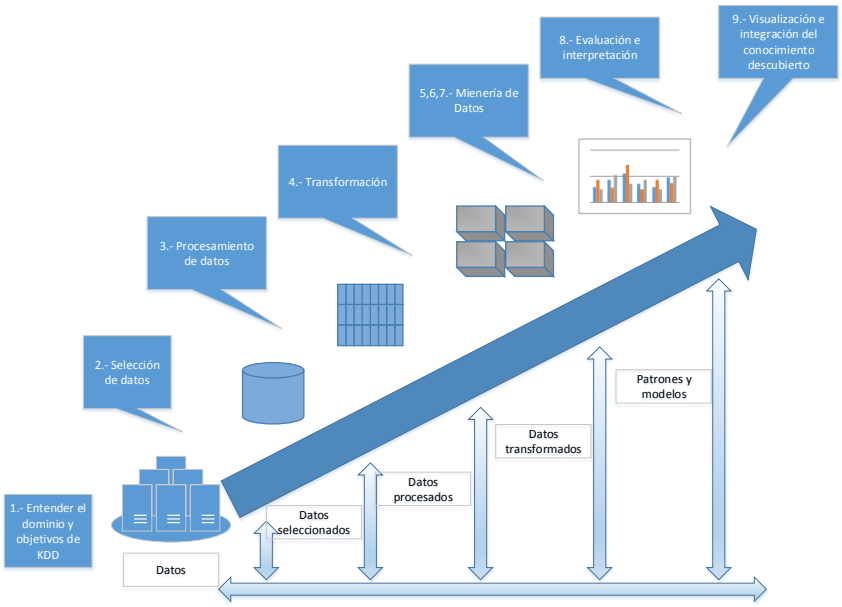


Figura 2.1: El proceso de descubrimiento de conocimiento en bases de datos.

debe hacer con las decisiones. Se deben definir los objetivos del proyecto de KDD.

2. **Selección y creación del conjunto de datos que se utilizará**. Una vez definidos los objetivos, se debe determinar los datos que se utilizarán en el proceso de KDD, esto corresponde a encontrar los datos que estén disponibles, buscar datos adicionales e integrarlos con el conjunto de datos que se considerará para el proceso.

3. **Procesamiento y limpieza**. En esta etapa se considera la confiabilidad de los datos, esto incluye el borrado de datos, como el manejo de los valores perdidos y la eliminación de valores atípicos.

4. **Transformación de los datos**. En esta etapa se transforman los datos, preparándolos para la Minería de Datos. Se utilizan métodos para la reducción de dimensiones, transformaciones (como la discretización de atributos numéricos).

5. **Elección de la técnica de Minería de Datos más apropiada**, como por ejemplo técnicas de clasificación, regresión o clustering. Esto depende en gran medida de los objetivos de KDD.

6. **Elección del algoritmo de Minería de Datos**, incluye seleccionar el método específico que se utilizará para la búsqueda de patrones, de acuerdo a la técnica elegida en la etapa anterior.

7. **Utilizar el algoritmo de Minería de Datos**. En esta etapa se alcanza la implementación del algoritmo de Minería de Datos.

8. **Evaluación**. En esta etapa se evalúa e interpreta los patrones extraídos con respecto a los objetivos definidos en el primer paso.

9. **Uso del conocimiento descubierto**, con el objetivo de apoyar la toma de decisiones

2.1.7. Minería de datos y sus técnicas

La Minería de Datos es el proceso de descubrir patrones y conocimientos interesantes a partir de grandes cantidades de datos [20]. Es una etapa en el proceso de descubrimiento de conocimiento en donde se escoge la(s) técnica(s) de Minería de Datos adecuada para cumplir con el objetivo del proceso de KDD. Éstas técnicas se pueden dividir en **técnicas de aprendizaje supervisado** y **técnicas de aprendizaje no supervisado**. En las técnicas basadas en aprendizaje supervisado se utiliza un conjunto de datos como datos de entrenamiento que están etiquetados y previamente clasificados. En las técnicas basadas en aprendizaje no supervisado los datos a analizar no están clasificados ni etiquetados, por lo que los algoritmos deben agrupar los datos automáticamente.

Dentro de las técnicas basadas en aprendizaje supervisado podemos encontrar:

- La **clasificación**, que es el proceso de encontrar un modelo que describe y distingue clases de datos. El modelo se deriva a partir de un conjunto de datos de entrenamiento, y es usado para predecir el nombre o etiqueta de la clase para objetos que no están clasificados, esto es, objetos a los que no se les conoce el nombre de la clase a la que pertenecen. Una forma de representar estos modelos es utilizando árboles de decisión que es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguno de sus hijos. Uno de los algoritmos de clasificación utilizados para construir un árbol de decisión es el algoritmo ID3 el cual pertenece a la familia TDIDT (Top-Down Induction od Decision Trees) y su objetivo es construir un árbol de decisión que explique cada instancia de entrada de la manera más compacta posible, estableciendo una secuencia dentro del árbol de decisión. ID3 en cada momento elige el mejor atributo dependiendo de una determinada heurística[3], por lo que se debe determinar que variables entregan información relevante para la solución del problema.

[3]Una heurística son reglas informales o intuitivas que señalan que acciones se pueden tomar cuando no es posible hacer uso de algoritmos.

- Las **reglas de asociación** son patrones que ocurren frecuentemente en los datos. Estos patrones corresponden a conjuntos de elementos frecuentes, esto se refiere a datos que a menudo aparecen juntos en un conjunto de datos transaccionales, como por ejemplo la leche y el pan a menudo se encuentran juntos en transacciones comerciales de un supermercado. Un algoritmo utilizado para encontrar patrones frecuentes es el algoritmo **apriori** [20] propuesto por R. Agrawal y R. Srikant en 1994, para minar conjuntos de datos frecuentes para reglas de asociación booleanas. Se basa en el conocimiento previo o apriori de los conjuntos de datos frecuentes.

- La **regresión** es una técnica que consiste en convertir datos en una función, cuando éstos son datos numéricos. Éste método se usa para predecir el valor de una variable de respuesta (variable dependiente) de una o más variables predictoras (variable independiente). La regresión lineal implica encontrar una recta que se ajuste a dos atributos (o variables) de tal manera que un atributo pueda utilizarse para predecir el otro.

Dentro de las técnicas basadas en aprendizaje no supervisado podemos encontrar:

- El **clustering** es una técnica de Minería de Datos que consiste en agrupar un conjunto de datos u objetos en diferentes sub-conjuntos (clusters), de tal forma que los objetos que pertenezcan a un cluster sean muy similares entre si y al mismo tiempo sean muy diferentes de los objetos pertenecientes a otros clusters. Cada cluster puede verse como una clase de objetos. Un algoritmo utilizado para realizar clustering es el algoritmo K-means, el cual particiona los datos en K grupos cada uno representado por su centro, se determinan las distancias[4] de cada objeto con todos los centros y se reagrupan los objetos en base a la distancia mínima a cada cluster, de esta forma se maximiza la similitud entre los objetos de un mismo cluster.

- La **inferencia estadística** es un conjunto de métodos estadísticos que permiten deducir como se distribuye la población en estudio partir de la información que proporciona una muestra. Se dedica a la generación de modelos y predicciones asociadas, teniendo en cuenta la aleatoriedad de las observaciones. Se utiliza la información entregada por la Estadística descriptiva [20], la cual se puede usar para identificar propiedades de los datos y destacar los valores que debieran ser tratados como ruido o atípicos. Se pueden observar tres áreas en estadística descriptiva, una de ellas están las medidas de tendencia central (que calculan el lugar del medio o centro de la distribución de los datos), la dispersión de los datos (útiles para identificar valores atípicos) y gráficos para inspeccionar visualmente los datos.

[4] En el algoritmo K-means la distancia corresponde a la distancia Euclidiana.

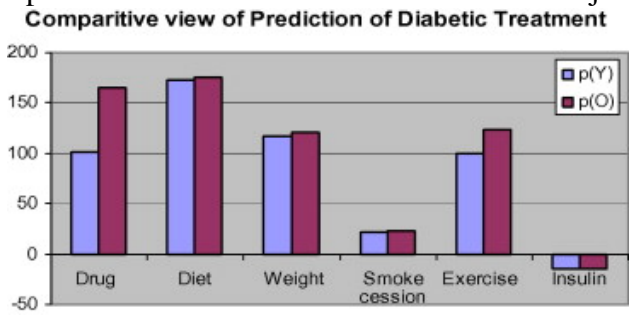
2.1.8. Minería de datos en cáncer y diabetes

En el campo de la salud se encuentran algunos trabajos realizados relacionados con Minería de Datos utilizándola como una herramienta para generar nuevo conocimiento en el área. En cuanto a los trabajos podemos encontrar a:

- **Application of data mining: Diabetes health care in young and old patients** [1] el cual es una investigación donde utilizan una técnica de Minería de Datos basado en regresión para identificar la efectividad de diferentes tipos de tratamientos para la diabetes en diferentes grupos etarios. El conjunto de datos utilizado es una base de datos del año 2005 acerca de el reporte de los factores de riesgo de enfermedades no transmisibles del Ministerio de Salud de Arabia Saudita, obtenida de forma gratuita desde la web de la Organización Mundial de Salud. La investigación se concentra en los datos de diabetes, donde se identifican seis tipos distintos de tratamiento: drogas, dieta, reducción de peso, dejar de fumar, ejercicio e insulina. Los grupos etarios identificados son los *jóvenes* y *mayores*, los jóvenes corresponden a personas que tienen entre 15 y 44 años de edad, separados en tres grupos de 15-24, 25-34 y 35-44; los mayores corresponden a personas que tienen entre 35 y 64 años de edad, separados en tres grupos de 35-44, 45-54 y 55-64, el grupo de 35 a 44 años de edad es compartido por ambos grupos etarios. Los resultados indican que los tratamientos basados en drogas son efectivos en ambos grupos etarios, pero son más efectivos en el grupo etario de los mayores. Con respecto a aplicar un tratamiento de dieta, los resultados indican que se recomienda realizar dieta en ambos grupos etarios. La reducción de peso es efectiva en ambos grupos etarios, según los resultados, sin embargo es un poco más efectiva en el grupo etario de los mayores. Dejar de fumar es un tratamiento efectivo en ambos grupos etarios. Realizar ejercicios físicos es más efectivo en pacientes del grupo etario de los mayores. Finalmente, según los resultados del estudio, la administración de insulina es un tratamiento inefectivo en ambos grupos etarios, el artículo explica que este resultado es debido a que la base de datos contiene a pacientes con diabetes tipo 1 y tipo 2, donde la diabetes tipo 1 es insulinodependiente (que necesita y se le administra insulina) y la diabetes tipo 2 es no-insulinodependiente (que no necesita ni se le administra insulina), por esta razón la predicción de la efectividad del tratamiento con insulina para la diabetes tiene un valor negativo para ambos grupos etarios. La Figura 2.2 muestra una tabla comparativa de la investigación realizada, donde $p(y)$ es la predicción de la efectividad del tratamiento para el grupo etario joven y $p(o)$ es la predicción de la efectividad del tratamiento para el grupo etario mayor.

- **Mining the information for structure based drug designing by relational database management notion** [2] el cual tiene por objetivo desarrollar un administrador de información que sea capaz de buscar en diferentes fuentes información relevante,

Figura 2.2: Comparación de la efectividad de tratamiento de jóvenes y mayores



mediante técnicas de minería de datos. Utiliza un banco de datos de documentos relacionados con casos de cánceres que incluye información textual y reportes además de documentos clínicos estructurados. La técnica de Minería de Datos utilizada es el agrupamiento automático de documentos (clustering) para agrupar los documentos que tienen información de cáncer relacionada, identificando el tamaño del tumor, los factores de riesgo, el tratamiento realizado, genética, detección de mutaciones, entre otros. El artículo menciona que con este estudio se puede proveer información que ayuda a diseñar drogas basadas en estructura.

- **Predicting breast cancer survivability using data mining techniques** [3] los autores realizaron un análisis de la predicción de la tasa de sobrevivencia de pacientes con cáncer de mamas utilizando tres técnicas de minería de datos: Naieve Bayes, redes neuronales, y el algoritmo de árboles de decisión C4.5 utilizando la herramienta Weka. La base de datos utilizada fue una nueva versión de Surveillance Epidemiology and End Results (SEER) con los datos del periodo 1973-2002 con 482.052 registros. En este estudio, la precisión de las tres técnicas de minería de datos son comparados y se muestran en la Tabla 2.4.

Técnica de clasificación	**Precisión**
Naive Bayes	84,5 %
Redes neuronales	86,5 %
C4.5	86,7 %

Tabla 2.4: Precisión de las técnicas de minería de datos.

El estudio muestra que los resultados preliminares son prometedores para la aplicación de técnicas de minería de datos en la predicción de sobrevivencia en bases de datos médicas. El porcentaje de predicción obtenidos por los tres algoritmos son comparados, sin embargo, el algoritmo C4.5 tiene mejor rendimiento que las otras dos técnicas.

- **Survival-Time classification of breast cancer patients** [22] Los autores han analizado si la quimioterapia puede prolongar el tiempo de vida de un paciente con cáncer de mama utilizando una técnica de minería de datos. El objetivo principal de este trabajo es intentar identificar a los pacientes con cáncer de mama quienes han prolongado su tiempo de vida con quimioterapia. El estudio se realiza a 253 pacientes con cáncer de mama con seis características de las cuales cinco son citológicas (promedio de área, error estándar de área, peor área, peor textura y peor perímetro) y una patológica (tamaño del tumor). Se utiliza el algoritmo *support vector machine*[5] para clasificar a los pacientes con cáncer de mama en tres grupos de pronóstico (Bueno, pobre, intermedio). Los resultados sugieren que los pacientes en el grupo Bueno no deberían recibir quimioterapia, mientras que aquellos que están en el grupo Intermedio sí deberían recibir quimioterapia.

2.1.9. Aplicaciones que utilizan Minería de Datos en sistemas de Salud

Como vimos en el punto anterior, existen diversos trabajos realizados donde utilizan la minería de datos como herramienta para generar nuevo conocimiento que apoye a la toma de decisiones en los sistemas de salud. En cuanto a aplicaciones que utilizan técnicas de minería de datos para procesar grandes cantidades de datos clínicos existe **Sistema Integral para la Atención Primaria de la Salud** (SIAPS) [6] desarrollado por el Centro de Informática Médica de la Universidad de las Ciencias Informáticas de Cuba. SIAPS posee un componente llamado **Sistema Clínico de Soporte para la Toma de Decisiones** el que inicialmente entrega información utilizando técnicas estadísticas y en el artículo [6] se menciona la inclusión de la minería de datos para aprovechar al máximo la información almacenada.

El componente utiliza la información de las Historias Clínicas Electrónicas que se encuentran almacenadas en un repositorio y enviados periódicamente a un Datamart. El caso de estudio mencionado analiza los datos de Hipertensión Arterial, para el cual utilizan dos algoritmos, el J48 dentro de la técnica supervisada de Árboles de Decisión y el Simple K-Means para el desarrollo de la técnica no supervisada de Agrupamiento.

2.1.10. SADEPI

SADEPI[6] es una herramienta que está en desarrollo para el DESS de Valparaíso que cuenta con el registro y análisis de casos de Diabetes, Causas de Muerte, Egresos Hospitalarios y casos de Cáncer. La información que entrega es mostrada mediante distintas formas como información geográfica donde, por ejemplo, se puede determinar la proporcionalidad

[5] Support vector machine es una técnica de clasificación basada en aprendizaje supervisado.

[6] `www.ssrv.cl/sadepi`

de los afectados de diabetes (sea el caso de compensado, descompensado, egreso hospitalario o defunción) o causas de muerte. Además cuenta con tablas y gráficos que entregan información descriptiva de los datos almacenados, los que además se pueden descargar para ayudar a los usuarios a crear informes.

Capítulo 3

Definición del Problema

3.1. Problema

La información es un recurso valioso que, cuando es precisa, puede apoyar a las personas en tomar mejores decisiones, por lo que contar con este recurso es imprescindible a la hora de hacer importantes decisiones.

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, aunque entrega información que le es útil para el encargado de epidemiología, no cuenta con un sistema que permita entregar información antes desconocida 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.

En resumen, el problema es que SADEPI no cuenta con un sistema de Minería de Datos con el que se pueda extraer información que le sea de utilidad al DESS de Valparaíso en la toma de decisiones, como medidas de prevención, control y mitigación.

3.2. Solución

Para dar solución al problema señalado anteriormente, se propone crear una sección de análisis de Minería de Datos que pertenezca al módulo de Vigilancia Diabetes y al

módulo Cáncer de SADEPI que permita la generación y visualización de información útil y no trivial utilizando técnicas de Minería de Datos.

Con este tipo de análisis sobre los datos se puede detectar por ejemplo patrones de comportamiento, relaciones entre los datos, asociación y dependencia de datos, como por ejemplo puede ser las relaciones de casos de cánceres o diabetes y su ubicación geográfica[1], 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.

Esta sección dentro de cada 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.

3.2.1. Importancia

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 (prevención, control y mitigación).
- 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.

3.2.2. Objetivos

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

[1] La ubicación geográfica corresponde a la del domicilio del paciente.

de Minería de Datos con los datos de cáncer y diabetes.

Objetivos específicos

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

- Analizar y comparar distintas técnicas de Minería de Datos, con el objetivo de seleccionar la técnica que mejor se adapte a la solución y descartar las que no entreguen información relevante.
- Establecer patrones de los datos almacenados, utilizando las técnicas de Minería de Datos seleccionadas.
- 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.

Capítulo 4

Especificación de Requerimientos

Los requerimientos son la descripción de las funcionalidades que un sistema que un sistema debe proporcionar. Los requerimientos funcionales corresponden a como el sistema se debe comportar en situaciones particulares, en la Tabla 4.1 se muestra la descripción de los requerimientos funcionales.

ID	Descripción
RF01	El módulo debe permitir al usuario seleccionar la fuente de datos (base de datos de cáncer o diabetes).
RF02	Los usuarios pueden seleccionar distintas técnicas de minería de datos a utilizar sobre los datos.
RF03	El módulo entrega un informe con el resultado realizado por la técnica de minería de datos.
RF04	El usuario puede seleccionar distintos parámetros, como la edad, el sexo, la ubicación geográfica, etc., para ejecutar una técnica de minería de datos.
RF05	El usuario puede seleccionar el lugar geográfico de la fuente de datos (servicio de salud, instituciones de una comuna).
RF06	El usuario puede descargar el informe generado.
RF07	El módulo debe permitir generar gráficos de los resultados que se obtengan.
RF08	El módulo debe permitir guardar resultados intermedios.

Tabla 4.1: Requerimientos funcionales.

Los requerimientos no funcionales son restricciones de los servicios o funciones que el sistema puede ofrecer. En la Tabla 4.2 se muestra la descripción de los requerimientos no funcionales.

ID	Descripción
RNF01	El informe podrá ser descargado en formato Microsoft Word ó PDF.
RNF02	Se contará con un manual de usuario.
RNF03	El módulo debe ser programado para que sea compatible con el sistema SA-DEPI.
RNF04	El usuario debe ser capaz de entender las funcionalidades del módulo luego de leer el manual de usuario.
RNF05	El tiempo de respuesta al ejecutar una técnica de minería de datos debe ser de a los más 10 segundos.
RNF06	Las funcionalidades deben ser modularizadas de forma que facilite la agregación, sustracción o modificación de éstas.

Tabla 4.2: Requerimientos No Funcionales.

4.1. Funciones del Sistema

En esta sección se presentan las principales funciones que tiene el sistema. En la Tabla 4.3 se muestra el nombre y descripción de las principales funciones, junto con un identificador y referencia al requerimiento funcional.

ID	Ref	Nombre	Descripción
F1	RF01	Seleccionar Fuente de Datos	Esta función se encarga de seleccionar la base de datos sobre la cual el usuario desee trabajar.
F2	RF04	Seleccionar Parámetros	Esta función es la encargada de seleccionar los datos que se utilizarán en el análisis de Minería de Datos.
F3	RF02	Seleccionar Técnica de Minería de Datos	Esta función permite al usuario seleccionar alguna técnica de minería de datos con la que se procesarán los datos.
F4	RF06	Descargar Informe	Esta función es la encargada de generar un informe para poder descargarlo en formato de Microsoft Word o PDF.
F5	RF07	Generar Gráficos	Esta función permite al usuario generar gráficos que permitan mostrar la información generada.

Tabla 4.3: Funciones del Sistema

4.2. Actores del Sistema

El módulo de Minería de Datos para el sistema SADEPI posee sólo un actor, que es el usuario, quien es el encargado de realizar análisis utilizando técnicas de Minería de Datos. La Figura 4.1 ilustra el modelo de usuario que se utiliza en este documento.

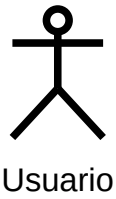


Figura 4.1: Usuario del Sistema

4.3. Casos de uso

En esta sección se muestra la interacción que tiene el usuario con el sistema. Los casos de uso documentan el comportamiento del sistema (acción y reacción) desde el punto de vista del usuario. A continuación se presentan los casos de uso extendido.

4.3.1. Diagrama general de caso de uso

En la Figura 4.2 se puede apreciar el diagrama general de caso de uso del módulo de Minería de Datos, donde se muestra la interacción del usuario con el sistema.

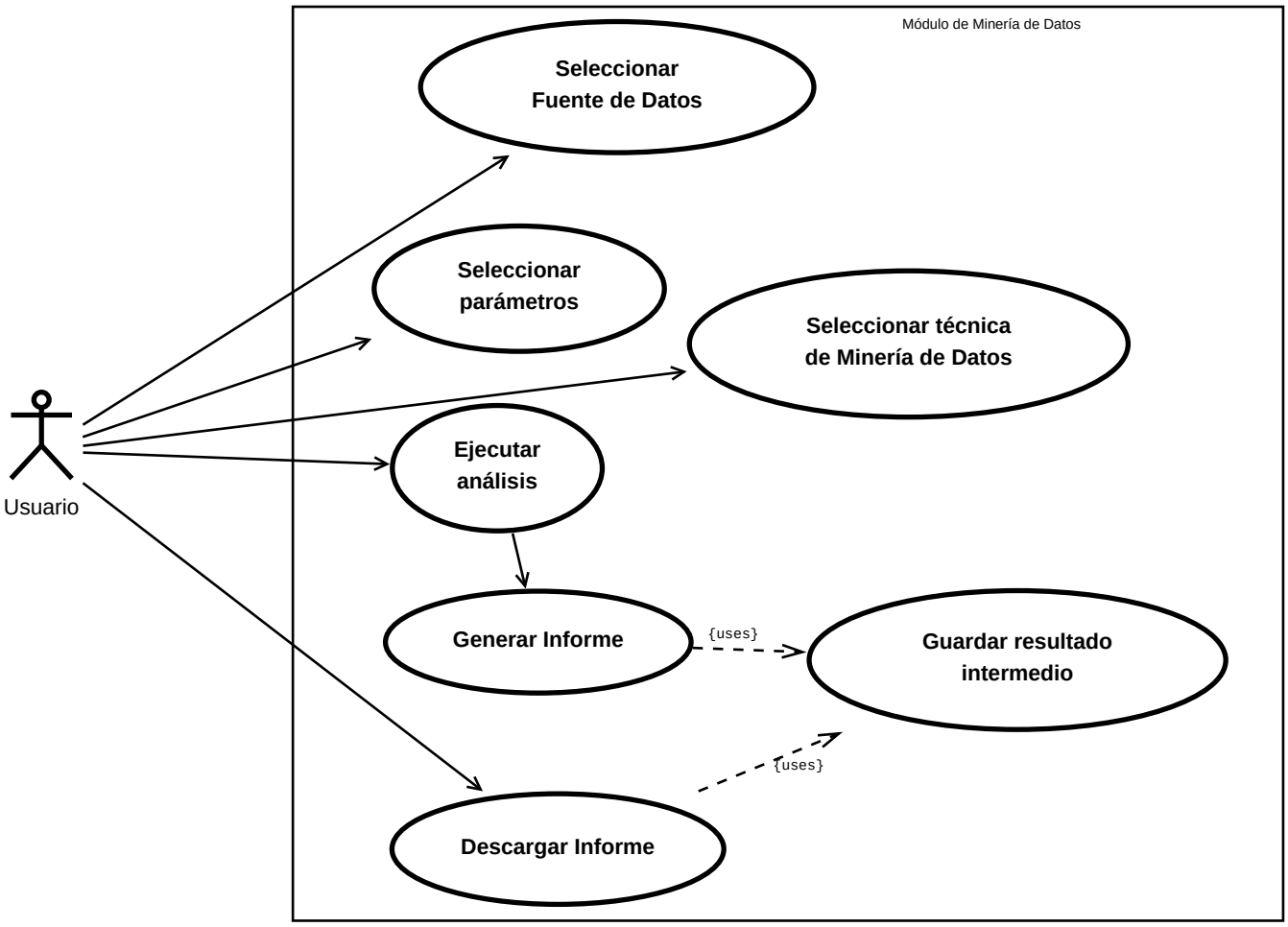


Figura 4.2: Diagrama de casos de uso.

4.3.2. Casos de Uso extendido

En la Tabla 4.4 se muestra el caso de uso extendido para el caso de uso de *Seleccionar Fuente de Datos*.

Caso de Uso	Seleccionar Fuente de Datos
Actores	Usuario
Propósito	Permite al usuario seleccionar la base de datos a utilizar, ya sea de cáncer o diabetes.
Pre-condición	El usuario debe estar registrado en el sistema SADEPI.
Resumen	El usuario puede seleccionar la base de datos que se usará al utilizar la técnica de Minería de Datos, la base de datos puede ser de casos de cáncer o diabetes.
Curso Normal de Eventos	
1.- Este caso de uso comienza cuando el usuario desee ejecutar un análisis de Minería de Datos.	
	2.- El módulo muestra las bases de datos que el usuario puede escoger.
3.- El usuario selecciona una base de datos, ya sea de cáncer o diabetes.	
4.- Finaliza el Caso de Uso.	
Curso Alternativo de Eventos	
3.- Si el usuario no desea continuar, puede cancelar.	

Tabla 4.4: Caso de Uso extendido Seleccionar Fuente de Datos.

En la Tabla 4.5 se muestra el caso de uso extendido para el caso de uso de *Seleccionar Técnica de Minería de Datos*.

Caso de Uso	Seleccionar Técnica de Minería de Datos
Actores	Usuario
Propósito	Permite al usuario seleccionar la técnica de Minería de Datos que se ejecutará.
Pre-condición	Haber seleccionado la fuente de datos en la que se ejecutará la técnica de Minería de Datos.
Resumen	El usuario puede seleccionar una técnica de Minería de Datos para para ejecutarla sobre una base de datos que haya seleccionado, bajo ciertos parámetros.
Curso Normal de Eventos	
1.- Este caso de uso comienza cuando el usuario desee seleccionar una técnica de Minería de Datos	
	2.- El módulo muestra las técnicas de Minería de Datos que el usuario puede elegir.
3.- El usuario selecciona una técnica de Minería de Datos.	
4.- Finaliza el Caso de Uso.	
Curso Alternativo de Eventos	
3.- Si el usuario no desea continuar, puede cancelar o volver al paso anterior.	

Tabla 4.5: Caso de Uso extendido Seleccionar técnica de Minería de Datos.

En la Tabla 4.6 se muestra el caso de uso extendido para el caso de uso de *Seleccionar Parámetros*.

Caso de Uso	Seleccionar Parámetros
Actores	Usuario
Propósito	Permite al usuario seleccionar los parámetros con los que se va a ejecutar una técnica de Minería de Datos en el sistema.
Pre-condición	Haber seleccionado la técnica de Minería de Datos.
Resumen	El usuario puede seleccionar los parámetros de la base de datos que se usarán para ejecutar una técnica de Minería de Datos, éstos parámetros pueden ser el sexo, la edad, la ubicación geográfica, etc., en general esta acción se utilizará para filtrar los datos de la base de datos seleccionada.
Curso Normal de Eventos	
1.- Este caso de uso comienza cuando el usuario desee seleccionar los parámetros que necesite para ejecutar la técnica de Minería de datos.	
	2.- El módulo muestra los parámetros que el usuario puede seleccionar.
3.- El usuario selecciona los parámetros.	
4.- Finaliza el Caso de Uso.	
Curso Alternativo de Eventos	
3 Si el usuario no desea continuar, puede cancelar o volver al paso anterior.	

Tabla 4.6: Caso de Uso extendido Seleccionar Parámetros.

En la Tabla 4.7 se muestra el caso de uso extendido para el caso de uso de *Ejecutar Análisis*.

Caso de Uso	Ejecutar Análisis.
Actores	Usuario.
Propósito	Permite al usuario ejecutar una técnica de Minería de Datos en el sistema.
Pre-condición	El usuario debe estar registrado en SADEPI y debe haber seleccionado una Base de Datos.
Resumen	El usuario puede seleccionar una técnica de Minería de Datos para ejecutarla sobre una base de datos que haya seleccionado, bajo ciertos parámetros.
Curso Normal de Eventos	
1.- Este caso de uso comienza cuando el usuario desee ejecutar un análisis de Minería de Datos.	
	2.- El módulo muestra las técnicas de Minería de Datos que el usuario puede elegir.
3.- El usuario selecciona una técnica de Minería de Datos y la ejecuta.	
	4.- El módulo muestra los parámetros que el usuario puede seleccionar.
5.- El usuario selecciona los parámetros.	
	6.- El módulo genera un informe y lo almacena como resultado intermedio.
7.- Finaliza el Caso de Uso.	
Curso Alternativo de Eventos	
(3,5) Si el usuario no desea continuar, puede cancelar o volver al paso anterior.	

Tabla 4.7: Caso de Uso extendido Ejecutar Análisis.

En la Tabla 4.8 se muestra el caso de uso extendido para el caso de uso de *descargar un informe*.

Caso de Uso	Descargar Informe
Actores	Usuario
Propósito	Permite al usuario descargar un informe en formato Microsoft Word o PDF con los resultados realizados.
Pre-condición	Debe haber al menos un resultado intermedio.
Resumen	Luego de haber realizado uno o más análisis con Minería de Datos, el usuario puede descargar un informe con los resultados obtenidos en el análisis.
Curso Normal de Eventos	
1.- Este caso de uso comienza cuando el usuario ha terminado de ejecutar técnicas de Mienría de Datos y quiera descargar los resultados obtenidos.	
	2.- El módulo verifica que exista al menos un resultado intermedio almacenado en la sesión del usuario y muestra los formatos en que puede descargar los resultados.
3.- El usuario selecciona un formato, ya sea Microsoft Word o PDF.	
	4.- El módulo genera el informe y pregunta al usuario el nombre del archivo y la dirección donde descargarlo.
5.- El usuario ingresa en nombre del archivo y la dirección donde se descargará.	
	6.- Se descarga el archivo al computador del usuario.
7.- Finaliza el Caso de Uso.	
Curso Alternativo de Eventos	
2.- Si no hay al menos un resultado intermedio almacenado no se puede generar un informe.	

Tabla 4.8: Caso de Uso extendido Descargar Informe.

4.4. Diagramas de Secuencia

En esta sección se presentan los diagramas de secuencia, este tipo de diagramas ayudan a identificar las comunicaciones que se producen dentro del módulo de Minería de Datos.

La Figura 4.3 muestra el diagrama de secuencia que corresponde al caso de uso *Seleccionar Fuente de Datos*.

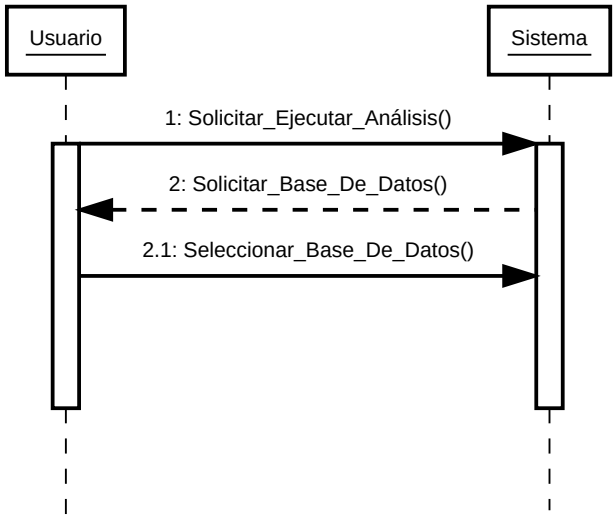


Figura 4.3: Diagrama de Secuencia: Seleccionar Fuente de Datos.

La Figura 4.4 muestra el diagrama de secuencia que corresponde al caso de uso *Seleccionar Técnica de Minería de Datos*.

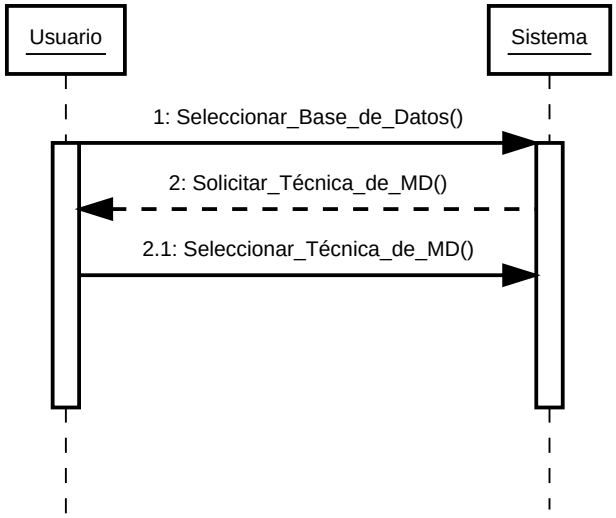


Figura 4.4: Diagrama de Secuencia: Seleccionar Técnica de Minería de Datos.

La Figura 4.5 muestra el diagrama de secuencia que corresponde al caso de uso *Seleccionar Parámetros*.

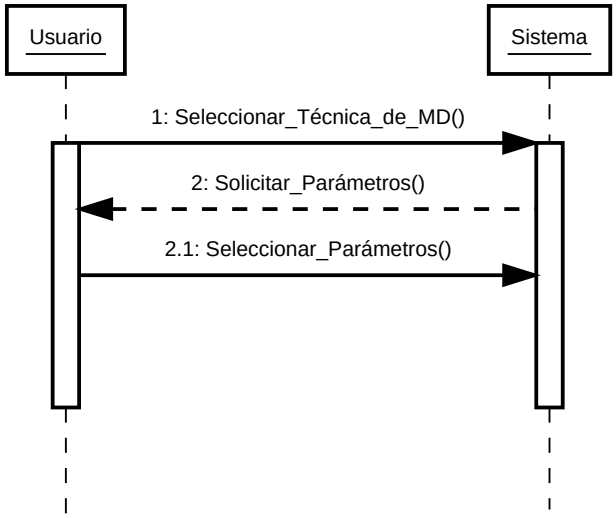


Figura 4.5: Diagrama de Secuencia: Seleccionar Parámetros.

La Figura 4.6 muestra el diagrama de secuencia para el caso de uso *ejecutar análisis* mediante el cual se observa la interacción que tiene el usuario con el sistema cuando necesita realizar un análisis con Minería de Datos.

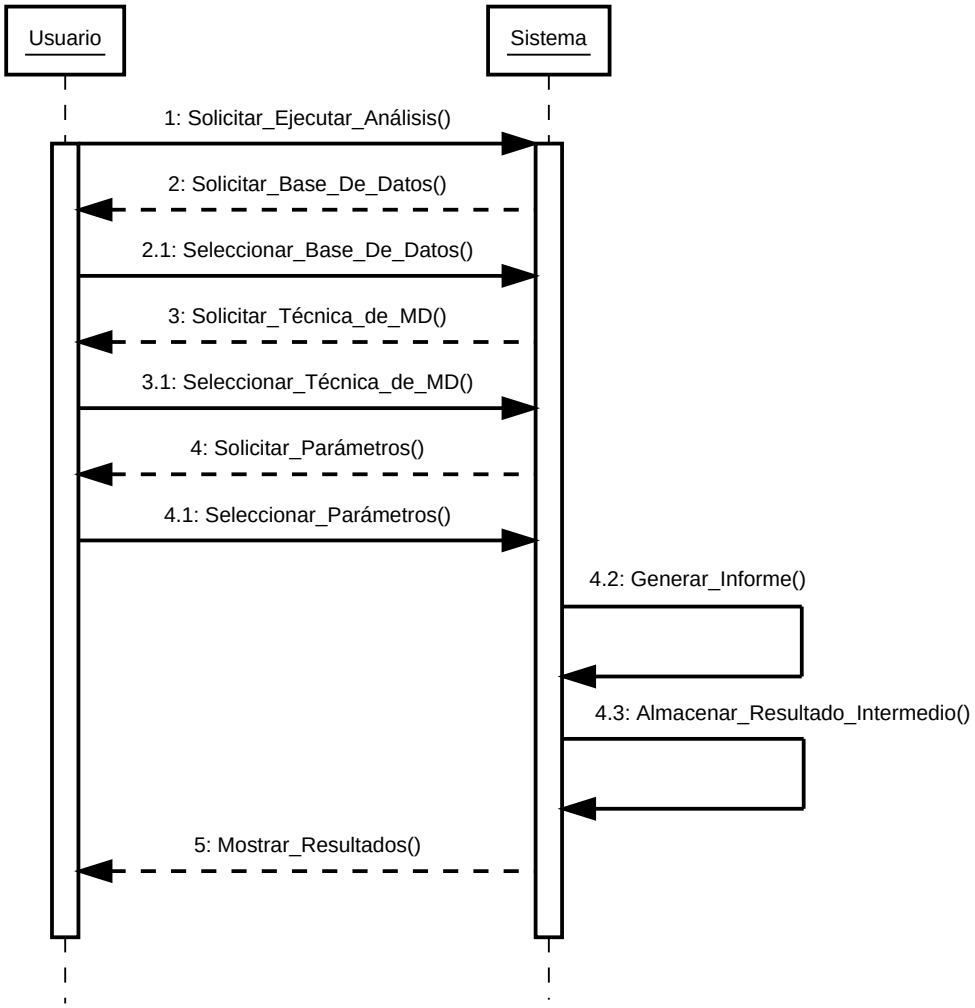


Figura 4.6: Diagrama de Secuencia: Ejecutar Análisis.

La Figura 4.7 muestra el diagrama de secuencia que corresponde al caso de uso *descargar informe*.

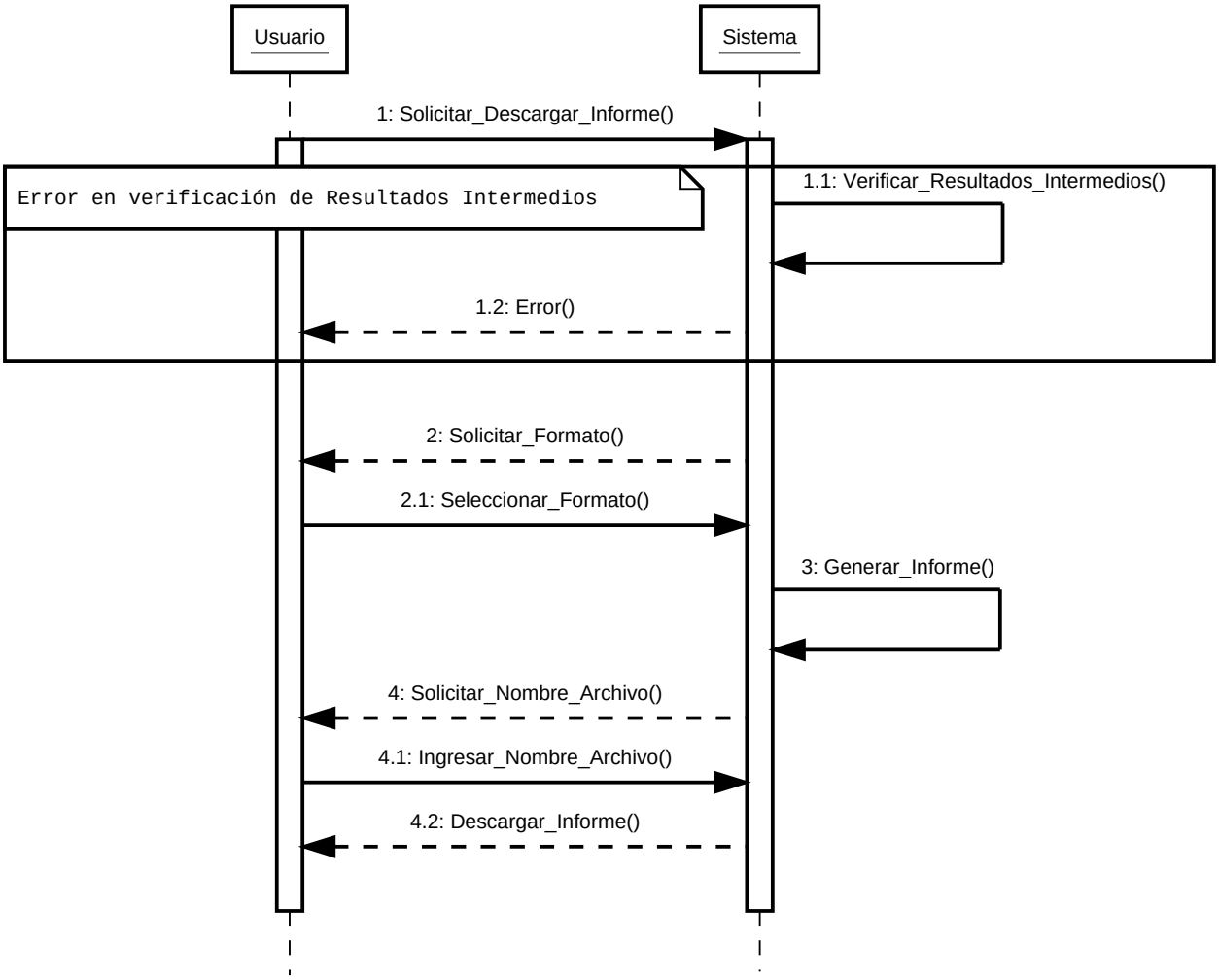


Figura 4.7: Diagrama de Secuencia: Descargar Informe.

4.5. Diagramas de Estado

Los diagramas de estado describen gráficamente los eventos que ocurren, sus transiciones y los estados que median entre esos eventos.

La Figura 4.8 muestra el diagrama de estado para el caso de uso *Seleccionar Fuente de Datos*.

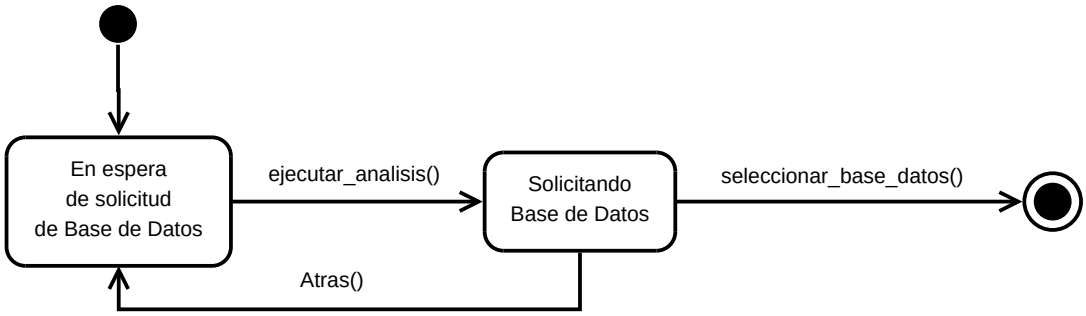


Figura 4.8: Diagrama de Estado: Seleccionar Fuente de Datos.

La Figura 4.9 muestra el diagrama de estado para el caso de uso *Seleccionar Técnica de Minería de Datos*.

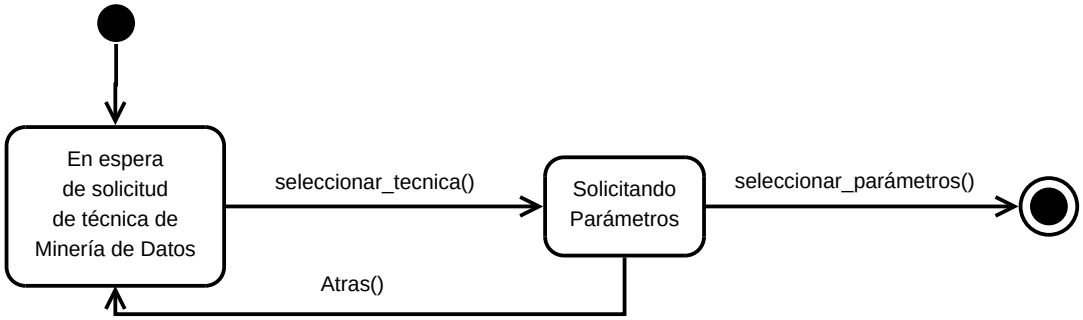


Figura 4.9: Diagrama de Estado: Seleccionar Técnica de Minería de Datos.

La Figura 4.10 muestra el diagrama de estado para el caso de uso *Seleccionar Parámetros*.

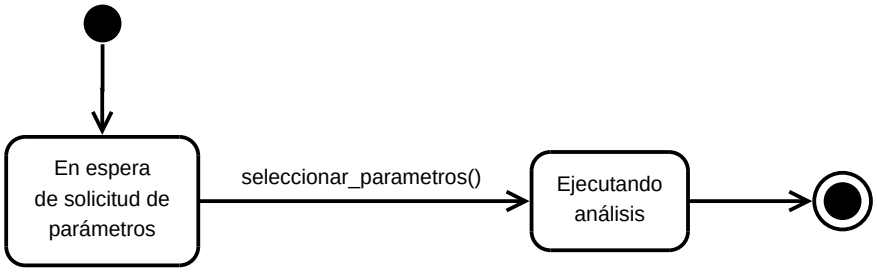


Figura 4.10: Diagrama de Estado: Seleccionar Fuente de Datos.

La Figura 4.11 muestra el diagrama de estado para el caso de uso *ejecutar análisis* donde se observa la secuencia de estados que tiene el sistema para este caso de uso.

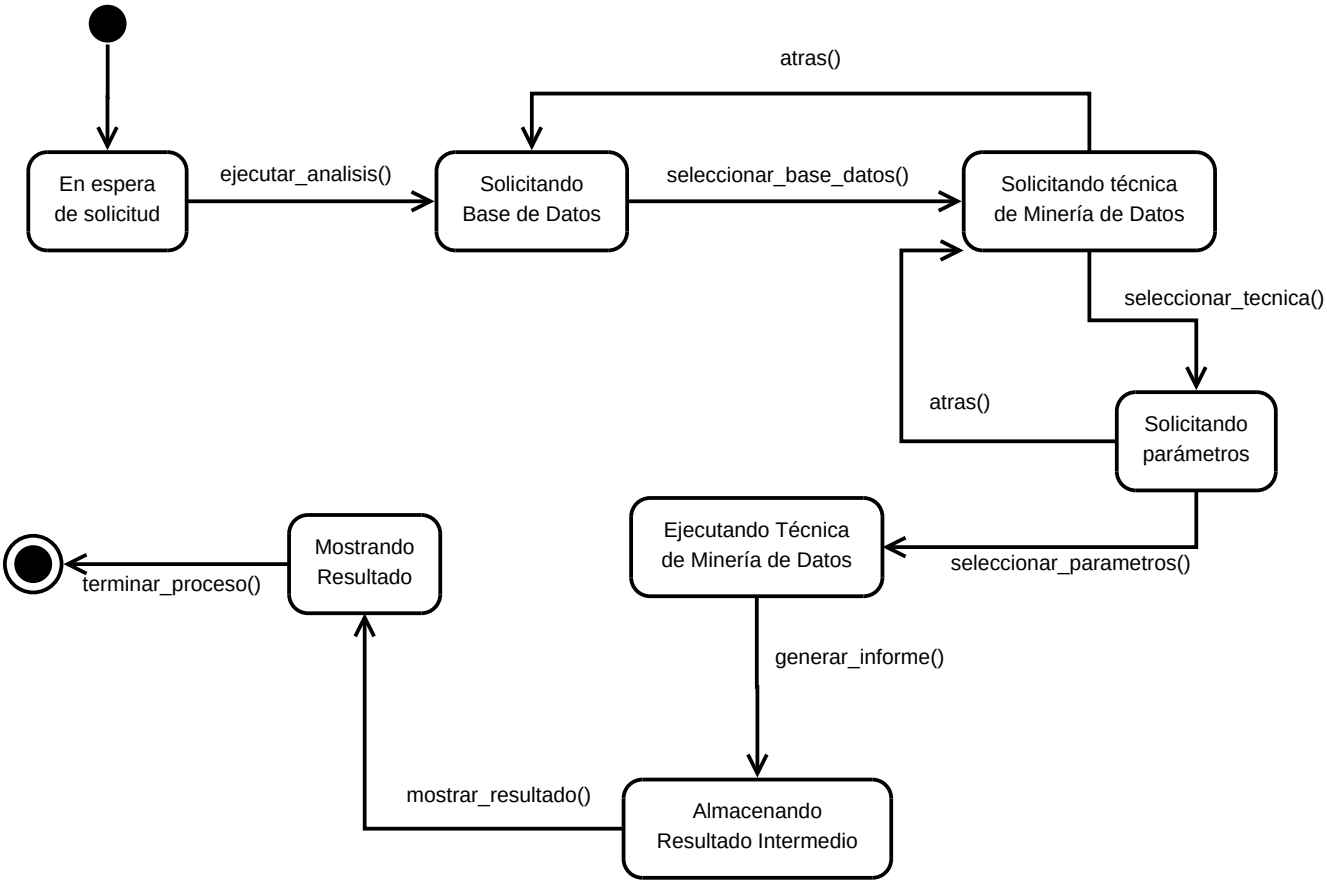


Figura 4.11: Diagrama de Estado: Ejecutar Análisis.

La Figura 4.12 muestra el diagrama de estado para el caso de uso *descargar informe*.

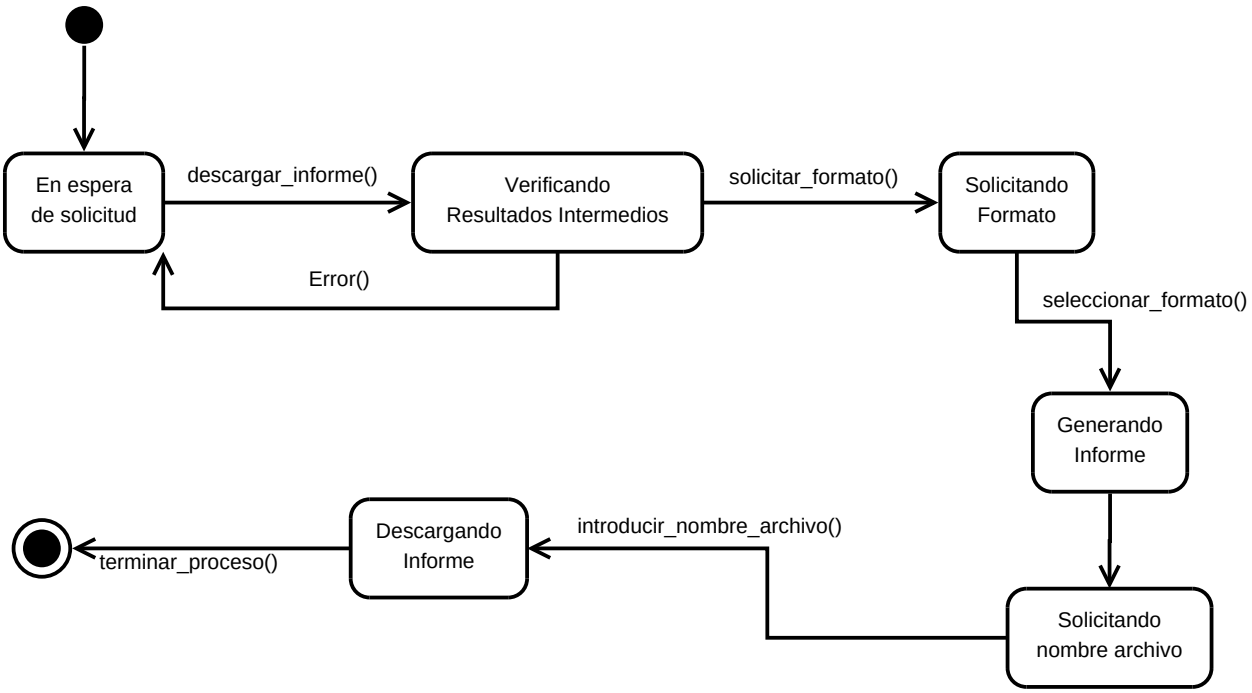


Figura 4.12: Diagrama de Estado: Descargar Informe.

4.6. Modelo conceptual

La Figura 4.13 muestra el modelo conceptual donde se da a conocer los conceptos más significativos del problema.

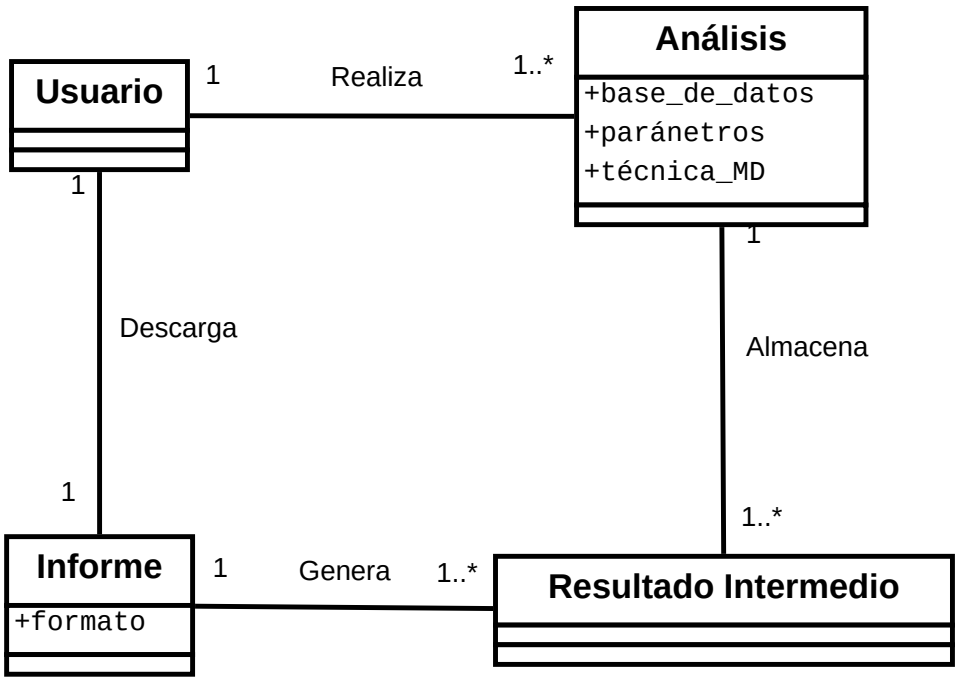


Figura 4.13: Modelo conceptual del sistema.

Bibliografía

[1] Abdullah A. Aljumah, Mohammed Gulam Ahamad, and Mohammad Khubeb Siddiqui. Application of data mining: Diabetes health care in young and old patients. *Journal of King Saud University - Computer and Information Sciences*, 25(2):127 – 136, 2013.

[2] R Balajee and MS Dhanarajan. Mining the information for structure based drug designing by relational database management notion. *Journal of Chemistry*, 6(4):1047–1054, 2009.

[3] Abdelghani Bellaachia and Erhan Guven. Predicting breast cancer survivability using data mining techniques. *Age*, 58(13):10–110, 2006.

[4] Silvia Franceschi Jérôme Vignat Freddie Bray David Forman Martyn Plummer Catherine de Martel, Jacques Ferlay. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *The Lancet Oncology*, 13:607–615, 2012/6/1.

[5] J. Larry Jameson Anthony S. Fauci Stephen L. Hauser Joseph Loscalzo Dan L. Longo, Dennis L. Kasper. *Harrison. Principios de Medicina Interna*. MCGRAW-HILL, 2008. 17 Edición.

[6] Frank Dávila Hernández and Yovannys Sánchez Corales. Técnicas de minería de datos aplicadas al diagnóstico de entidades clínicas. *Revista Cubana de Informática Médica*, 4:174 – 183, 12 2012.

[7] Gobierno de Chile. Metas 2011 - 2020. `http://web.minsal.cl/portal/url/item/c4034eddbc96ca6de0400101640159b8.pdf`. Último acceso 15-04-2014.

[8] Ministerio de la Protección Social. Instituto Nacional de Cancerología E.S.E. `http://www.cancer.gov.co/documentos/Cartillas/Elcancer.pdf`. Último acceso 15-04-2014.

[9] Organización Mundial de la Salud. `http://www.who.int/mediacentre/factsheets/fs297/es/`. Último acceso 15-04-2014.

[10] Organización Mundial de la Salud. http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx. Último acceso 15-04-2014.

[11] Organización Mundial de la Salud. http://www.who.int/mediacentre/factsheets/fs312/es/. Último acceso 15-04-2014.

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

[13] Ministerio de Salud. Vigilancia de diabetes mellitus tipo 2 en chile, reporte nacional. http://epi.minsal.cl/epi/html/presenta/Taller2011/Dia3/08_VENT_DM.pdf. Último acceso 15-04-2014.

[14] Ministerio de Salud. Vigilancia epidemiologica de cancer. http://epi.minsal.cl/epi/html/presenta/Taller2011/Dia3/03_Vigilancia_de_Cancer.pdf. Último acceso 15-04-2014.

[15] Ministerio de Salud. Primer informe de registros poblacionales de cáncer de chile quinquenio 2007-2007. http://epi.minsal.cl/epi/0notransmisibles/cancer/INFORME%20RPC%20CHILE%202003-2007,%20UNIDAD%20VENT,%20DEPTO.EPIDEMIOLOGIA-MINSAL,13.04.2012.pdf, 2012. Último acceso 15-04-2014.

[16] DEIS. Defunciones y mortalidad por causas. http://www.deis.cl/defunciones-y-mortalidad-por-causas/. Último acceso 29-04-2014.

[17] Ministerio de Salud Departamento de Estadísticas e Información de Salud. Indicadores básicos de salud chile 2011. http://deis.minsal.cl/deis/indicadores/Folleto_IBS_2011.pdf. Último acceso 15-04-2014.

[18] International Diabetes Federation. Idf diabetes atlas, 6th edn. http://www.idf.org/diabetesatlas. Último acceso 15-04-2014.

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

[20] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Morgan kaufmann, 2012. Third Edition.

[21] César Ferri Ramírez José Hernández Orallo, María José Ramírez Quintana. *Introducción a la minería de datos*. Pearson, 2004.

[22] Y.-J. Lee, O. L. Mangasarian, and W. H. Wolberg. Survival-time classification of breast cancer patients. *Comput. Optim. Appl.*, 25(1-3):151–166, March 2003.

[23] Oded Maimon and Lior Rokach, editors. *Data Mining and Knowledge Discovery Handbook, 2nd ed*. Springer, 2010.

[24] SEREMI región de Valparaíso. `http://seremi5.redsalud.gob.cl/?page_id=70`. Último acceso 15-04-2014.

[25] SEREMI región de Valparaíso. `http://seremi5.redsalud.gob.cl/?p=1013`. Último acceso 15-04-2014.

[26] Dra. M Teresa Valenzuela. `http://www.sabin.org/sites/sabin.org/files/oct21_1000valenzuela.pdf`. Último acceso 15-04-2014.