PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

ESTUDIO DEL COMPORTAMIENTO DE LA MORTALIDAD EN CHILE BASADO EN LA MINERÍA DE DATOS Y ANÁLISIS PREDICTIVOS

DAVID MATIAS CORNEJO ROJAS
ALEXANDER PAOLO GONZÁLEZ ILUFI
Profesor guía: Pamela Hermosilla Monckton

INFORME FINAL DEL PROYECTO
PARA OPTAR AL TÍTULO PROFESIONAL DE
INGENIERO CIVIL EN INFORMÁTICA

ABRIL 2017

Dedicatoria

En primer lugar, agradezco a mi familia por el sacrificio realizado durante 6 años. Agradezco también a cada una de las personas que fueran partícipe de este proceso académico y me ayudaron a ser una mejor persona. Agradezco con quienes compartí tardes de estudios y que fueron de mucha ayuda para sobrellevar las asignaturas. Finalmente dedico los logros obtenidos a quienes no pudieron estar ahí físicamente, en especial mis bisabuelos. Salud.

David Cornejo Rojas

Dedico estos agradecimientos a mi familia, amigos, y a las personas esenciales en mi vida, que fueron mi soporte y siempre me brindaron su apoyo en todos los aspectos posibles, acompañándome durante esta larga transición, desde el ser un estudiante en busca de un futuro mejor, hasta convertirme en un profesional. Gracias totales.

Alexander González Ilufi

Índice

Lista de figuras

Resumen

La creciente tasa de mortalidad en el país representa un relevante problema hoy en día. Cada año fallecen más personas y con el paso de los años las causas también aumentan. Existen muchas causas de fallecimiento, por lo que es de vital importancia analizar estas causas y las variables que hay detrás para poder comprender como se comporta la mortalidad en Chile.

Es por ello que, en el presente trabajo, se propone una metodología para encontrar, analizar y medir los factores que predicen si una persona fallece o no basándonos en la información disponible, usando técnicas de minería de datos, así como modelos estadísticos y de base de datos. Por lo que cada concepto relacionado a la minería de datos y que necesita ser entendido será explicado en la primera parte del informe, y posterior a eso se describirá todo lo que implica la implementación de la solución final.

Finalmente, este informe será concluido con la interpretación de los resultados conseguidos, acompañando de tablas e imágenes para ofrecer una lectura clara de lo logrado.

Palabras claves: minería de datos, aprendizaje automático, selección de atributos, algoritmos de clasificación, ingeniería informática.

Abstract

The rising mortality rate in the country represents a relevant problem today. Every year more people die and over the years causes also increase. There are many death causes, so it is vitally important to analyze these causes to understand how mortality behaves in Chile.

Therefore, in the present work, a methodology is proposed to find, analyze and measure the factors that predict whether a person dies based on available information, using data mining techniques, as well as statistical and base models of data. So, that each concept related to data mining and that needs to be understood will be explained in the first part of the report, and after that will describe everything that implies the implementation of the final solution.

Finally, this report will conclude with the interpretation of the results, accompanied by tables and images to provide a clear reading of what has been achieved

Keywords: data mining, machine learning, selecting variables, classification algorithms, computer engineering.

1. Introducción

En la actualidad el uso adecuando de los datos y la información se traduce en dinero para las organizaciones tanto en la banca privada como en la banca pública. Y por cada segundo que pasa, la recolección de datos e información y su acumulación, aumenta de forma exponencial y automatizada. En este contexto, darle un significado a la información y a los datos almacenados en las grandes bases de datos de las organizaciones, permite generar conocimiento valioso, que les brinda una gran ayuda en los procesos de toma de decisiones, y junto con ello una ventaja competitiva. Es así como la minería de datos, nos brinda las técnicas y herramientas necesarias para lograr éxito en generar tal conocimiento.

Por consiguiente, en el presente informe se darán a conocer los conceptos más importantes relacionados con la minería de datos y orientados a los objetivos definidos en los párrafos posteriores, que serán los conceptos bases para la implementación de las técnicas y las herramientas aplicadas en una base de datos. Se busca establecer en primera instancia las bases que permitan conocer y estar al tanto de las herramientas a utilizar, para posteriormente implementarlas y así obtener resultados que apoyen a la toma de decisiones.

Finalmente, se especificarán las técnicas o pasos ejecutados con la herramienta seleccionada, que permiten poder realizar las pruebas, junto con los resultados.

2. Marco general del proyecto

2.1. Situación en Estudio

La muerte es algo que afecta a cualquier tipo de sociedad y a cada individuo eventualmente, y si bien todos crecemos en distintas condiciones, existen determinadas variables que son semejantes para todos dentro de alguna zona geográfica. Los problemas surgen cuando existe información de la población, pero no es explotada para fines preventivos. Dado lo anterior, en Chile está disponible mucha información relacionada a la salud, así como del concepto de mortalidad y todo lo que gira en torno a él. Por lo que el estudio que se llevará a cabo y expuesto en este informe, apunta a determinar las variables que más impacto tienen en cada individuo y que causaron su fallecimiento, como también quienes bajo condiciones similares no fallecieron. Y con estos resultados, en algún caso poder predecir la probabilidad de muerte de alguna persona con fines preventivos.

2.2. Definición de Objetivos

2.2.1. Objetivo General

El objetivo principal del proyecto es, a través la utilización de métodos de selección de atributos y algoritmos predictivos aplicados sobre un set de datos determinado, analizar el comportamiento de la mortalidad en Chile. En este sentido, mediante la comparación de diversos métodos de selección, así como de algoritmos predictivos, se pretende mejorar la precisión del estudio con el fin de obtener resultados realistas.

2.2.2. Objetivos Específicos

1. Investigar y analizar técnicas, algoritmos, y herramientas de minería de datos para determinar las que se utilizarán posteriormente.
2. Realizar un modelo de DW que represente de manera óptima la información con la que se cuenta.
3. Aplicar métodos de selección de atributos para definir aquellas variables que son significativas para el estudio.
4. Usando la información obtenida en (3), emplear algoritmos predictivos.
5. Validar los resultados con la(s) persona(s) encargadas de la base datos.
6. Realizar posibles ajustes a los modelos.

2.3. Orientación del proyecto

Este proyecto consta de dos grandes etapas identificables. La primera, está orientada a esclarecer todo lo necesario para una implementación correcta. En este sentido, este informe apunta hacia la predicción dentro del campo de la minería de datos, por lo que en el estado del arte se definirán las principales herramientas que nos ayudarán en ámbitos predictivos; especificando aquellos algoritmos y técnicas que serán usados. La segunda etapa, por

consiguiente, estará orientada a todo lo que tiene que ver con la implementación de lo definido anteriormente, con una muestra de resultados en ámbitos predictivos.

3. Marco teórico

3.1. Conceptos básicos

Para entender la minería de datos debemos entender cómo funciona el proceso al cual pertenece y los conceptos básicos que lo rodean. Es por ello, que se procederá a definir los conceptos básicos de este contexto.

Dato(s): son la mínima unidad semántica, que por sí solos son irrelevantes como apoyo en la toma de decisiones. Por ejemplo: El número 3000.

Información: la información se puede definir como un conjunto de datos procesados y que poseen un significado (relevancia, propósito y contexto), y que por lo tanto nos sirven para disminuir la incertidumbre en el proceso de toma de decisiones. Por ejemplo: Las ventas del mes de noviembre fueron de 3000.

Conocimiento: el conocimiento es una mezcla de experiencia, valores, información y know.how que sirve como marco para la incorporación de nuevas experiencias e información, y que es útil para todos los niveles dentro de las organizaciones debido a su relevancia en el proceso de toma de decisiones. A su vez, el conocimiento deriva de la información, así como la información deriva de los datos. Por ejemplo, Las ventas del mes de noviembre fueron de 3000. Noviembre es el mes de menos ventas.

Descubrimiento de conocimiento en bases de datos (*Knowledge Discovery in Databases*, *KDD*): proceso mediante el cual se extrae la información necesaria para que sea utilizada por una organización y que facilite la toma de decisiones para obtener beneficios. El proceso consta de diversas etapas:

i. Selección: Se determina la fuente de datos y el tipo de información con el que se trabajará. Se conocen las variables significativas para darle solución a la problemática.
ii. Pre-procesamiento: Se realiza un proceso de DQ para que la información quede manejable para las fases posteriores (se mitiga el ruido, se eliminan datos anómalos o fuera de rango, entre otros)
iii. Transformación: Se realiza transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos adecuada para el problema. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente.
iv. Minería de Datos (Data Mining, DM): Mediante la aplicación de algún algoritmo o técnica de DM se establecen las tendencias y patrones previamente desconocidos, potencialmente útiles y comprensibles.

v. Interpretación y Evaluación: Se identifican los patrones obtenidos y que son realmente interesantes para la organización. Además, se realiza una evaluación de los resultados obtenidos. En caso de no ser óptimos, se vuelve a aplicar técnicas de DM en busca de mejores resultados.

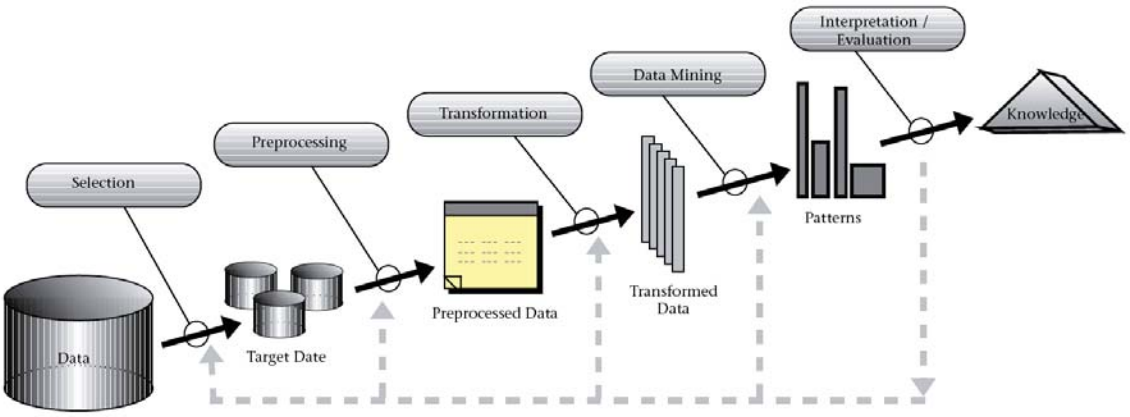

Figura 3.1. Proceso KDD. [1]

Inteligencia de Negocios (*Business Intelligence, BI*): es el conjunto de estrategias y aspectos relevantes enfocados a la creación de datos en información y a su vez la información en conocimiento para apoyar la toma de decisiones. [2]

Minería de Datos (*Data Mining, DM*): campo de las ciencias de la computación que comprende el proceso de detección de patrones en grandes volúmenes de datos.

Aprendizaje Supervisado (*Machine Learning, ML*): tipo de método predictivo utilizado para solucionar la problemática.

Base de Datos (*Data Base, DB*): en una base de datos operacional se almacenan todas las transacciones de la organización, tanto datos útiles como no útiles.

Almacén de Datos (*DataWarehouse, DW*): en un DW se almacena toda la información de interés para una organización que luego queramos analizar.

Proceso Extracción, Transformación y Carga (*Extract, Transformation and Load Process, ETL*): proceso que permite a las organizaciones mover datos desde diversas fuentes (*Extract*), reformatearlos, limpiarlos (*Transformation*) y cargarlos en otra base de datos o almacen de datos para apoyar el proceso de negocios.

Calidad de Datos (*Data Quality, DQ*): la calidad de los datos es una percepción o una evaluación de la aptitud de los datos que cumplan un propósito en un contexto determinado.

Datos discretos: son datos que solo pueden tener ciertos valores.

Datos continuos: son datos que pueden tomar cualquier valor (dentro de un rango determinado).

Entropía: mide la incertidumbre de una fuente de información.

Ruido: se refiere a aquellos datos que contienen errores o valores anómalos.

Oversampling: Es una técnica que permite balancear la instancia en menor cantidad de una variable de clasificación.

3.2. Minería de datos

El proceso de DM es en el cual se interpretan los datos y se extraen patrones (información) que se encuentran ocultos o son desconocidos a simple vista desde grandes DB, permitiendo a partir de estos predecir futuras tendencias y comportamientos, con el fin de convertir dicha información en algo útil que represente una ventaja (potencialmente económica) para una organización. Mediante un conjunto de tecnologías, técnicas o métodos de aplicación, el proceso de DM ayuda a las organizaciones a enfocar de mejor manera sus objetivos y alcances a partir del análisis e interpretación de su propio DW.

DM es la fase de análisis dentro del proceso KDD, el cual consta de diversas etapas en las que se trabaja sobre la DB de tal manera de ajustarla dependiendo del propósito del ámbito de trabajo.

3.2.1. Importancia de DM

El proceso de DM se hace necesario debido a:

- A medida que el volumen de la DB se incrementa, las personas que entienden esta información decrece.
- La información esta explicita, pero lo realmente significativo para tomar ventaja de esta información es lo que está implícito u oculto.
- En algunas situaciones, el acceso a los datos no es sencillo.
- Se perfecciona la estrategia organizacional y se obtienen beneficios a partir del valor agregado que este proceso entrega

3.2.2. Modelos de DM

Los modelos de DM se clasifican en dos grandes grupos, dependiendo del propósito del ámbito de trabajo: por un lado, están el modelo descriptivo, el cual describe el

comportamiento de la DB, por lo que se establecen correlaciones, agrupamientos y asociaciones entre los datos; por el otro está el modelo predictivo, cuyo objetivo es predecir el comportamiento futuro de la DB mediante el análisis de tendencias de la información histórica de la misma.

A su vez cada uno de estos métodos posee variadas técnicas, las cuales se describen a continuación:

i. Modelo Descriptivo
   - Aprendizaje no Supervisado: La DB no tiene asociaciones, el objetivo es detectar regularidades en los datos (ya sea agrupaciones, asociaciones, valores anómalos) Por ejemplo: Responder a preguntas del tipo ¿Cuantos grupos hay? ¿De cuántos datos están conformados?

ii. Modelo Predictivo
   - Aprendizaje Supervisado: Cada observación incluye un valor asociado a la clase correspondiente. Se aprende un clasificador.
     Por ejemplo: 9-7 = 2; 7-4 = 3; 4+0 =?
   - Predicción Secuencial: Las observaciones están ordenadas secuencialmente. El objetivo es predecir el siguiente valor en la secuencia. Por ejemplo: 2,4,6,8...?
   - Interpolación: Es una función continua aplicada en varias dimensiones.

3.2.3. Clases de DM

Dependiendo del propósito del análisis, se pueden determinar diversas clases (o funciones) de DM, las cuales tienen sus respectivas respuestas. Esto ayuda al momento de tener clara la respuesta esperada, ya que para cada clase de DM existen distintos algoritmos a utilizar. Las clases que abarca el proceso de DM son:

i. Clasificación: Utilizada para aquellos problemas donde se pretende identificar grupos de información en términos de sus atributos. Los problemas de clasificación constituyen aprendizaje supervisado, donde el conjunto de clasificación de salida pertenece a alguna de las clases identificadas en el conjunto de entrada. El propósito de la clasificación es encontrar algún tipo de relación entre los atributos de entradas y el conjunto de salida, con el fin de que ese conocimiento sea utilizado para predecir la clase de un nuevo objeto desconocido.
ii. Regresión: Se encarga de establecer relaciones entre series de objetos. Al igual que la clasificación, los problemas de regresión constituyen un aprendizaje supervisado, por lo que su objetivo principal se basa en la predicción. A diferencia de la clasificación, el conjunto de salida es un valor numérico continuo o un vector, en vez de una clasificación discreta. Entonces, dando un objeto es posible predecir uno de sus

atributos por medio de otros atributos, utilizando el modelo construido. La predicción de valores numéricos se puede realizar por métodos estadísticos.

iii. Problemas Temporales: Es lo mismo que un problema de regresión, pero agregando el factor de tiempo. En algunos problemas se hace imperante la necesidad de conocer el tiempo en que se desarrolla cierta información (por ejemplo, la estacionalidad de los frutos durante el año), por lo que esta información es útil para determinar tendencias o comportamientos que abarquen un periodo específico de tiempo.

iv. Agrupamiento: El objetivo principal es buscar comportamientos, características o propiedades similares en la DB para crear grupos o subconjuntos. El problema de agrupamiento es del tipo no supervisado. Se utiliza para construir subconjuntos representativos, analizar correlación entre los atributos, es decir, el conjunto de salida (o subconjuntos de salida) describen de manera óptima la DB según los criterios o propiedades establecidas.

v. Asociación: Utilizada en problemas para reconocer que la presencia (o ausencia) de un conjunto de datos implica la presencia (o ausencia) de otro conjunto de datos. Es del tipo no supervisado.

vi. Secuencialidad: El propósito es reconocer la secuencia existente entre dos conjuntos de datos. Constituye un problema no supervisado.

vii. Dependencia: Consiste en la construcción de un modelo que describe la dependencia existente entre atributos o propiedades de la DB. Esta dependencia es del tipo "causa-efecto" (por ejemplo: si cierto atributo es verdadero, en consecuencia, otro atributo en respuesta también será verdadero).

Los problemas de clasificación, regresión y series temporales pertenecen al modelo predictivo, mientras que los problemas de agrupamiento, asociación, secuencialidad, dependencia son utilizados en el modelo descriptivo.

3.2.4. Modelo Descriptivo

Objetivo

Mediante la utilización del modelo descriptivo se busca describir el comportamiento de la DB. Esto se realiza a través de algoritmos que establecen reglas de:

- Asociación
- Dependencia
- Agrupamiento
- Secuencialidad

Métodos

Los métodos (algoritmos) no supervisados que establecen dichas reglas son:

- *k-NN* (*Nearest Neighbour*): Regla del vecino más cercano. Según el conjunto de datos, se conecta cada punto con el punto más cercano a él, de esta forma se generan los grupos.
- *k-means Clustering*: Se utiliza para encontrar los k puntos más densos en un conjunto de datos
- Centroides
- SOM (*Self-Organizing Maps*) o Redes Kohonen
- Cobweb
- Algoritmo AUTOCLASS

3.2.5. Modelo Predictivo

Objetivo

El objetivo principal de utilizar el modelo predictivo es, además de describir el comportamiento de la DB, predecir tendencias futuras a priori desconocidas.

La extracción de patrones ayuda a determinar potenciales oportunidades o riesgos para la organización. Los modelos predictivos establecen relaciones entre diversos factores y condiciones significativas dentro de la DB, por lo que el análisis del resultado de dicho proceso se considera crítico para la toma de decisiones de la organización.

Métodos

Al igual que el modelo descriptivo, el modelo predictivo también posee diversas técnicas y algoritmos, entre los cuales se destacan:

- *k-NN*
- *K-means Clustering*
- *Perceptron Learning*
- Redes Neuronales Multicapa (*MultiLayer Perceptron*)
- *Radial Basis Function*
- Árboles de Decisión (*Decision Trees*)

- *Naive Bayes Classifiers*
- *Center Splitting*

3.3. Estado del arte

3.3.1. Modelamiento

En ámbitos de inteligencia de negocios (*Business Intelligence*) existe una serie de necesidades con respecto al almacenamiento de la información (accesibilidad, disponibilidad, entre otros), por lo que se vuelve imperante la creación un almacén de datos (*Datawarehouse*), ya que estos son capaces de:

- Centralizar la información en un solo lugar
- Administrar grandes cantidades de información
- Guardar información histórica
- Condensar y agregar información nueva
- Controlar el acceso a la información
- Ayuda a toma de decisiones

Los almacenes de datos están diseñados para agilizar la consulta de una gran cantidad de datos (especialmente OLAP, *On-Line Analytical Processing*), para esto se utilizan bases de datos multidimensionales. Las bases de datos multidimensionales permiten tener un acceso flexible a los datos, para así poder analizar y explorar las relaciones existentes y así lograr resultados que faciliten la toma de decisiones.

En estas bases la información se almacena través de tablas de hechos (o *tabla fact*) y tablas de dimensión, en los cuales los hechos contienen los datos del estudio, es decir, son el objeto de los análisis y las dimensiones representan factores que describen a los hechos, es decir, proveen el medio para analizar el contexto del negocio. Estas tablas se relacionarán dependiendo el esquema que se pretende diseñar, los cuales son:

- Esquema Estrella
- Esquema Copo de nieve
- Esquema Constelación

Esquema Estrella

El esquema consta de una tabla de hechos central que contiene los principales datos de análisis, rodeada por las tablas de dimensiones. La relación existente entre hechos y dimensiones se establece mediante la inclusión de las claves primarias de las dimensiones como claves foráneas en la tabla de hechos. Este esquema es el más simple de interpretar, además de optimizar el tiempo de respuesta de las consultas de los usuarios. Sin embargo, su carga es lenta, así como su construcción.

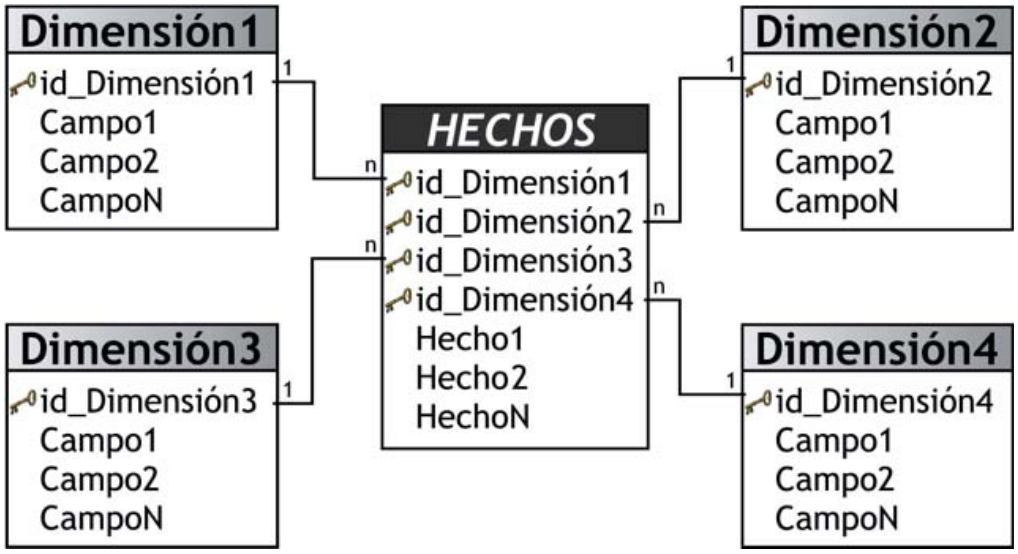

Figura 3.2. Esquema Estrella [3]

Esquema Copo de nieve

Este esquema es una extensión del modelo estrella cuando las dimensiones se organizan en jerarquías de dimensiones. Existe una tabla de hechos central que está relacionada con una o más tablas de dimensiones, quienes a su vez también están relacionadas con otras tablas de dimensiones.

Este esquema presenta una estructura flexible, ya que segrega los datos de las tablas de dimensiones ya existentes y provee un mayor grado de análisis. Sin embargo, al ser más complejo existe una baja de rendimiento debido a la creación de más tablas de dimensiones y más relaciones entre ellas, por lo que las consultas de usuarios se tornan más lentas.

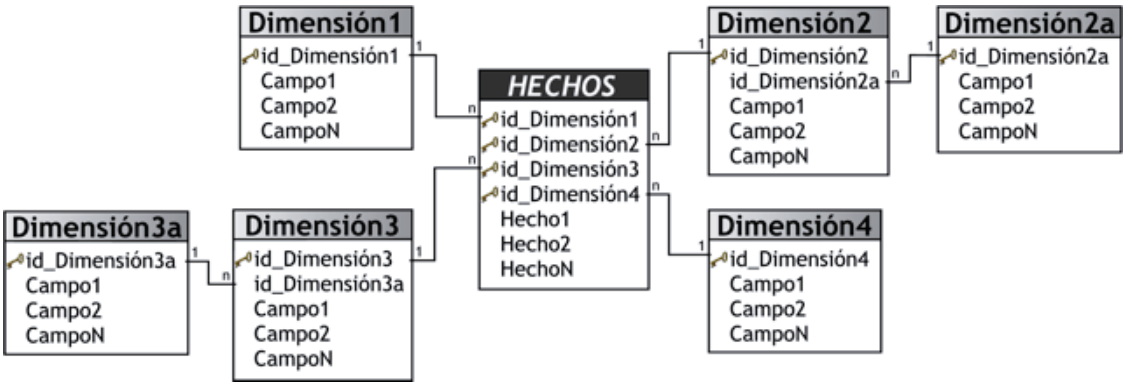

Figura 3.3. Esquema Copo de Nieve [3]

Esquema Constelación

Este modelo está compuesto por una serie de esquemas en estrella, y tal como se puede apreciar en la siguiente figura, está formado por una tabla de hechos principal ("HECHOS_A") y por una o más tablas de hechos auxiliares ("HECHOS_B"), las cuales pueden ser resúmenes de la tabla principal. Dichas tablas yacen en el centro del modelo y están relacionadas con sus respectivas tablas de dimensiones. No es necesario que las diferentes tablas de hechos compartan las mismas tablas de dimensiones, ya que, las tablas de hechos auxiliares pueden vincularse con solo algunas de las tablas de dimensiones asignadas a la tabla de hechos principal, y también pueden hacerlo con nuevas tablas de dimensiones.

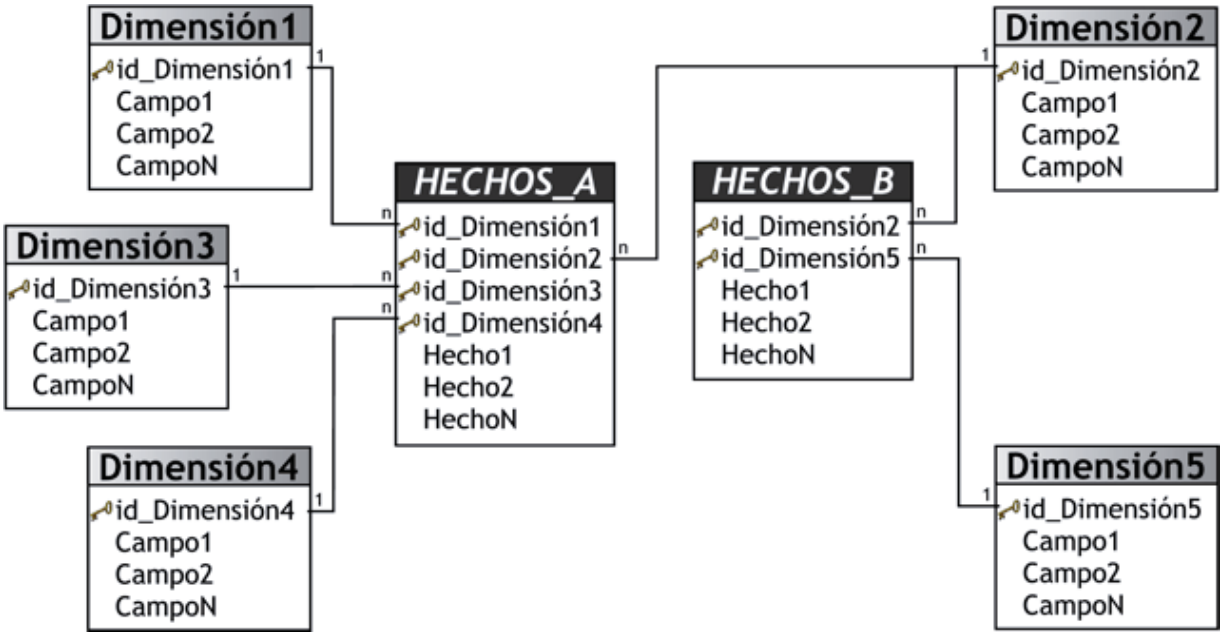

Figura 3.4. Esquema Constelación [3]

Su diseño y características son bastante similares a las del esquema estrella, pero posee una serie de diferencias con el mismo, que son precisamente las que lo destacan. Este esquema permite tener más de una tabla de hechos, por lo cual se podrán analizar más aspectos claves del negocio con un mínimo esfuerzo adicional de diseño, además de esto contribuye a la reutilización de las tablas de dimensiones, ya que una misma tabla de dimensión puede utilizarse para varias tablas de hechos.

3.3.1. Métodos de Selección de Atributos

Luego del realizar el proceso de modelamiento del DW o de haber recopilado la mayor cantidad de información para el estudio, es preciso saber que variables son más determinantes para la obtención de mejores resultados a posteriori, por lo que se necesitará realizar un proceso de filtro usando métodos de selección de atributos (*Feature Selection*).

Algunos atributos pueden ser redundantes, irrelevantes o incluso puede resultar complejo poder interpretarlos, por lo que la selección de atributos nos ayuda a:

- Reducir el número de atributos, lo que afecta directamente a la complejidad del modelo con respecto a los datos disponibles.
- Cuantos menos atributos, más fácil poder interpretar el modelo.
- Mejorar la capacidad de generalización (eliminando atributos redundantes e irrelevantes).
- Acelerar el proceso de aprendizaje.

Algunos algoritmos son capaces de descartar características (como C4.5, arboles de decisión), pero también existen algoritmos de selección de atributos que se pueden usar para pre procesar los datos. Estos algoritmos se categorizan en dos apartados: [4]

- Ranking: Evaluación y ordenación de atributos de manera individual y eliminación de los menos valorados
- Selección de Subconjunto: Búsqueda del subconjunto de atributos más interesante

Ranking

Dado un subconjunto X_1, X_2…..X_n, se evalúa cada X_i (i = 1,2…n) de manera independiente, calculando medidas de correlación del atributo con la clase. Un atributo X_i esta correlacionado con la clase, si al conocer su valor este logra predecir la clase con cierta probabilidad. (Por ejemplo, la enfermedad de una persona esta correlacionada con el hecho de que está viva o muera). Existen varios criterios para evaluar a los atributos (una vez evaluados los atributos, se quitan los *k* peores:

- Entropía (*Information Gain*)
- Chi.cuadrado (*Chi-square*)
- Información Mutua (*Mutual Information*)

La ventaja de esta evaluación es su rapidez, sin embargo, no elimina atributos redundantes ni tampoco detecta aquellas variables que funcionan bien de manera conjunta, pero mal de manera separada.

Selección de Subconjunto

Este método recorre un espacio de búsqueda en el cual va evaluando subconjuntos completos de atributos con el fin de determinar cuál(es) es (son) el (los) más relevante(s). No se recorre el espacio entero, sino solo aquellos subconjuntos más prometedores, los cuales se evalúan de manera conjunta. Para esto hay que definir:

- Una manera de moverse por el espacio de búsqueda de subconjuntos de atributos:
  - Hacia delante (*Forward Propagation*)
  - Hacia atrás (*Back Propagation***)**
- Una medida para evaluar subconjuntos de atributos:
  - CFS (*Correlation Feature Selection*)
  - Wrapper

3.3.2. Técnicas y algoritmos predictivos

A modo de centrarse en lo práctico del proyecto, los fundamentos que definen a los árboles de decisión y al Perceptrón multicapa (técnicas y algoritmos predictivos) que aplicaremos, se encuentran en el Anexo A.

3.3.2. Software para minería de datos

La información detallada de los softwares se encuentra en el Anexo B.

4. Solución Propuesta

En este capítulo se presentará el diseño de la solución del problema, el cual será la base para la posterior implementación de las técnicas de DM. Así, se definirán los componentes y características de la solución propuesta.

A su vez en este capítulo, luego de que se realizará un análisis de las herramientas, se dejará establecido aquella con la que se pretende trabajar. En este aspecto, se llega a la decisión de que Weka será aquella herramienta que nos ayudará en la solución. Principalmente, se hace esta elección porque es un software libre, portable, fácil de usar, rápido en el procesamiento de información, y como se vio anteriormente, posee las técnicas definidas que se requieren en la aplicación.

4.1. Descripción General

A continuación, se define el diseño de la solución propuesta, con las etapas a realizar en cada una de ellas hasta lograr la obtención de resultados:

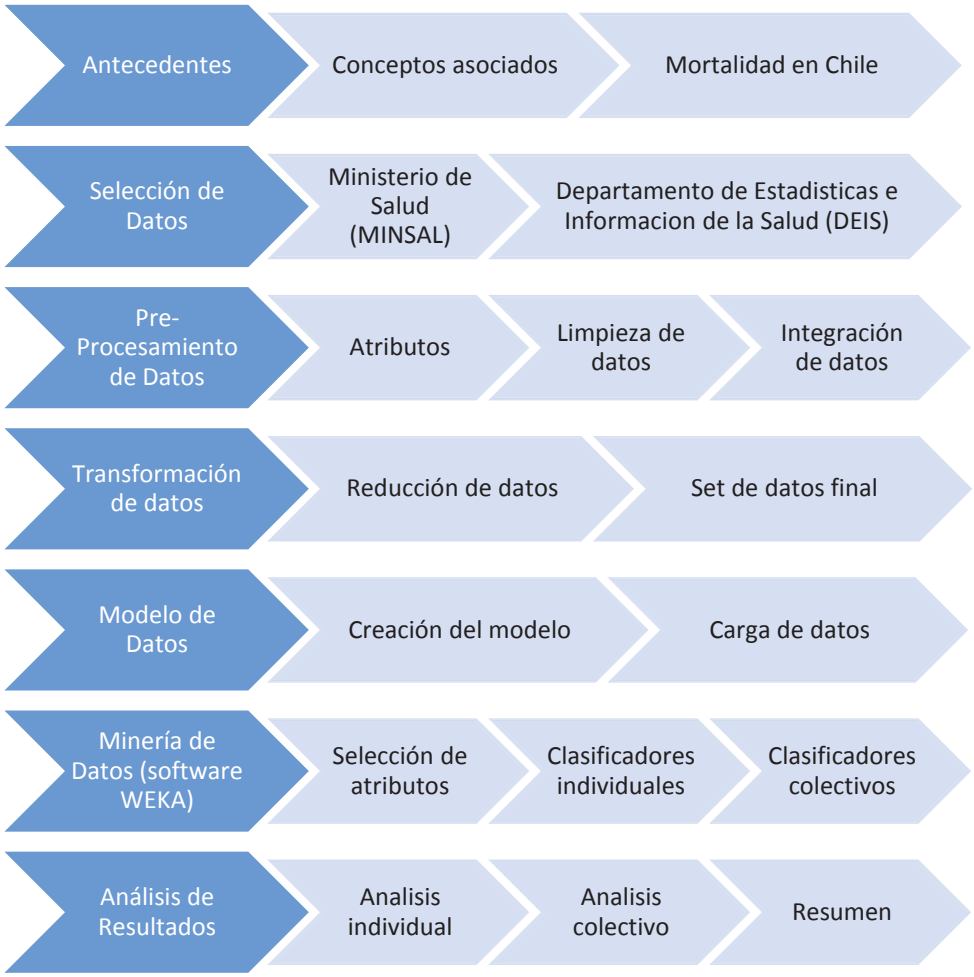

Figura 4.1Etapas de la solución propuesta

4.2. Antecedentes a considerar

Los antecedentes que plasman la información respecto de la temática de la mortalidad en Chile, se encuentran en el Anexo C.

4.3. Selección de Datos

La información que será utilizada para aplicar las técnicas fue seleccionada desde el Departamento de Estadísticas e Información de la Salud (DEIS), el cual es un departamento perteneciente al Ministerio de Salud (MINSAL). [19]

Este departamento es un referente técnico-estadístico y estratégico en la producción de información y estadísticas de la Salud a nivel nacional, por lo que posee la información necesaria con respecto a la problemática propuesta.

La información provista por el sitio se categoriza en tres apartados:

- Defunciones y mortalidad general y por grupos de edad
- Defunciones y mortalidad por causas
- Series y gráficos de mortalidad

Cada uno de los apartados posee una serie de archivos (en su mayoría con extensión .xls), por lo que, para efectos de selección de datos, se toman aquellos archivos que muestren información histórica.

El sitio web del DEIS, a su vez dispone de bases de datos accedidas por medio de un formulario, que entregan un registro detallado de cada persona, categorizado por los siguientes apartados:

- Bases de datos REM
- Bases de datos Algunas Atenciones (REM)
- Bases de datos Defunciones
- Bases de datos REMSAS
- Bases de datos Egresos
- Bases de datos Nacimientos
- Bases de datos Fetales
- Bases de datos Atenciones de Urgencia
- Bases de datos ENO
- División Político Administrativa
- Bases de Datos Notificación de Tuberculosis

4.4. Pre-procesamiento de los Datos

El primer paso para realizar el estudio es obtener los datos desde el DEIS, lugar donde se encuentran las series y archivos necesarios para diseñar la solución. La información disponible es desde 1990 hasta el año 2014, por lo que se tomará información dentro de este rango:

Los sets de datos vienen dados por:

- Series Principales (causas de muerte según sexo y edad), Chile 1997-2012.
- Bases de datos Egresos, 2001-2014
- Bases de datos Defunciones, 1997-2014

Se decidió trabajar con las bases de datos correspondientes a las defunciones y los egresos, porque manejan variables iguales, y porque los egresos hospitalarios, disponen de la variable que indica si la persona resultó fallecida o viva. A su vez se usará una serie temporal para efectos comparativos. El objetivo principal es combinar los sets de datos según los aspectos relevantes para el estudio realizado, de manera de integrar de la mejor forma posible la información provista.

4.4.1. Atributos

Los sets de datos contienen una gran cantidad de atributos, entre las cuales una de las más significativas es el tipo de enfermedad, cuyas vienen codificadas. A continuación, se detallan en rangos las causas de muerte más significativas en el país, con su respectivo código:

- A00-B99: Ciertas enfermedades infecciosas y parasitarias
- C00-C97: Tumores Malignos
- D00-D48: Tumores in situ, tumores benignos y tumores de comportamiento incierto
- E00-E90: Enfermedades endocrinas, nutricionales y metabólicas
- F00-F99: Demencias, trastornos mentales y del comportamiento debidos al uso de sustancias psicoactivas
- G00-G99: Trastornos neurológicos
- I00-I99: Enfermedades del sistema circulatorio
- J00-J99: Enfermedades del sistema respiratorio
- K00-K93: Enfermedades del sistema digestivo
- M00-M99: Enfermedades del sistema osteomuscular y del tejido conjuntivo
- P00-P96: Ciertas afecciones originadas en el periodo perinatal
- Q00-Q99: Malformaciones congénitas, deformidades y anomalías cromosómicas
- R00-R99: Síntomas, signos y hallazgos anormales clínicos y de laboratorio, no clasificados en otra parte
- V01-V89: Accidentes de transporte terrestre
- W00-W84: Accidentes no intencionales
- X00:Y09: Accidentes intencionales
- Y10:Y34: Eventos de intención no determinada

Otra variable realmente importante en este estudio, es la condición de egreso de las personas que fueron atendidos en algún reciento hospitalario del país. Dicha variable nos dice si la persona egreso fallecida o viva, lo que, para esta investigación es de suma relevancia pues se convierte en la variable de clasificación. (Para efectos de la aplicación y el uso de WEKA, el nombre de esta variable será usado como "CLASS") [20]

- COND_EGR: Condición al egreso, los valores aceptados son: 1 = Vivo 2 = Fallecido

Las demás variables como la edad de la persona, sexo, región, comuna, ocupación, previsión, y entre muchas otras, que no son menos importantes de las explicadas anteriormente, están detalladas en el anexo D.

En este anexo, se encuentra la información de cada diagnóstico con su respectivo código (también para las causas externas). En cuanto a las demás variables que provienen desde los sets de datos mencionados, se han seleccionado las que puedan agregar valor en la fase de aplicación, dichas variables aparecen en el Anexo E. Los esquemas de registros son dispuestos por el DEIS y pueden ser obtenidos en la página web del mismo.

4.4.2. Limpieza de Datos

Por ser bases de datos públicas del Ministerio de Salud, cada variable ha sido definida de manera precisa, declarando el nombre de la variable, el tipo de dato, el largo y la descripción de la misma. Este sentido, el DEIS se ha encargado de que cada valor presente en sus bases de datos coincida con lo definido en sus esquemas, por lo que no hay errores en sus valores.

A pesar de lo anterior, los sets de datos si han necesitado cierto grado de limpieza. El proceso de limpieza de datos se encarga de descubrir, corregir o eliminar los datos erróneos que se encuentren en la base de datos, con el fin de eliminar el ruido y así asegurar datos de calidad para su posterior uso. Para esto es necesario realizar:

- Estandarizar los datos de tal manera que todos sean del tipo correspondiente (evitar que una variable de tipo numérico posea letras, por ejemplo).
- Identificar valores fuera de rango y modificarlos según se requiera.
- Evitar valores nulos o "vacíos" que puedan interferir con la utilización de los métodos.
- Verificar consistencia entre los sets de datos, de manera de evitar valores duplicados.

4.4.3. Integración de Datos

Es necesario integrar los diversos sets de datos de manera tal que decante en un solo set de datos único y completo, por lo que la integración debe ser un proceso riguroso con el fin de evitar la inconsistencia, así como la redundancia. Como las bases de datos poseen variables iguales, por ejemplo, en el caso de los diagnósticos, sexo, edad, región, entre otros; hace que la integración de los registros adquiera sentido y esto pueda entregar resultados aún mejores. A continuación, se procederá a especificar lo realizado:

- Integración: Como el formato de los archivos es el mismo, la integración de los registros se vuelve un proceso relativamente fácil. Primero se tomó el conjunto de datos de las defunciones y se copió su contenido a un nuevo archivo. Luego, se crearon columnas dentro del nuevo archivo, con las variables de los egresos que no estuviesen en el archivo (ya que como se dijo antes estos conjuntos manejan variables que son las mismas). Finalmente, se copia el contenido de los egresos en las columnas correspondientes a cada variable, y queda un solo archivo.

4.5. Transformación de los Datos

Finalmente se procede a la etapa de transformación de los datos. Dicha etapa se realiza en torno a la reducción de datos, con el fin de converger en un set de datos representativo, por lo que se sintetizan los datos, se reduce la dimensión de la data, es decir, se eliminan atributos que se consideran irrelevantes para el estudio.

- Identificación, selección y eliminación de variables: Del nuevo archivo generado resultante de la integración, se han seleccionado las variables más representativas o significativas, que son aquellas que pueden entregar un valor importante a la investigación y no entorpecerla. Este subconjunto queda plasmado en el Anexo E. Las demás variables se han eliminado para disminuir la data y el tiempo de ejecución durante su uso en WEKA.

4.6. Creación y Carga en Modelo de Datos.

Luego de transformar la información e integrarla en un solo repositorio, se procede a crear un modelo de datos representativo usando modelamiento de DW y bases de datos multidimensionales. Este modelo se creará utilizando software MySQL con el propósito fundamental de cargar la información aquí para su posterior análisis (cabe destacar que cada set de datos corresponde a un año en particular, por lo que con el modelo se busca tener toda la información histórica almacenada allí).

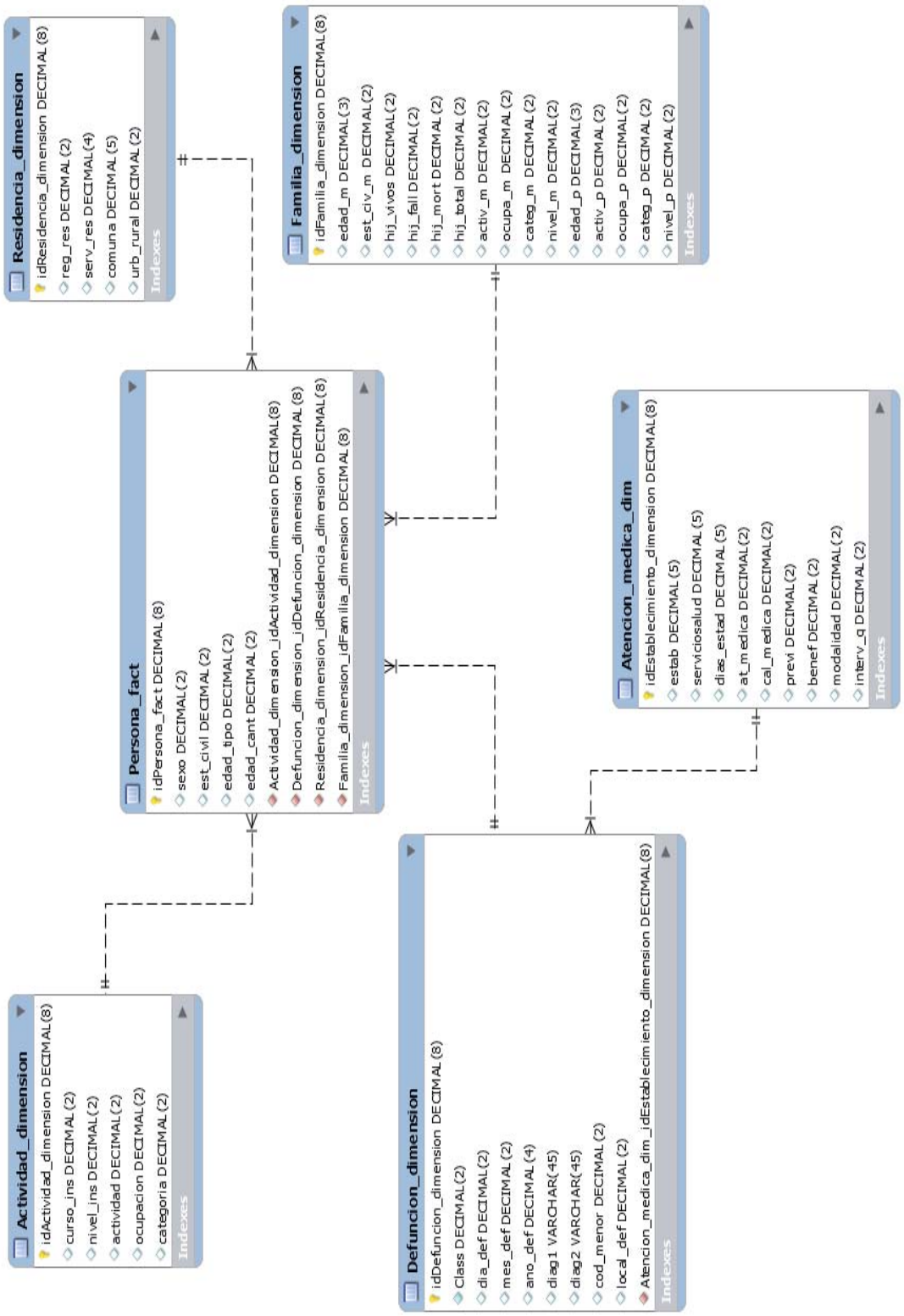

Figura 4.2. Modelo Multidimensional

4.7. Aplicación herramienta de DM

Se han anexados los procedimientos y pruebas preliminares que se realizaron en WEKA (filtros y conversiones de datos), en ellos se especifica con que filtros se ha trabajado ya que también fueron usados en las pruebas finales con tal de evitar complicaciones.

Utilizando los atributos que se mencionan en el anexo E, se aplicaran diversos filtros sobre el set de datos (filtros explicados con mayor detalle en anexo F), para luego realizar el proceso de selección de atributos. Esto para poder determinar, en base a la concordancia de los resultados, los atributos más significativos.

4.7.1. Selección de atributos

Los métodos de selección de atributos trabajan sobre subconjuntos del conjunto total de atributos y tratan de medir la calidad y la relevancia de dichos subconjuntos. Se trata no solo de intentar evaluar la relevancia de cada atributo por separado sino del subconjunto como un todo ya que dos atributos tomados de forma conjunta pueden aportar aquella información más relevante de la que aportan considerándolos por separado.

Al momento de la selección de atributos, se debe elegir el evaluador de atributos y método de búsqueda. En base a esto las combinaciones fueron las siguientes:

- Evaluador de atributos
    - CfsSubsetEval: Elabora un ranking de subconjuntos de atributos de acuerdo a su correlación basada en la evaluación de una función de evaluación heurística. Dicha función de evaluación se base en el cálculo de la correlación estadística, buscando atributos que están muy poco correlacionados entre ellos, pero tienen una buena correlación con la clase. Las características irrelevantes por tanto son ignoradas por definición ya que ellas mantendrán una muy baja o nula correlación con la clase. La información redundante por otra parte será penalizada ya que el atributo redundante tendrá una alta correlación con una o varias de las características restantes.
    - ConsistencySubsetEval: Evalúa el valor de un subconjunto de atributos por el nivel de consistencia en los valores de la clase cuando las instancias de training se proyectan en el subconjunto de atributos. La consistencia de un subconjunto nunca puede puede ser inferior a la del conjunto completo de atributos.
    - ClassifierSubsetEval: Evalua los subconjuntos de atributos en los datos de training. Utiliza un clasificador para estimar el "merito" de un conjunto de atributos.
    - ChiSquaredAttributeEval: Trabaja sobre atributos individuales, lo que hace es realizar el test χ 2 con respecto a la clase para cada atributo y esos son los valores que utiliza para crear el ranking. El algoritmo Chi-cuadrado para selección de atributos se basa en el cálculo del estadístico χ^2.
    - InfoGainAttributeEval: Está basado en el principio de ganancia de información. Evalúa el valor de un atributo midiendo la ganancia de información con respecto a la clase. Anteriormente discretiza los atributos numéricos.

- CorrelationAttributeEval: Se refiere al uso del coeficiente de correlación de Pearson para calcular la correlación que existe entre cada atributo y la variable de clasificación y seleccionar solo aquellos atributos con una correlación alta.
- GainRatioAttributeEval: Evalúa el valor de un atributo midiendo la relación de ganancia con respecto a la clase.
- OneRAttributeEval: evalúa la calidad de cada atributo utilizando el clasificador OneR, el cual usa el atributo de mínimo error para predecir, discretizando los atributos numéricos.

- Métodos de búsqueda:
  - GreedyStepwise: Una búsqueda voraz sin backtraking, es decir, coge el mejor de todos, luego la mejor pareja que lo incluye, luego el mejor trío que incluye a los anteriores y así hasta que la solución ya no mejora.
  - BestFirst: Igual que Greedy es una búsqueda en profundidad, pero aplicando vuelta atrás (en inglés backtraking) hasta un límite de retrocesos. Básicamente la búsqueda se desarrolla usando un árbol y consiste en ir eliminando atributos hasta llegar a un cierto número de atributos (predeterminados por el usuario). Con este procedimiento se logran ahorros en tiempo de procesamiento, y al mismo tiempo garantizando que la solución es óptima.
  - Ranker: Se utiliza cuando evaluamos atributos por separado (de uno en uno)

A continuación, se muestran los atributos obtenidos por cada método de selección empleado

Métodos de selección de atributos						
Cfs	Consistency	Classifier	ChiSquared	GainRatio	InfoGain	OneR
Edad Interv_Q Diag1	Edad Reg_res Dias_Estad Previ Benef Modalidad Interv_Q Diag1 Sexo Diag2	Edad Diag1 Interv_Q Previ	Edad Diag1 Dias_Estad Interv_Q Benef	Edad Interv_Q Diag1 Dias_Estad Modalidad	Edad Diag1 Dias_Estad Interv_Q Benef	Diag2 Diag1 Edad Servicio_salud Reg_Res

Tabla 4.7.1 Atributos obtenidos de los métodos de selección de atributos

El entendimiento de los resultados provenientes del proceso de selección de atributos, nos permite interpretar de mejor manera la importancia de cada uno de estos atributos que se han usado en la aplicación de algoritmos. De esta manera, se puede reflexionar sobre qué cosas se podrían realizar sabiendo la gran incidencia que tienen en la mortalidad.

Basándonos en la tabla adjunta se determinan los atributos de mayor influencia ordenados por su frecuencia de selección: Edad (7), Diag1(7), Interv_Q(6), Dias_Estad(4) y Benef(3), además de otros dos que consideramos relevantes (Sexo y Diag2). Cada uno de estos atributos se describe a continuación:

- Diagnóstico: Es el atributo que más incidencia tiene en la muerte de una persona. Ya sea mirado desde un análisis simple, hasta un análisis más complejo. Y es que se hace obvia la implicancia del diagnóstico en la mortalidad, no hay duda de eso, porque si es analizado desde un modelo más complejo, es que, por ejemplo, todos y cada uno de los algoritmos en el proceso de selección de atributos lo arrojaron en el primer lugar.
- Edad: La edad es una característica de las personas que dice mucho, con la edad podemos segmentar a las personas en grupos y con ello encontrar relaciones que expliquen algún comportamiento. Debido a esto es que este atributo fue seleccionado por todos los algoritmos de selección como uno de los más importantes para la clasificación.
- Beneficio: Este atributo tiene una dependencia directa con el atributo Previsión, y solamente cuando la Previsión es Fonasa, en donde explica los tramos existentes en dicha previsión, que van desde el A hasta el D. Y esta selección del atributo Beneficio, se explica porque las personas que tienen previsión de salud Fonasa, son mucho más que las que tienen otra previsión como Isapre, Capredena, Dipreca, entre otras. Por ello, es que en el proceso de selección obtuvo un valor importante, y es que este atributo explica de mejor manera la variable de decisión.
- Intervención quirúrgica: En temas de salud, lo que puede cambiar entre que una persona viva o muera, es el hecho de que hayan recibido intervención quirúrgica. Lo anterior, fue satisfactoriamente identificado como relevante en esta investigación.
- Días de estado: En temas de mortalidad, el tiempo de estancia en un establecimiento de salud puede hacer la diferencia entre aumentar o disminuir las probabilidades de supervivencia. Esta relación con la variable de clasificación, fue correctamente identificada, lo que conlleva a que el análisis final tenga sentido.

Atributos considerados:

- Sexo: Este atributo es esencial en temas de salud, por el simple hecho de que existen diagnósticos que afectan a solo a mujeres y otros solamente a los hombres. Y como característica permite esclarecer los resultados para que sean más confiables.
- Diagnóstico número dos: Al igual que el diagnóstico uno, tiene una gran incidencia en la variable de clasificación. A diferencia del diagnóstico uno, no todos los registros tenían definido un diagnóstico dos, haciendo que perdiese importancia en análisis de los registros.

Teniendo en consideración estos atributos se pueden plasmar las siguientes reflexiones:

1. Las bases de datos del gobierno poseen muchos atributos innecesarios, que solo causan problemas en el trato de la data, porque no tienen representatividad ni valor, por lo que

en futuros procesos de documentación se recomienda reevaluar los esquemas de datos para disminuir data superflua.

2. Existen muchas personas que comparten características similares, como el mismo diagnóstico, edad, región, entre otras características, y que solo difieren en la variable de decisión, es decir, si la persona vivió o murió. A pesar de que gran parte de los atributos son confiables, existe una falta de información que podría ayudar a tener mejores resultados, por ejemplo, cual medicamento fue usado para tratar su diagnóstico.

Con estos atributos se procede a realizar los procesos de clasificación individuales y colectivos para posteriormente analizar los resultados obtenidos. Aquellos atributos que no fueron seleccionados serán obviados para los casos de estudio a realizar, con el fin de mejorar la significancia, la calidad y la rapidez de ejecución de los clasificadores.

4.7.2. Clasificadores

Teniendo una vez los atributos más relevantes para los casos de estudio, se realizan diversos experimentos, los cuales se clasifican en:

- Clasificadores Individuales: Se procesa la información usando un algoritmo de clasificación en particular, para luego analizar el resultado de cada uno de ellos y determinar la significancia que tiene para el estudio. Los algoritmos utilizados son: ID3, j48 (Árboles de Decisión); BayesNet, NaiveBayes (Redes Bayesianas); MultiLayer Perceptron (Perceptrón Multicapa).
- Comparación de Clasificadores: Utilizando diversos algoritmos sobre el set de datos, se comparan para medir el performance de los algoritmos con la clase y determinar cuál es el más apropiado o eficaz de los algoritmos. Para este experimento se utilizó: NaiveBayes (Redes Bayesianas); ZeroR, OneR (Reglas de asociación); j48 (Árbol de Decisión).

Cabe destacar que la selección de algoritmos a utilizar en ambas pruebas se basó en la capacidad de procesamiento de software y hardware para el nivel de datos que se poseen, por lo que la utilización de algún otro algoritmo (máquinas de soporte vectorial, por ejemplo) no fue posible debido a la limitación de recursos con los que se contaba al momento de las pruebas. Para poder llevar a cabo las pruebas, se redujo el número de registros (fundamentada en los recursos de software y hardware mencionados con anterioridad), por lo que la configuración final es:

- Número de registros: 348.711
- Número de atributos: 7
- Años: 2013-2014

4.7.3. Clasificadores Individuales

Descripción indicadores

Al momento de realizar la clasificación, el algoritmo entrega una serie de indicadores, los cuales sirven para determinar la eficacia, relevancia, entre otros aspectos del clasificador, lo que nos ayuda al posterior análisis de resultados. Entre los indicadores arrojados por los algoritmos se encuentran:

- **F-score:** Es la medida de precisión que tiene un test, se considera una medida armónica que combina valores de precisión y exhaustividad.
- **Curva ROC:** Es una representación gráfica de la sensibilidad frente a la especificidad (razón de verdaderos) para un sistema binario según varia el umbral de discriminación, o bien es la representación de la razón o ratios de verdaderos positivos frente a la razón o ratio de falsos positivos según varia el umbral de discriminación.
  Interpretación: "se puede interpretar como la probabilidad de que un clasificador ordenara o puntuara una instancia positiva elegida aleatoriamente más alta que una negativa".
  - 0-5 -0.6 = Malo
  - 0-6-0.75 = Regular
  - 0.75-0.9 = Bueno
  - 0.9-0.97 = Muy Bueno
  - 0.97-1 = Excelente
- **Kappa Statistic:** Ajusta el efecto del azar en proporción de la concordancia observada. En este caso dio 0.xxx, se considera bueno porque es sobre 0 y tiende al 1.
- **TP Rate:** Tasa de verdaderos positivos.
- **FP Rate:** Tasa de falsos positivos.
- **Precisión:** Tasa de precisión del modelo.
- **Recall:** Sensibilidad, se refiere a la fracción de ejemplos de la clase de todo el conjunto que se clasifican correctamente.
- **F-Measure:** Es una medida que combina el Recall y la precisión.

Algoritmo ID3

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      326997               93.7731 %
Incorrectly Classified Instances     20828                5.9729 %
Kappa statistic                          0.2336
Mean absolute error                      0.0862
Root mean squared error                  0.2173
Relative absolute error                 75.3435 %
Root relative squared error             90.986  %
UnClassified Instances                 886                0.2541 %
Total Number of Instances           348711

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,990    0,831    0,948      0,990    0,969      0,275    0,859     0,985     1
               0,169    0,010    0,526      0,169    0,256      0,275    0,854     0,339     2
Weighted Avg.  0,940    0,781    0,923      0,940    0,925      0,275    0,859     0,945

=== Confusion Matrix ===

     a      b   <-- classified as
 323408   3236 |      a = 1
  17592   3589 |      b = 2
```

Figura 4.3 Algoritmo ID3

Algoritmo j48

```
=== Summary ===

Correctly Classified Instances      329494               94.4891 %
Incorrectly Classified Instances     19217                5.5109 %
Kappa statistic                          0.2968
Mean absolute error                      0.0845
Root mean squared error                  0.2055
Relative absolute error                 73.3691 %
Root relative squared error             85.6566 %
Total Number of Instances           348711

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,993    0,791    0,951      0,993    0,971      0,351    0,917     0,994     1
               0,209    0,007    0,660      0,209    0,318      0,351    0,917     0,467     2
Weighted Avg.  0,945    0,743    0,933      0,945    0,931      0,351    0,917     0,962

=== Confusion Matrix ===

     a      b   <-- classified as
 325022   2307 |      a = 1
  16910   4472 |      b = 2
```

Figura 4.4 Algoritmo j48

BayesNet

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      316940               90.889  %
Incorrectly Classified Instances     31771                9.111  %
Kappa statistic                          0.3279
Mean absolute error                      0.1119
Root mean squared error                  0.25
Relative absolute error                 97.2128 %
Root relative squared error            104.2193 %
Total Number of Instances           348711

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,939    0,553    0,963      0,939    0,951      0,333    0,891     0,992     1
               0,447    0,061    0,324      0,447    0,376      0,333    0,891     0,311     2
Weighted Avg.  0,909    0,523    0,924      0,909    0,916      0,333    0,891     0,950

=== Confusion Matrix ===

     a      b   <-- classified as
 307380  19949 |      a = 1
  11822   9560 |      b = 2
```

Figura 4.5 Algoritmo BayesNet

NaiveBayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      316940               90.889  %
Incorrectly Classified Instances     31771                9.111  %
Kappa statistic                          0.3279
Mean absolute error                      0.1119
Root mean squared error                  0.25
Relative absolute error                 97.2128 %
Root relative squared error            104.2193 %
Total Number of Instances           348711

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,939    0,553    0,963      0,939    0,951      0,333    0,891     0,992     1
               0,447    0,061    0,324      0,447    0,376      0,333    0,891     0,311     2
Weighted Avg.  0,909    0,523    0,924      0,909    0,916      0,333    0,891     0,950

=== Confusion Matrix ===

     a      b   <-- classified as
 307380  19949 |      a = 1
  11822   9560 |      b = 2
```

Figura 4.6 Algoritmo NaiveBayes

MultiLayer Perceptron

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      328154               94.1049 %
Incorrectly Classified Instances     20557                5.8951 %
Kappa statistic                          0.2212
Mean absolute error                      0.0911
Root mean squared error                  0.213
Relative absolute error                 79.095  %
Root relative squared error             88.7995 %
Total Number of Instances           348711

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,993    0,847    0,947      0,993    0,969      0,275    0,895     0,991     1
               0,153    0,007    0,572      0,153    0,241      0,275    0,895     0,381     2
Weighted Avg.  0,941    0,796    0,924      0,941    0,925      0,275    0,895     0,954

=== Confusion Matrix ===

      a      b   <-- classified as
 324884   2445 |      a = 1
  18112   3270 |      b = 2
```

Figura 4.7 Algoritmo MultiLayer Perceptron

4.7.4. Comparación de Clasificadores

Los algoritmos (ZeroR, OneR, j48 y BayesNet) utilizados se comparan de acuerdo a diversos patrones, con el fin de determinar cuál es mejor en cada uno de los aspectos que consideramos relevantes para que el procesamiento fuera exitoso. Los aspectos a considerar son:

- Instancias correctas.
- Tasa de Verdaderos Positivos
- Tasa de Falsos Positivos
- Curva ROC
- F-Score

Los resultados obtenidos para cada aspecto considerado se resumen en la siguiente tabla:

Aspecto	Algoritmo			
	ZeroR	OneR	J48	BayesNet
Instancias correctas	93.87%	93.87%	**94.06%**	90.89%
Verdaderos Positivos	**1.00**	**1.00**	**1.00**	0.94
Falsos Positivos	**1.00**	**1.00**	0.89	0.55
Curva ROC	0.50	0.50	0.84	**0.89**
F-Score	**0.97**	**0.97**	**0.97**	0.95

Tabla 4.7.4 Tabla comparación de algoritmos

4.7.5. Análisis de Resultados

Clasificadores individuales

Según los resultados obtenidos para cada uno de los algoritmos, es posible analizar que:

- ID3: A pesar de ser el algoritmo de árbol de decisión mas básico, obtuvo un 93.771% de instancias bien clasificadas. Y pese a su simplicidad, respecto de otros algoritmos, se considera que consiguió un buen nivel de confirmidad en cuanto a instancias bien clasificadas. Pero a pesar de ello, su curva ROC solo alcanza un 0.5, lo que implica que a nivel global el ID3 pierda validez.
- J48: Como se ha mencionado en mas de una oportunidad en esta investigación, J48 es una extensión del ID3, por lo que la lógica es muy similar con que trabaja su versión anterior. Dado lo anterior, se condice que el resultado de aplicar este algoritmo sea mayor, llegando a un 94.06% de instancias bien clasificadas. En relación a otras medidas obteninas por J48, su curva ROC alcanza buenos niveles, así como el error abosulto que a no supera el 1%, por lo que permiten asegurar un nivel alto de confianza y conformidad con este algoritmo para efectos de aplicación.
- BayesNet: Esta red bayesiana supera el 90% de instancias bien clasificadas, pero pierde fuerza en cuanto a medidas de error se refiere, ya que alcanza porcentajes demasiado altos, a pesar de tener un ROC alto. Por lo que, pierde mucha credibilidad como clasificador.
- NaiveBayes:Al igual que el anterior, trabaja con una lógica muy parecida, y por ende los resultados logrados son muy similares. Dado lo anterior, en terminos de calidad de clasificador, también pierde mucha credibilidad como clasificador.

- MultiLayer Perceptron: El MLP alcanzó un 94.1049% de instancias bien clasificadas, resultado propiamente no muy alto para un algoritmo de su embargadura. En esa línea, bastaría con ir ajustando los parámetros que configuran su aplicación, como lo es la cantidad de nodos en la capa oculta, para ir mejorando sus resultado. Y en cuanto a sus otras medidas, su curva ROC y medidas de error, son estables haciendo que sea un clasificador que tiene peso.

Comparación de Algoritmos

Asimismo, para los resultados de cada aspecto considerado, se determina que:

- J48 destaca entre los demás algoritmos al tener buenos resultados para cada aspecto, por lo que a grandes rasgos es el algoritmo más eficaz y preciso para predecir la mortalidad.
- Pese a sus buenos resultados, los demás algoritmos tampoco son considerados "malos" (excepto para el caso de las reglas de asociación ZeroR y OneR para el análisis de la curva ROC, en el cual sus resultados no fueron los esperados), ya que en su mayoría son indicadores por sobre 0.85, lo que se considera bueno para los casos de estudio.
- Todos los algoritmos poseen un buen indicador F-Score y Verdaderos Positivos, es por ello que se puede afirmar que los algoritmos son bastante eficaces para predecir correctamente la clase.
- BayesNet resultó ser uno de los algoritmos con menores resultados (aunque su diferencia no es tan grande con los demás algoritmos), sin embargo posee el mayor índice de Curva ROC, lo que implica que tiene un buen rendimiento como clasificador.
- Como conclusión de análisis, se logra determinar que todos los algoritmos son considerados buenos o muy buenos (j48), lo que a su vez significa que el proceso de selección de atributos tuvo bastante relevancia para conseguir buenos resultados en las pruebas de clasificación.

5. Conclusiones

Para concluir, se definen los aspectos tanto positivos como negativos con respecto al trabajo realizado:

Aspectos positivos

- Problemática real: la mortalidad es un tema relevante y considera muchos aspectos de nuestra sociedad, por lo que la investigación realizada, así como el tratamiento de la información resultó ser positiva, ya que se logró generar conocimiento e interpretar a mayor escala la envergadura del tema.
- Precursor: considerando la importancia del problema expuesto, esta investigación se considera como un punto inicial en lo que respecta a la salud, con la idea de que a futuro se logre ahondar aún más en este tema y se observen resultados significativos en lo que a la mortalidad respecta.
- Buen enfoque: Si bien la complejidad es alta, por el hecho de manejar una gran cantidad de información, se logra dar un sentido coherente al tratamiento de los datos, al enfoque de la investigación y al análisis de resultados.
- Optimizar: Con el análisis de resultados realizado, es posible determinar a grandes rasgos que características son las más relevantes para el estudio, lo que podría ayudar a optimizar la información provista por el DEIS, con el fin de aumentar la significancia de los archivos y eliminar información que no es relevante.
- Conocimientos adquiridos: personalmente, creemos que realizar esta investigación nos ayudó mucho en nuestro proceso de aprendizaje y en darnos cuenta que no nos equivocamos al haber escogido este tema. Esta investigación nos genera una base solvente para enfrentar los problemas que vendrán en el futuro para nosotros en el aspecto laboral.

Aspectos negativos

- Limitación de recursos: Debido a la gran cantidad de registros y características, fue imposible realizar el estudio a cabalidad, ya que los programas utilizados no soportan tal cantidad de datos, además de consumir muchos recursos de hardware de la computadora, por lo que el proceso se ve ampliamente ralentizado.

6. Trabajos Futuros

Con respecto a lo realizado en esta investigación, se proponen como trabajos futuros que ayuden y complementen la información provista en este informe:

- Experimentos históricos: Teniendo en cuenta que no se logró realizar una investigación histórica debido a la limitación de recursos para trabajar con la gran cantidad de datos provistos, se espera que a futuro continúe la investigación utilizando computadoras que si cumplan los estándares para el tratamiento de la información.
- Optimización del DEIS: Con los resultados obtenidos se podría realizar una reestructuración de los archivos provisto por el DEIS, con el fin de que estos aumenten la significancia para las personas.
- Eliminar especificidad: Para experimentos que se apoyen en algoritmos visuales para su entendimiento y descripción (Arboles de Decisión, por ejemplo), es necesario transformar la especificidad de algunas variables en datos más generales (como es el caso del atributo Establecimiento), para así reducir la complejidad del modelo a visualizar.
- Validar el modelo: Aunque las variables utilizadas fueron halladas relevantes para el modelo, no se han concluido en relaciones semánticas precisas con los casos de mortalidad en el país. Es de esperar que los resultados obtenidos en esta investigación puedan ser observables a futuro para validar la solución propuesta.

7. Referencias

[1] U. Fayyad, G. Piatetsky-Shapiro y P. Smyth, "From Data Mining to Knowledge Discovery: An overview" in Advances in Knowledge discovery and Data Mining, Cambridge: MIT Press, 1996, pp. 1-36.

[2] L. D. X. Lian Duan, Business Intelligence for Enterprise Systems: A Surve, vol. 8, IEEE Tranactions on Industrial Informatics , 2012, pp. 679-687.

[3] D. Bernabeu, «Dataprix,» 2009. [En línea]. Available: http://www.dataprix.com/datawarehouse-manager.

[4] R. A. Mur, «Aprendizaje Automático para el Análisis de Datos,» Madrid.

[5] M.-l. L. Fang Lie, Implementation of Map Matching for GPS Using Decision Tree Method, Beijing: Beijing: computer engineering, 2008.

[6] P. A. P. Castaño, Minería de Uso para la identificación de patrones, 2009.

[7] A. a. M. J, Building classification trees using the total uncertainty criterion, International Journal of Intelligent Systems, 2003.

[8] J. Quinlan, Induction of Decision Tree Machine Learning, 1986.

[9] W. L. L. N. Wang Xiaohu, An Application of Decision Tree Based on ID3, 2012.

[10] F. J. Palacios, Redes Neuronales con GNU/Linux, 2003.

[11] K. B. a. N. B. A. Jovic, An overview of free software tools for general data mining, 2014.

[12] «RapidMiner Open Source Data Science Platform,» [En línea]. Available: https://rapidminer.com.

[13] «The R Project for Statical Computing,» [En línea]. Available: https://www.r-project.org.

[14] «Weka, A Machine Learning Group at the University of Waikato,» [En línea]. Available: http://www.cs.waikato.ac.nz/ml/weka.

[15] «Orange Data Mining,» [En línea]. Available: http://orange.biolab.si.

[16] «KNIME, Open for Navigation,» [En línea]. Available: https://www.knime.org.

[17] M. d. S. –. I. N. d. E. –. S. d. R. C. e. Identificación, «Estadísticas Vitales, Anuario 2013,» 2015.

[18] M. d. Salud, «Indicadores Básicos de Salud,» 2014.

[19] «Departamento de Estadísticas e Información de Salud,» [En línea]. Available: http://www.deis.cl/estadisticas-mortalidad/.

[20] J. G. H. José Manuel Molina López, Técnicas de análisis de datos, Aplicaciones prácticas utilizando Microsoft Excel y Weka, 2004.

8. Anexos

A: **Algoritmos de clasificación**

B: **Descripción de Software para Minería de Datos**

C: **Antecedentes**

D: **Esquema de registros de egresos hospitalarios**

E: **Esquema de atributos a usar**

F: **Procedimientos realizados en software WEKA**

G: **Esquema de registros de defunciones (DEIS)**