Ezequiel Agustín Baldizzoni

Un proceso de transformación de datos para proyectos de explotación de información

Trabajo Final Integrador presentado para la obtención del título de Licenciado en Sistemas

Director de Trabajo Final Integrador

Ramón García Martínez

UN PROCESO DE TRANSFORMACIÓN DE DATOS PARA PROYECTOS DE EXPLOTACIÓN DE INFORMACIÓN

Alumno

APU Ezequiel Baldizzoni

Directores

Dr. Ramón GARCIA MARTINEZ

Mg. Darío RODRIGUEZ

TRABAJO FINAL PRESENTADO PARA OBTENER EL GRADO
DE
LICENCIADO EN SISTEMAS

DEPARTAMENTO
DE DESARROLLO PRODUCTIVO Y TECNOLÓGICO
UNIVERSIDAD NACIONAL DE LANUS

FEBRERO, 2013

RESUMEN

La exploración y análisis, en forma automática o semi-automática, de grandes volúmenes de información para la detección de patrones se enmarca en los procesos de explotación de información para lo cual utiliza algoritmos de minería de datos (Data Mining). Dentro de un proceso normal de explotación de información, existe una de las etapas fundamentales llamada transformación de datos que se ocupa de la preparación de los datos con el propósito de entregar a las etapas posteriores del proceso un conjunto de datos de calidad y de esta forma concluir con resultados exactos. Esta etapa normalmente consume aproximadamente un 60% del esfuerzo de desarrollo. Dado al gran esfuerzo necesario se propone un proceso de transformación de datos.

ABSTRACT

The exploration and analysis, automatically or semi-automatic, large volumes of information to detect patterns is part of the operations process information for which it uses data mining algorithms (Data Mining). In a normal process operation of information, there is a call key stage of data processing which deals with the preparation of data for the purpose of delivering to the later stages of the process of operation of a data set information quality and thus conclude with accurate results. This stage usually consumes about 60% of the development effort. Given the large effort required proposes a data transformation process.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

NOMENCLATURA

%N	Porcentaje de datos perdidos en una variable.
ABR	Abreviaturas: truncamiento de uno o más tokens.
CP	Categoría de datos perdidos.
ERR	Errores ortográficos.
ESP	Espacios en blanco: eliminación o adición de espacios en blanco.
FR	Fuera de rango o ruidoso.
GD	Gráfico de dispersión.
IQR	Rango intercuartíl.
MAR	Datos perdidos al azar.
MCAR	Datos perdidos completamente al azar.
MOA	Análisis de ruidosos de Mahalanobis (Mahalanobis Outlier Analysis).
NMAR	Datos perdidos no al azar.
NULL	Ausencia de valor.
OUTLIERS	Datos ruidosos, fuera de rango o atípicos.
PSF	Prefijos/sufijos sin valor semántico presencia de subcadenas al principio y/o al final.
SE	Series encontradas.
SP	Situación del problema.
SPSS	Paquete estadístico para las ciencias sociales del ingles Statistical Package for the Social Sciences
SRSWOR	Muestreo aleatorio simple sin remplazo del ingles Sampling Random Sample WithOut Replacement
SRSWR	Muestreo aleatorio simple con remplazo del ingles Simple Random Sampling With Replacement
STATA	Estadística y datos del ingles statistics & data
TANAGRA	Herramienta para proyectos de explotación de información.
TD	Tamaño de ancho de las variables.
TDR	Tokens en desorden.
TFL	Tokens faltantes: eliminación de uno o más tokens.
TOKEN	Conjunto de cadenas de caracteres separadas por caracteres especiales, como por ejemplo espacios en blanco, puntos y comas.
TR	Tamaño total de registros de los datos.

1. INTRODUCCION

En este Capítulo se plantea el contexto del trabajo final de licenciatura (sección 1.1), se establece su objetivo (sección 1.2), y se resume la visión general del trabajo final de licenciatura (sección 1.3).

1.1. CONTEXTO DEL TRABAJO FINAL DE LICENCIATURA

Los proyectos de explotación de información son hoy en día una de las herramientas más importantes en las organizaciones y son utilizadas para tomar decisiones de negocio. La minería de datos es un tipo de técnica para extraer información de los datos que las organizaciones fueron almacenando a lo largo de sus vidas. Dado que este tipo de tecnología es de gran ayuda, se puede decir que es importante reduciendo las dificultades que esta pueda conllevar. También es de destacar que cada una de las etapas de un proyecto de explotación de información consume un esfuerzo distinto según las dificultades por las que atraviesa. Las etapas de un proyecto de explotación de información pueden ser las siguientes:

- **Análisis de requerimientos**, esta etapa va a cubrir el análisis de las necesidades de la organización y es donde se va a decidir qué datos van a ser necesarios para el resto del proceso.
- **Selección del conjunto de datos**, tanto en lo que se refiere a las variables objetivo (aquellas que se quiere predecir, calcular o inferir), como a las variables independientes (las que sirven para hacer el cálculo o proceso), como posiblemente al muestreo de los registros disponibles.
- **Análisis de los datos**, se analizan los datos obtenidos y se decide si son los correctos o si hay que volver a la primera etapa ya que puede ser esta la incorrecta.
- **Transformación de los datos**, se ejecutara una serie de pasos que mejoraran la calidad de los datos para luego poder ser utilizados por la siguiente etapa.
- **Selección y ejecución de las técnicas de minería de datos**, según el modelo seleccionado se ejecutaran las distintas herramientas o técnicas de minería de datos.
- **Análisis de resultados**, Una vez ejecutadas las técnicas de minería de datos se analizan los resultados arrojados en busca de conocimiento.

Los proyectos de explotación de información se encargan de efectuar distintos pasos con el fin de analizar, de forma automática o semi-automática, grandes volúmenes de datos y de esta forma extraer patrones de interés para el negocio que hasta este momento eran desconocidos.

Hay que reconocer que reducir el tiempo y el esfuerzo de este tipo de proyectos da como resultado una mayor cantidad de conocimiento adquirido de los datos y un menor costo del proyecto.

La etapa de transformación de los datos consume el 60% del esfuerzo total del proyecto por lo que mejorar el rendimiento en la etapa mencionada, sería de gran ayuda para reducir el esfuerzo total del proceso.

1.2. OBJETIVO DEL TRABAJO FINAL DE LICENCIATURA

Este trabajo final de licenciatura tiene como objetivo mostrar un proceso de transformación de datos para proyectos de explotación de información, que propone una división en actividades y propone una serie de técnica para estas.

Para llevarlo a cabo, es necesario dividir esta etapa de proyectos de explotación de información en una cantidad finita de actividades que deben ir ejecutándose secuencialmente en una serie de ciclos que garantizarán la salida a esta etapa.

Se plantea a su vez una serie de técnicas que se utilizaran en cada actividad con el fin de garantizar también que cada entrada/salida sea la correcta.

Cada una de las técnicas se dividirá en distintos pasos los cuales deben ir siendo cumplidos secuencialmente hasta transformar cada entrada de las actividades en salida.

Por último se propone una organización de documentación para cada actividad con el fin de aportar a la gestión del conocimiento de la organización y de esta forma, disminuir la dificultad de este proceso a medida que la organización lo ejecuta una mayor cantidad de veces.

De esta forma es posible afirmar que el objetivo general de este trabajo, es optimizar la tarea de transformación de los datos en proyectos de explotación de información.

1.3. VISIÓN GENERAL DEL TRABAJO FINAL DE LICENCIATURA

Este trabajo final de licenciatura se encuentra estructurado en siete capítulos que se describen a continuación.

En el Capítulo Introducción se plantea el contexto del trabajo, luego se establece su objetivo y se resume dando una visión general del trabajo final de licenciatura.

En el Capítulo Estado de la Cuestión se desarrolla una investigación sobre distintas teorías y técnicas que son concurrentes con los objetivos de este trabajo. Dentro del mismo se presentaran las distintas teorías que encuadran dentro de cada actividad: enriquecer los datos, obtener y ejecutar los casos testigo, determinar y aplicar las estructura de los datos, construir el modelo de entrada y por último la inspección de los datos.

En el Capítulo Descripción del Problema se presenta el problema de investigación partiendo de las dificultades que hoy en día poseen las organizaciones al momento de ejecutar proyectos de

explotación de información sobre los repositorios de datos que se almacenan desde los sistemas existentes o deprecados. En primer lugar se describe la identificación del problema de investigación, luego se caracteriza el problema abierto y se concluye con un sumario de investigación.

En el Capítulo Solución se presenta una serie de cuestiones generales sobre la solución, se describe la propuesta de proceso de transformación de datos para proyectos de explotación de información, la estructura general del proceso y las actividades.

En el Capítulo Casos de Validación se presentan una serie de casos con el fin de describir como es la ejecución del proceso de transformación de datos en proyectos de explotación de información específicos, a los efectos de implementar las tareas correspondientes a cada actividad del proceso y evaluar su desempeño en cada uno de sus pasos. Se analiza un caso de validación correspondiente a un banco el cual busca mejorar la cartera de depósitos a largo plazo, otro de estos es el caso de validación para un conjunto de datos de pacientes indios con problemas hepáticos y por último el caso de validación sobre un conjunto de datos para el diagnóstico de cáncer de mama.

En el Capítulo Conclusiones se presentan los aportes de este trabajo y se destacan las futuras líneas de investigación que se consideran de interés en base al problema abierto que se presenta en este trabajo final de licenciatura.

En el Capítulo Referencias se listan todas las publicaciones consultadas para el desarrollo de este trabajo final de licenciatura.

2. ESTADO DE LA CUESTIÓN

En este capítulo se presenta el estado de la cuestión sobre distintas teorías y técnicas que son concurrentes con los objetivos de este trabajo. Dentro del mismo se presentarán las distintas teorías que encuadran dentro de cada actividad: enriquecer los datos (sección 2.1), obtener y ejecutar los casos testigo (sección 2.2), determinar y aplicar las estructura de los datos (sección 2.3), construir el modelo de entrada (sección 2.4) y por último la inspección de los datos (sección 2.5).

2.1. ENRIQUECER LOS DATOS

En la primera etapa, la de enriquecimiento de los datos, es posible afirmar que no existe una teoría específica que abarque esta tarea, dado que es una tarea de análisis y depende de la formación de la persona encargada del mismo. También, depende en gran medida, del modelo utilizado en el proceso de explotación de información del proyecto.
Por estas razones no se va a desarrollar una teoría específica para esta etapa del proceso.

2.2. OBTENER Y EJECUTAR LOS CASOS TESTIGO

Esta es la etapa en la que se decide si los datos cumplen con las características mínimas, para entregar un resultado útil al proyecto de explotación de información que lo ejecuta.
La obtención de los casos testigo se puede lograr mediante entrevistas al personal de la organización que estén involucrados con los datos necesarios o con el proyecto en sí.
Una vez entrevistado al personal involucrado, se generan los casos testigo que serán la guía para decidir qué datos y con qué formato serán necesarios.
Por último es necesario impactar están información en algún tipo de soporte para luego comprobar que los datos cumplen con lo especificado.

Item	Descripción	Esperado	Encontrado	observaciones
1	**Columna nombre y apellido**			
1.1	El formato es apellido, nombres	**SI**		
1.2	Los nombres comienzan con mayúsculas	**SI**		
1.3	Existe nombre y apellido	**SI**		
2	**Columna Dirección**			
2.1	Esta separado dirección de numero	**SI**		
2.2	Esta el barrio	**SI**		
2.3	Esta la provincia	**SI**		
2.4	Hay datos perdidos	**NO**		
3	**Columna jefe**			
3.1	Hay datos nulos	**NO**		
4	**Tabla relaciones**			
4.1	**Columna relación**			
4.2.1	Hay datos perdidos	**NO**		
4.2.2	La relación es correcta	**SI**		

Tabla 2.1. *Ejemplo de lista de chequeo.*

Una técnica muy utilizada para este tipo de tareas es la utilización de listas de chequeo (en ingles check list) que consta de una tabla que muestra, al personal que ejecuta las pruebas, una lista con ítems que deben ser analizados para determinar si los datos son los correctos para el resultado final.
Estas listas no tendrán la necesidad de chequear si los datos son de calidad, solo se carga en su interior los ítems que determinen si las variable, atributos, datos, tipos, etc. son los correctos para el resultado final (cuentan con las necesidades mínimas).
Una lista de chequeo consta de varias columnas las cuales pueden ser número de identificación del ítem, descripción (alguna descripción de lo que hay que chequear), resultado esperado, resultado encontrado, observaciones, etc. Un ejemplo de lista de chequeo se puede ver en la Tabla 2.1.
Una lista de cheque correcta debe detallar uno por uno distintos aspectos que se deben analizar, comprobar, verificar, etc.

2.3. DETERMINAR Y APLICAR LA ESTRUCTURA DE LOS DATOS

En esta actividad se debe efectuar un trabajo de análisis y transformación de los datos de tal manera de que si estos provienen de diferentes fuentes, unificarlos para que de varias fuentes quede un solo repositorio. También hay que lograr que a partir de las relaciones existentes entre los objetos de una fuente de datos se obtenga un solo objeto que contenga todos los anteriores. Partiendo de los valores del objeto se une a este por medio de las relaciones con los demás objetos, se une al objeto principal hasta llegar a tener un solo objeto que contenga todos los datos. Aquí es donde se escoge dejar de lado alguna variable que no sea utilizada en adelante.
Luego de esta tarea es necesario documentar todo criterio de estructuración utilizado para, de aquí en más, no necesitar un análisis extenso y de esta forma facilitar los futuros trabajos.
Posterior a la técnica manual se puede ejecutar una técnica asistida por computadora para mejorar el resultado y siendo monitoreada por un ser humano se obtendrán resultados mucho más valiosos que los obtenidos en un principio con las técnicas manuales.
Una técnica asistida por computadora para esta etapa es la técnica de agrupamiento (en ingles clustering). Agrupamiento es una técnica usada para lograr grupos multidimensionales. Con el fin de determinar que cada grupo utilice el concepto de distancia euclideana. A partir de un elemento calcula esta distancia y según el valor que se obtenga cómo resultado determina si pertenece a algún grupo o no. En este caso se utiliza para hacer un pre análisis de los datos pero también puede ser utilizada para técnicas post análisis.
Por ejemplo, en un espacio bidimensional, la distancia euclidiana entre dos puntos P1 y P2, de coordenadas (x1, y1) y (x2, y2) respectivamente se muestra en la formula de la Figura 2.1.

$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Figura 2.1. *Formula de distancia euclideana para espacios bidimensionales.*

2.4. CONSTRUIR EL MODELO DE ENTRADA DE DATOS

Esta etapa es la más costosa de todas ya que es la que consume más del 80% del esfuerzo de todo el proceso de transformación de datos. Aquí se mencionan las teorías que se utilizan para detectar problemas en los datos y repararlos, modificar sus tipos, el ancho y profundidad. Primero se abordará el tratamiento de los valores nulos o vacíos (sección 2.4.1), luego el tratamiento de duplicados (sección 2.4.2), el tratamiento de valores ruidosos (sección 2.4.3), la normalización (sección 2.4.4), tratamiento de series (sección 2.4.5), reducción del ancho de los datos (sección 2.4.6) y por último la reducción de la profundidad de los datos (sección 2.4.7).

2.4.1. TRATAMIENTO DE LOS VALORES NULOS O VACÍOS

La existencia de una cantidad considerable de valores nulos en una variable dificulta el análisis de los datos, ya que usualmente no permite la aplicación de las técnicas existentes que posibilitan el descubrimiento de conocimiento.

Hay una serie de enunciados que pueden agrupar a las distintas causas por las cuales estos valores son extraviados. Ejemplos de estos son:

- Los datos faltantes son parte del dominio de la variable.
- Los datos se han perdido.
- La existencia potencial de desviaciones en el análisis atribuible a las diferencias sistemáticas entre los datos observados y los datos perdidos.
- Problemas en la utilización de hardware disponible.
- Inconsistencia con otros registros de datos que son borrados.
- Datos que nunca fueron ingresados.
- Los datos incompletos.

En la práctica, la eficacia de las técnicas de tratamiento de valores nulos está directamente relacionada con la razón por la cual tuvo su origen el valor perdido. Si existe alguna información acerca de ella, es posible encontrar una regla para completar estos valores, por el contrario, si no se cuenta con dicha información, es necesario aplicar técnicas de evaluación de los valores perdidos que encuentren algún patrón que permita ya sea completarlos o descartarlos (en el caso que no afecten el análisis); decisión que depende en gran medida del tipo del valor perdido y la importancia

del registro en la base de datos [Allison, 2001]. A continuación se describen las categorías de los datos perdidos (sección 2.4.1.1) y el tratamiento de los datos perdidos (sección 2.4.1.2).

2.4.1.1. CATEGORÍAS DE DATOS PERDIDOS

Los datos faltantes o perdidos pueden ser categorizados en tres tipos: Datos Completamente Ausentes al azar (Missing Completely at random MCAR) (sección 2.4.1.1.1), Datos Ausentes al azar (Missing at Random MAR) (sección 2.4.1.1.2), Datos Ausentes No al Azar (Not Missing at random NMAR) (sección 2.4.1.1.3) [Little y Rubin, 1987].

2.4.1.1.1. Datos Completamente Ausentes al azar (Missing Completely at random MCAR)

Corresponden a variables con datos perdidos que no poseen relación alguna a los valores de otros registros dentro de la misma variable ni a los valores de otras variables. Cuando ocurre esto, las distribuciones de probabilidad de los datos faltantes y de todos los datos son idénticas.

2.4.1.1.2. Datos Ausentes al azar (Missing at Random MAR)

Corresponde a variables con datos perdidos que tiene alguna relación con otras variables, es decir, que si se tiene una variable Y con valores faltantes, y otra variable X, se dice que los datos son MAR si: $P(Y = NULL \mid X,Y) = (Y = NULL \mid X)$. Dicho de otra forma, el dato perdido puede predecirse a partir de otros datos (existentes) para el registro.

2.4.1.1.3. Datos Ausentes No al Azar (Not Missing at random NMAR)

Corresponde a variables con datos perdidos donde el mismo dato perdido determina en sí mismo por qué es desconocido.

2.4.1.2. TRATAMIENTO DE LOS DATOS PERDIDOS

Según [Farhangfar *et al.*, 2008] se dispone de una serie de técnicas para abordar los problemas de valores nulos o faltantes, descartar los registros con datos faltantes (sección 2.4.1.2.1), imputar los datos faltantes (sección 2.4.1.2.2), imputar los datos faltantes con otro valor (sección 2.4.1.2.3).

2.4.1.2.1. Descartar los registros con datos faltantes

Este método es práctico sólo cuando los datos contienen una relativamente baja cantidad de registros con datos perdidos y cuando el análisis de todos los datos no produce un sesgo importante por no utilizar estos registros.

Como métodos para descarte de datos faltantes podeos nombrar a Listwise Deletion que elimina todo aquel registro que contenga datos faltantes en cualquier variable y Pairwise Deletion solo

elimina los registros que tengan datos perdidos en las variables poco relevantes o que no son necesarias para el análisis [Jöreskog, 2005]. Según sea el caso se puede plantear la eliminación de variables o de registros: Eliminación de variable (sección 2.4.1.2.1.1) y eliminación de registros (sección 2.4.1.2.1.2).

2.4.1.2.1.1. Eliminación de variable

La eliminación de datos perdidos es una medida poco recomendable debido a la pérdida de información que genera, sin embargo puede ser una buena opción en caso que los datos perdidos sean de naturaleza MCAR y no puedan ser imputados fehacientemente [Scheffer, 2002].
En el caso que se esté pensando en eliminar una columna se debe tener en consideración la cantidad de datos perdidos que posee en total y si existe o no alguna relación con otra variable. Sería recomendable eliminar la columna en el caso que la cantidad de valores perdidos supere un umbral mínimo que permita análisis (por ejemplo que posea más de la mitad de datos perdidos probablemente no genere información fidedigna), o cuando la variable sea MCAR y no pueda ser deducida de ninguna forma con los datos existentes, por lo que imputarlos generaría un mayor costo de error que por pérdida de información.

2.4.1.2.1.2 .Eliminación de registros

Al eliminar registros en primer lugar hay que ser cuidadoso con la proporción de las clases que están evaluando. Si el problema a resolver constituye uno de clases desbalanceadas, la eliminación de un registro del cual hay pocas filas puede ser una gran pérdida de información inclusive si posee gran parte de sus atributos con datos perdidos o vacios. Por ello se debe tener cuidado con qué tipo de registro se está eliminando y tomar medidas de acuerdo a la información relativa que se pierde con su eliminación. Ahora, se recomienda eliminar registros siempre y cuando posean una gran cantidad de valores perdidos o blancos y correspondan a una pequeña cantidad dentro del total de los datos.

2.4.1.2.2. Imputar los datos faltantes

Este método es aplicable cuando la cantidad de atributos con datos faltantes es relativamente pequeña en relación al número de registros que presentan dicha condición. Existen dos grandes tipos de técnicas que pueden ser agrupadas en dos grupos, Imputación Simple (Single Imputation) (sección 2.4.1.2.2.1) e Imputación Múltiple (Multiple Imputation) (sección 2.4.1.2.2.2)

2.4.1.2.2.1. Imputación Simple (Single Imputation)

Imputación Simple es quizás el enfoque más utilizado en la práctica. En este método se estima el dato faltante usando otros datos relacionados que estén disponibles. Esto puede lograrse de varias formas, entre ellas [Farhangfar *et al.*, 2008]: Imputación por el promedio (sección 2.4.1.2.2.1.1), Imputación por la moda (sección 2.4.1.2.2.1.2), Imputación Hot Deck (sección 2.4.1.2.2.1.3), Imputación por regresión (sección 2.4.1.2.2.1.4).

2.4.1.2.2.1.1. Imputación por el promedio

Reemplazar los datos faltantes a través de la imputación por el promedio (en el cual se reemplaza el valor faltante de acuerdo al valor promedio de un grupo apropiadamente definido de valores disponibles). Imputación Simple a través de la imputación por el promedio posee tres limitaciones potenciales [Farhangfar *et. al.*, 2007]:

- Disminuye la variabilidad inherente al conjunto original de datos, particularmente en el caso que el mismo valor promedio sea utilizado para reemplazar varios datos faltantes.
- Es dependiente de las elecciones de grupos de datos.
- Si existen datos fuera de rango (outliers) entre los conjuntos de datos candidatos para obtener el promedio que se empleará para realizar la imputación, este valor puede generar una importante desviación o diversificación en el resultado.

2.4.1.2.2.1.2. Imputación por la moda

Reemplazar los datos faltantes a través de la imputación por la moda (en el cual se reemplaza el valor faltante de acuerdo a la moda de un grupo apropiadamente definido de valores disponibles). Dado que el promedio es afectado por la presencia de valores fuera de rango, parece natural usar la mediana en vez de la media con el fin de asegurar robustez [Acuña y Rodríguez, 2009].

2.4.1.2.2.1.3. Imputación Hot Deck.

Imputación Hot Deck (Hot Deck Imputation), en este caso para cada registro que contiene datos perdidos se busca el registro más parecido que no tenga datos perdidos y de esta forma, el dato perdido se imputa con el valor del dato existente en dicho registro [Nisselson *et. al.*, 1983].

2.4.1.2.2.1.4. Imputación por Regresión

Imputación por Regresión utiliza modelos de regresión que a partir de datos de otras variables puede predecir las observaciones faltantes [Farhangfar *et. al.*, 2007].

2.4.1.2.2.2. Imputación Múltiple (Multiple Imputation)

En Imputación Múltiple los valores perdidos de cualquier variable son estimados usando los valores existentes en otras variables. Los valores estimados (o imputados) sustituyen a los valores faltantes con lo cual se obtiene un conjunto de datos completo denominado "conjunto de datos imputados". Este proceso es realizado varias veces, produciendo varios conjuntos de datos imputados (de aquí el nombre de Multiple Imputation) [Farhangfar *et al.*, 2008]. Se realizan análisis estadísticos sobre cada uno de los conjuntos de datos imputados, obteniéndose múltiples resultados. Estos resultados posteriormente son combinados para producir un análisis final.

2.4.1.2.3. Reemplazar los datos faltantes con otro valor

Este método tiende a producir serios problemas de inferencia. No se entrará en detalle en este método.

Sugerencias para tratar el problema de los datos perdidos [Scheffer, 2002]:

- No usar imputación por el promedio a menos que el dato sea MCAR.
- Eliminar cuidadosamente verificando antes que el dato es MCAR.
- Imputación simple trabaja bien para datos faltantes MAR, siempre que menos del 10% de ellos sean datos nulos.
- Si se debe usar imputación simple, use EM o Regresión.
- Si las estructuras de varianza en los datos son importantes, no use el método de Eliminación o imputación simple si más del 5% de los datos están perdidos.
- Imputación múltiple opera correctamente para casos sobre el 25% de los datos perdidos.
- Para NMAR, sólo se podría usar imputación múltiple, y preferiblemente con niveles de datos perdidos menores a 25%.
- Siempre que sea factible, usar imputación múltiple debido a sus características.

2.4.2. TRATAMIENTO DE DUPLICADOS

Los duplicados son un problema no menor en los datos de las compañías. Esto se puede deber a errores ortográficos, mal entendimiento de las personas, mal entendimiento de textos que son ingresados a mano, etc. Un error de este tipo puede ser el ingreso de un usuario con un identificador y un nombre y otro usuario con otro identificador con el mismo nombre (siendo ambos la misma persona) y otro puede ser ingresar nombres, direcciones, documentos, etc., con errores (por ejemplo Jorge y Jorje). En esta sección se describen una serie de teorías sobre el tratamiento de duplicados, abordando: detección de duplicados (sección 2.4.2.1) y funciones de similitud (sección 2.4.2.2).

2.4.2.1. DETECCIÓN DE DUPLICADOS

El método de detección de duplicados se ejecuta de la siguiente manera [Winkler, 1990]:

- Se define un umbral real $\Phi \varepsilon [0,1]$.
- Se compara cada registro de la variable con el resto.
- Si la similitud entre una pareja de registros es mayor o igual que Φ, se asumen duplicados; es decir, se consideran representaciones de una misma entidad real.

2.4.2.2. FUNCIONES DE SIMILITUD

Actualmente existen diversas funciones de similitud, las cuales pueden ser clasificadas en dos categorías: basadas en caracteres y basadas en tokens [Elmagarmid *et. al.*, 2007]. Es posible dividir esta sección en funciones de similitud basada en caracteres (sección 2.4.2.2.1) y Funciones de similitud basadas en tokens (sección 2.4.2.2.2).

2.4.2.2.1. Funciones de similitud basadas en caracteres

Estas funciones de similitud consideran cada cadena como una secuencia ininterrumpida de caracteres. En esta sección se cubren las siguientes: Distancia de edición (sección 2.4.2.2.1.1), Distancia de brecha afín (sección 2.4.2.2.1.2), Similitud Smith-Waterman (sección 2.4.2.2.1.3) y Similitud de q-grams (sección 2.4.2.2.1.4).

2.4.2.2.1.1 Distancia de edición.

La distancia de edición entre dos cadenas A y B se basa en el conjunto mínimo de operaciones de edición necesarias para transformar A en B (o viceversa) [Levenshtein, 1966]. Esto se puede utilizar cuando existen errores de ortografía dado que la distancia de edición y otras funciones de similitud tienden a fallar identificando cadenas equivalentes que han sido demasiado truncadas.

2.4.2.2.1.2. Distancia de brecha afín.

La distancia de brecha afín ofrece una solución al penalizar la inserción/eliminación de k caracteres consecutivos (brecha) con bajo costo, mediante una función afín $\rho(k) = g + h \cdot (k - 1)$, donde g es el costo de iniciar una brecha, k el costo de extenderla un carácter, y h + g [Gotoh, 1982].
Se suele utilizar cuando hay abreviaciones en las cadenas o cuando hay un gran volumen de datos y además existen prefijos/sufijos sin valor semántico.

2.4.2.2.1.3. Similitud Smith-Waterman.

El modelo original de Smith y Waterman [Smith y Waterman, 1981] define las mismas operaciones de la distancia de edición, y además permite omitir cualquier número de caracteres al principio o

final de ambas cadenas. Esto lo hace adecuado para identificar cadenas equivalentes con prefijos/sufijos que, al no tener valor semántico, cuando existe una o más palabras (tokens) que no se encuentran en alguna de las dos cadenas o cuando existen espacios en blanco inútiles.

2.4.2.2.1.4. Similitud de q-grams

Un q-gram, también llamado n-gram, es una subcadena de longitud q [Yancey, 2006]. El principio tras esta función de similitud es que, cuando dos cadenas son muy similares, tienen muchos q-grams en común. Se utiliza cuando hay múltiples problemas de similitud o cuando las palabras (tokens) están desordenadas.

2.4.2.2.2. Funciones de similitud basadas en tokens

Estas funciones de similitud consideran cada cadena como un conjunto de subcadenas separadas por caracteres especiales, como por ejemplo espacios en blanco, puntos y comas. Esto es, como un conjunto de tokens, y calculan la similitud entre cada pareja de tokens mediante alguna función de similitud basada en caracteres. Solo se nombrara la función de coseno TF-IDF (sección 2.4.2.2.2.1).

2.4.2.2.2.1 Similitud coseno TF-IDF

Es una función que representa a los tokens como vectores y calcula su distancia mediante los cósenos de sus ángulos [Cohen, 1998].
Esta función produce altos valores de similitud para cadenas que comparten muchos tokens poco comunes (con alto poder discriminante).
La similitud coseno TF-IDF no es eficiente bajo la presencia de variaciones a nivel de caracteres, como errores ortográficos o variaciones en el orden de los tokens.

2.4.3. TRATAMIENTO DE LOS VALORES RUIDOSOS

Los valores ruidosos son valores fuera del rango normal de la variable y son detectados mediante funciones especiales. Estos valores por lo general son valores mal ingresados y presentan una distancia con los demás datos del conjunto. Existen una serie de teorías que describen como son usadas para descubrirlos y según sea el caso también existen funciones para tratarlos. Es posible nombrar: prueba de Grubbs (sección 2.4.3.1), prueba de Dixon (sección 2.4.3.2), prueba de Tukey (sección 2.4.3.3), análisis de valores ruidosos de Mahalanobis (sección 2.4.3.4) y detección de valores ruidosos mediante regresión simple (sección 2.4.3.5).

2.4.3.1. PRUEBA DE GRUBBS

Este método fue planteado por Frank E. Grubbs desde el año 1969 [Grubbs, 1969] y también es conocido como el método ESD (Extreme Studentized Deviate). La prueba de Grubbs se utiliza para detectar valores atípicos en un conjunto de datos univariantes y se basa en el supuesto de normalidad. Es decir, primero debe verificarse que sus datos pueden aproximarse razonablemente a una distribución normal antes de aplicar la prueba. Es especialmente fácil de seguir y sirve para detectar un valor atípico a la vez [Iglewicz y Hoaglin, 1993].

Esta técnica es muy fácil de usar y funciona bien bajo una variedad de condiciones incluyendo tamaños de muestra muy grandes, recordando que los datos deben provenir de una distribución normal.

2.4.3.2. PRUEBA DE DIXON

La prueba de Dixon permite determinar si un valor sospechoso de un conjunto de datos es un outlier. El método define la relación entre la diferencia del mínimo/máximo valor y su vecino más cercano y la diferencia entre el máximo y el mínimo valor aplicado [Li y Edwards, 2001].

Los datos deben provenir de una distribución normal. Si se sospecha que una población lognormal subyace en la muestra, la prueba puede ser aplicada al logaritmo de los datos. Antes de realizar el procedimiento es importante definir las hipótesis (si el valor sospechoso se encuentra al inicio o al final del conjunto de datos) y determinar la distribución de la que provienen los datos (normal o lognormal) [Davis y McCuen, 2005].

2.4.3.3. PRUEBA DE TUKEY

El diagrama conocido como diagrama de cajas y bigotes (Box and Whiskers Plot o simplemente BoxPlot) es un gráfico representativo de las distribuciones de un conjunto de datos creado por Tukey en 1977, en cuya construcción se usan cinco medidas descriptivas de los mismos: mediana, primer cuartil (Q1), tercer cuartil (Q3), valor máximo y valor mínimo [Tukey, 1977].

Está compuesto por un rectángulo o caja la cual se construye con ayuda del primer y tercer cuartil y representa el 50% de los datos que particularmente están ubicados en la zona central de la distribución, la mediana es la línea que atraviesa la caja, y dos brazos o bigotes son las líneas que se extienden desde la caja hasta los valores más altos y más bajos.

2.4.3.4. ANÁLISIS DE VALORES RUIDOSOS DE MAHALANOBIS

El Análisis de Valores ruidosos de Mahalanobis (Mahalanobis Outlier Analysis – MOA), es un método basado en una distancia, llamada distancia de Mahalanobis (DM). Esta distancia es calculada con base en la varianza de cada punto. Ésta describe la distancia entre cada punto de datos

y el centro de masa. Cuando un punto se encuentra en el centro de masa, la distancia de Mahalanobis es cero y cuando un punto de datos se encuentra distante del centro de masa, la distancia es mayor a cero. Por lo tanto, los puntos de datos que se encuentran lejos del centro de masa se consideran valores atípicos [Matsumoto *et. al.*, 2007].

2.4.3.5. DETECCIÓN DE VALORES RUIDOSOS MEDIANTE REGRESIÓN SIMPLE

El análisis de regresión es una importante herramienta estadística que se aplica en la mayoría de las ciencias. De muchas posibles técnicas de regresión, el método de mínimos cuadrados (LS) ha sido generalmente la más adoptada por tradición y facilidad de cálculo. Este método a través de unos cálculos, aproxima un conjunto de datos a un modelo, el cual puede ser lineal, cuadrado, exponencial, entre otros. Es decir, es una técnica de optimización, que intenta encontrar una función que se aproxime lo mejor posible a los datos. La diferencia entre el valor observado y el valor obtenido del modelo de regresión se denominan residuos o suma de cuadrados y el objetivo es tratar de minimizar este valor y así obtener el mejor ajuste [Rousseeuw y Leroy, 1996].

2.4.4. NORMALIZACIÓN

Cuando se habla de normalizar no se está hablando de normalización de bases de datos sino de normalización de variables o atributos. Normalizar es transformar una variable aleatoria que tiene alguna distribución en una nueva variable aleatoria con distribución normal o aproximadamente normal.

Existen varias técnicas de normalización [Botía, 2010]: Normalización mínimo – máximo (sección 2.4.4.1), Normalización a media cero (sección 2.4.4.2) y Normalización de escalado decimal (sección 2.4.4.3).

2.4.4.1. NORMALIZACIÓN MÍNIMO – MÁXIMO

Esta técnica utiliza la fórmula mostrada en la Figura 2.2.

Ejecuta una transformación lineal de los datos originales. Con base en los valores mínimo y máximo (Max_a y Min_a) de un atributo A y tomando un rango de variación ($RanMax_a$ y $RanMin_a$), se calcula un valor de normalización v' con base en el valor v [Botía, 2010].

$$v' = \frac{v\ Min_a}{Max_a - Min_a}(RanMax_a - RanMin_a) - RanMin_a$$

Figura 2.2. *Formula de normalización mínimo – máximo.*

2.4.4.2. NORMALIZACIÓN A MEDIA CERO

Los valores para un atributo A son normalizados basados en la media y la desviación estándar A (μ_a y σ_a). Un valor v de A es normalizado a v' con el cálculo de la función de la Figura 2.3 [Botía, 2010].

$$v' = \frac{v - \mu_a}{\sigma_a}$$

Figura 2.3. *Formula de normalización a media cero.*

2.4.4.3. NORMALIZACIÓN DE ESCALADO DECIMAL

Normaliza moviendo los puntos decimales de los valores del atributo A. El número de puntos decimales movidos depende del máximo valor absoluto de A, j es el entero más pequeño de Max(|v'|) < 1. Un valor v de A es normalizado a v' con el cálculo de la formula de la Figura 2.4 [Botía, 2010].

$$v' = \frac{v}{10^j}$$

Figura 2.4. *Formula de normalización de escalado decimal.*

Es de notar, que la normalización puede cambiar los datos originales un poco, especialmente los dos últimos métodos mencionados. También es necesario guardar los parámetros como la media o desviación estándar para uso futuro y que se pueda normalizar de manera uniforme.

2.4.5. TRATAMIENTO DE SERIES

En el momento de determinar qué hacer con las series de datos es importante entender que las series o sucesiones de datos son un conjunto de datos que tienen características (patrones) que los relacionan con otros datos formando series o sucesiones. Cada sucesión de datos es un conjunto de registros relacionados en los datos. Estos conjuntos comúnmente están relacionados con una variable tiempo [Neftalí, 2006].

Un patrón secuencial consiste de una serie de registros que caracterizan a un conjunto. Se puede decir que el problema de encontrar estos patrones es minimizar la intervención del ser humano.

Para la determinación de patrones se puede utilizar un algoritmo propuesto por Quest de IBM, el cual conduce a la solución utilizando una serie de pasos [Neftalí, 2006].

- Ordenamiento. Convierte la base de datos en sucesiones.
- Ítem. Se encuentra el conjunto de todos los ítems L.
- Transformación. Se necesita determinar repetidamente si en un conjunto dado de sucesiones grandes existe una sucesión de clientes. Para hacer esta prueba rápidamente, transforma cada sucesión de cliente en una representación alternativa. En una sucesión de cliente

transformada, cada transacción es reemplazada por el conjunto de todos los ítems contenidos en esa transacción. Si una sucesión de cliente no contiene ningún ítem, esta sucesión es desechada de la base de datos transformada. Sin embargo, todavía contribuye en el conteo total de clientes. una sucesión de cliente se representa ahora por una lista de conjuntos de los ítems.

- Sucesión. Se utiliza el conjunto de ítems para encontrar las sucesiones deseadas.
- Máxima. Encuentra las sucesiones máximas entre el conjunto de sucesiones grandes. En ciertos algoritmos esta fase es combinada con la fase de sucesión para reducir el tiempo al contar las sucesiones no máximas.

2.4.6. REDUCIR EL ANCHO DE LOS DATOS, ES DECIR LA CANTIDAD DE COLUMNAS

En ocasiones puede suceder que ciertas variables o columnas no son necesarias para el propósito del proceso por razones específicas dado que cada uno de sus valores arrojan resultados muy similares. Por esta razón es necesario descartar estas variables con el objetivo de reducir los tiempos de cálculo de los algoritmos ejecutados en la etapa de modelado.

2.4.7. REDUCIR LA PROFUNDIDAD DE LOS DATOS, ES DECIR LA CANTIDAD DE REGISTROS

Con el fin de reducir la cantidad de registros se pueden definir algunas técnicas estadísticas. En la estadística, la teoría de muestreo, también conocido como estimación estadística, o el método representativo, se ocupa del estudio de los métodos adecuados de selección una muestra representativa de una población, con el fin de estudiar valores estimativos que caractericen a los miembros de una población [Neyman, 1934]. Dado que las características estudiadas sólo pueden ser estimadas a partir de la muestra, se calculan intervalos de confianza para dar el rango de valores dentro del cual el valor real caerá, con una probabilidad dada. Hay una cantidad de métodos de muestreo. Algunos métodos parecen ser más adecuado que otros. Por ejemplo: Muestreo aleatorio simple (sección 2.4.7.1), Muestreo aleatorio sistemático (sección 2.4.7.2), Muestreo estratificado (sección 2.4.7.3) y Muestreo aleatorio por conglomerado (sección 2.4.7.4).

2.4.7.1 MUESTREO ALEATORIO SIMPLE

Consiste en seleccionar elementos aleatorios, de la población, P, a ser estudiada. El método de selección simple puede ser con reemplazo (SRSWR) o sin remplazo (SRSWOR). Para poblaciones muy grandes, sin embargo, SRSWR y SRSWOR son equivalentes. Para el muestreo simple

aleatorio, las probabilidades de inclusión de los elementos pueden o no ser uniforme. Si las probabilidades no son uniformes, se obtiene una muestra aleatoria ponderada [Neyman, 1934].

2.4.7.2 MUESTREO ALEATORIO SISTEMATICO

En este caso se elige el primer individuo al azar y el resto viene condicionado por aquél. Este método es muy simple de aplicar en la práctica y tiene la ventaja de que no hace falta disponer de un marco de encuesta elaborado. Puede aplicarse en la mayoría de las situaciones, la única precaución que debe tenerse en cuenta es comprobar que la característica que se estudia no tenga una periodicidad que coincida con la del muestreo [Casal y Mateu, 2003].

2.4.7.3 MUESTREO ESTRATIFICADO

Se divide la población en grupos en función de un carácter determinado y después se muestrea cada grupo aleatoriamente, para obtener la parte proporcional de la muestra. Este método se aplica para evitar que por azar algún grupo esté menos representado que el resto [Jordi y Enric, 2003].

2.4.7.4 MUESTREO ALEATORIO POR CONGLOMERADOS.

Se divide la población en varios grupos de características parecidas entre ellos y luego se analizan completamente algunos de los grupos, descartando los demás. Dentro de cada conglomerado existe una variación importante, pero los distintos conglomerados son parecidos. Requiere una muestra más grande, pero suele simplificar la recogida de muestras. Frecuentemente los conglomerados se aplican a zonas geográficas [Jordi y Enric, 2003].

2.5. INSPECCIÓN DE LOS DATOS

En la etapa de inspección de los datos, las teorías involucradas dependen de cada proyecto de explotación de información ya que trata de probar los datos utilizando los algoritmos o procesos de modelado de cada proyecto. Por estas razones no se va a desarrollar una teoría específica para esta etapa del proceso.

3. DESCRIPCIÓN DEL PROBLEMA

El presente capítulo presenta el problema de investigación partiendo de las dificultades que hoy en día poseen las organizaciones al momento de ejecutar proyectos de explotación de información sobre los repositorios de datos que se almacenan desde los sistemas existentes o deprecados.
En primer lugar se describe la identificación del problema de investigación (sección 3.1), luego se caracteriza el problema abierto (sección 3.2) y se concluye con un sumario de investigación (sección 3.3).

3.1. IDENTIFICACIÓN DEL PROBLEMA DE INVESTIGACIÓN

Las empresas suelen generar grandes cantidades de información sobre sus procesos productivos, desempeño operacional, mercados y clientes. Pero el éxito de los negocios depende por lo general de la habilidad para ver nuevas tendencias o cambios en los datos.
La aplicación de proyectos de explotación de información sobre los datos puede identificar tendencias y comportamientos, no sólo para extraer información, sino también para descubrir las relaciones en bases de datos que pueden identificar comportamientos.
La limpieza de datos dentro de un proyecto de explotación de información es una de las tareas más costosas y se calcula que consume un 60% del total del esfuerzo de ejecución del proyecto [Merlino, 2004].
Se puede afirmar que la limpieza de datos, es un trabajo sumamente tedioso y que pocas veces se puede automatizar totalmente, debido al desconocimiento de las combinaciones que se puedan llegar a producir en grandes volúmenes de datos [Merlino, 2004].
Con el fin de disminuir el esfuerzo en el trabajo de la limpieza de los datos se presenta como primer problema el de la organización. Al no haber un proceso de limpieza de datos disponible, la tarea de limpieza de datos suele ser desorganizada y además no cuenta con tareas especificas para asegurar la calidad.
Por otra parte, al no contar con un proceso, tampoco existe un circuito de mejora a partir de la gestión del conocimiento sobre la limpieza de los datos. Esto mejoraría sustancialmente el proceso a medida que se ejecutan distintos procesos de limpieza de datos a través del tiempo.

3.2. PROBLEMA ABIERTO

El problema abierto que se identifica en la presente sección, consiste en que la gran cantidad del esfuerzo que conlleva el proceso de explotación de información, parte de la falta de procesos de

limpieza de datos organizados y documentados para lograr el fin, asegurando la calidad de los datos y reduciendo los esfuerzos de la ejecución de esta tarea, larga y tediosa, en todo proyecto de explotación de información.

Por otra parte la falta de documentación de procesos anteriores que ayudarían en gran medida a procesos futuros.

3.3. SUMARIO DE INVESTIGACIÓN

De lo expuesto precedentemente surgen las siguientes preguntas de investigación:

Pregunta 1: ¿Se puede plantear un proceso de limpieza de datos divido en actividades que abarque toda las tarea de transformación de datos en proyectos de explotación de información? En caso afirmativo: ¿Cuales son las actividades?

Pregunta 2: ¿En caso de existir la posibilidad de dividir en actividades, se puede identificar una serie de técnicas genéricas que se ejecuten en cada una de las actividades? De ser posible: ¿Cuáles son las técnicas y como deben ser ejecutadas dentro de cada actividad?

Pregunta 3: ¿En caso de existir estas técnicas, se puede diferenciar una entrada y una salida a cada una de estas de tal forma de diferenciarlas dentro del proceso? De ser posible: ¿Cuáles son las entradas/salidas de cada técnica?

Pregunta 4: ¿Se puede generar una documentación de tal forma de poder colaborar con la gestión del conocimiento? De ser posible: ¿Como seria el proceso de documentación de cada actividad?

4. SOLUCIÓN

En este capítulo se presenta: Cuestiones generales sobre la solución (sección 4.1), una propuesta de proceso de transformación de datos para proyectos de explotación de información (sección 4.2), la estructura general del proceso (sección 4.3) y sus actividades (sección 4.4).

4.1. CUESTIONES GENERALES

En función del análisis realizado en el capítulo 3 correspondiente a la Descripción del Problema, se considera de interés citar nuevamente el problema abierto que se aborda en este trabajo, recordando que el mismo se focaliza en una de las tareas más costosas y se calcula que consume un 60% del total del esfuerzo de la ejecución de proyectos de explotación de información [Merlino, 2004].

Puede afirmarse que esta tarea, es un trabajo sumamente tedioso y que pocas veces se puede automatizar totalmente, debido al desconocimiento de las combinaciones que se puedan llegar a producir en grandes volúmenes de datos [Merlino, 2004].

La solución que se propone en este trabajo consiste en el desarrollo de un procesos que ayude a disminuir este esfuerzo con el fin de hacer menos costosos los proyectos de explotación de información y de esta forma poder ejecutarlo con mayor frecuencia dentro de las organizaciones.

Por medio de un proceso las tareas de transformación de datos se organizan por tipos y propósitos y se ejecutan una a una hasta lograr resultados de calidad para los pasos posteriores de los proyectos de explotación de información.

4.2. PROPUESTA DE PROCESO DE TRANSFORMACIÓN DE DATOS PARA PROYECTOS DE EXPLOTACIÓN DE INFORMACIÓN

La propuesta de proceso estará formada por una serie de actividades que cumplan funciones específicas y bien divididas.

Cada una de estas actividades tendrá en sí que ejecutar una cierta técnica de tal forma de cumplir con su cometido.

Cada técnica deberá cumplir con una serie de pasos con el fin de transformar los productos de entrada de las actividades en productos de salida.

Cada actividad dependerá de la salida de alguna de las actividades anteriores por lo que su cumplimiento correcto es fundamental.

Cada actividad a su vez aportará de alguna forma a la gestión del conocimiento de la compañía con el fin de disminuir la dificultad de las futuras ejecuciones del proceso.

4.3. ESTRUCTURA GENERAL DEL PROCESO

Para lograr el objetivo se plantea primeramente un proceso que abarque todo lo necesario. El proceso contará con una serie de actividades fundamentales que son las siguientes:

- Enriquecer los datos.
- Obtener y ejecutar de los casos testigo.
- Determinar y aplicar la estructura de los datos.
- Construir el modelo de entrada de datos.
- Inspeccionar los datos.

Mediante la tabla 4.1 se puede observar cómo se distribuyen las técnicas, entradas y salidas de cada actividad del proceso.

Actividad	Productos de Entrada		Técnica de transformación	Productos de Salida	
	Entrada	Representación		Salida	Representación
Enriquecer los datos	• Datos posiblemente sucios (DPS) • Información sobre el Proyecto de Explotación de Información (IPEI)	• Archivo plano, SQL, XLS, Access, etc. • Documento formateado	Técnica de Enriquecimiento de los datos (TED)	• Datos Sucios (DS) • Documento de Solución (DO-SO) • Documento de Técnicas de Modelado (DO-TM)	• Archivo plano, SQL, XLS, Access, etc. • Documento formateado • Documento formateado
Obtener y ejecutar de los casos testigo	• Datos Sucios (DS) • Documento de Solución (DO-SO) • Documento de Técnicas de Modelado (DO-TM)	• Archivo plano, SQL, XLS, Access, etc. • Documento formateado • Documento formateado	Técnica de obtención y ejecución de los casos testigo (TOECT)	• Datos Válidos (DV) • Documento de Listas de Chequeo (DO-LC)	• Archivo plano, SQL, XLS, Access, etc. • Documento formateado
Determinar y aplicar la estructura de los datos	• Datos Válidos (DV) • Documento de Solución (DO-SO) • Documento de Técnicas de Modelado (DO-TM)	• Archivo plano, SQL, XLS, Access, etc. • Documento formateado • Documento formateado	Técnica de Determinar y Aplicar la Estructura de los Datos (TDAED)	• Documento de Integración (DO-IN) • Datos Integrados (DI)	• Archivo SQL. • Documento formateado
Construir el modelo de entrada de datos	• Datos Integrados (DI)	• Archivo SQL.	Técnica de Construir el Modelo de Entrada de Datos (TCMED)	• Documento de Estructuración de los datos (DO-ES) • Datos Estructurados (DE)	• Archivo SQL. • Documento formateado
Inspeccionar los datos	• Datos Estructurados (DE) • Documento de Estructuración de los datos (DO-ES) • Documento de Integración (DO-IN) • Documento de Solución (DO-SO) • Documento de Técnicas de Modelado (DO-TM) • Documento de Listas de Chequeo (DO-LC)	• Archivo SQL. • Documento formateado • Documento formateado • Documento formateado • Documento formateado • Documento formateado	Técnica de Inspección de los datos (TIND)	• Datos de calidad (DC)	• Archivo SQL.

Tabla 4.1. *Distribución de tareas, entrada y salida de cada actividad.*

Estas actividades están organizadas mediante un circuito organizado de tal forma de que cada una se ejecute en un momento específico dado que estas dependen de la anterior. El circuito propuesto se muestra en la figura 4.1.

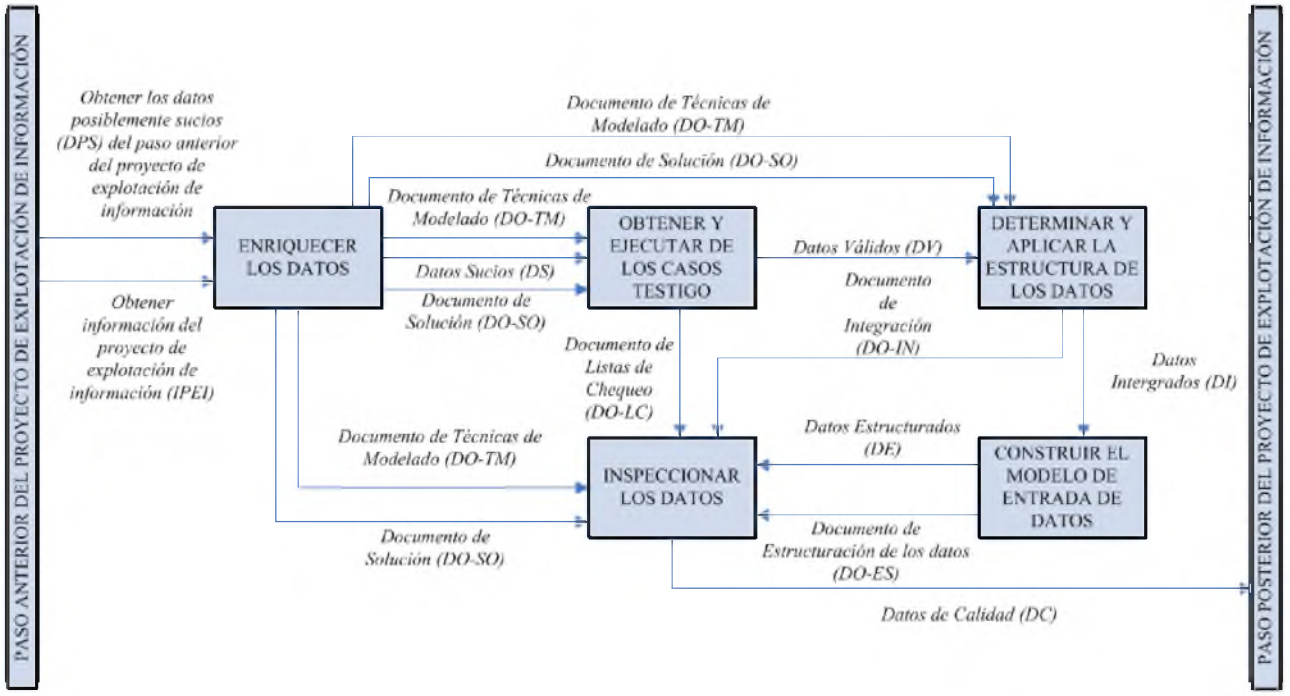

Figura 4.1. *Diagrama de flujo del proceso completo.*

4.4. ACTIVIDADES

En esta sección se presenta las actividades del proceso las cuales están dividas en Enriquecer los datos (sección 4.4.1), Obtención y ejecución de los casos testigo (sección 4.4.2), Determinar y aplicar la estructura de los datos (sección 4.4.3), Construir el modelo de entrada de datos (sección 4.4.4) e Inspeccionar los datos (sección 4.4.5).

4.4.1. ENRIQUECER LOS DATOS

En esta actividad se reciben los datos que se obtuvieron en momentos previos al proceso actual por ende hay que abstraerse de cómo se obtienen dado que para cada organización esto puede ser más o menos engorroso.

También se obtiene la información sobre el proyecto de explotación de información.

Una vez con los datos en el poder, el primer paso, para la preparación de datos es conocer el problema a resolver, o al menos hacia qué objetivo se quiere llegar. Sin esto resulta imposible afirmar que los datos con los que se cuenta son los correctos para continuar con el proceso. También es necesario conocer la forma en que se debe presentar la información al modelo seleccionado para la explotación de datos.

Para determinar que transformaciones es necesario realizar y como se debe presentar, el equipo de trabajo se debe plantear dos preguntas que serán de guía para esta actividad:

- ¿Qué solución deben obtener?

- ¿Qué técnica de explotación se utilizarán?

Luego de analizar la información y teniendo respuesta a las preguntas antes generadas, se puede continuar con el proceso actual.

En la tabla 4.2 se muestra la serie de pasos que determinan la técnica de enriquecimiento de los datos.

Esta técnica se puede describir mediante el gráfico de la figura 4.2.

Técnica de Enriquecimiento de los datos (TED)	
Entradas:	Datos posiblemente sucios (DPS)
	Información sobre el Proyecto de Explotación de Información (IPEI)
Salidas:	Datos Sucios (DS)
	Documento de Solución (DO-SO)
	Documento de Técnicas de Modelado (DO-TM)
Paso 1.	Conocer el Problema a Resolver
Paso 2.	Analizar Solución a Obtener
Paso 3.	Generación de Documento de Solución
Paso 4.	Analizar Técnicas de Modelado a Utilizar
Paso 5.	Generación de Documento de Técnicas de Modelado

Tabla 4.2. *Técnica de Enriquecimiento de los datos (TED)*

Paso 1. Conocer el Problema a Resolver: Una vez obtenida la información sobre el proyecto de explotación de información se procede a analizar todo lo relacionado con las técnicas de modelado, los datos obtenidos y la forma en que deben presentarse sobre cada una de las técnica de modelado mencionadas.

Paso 2. Analizar Solución a Obtener: En este caso se debe inferir en el resultado que es necesario obtener para poder tener un acercamiento sobre los datos requeridos para el este propósito en especial.

Paso 3. Generación de Documento de Solución: Se genera un documento donde conste todo lo obtenido del paso anterior.

Paso 4. Analizar Técnicas de Modelado a Utilizar: Se analiza las técnicas de modelado a utilizar en los pasos posteriores del proyecto de explotación de información, centrándose en todo lo relacionado con los datos de entrada para cada una, con el fin de detallar cuáles van a ser los requerimientos de los datos de entrada para dichas técnicas.

Paso 5. Generación de Documento de Técnicas de Modelado: En este paso se genera la documentación de toda la información obtenida en el paso anterior.

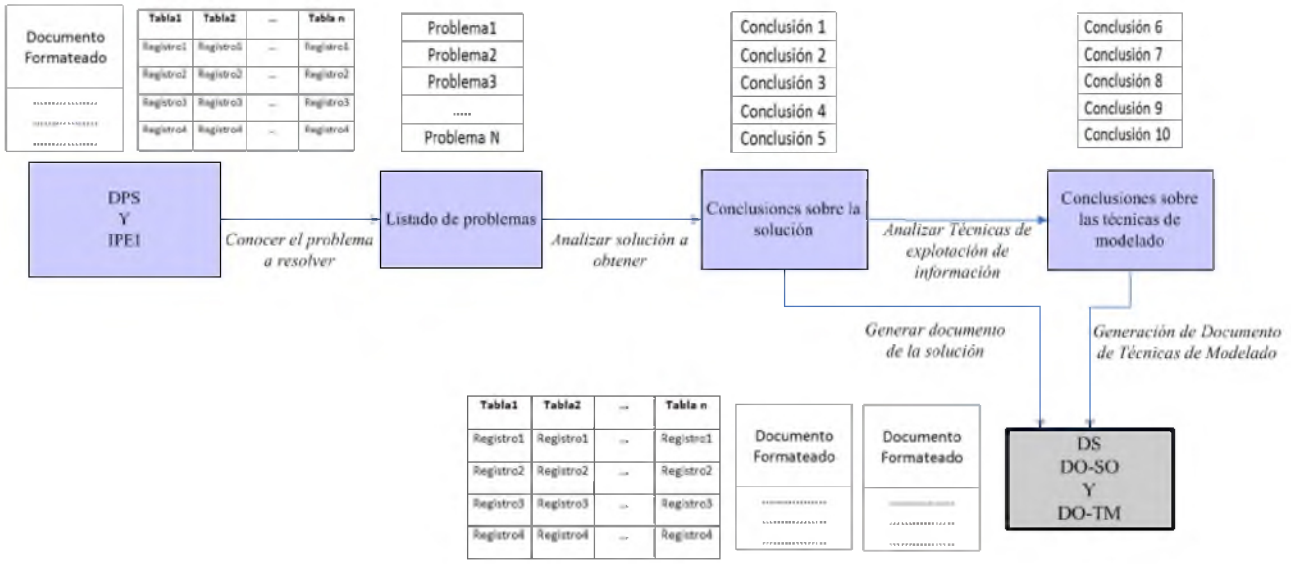

Figura 4.2. *Diagrama de flujo de la técnica de Enriquecimiento de los datos.*

4.4.2. OBTENCIÓN Y EJECUCIÓN DE LOS CASOS TESTIGO

La obtención de los casos testigo puede convertirse en un proceso muy tedioso dado que esto nos permitirá definir si el modelo al que lo aplicaran es viable o no en relación al conjunto de datos que se obtuvieron del paso anterior.

Estos casos son listados de ítems a tener en cuenta de los datos, definen cuales son los atributos a tener en cuenta, formato, tamaño, etc.

La ejecución de los casos es simplemente efectuar pruebas sobre los datos y los resultados son los que deciden si estos datos son los correctos para continuar el proceso.

Si se encuentran problemas en los datos se deberá ejecutar nuevamente la actividad de enriquecimiento.

En la tabla 4.3 se muestra la serie de pasos que determinan la técnica de obtención y ejecución de los casos testigo. Esta técnica se puede describir mediante el gráfico de la figura 4.3.

Paso 1. Planteo de los casos testigo: Partiendo de los documentos DO-SO y DO-TM se genera una serie de ítems que sirven como pautas a tener en cuenta con el fin de determinar la viabilidad del proceso.

Paso 2. Generar Lista de Chequeo: Confeccionar la listas de chequeo partiendo de las pautas obtenidas en el paso anterior

Paso 3. Test de los datos: Se procede a efectuar el test de los datos partiendo de las pautas de las listas de chequeo rellenando los campos de la misma y obtener conclusiones.

Paso 4. Documentar conclusiones: Se pondera los resultados con el fin de hacer una evaluación de viabilidad del proceso con respecto a requerimiento de los datos y a las entradas de las técnicas de modelado. Se da como validados los datos para continuar con el proceso.

	Técnica de Obtención y ejecución de los casos testigo (OECT)
Entradas:	Datos Sucios (DS)
	Documento de Solución (DO-SO)
	Documento de Técnicas de Modelado (DO-TM)
Salidas:	Documento de Listas de Chequeo (DO-LC)
	Datos Válidos (DV)
Paso 1.	Planteo de los casos testigo.
Paso 2.	Generar Lista de Chequeo.
Paso 3.	Test de los datos.
Paso 4.	Documentar conclusiones.

Tabla 4.3. *Técnica de Obtención y ejecución de los casos testigo (OECT)*

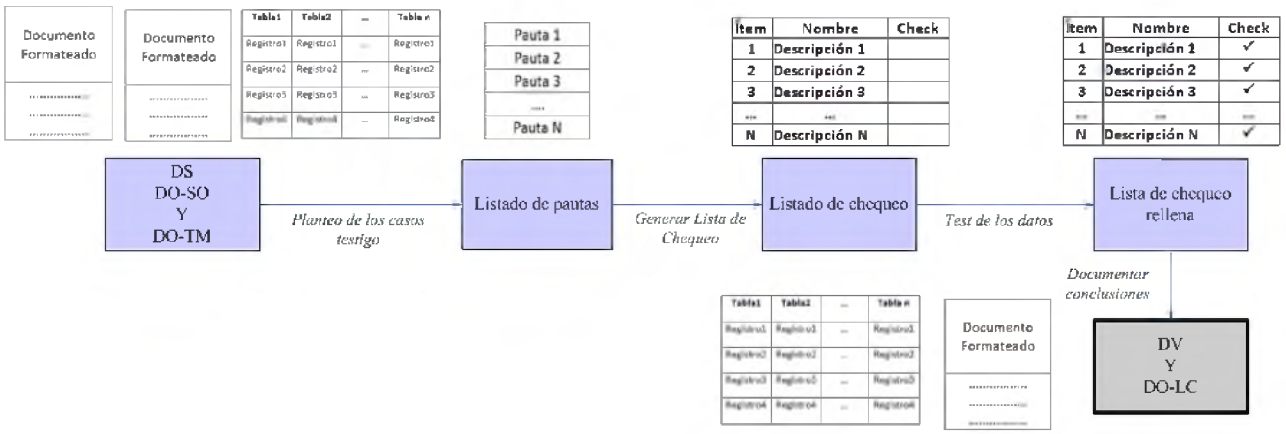

Figura 4.3. *Diagrama de flujo de la técnica de obtención y ejecución de los casos testigo*

4.4.3. DETERMINAR Y APLICAR LA ESTRUCTURA DE LOS DATOS

Para poder entender este concepto es necesario definir el término conjunto de datos. El mismo abarca a los datos que serán utilizados por proceso de explotación de información.

La estructura de datos hace referencia a la forma en que las variables se relacionan unas con otras en los conjuntos de datos. Es en esta estructura donde se buscarán relaciones y patrones de comportamiento.

Esta actividad es necesaria dado que los datos pueden provenir de diferentes fuentes y tener distintos formatos o incluso estar en distintos lugares.

Para esta actividad se pueden utilizar técnicas manuales y asistidas por computadora para lograr una única estructura y a su vez los datos más completos.

En la tabla 4.4 se muestra la serie de pasos que determinan la técnica de determinación y aplicación de la estructura de los datos. Esta técnica se puede describir mediante el gráfico de la figura 4.4.

Técnica de Determinación y Aplicación de la Estructura de los Datos (TDAED)	
Entradas:	Datos Válidos (DV)
	Documento de Solución (DO-SO)
	Documento de Técnicas de Modelado (DO-TM)
Salidas:	Documento de Integración (DO-IN)
	Datos Integrados (DI)
Paso 1.	Determinar las fuentes de los datos.
Paso 2.	Determinar las relaciones.
Paso 3.	Unificar Tipos de datos.
Paso 4.	Unificar Rangos de variables.
Paso 5.	Generar documento de integración.

Tabla 4.4. *Técnica de Determinación y Aplicación de la Estructura de los Datos (TDAED)*

Paso 1. Determinar las fuentes de los datos: En este paso se determina de que tipo de fuentes provienen los datos para comenzar con la integración.

Paso 2. Determinar las relaciones: Una vez que se entiende las fuentes se comienza la integración partiendo de las relaciones entre las distintas fuentes de datos para determinar que tablas están relacionadas para concluir con una tabla única.

Paso 3. Unificar Tipos de datos: Si los datos unificados son de distintos tipos hay que unificar criterios y determinar un solo tipo de datos.

Paso 4. Unificar Rangos de variables: Si hay variables discretas se debe unificar los rangos de las mismas hasta obtener un solo rango de variable.

Paso 5. Generar documento de integración: Se reúnen todos los criterios utilizados para la integración y se almacenan en el documento de integración.

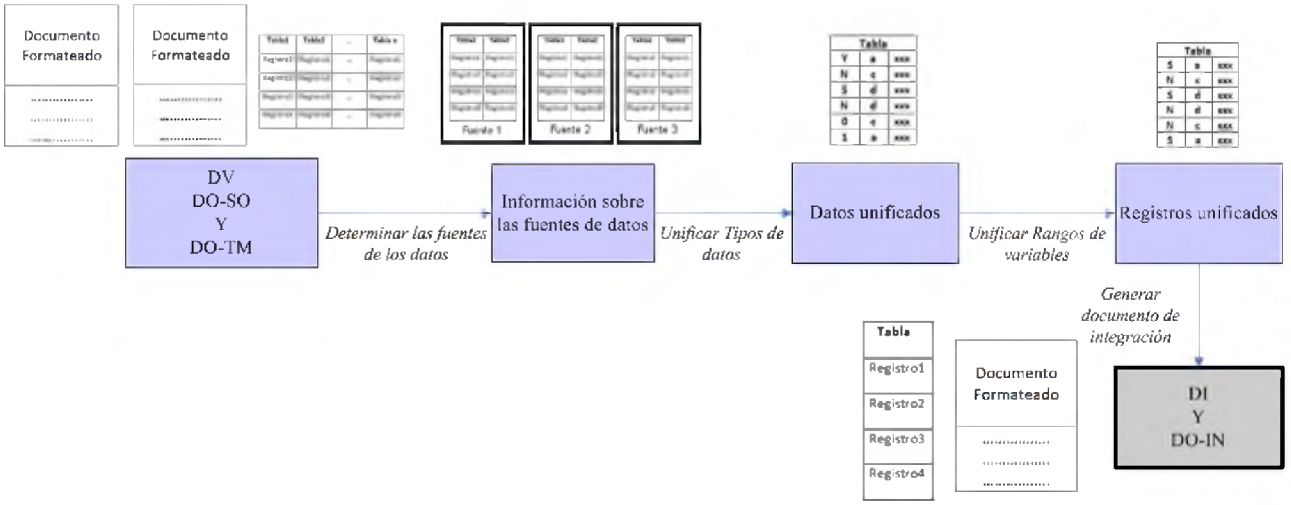

Figura 4.4. *Diagrama de flujo de la técnica de determinación y aplicación de la estructura de datos*

4.4.4. CONSTRUIR EL MODELO DE ENTRADA DE DATOS

Hasta esta actividad el proceso se centra en obtener y conocer los datos disponibles y se han adaptado las diferentes fuentes de datos. Lo que debería suceder en esta actividad es determinarán los procesos que se seguirán para el modelado de los datos, entre los cuales es posible nombrar:

- Tratamiento de los valores nulos o vacíos.
- Eliminación de duplicados.
- Tratamiento de los valores ruidosos,
- Tratamiento de series (las más comunes de tiempo).
- Reducir el ancho de los datos, es decir la cantidad de columnas.
- Reducir la profundidad, la cantidad de registros.

Esta actividad es la que modela los datos de tal forma de darles una calidad suficiente para continuar con la siguiente actividad, que será la de evaluar el resultado. La calidad de los datos es tan importante para el proceso de explotación de información, dado que de no contar con esta el proceso podría tomar caminos diferentes y hasta incluso caminos inexistentes.

En la tabla 4.5 se muestra la serie de pasos que determinan la técnica de construcción del modelo de entrada de datos. Esta técnica se puede describir mediante el gráfico de la figura 4.5.

Paso 1. Efectuar análisis iniciales: En este paso se obtienen una serie de valores que determinan la calidad de los datos. Cada una de las fases de transformación necesita de estos valores por lo que son de suma importancia para continuar con la técnica. Ejemplos son: Cantidad de datos perdidos por variable, patrón de datos perdidos, valores fuera de rango, cantidad total de registros, cantidad total de variables, tipos de datos de cada variable, etc. También se

generan una serie de gráficos que, mediante el análisis, determinan información necesaria para continuar con la técnica.

Técnica de Construcción del Modelo de Entrada de Datos (TCMED)	
Entradas:	Datos Integrados (DI)
Salidas:	Documento de Estructuración de los datos (DO-ES)
	Datos Estructurados (DE)
Paso 1.	Efectuar análisis iniciales.
Paso 2.	Ejecutar las distintas fases de transformación.
Paso 3.	Generar documento de Estructuración.

Tabla 4.5. *Técnica de Construcción del Modelo de Entrada de Datos (TCMED)*

Paso 2. Ejecutar las distintas fases de transformación: Este paso de dividirá en fases de transformación. Cada fase efectuara uno de los tipos de transformación planteados en esta tesis como ser tratamiento de valores faltantes o nulos, tratamiento de valores fuera de rango, normalización, etc.

Paso 3. Generar documento de Estructuración: En este paso se generara un documento donde conste todos los valores obtenidos en el paso 1 y todo lo relacionado con la ejecución del paso 2.

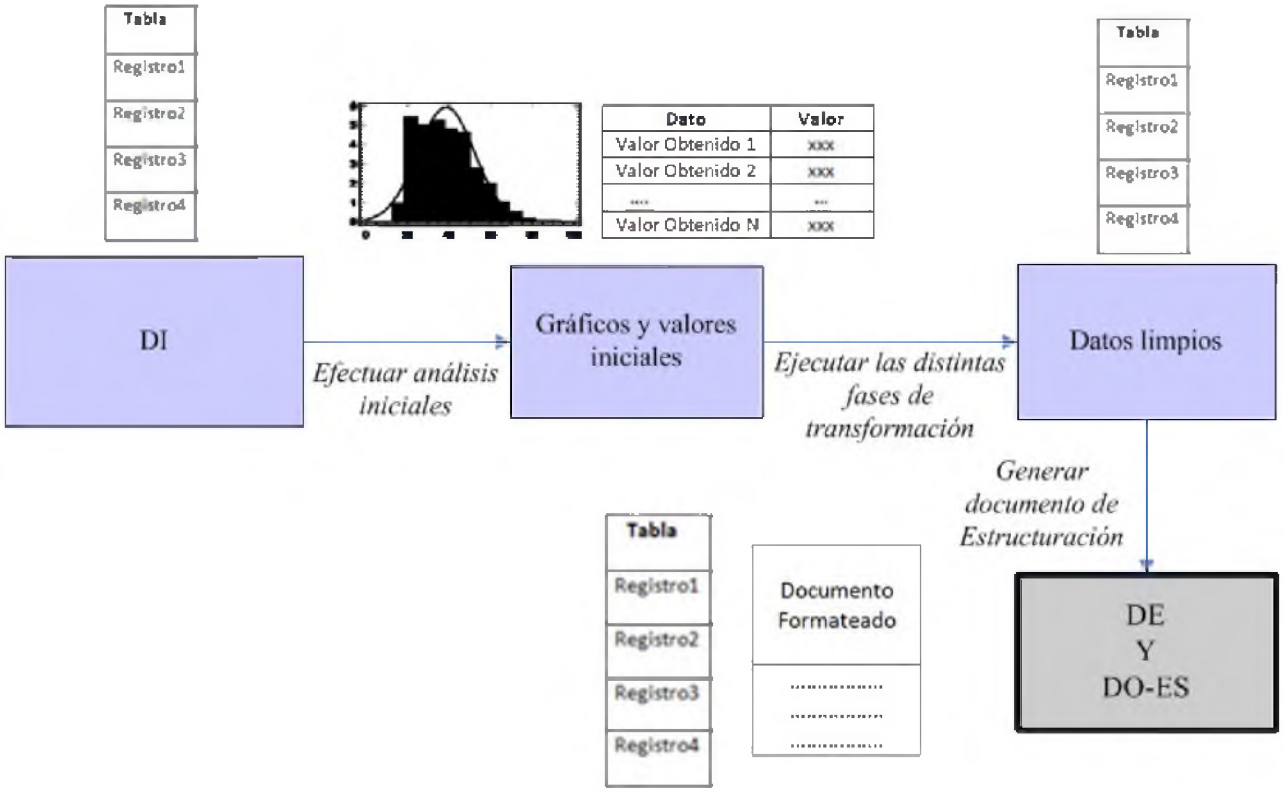

Figura 4.5. *Diagrama de flujo de la técnica de construcción del modelo de entrada de datos*

4.4.5. INSPECCIONAR LOS DATOS

En esta actividad se procede a analizar los datos resultantes de la actividad de construcción del modelo de entrada de los datos, para evaluar si estos datos resultantes son los convenientes para entregar al próximo paso del proyecto de explotación de información, dado que los mismos son los que harán viable el modelo elegido.

En la tabla 4.6 se muestra la serie de pasos que determinan la técnica para inspeccionar los datos.

Esta técnica se puede describir mediante un grafico (Figura 4.6).

Técnica de Inspección de los datos (TIND)	
Entradas:	Datos Estructurados (DE)
	Documento de Estructuración de los datos (DO-ES)
	Documento de Integración (DO-IN)
	Documento de Solución (DO-SO)
	Documento de Técnicas de Modelado (DO-TM)
	Documento de Listas de Chequeo (DO-LC)
Salidas:	Datos de calidad (DC)
Paso 1.	Efectuar la inspección de los datos.
Paso 2.	Actualizar el repositorio del conocimiento de la compañía.
Paso 3.	Preparar los datos para el siguiente paso del proyecto de EI.

Tabla 4.6. *Técnica de Inspección de los datos (TIND)*

Paso 1. Efectuar la inspección de los datos: Se efectúa una inspección completa de los datos para detectar problemas de calidad. Si es posible se ejecuta los algoritmos de modelado a un muestrario de registros con el fin de detectar problemas en la entrada de dichos algoritmos.

Paso 2. Actualizar el repositorio del conocimiento de la compañía: Se recopila toda la documentación generada en el proceso y se actualiza el repositorio de conocimiento de la compañía con el fin de dar un valor agregado a las próximas ejecuciones.

Paso 3. Preparar los datos para el siguiente paso del proyecto de EI: Se prepara para la entrega de los datos con una calidad óptima al siguiente paso del proyecto de explotación de información.

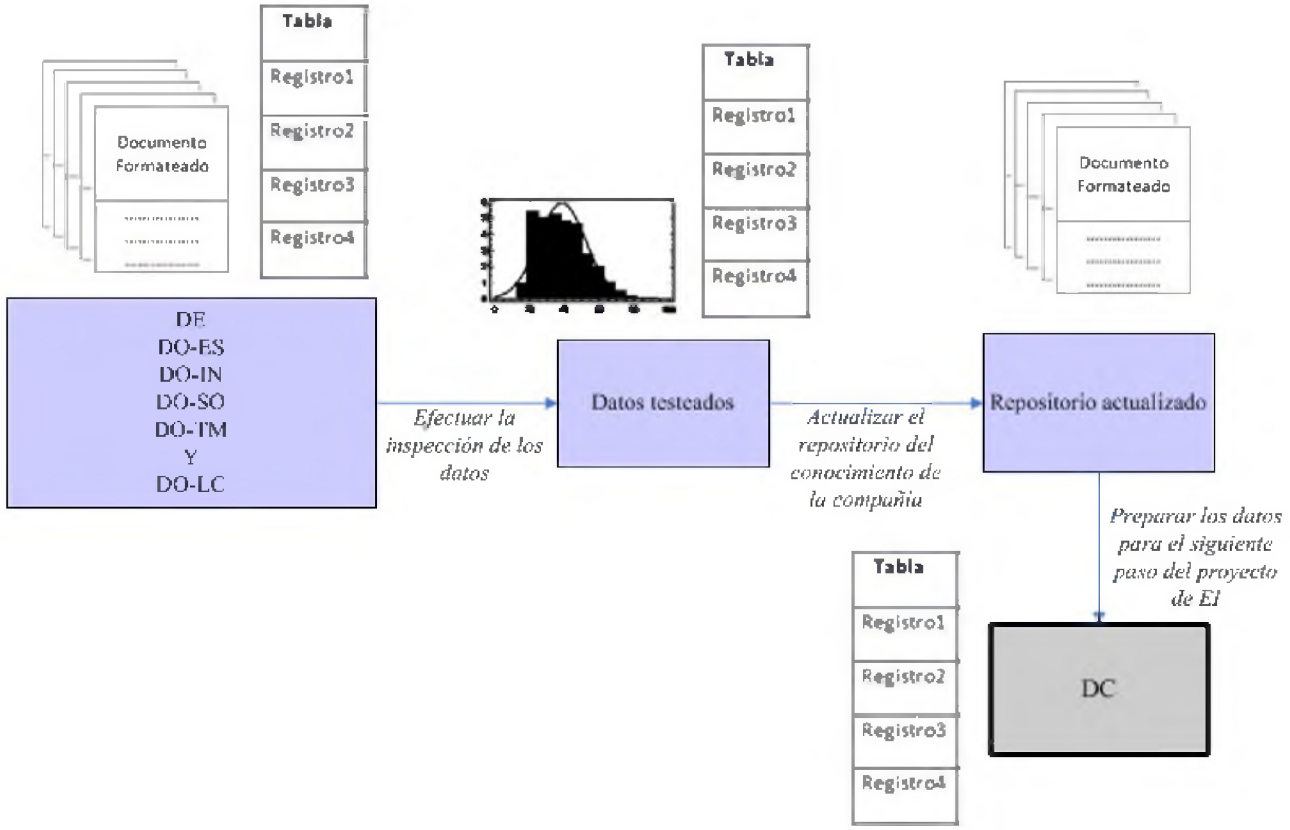

Figura 4.6. *Diagrama de flujo de la técnica de inspección de los datos*

5. CASOS DE VALIDACIÓN

En este capítulo se presenta la ejecución del proceso de transformación de datos en proyectos de explotación de información en casos de validación específicos, a los efectos de implementar las tareas correspondientes a cada actividad del proceso y evaluar su desempeño en cada uno de sus pasos.

Se analiza un caso de validación correspondiente a un banco el cual busca mejorar la cartera de depósitos a largo plazo (sección 5.1), otro de estos es el caso de validación para un conjunto de datos de pacientes indios con problemas hepáticos (sección 5.2) y por último el caso de validación sobre un conjunto de datos para el diagnóstico de cáncer de mama (sección 5.3).

5.1. CASO DE VALIDACIÓN: PREDICCIÓN DE POTENCIALES CLIENTES DE DEPÓSITOS A LARGO PLAZO

En esta sección se analiza el caso de validación correspondiente a la predicción de potenciales clientes de depósitos a largo plazo.

El problema parte de la necesidad de nuevos objetivos para el banco tratando de utilizar los datos almacenados en el sistema para buscar potenciales clientes que se suscriban a productos para mejor las ventas a partir de clientes activos. El departamento de Marketing propone buscar en los datos potenciales clientes para los productos depósito a largo plazo. La idea fundamental es partir de los datos obtenidos durante un determinado tiempo de movimientos de clientes activos y de esta manera predecir si el cliente va a suscribirse a un depósito a largo plazo.

Se ejecutará el proceso de transformación de datos propuesto ya que el problema no es más que un problema de explotación de información. Para comenzar se aplica la primer actividad, la actividad de enriquecimiento de los datos (sección 5.1.1), continúa aplicando la actividad de obtención y ejecución de los casos testigo (sección 5.1.2), luego se aplica la actividad para determinar y aplicar la estructura de los datos (sección 5.1.3), más adelante se aplica la actividad de construir el modelo de entrada de los datos (sección 5.1.4) y por último se aplica la actividad de inspección de los datos (sección 5.1.5).

5.1.1. APLICACIÓN DE LA ACTIVIDAD DE ENRIQUECIMIENTO DE LOS DATOS

En esta sección se aplica a este caso de validación la técnica utilizada para la ejecución de la actividad de enriquecimiento de los datos. Para esto se deben efectuar los diferentes pasos de la técnica: Paso 1: Conocer el Problema a Resolver (sección 5.1.1.1), Paso 2: Analizar Solución a

Obtener (sección 5.1.1.2), Paso 3: Generación de Documento de Solución (sección 5.1.1.3), Paso 4: Analizar Técnicas de Modelado a Utilizar (sección 5.1.1.4) y Paso 5: Generación de Documento de Técnicas de Modelado (sección 5.1.1.5).

5.1.1.1. PASO 1: CONOCER EL PROBLEMA A RESOLVER

En este paso se procede a efectuar el análisis del problema planteado. Para el caso de validación es el personal del área de Marketing quien propone un trabajo de investigación de los datos del banco en busca de potenciales clientes para depósitos a largo plazo. Estos depósitos son necesarios para que el banco cuente con el dinero necesario para poder crear nuevos productos del tipo de crédito los cuales no esteran en el alcance de este proyecto. Este problema puede transformarse en un problema de explotación de información.

5.1.1.2. PASO 2: ANALIZAR SOLUCIÓN A OBTENER

Dado que la solución a obtener es un listado de potenciales clientes para los depósitos a largo plazo, es necesario obtener como salida del proyecto de explotación de información un listado de potenciales clientes para depósitos a largo plazo. Para esto es necesario aplicar a los datos las técnicas de modelado de árboles de decisión (mas precisamente el algoritmo C4.5). A partir de esto, se evidencia que es necesario tener como datos de entrada, variables continuas y discretas; y como variable de salida del tipo discreta.

El proyecto plantea buscar estos potenciales clientes basándose en campañas anteriores por lo que se tomará como variable de salida a la variable “y” la cual representa con una “Y” los clientes que se suscribieron y con una “N” los que no. El resto de las variables se tomará coma variables de entrada.

5.1.1.3. PASO 3: GENERACIÓN DE DOCUMENTO DE SOLUCIÓN

En este documento se describe todo lo planteado en el paso 1 y en el paso 2 con el fin de aportar a la gestión del conocimiento datos de este proceso de transformación de datos. En el mismo se ejecutará un algoritmo de árboles de decisión con el fin de detectar reglas de pertenencia a grupos que se dividirán en dos: Los que “SI” se suscribieron a depósitos a largo plazo y los que “NO” se suscribieron a depósitos a largo plazo.

La ejecución de este algoritmo devolverá un árbol con reglas de pertenencia al “SI” y al “NO” y con esto es posible predecir potenciales clientes. Como lo único que nos interesa son los que “SI” se suscriban a los depósitos a largo plazo, solo se hará una extracción de este grupo dado que el grupo de la negación declara que el cliente no aceptará la propuesta realizada por el personal que lo contacta. De esta forma también se disminuye la cantidad de resultados obtenidos.

Lo siguiente es analizar las técnicas de modelado.

5.1.1.4. PASO 4: ANALIZAR TÉCNICAS DE MODELADO A UTILIZAR

Para el proyecto actual de explotación de información solo se utilizará el algoritmo C4.5 que es un algoritmo de árboles de decisión.

Los tipos de datos de entrada son uno o más valores continuas y discretas; y variable de salida es solo una variable del tipo discreta.

Este tipo de algoritmos construye un árbol de decisión con una serie de reglas que separan los que pertenecen a un grupo o a otro (lo que se llama generalmente como reglas de partencia a grupo).

5.1.1.5. PASO 5: GENERACIÓN DE DOCUMENTO DE TÉCNICAS DE MODELADO

En este documento se describe todo lo necesario sobre las técnicas de modelado con el fin de aportar a la gestión del conocimiento, datos sobre estos algoritmos. El formato y contenidos se deben especificar en cada organización y no se presenta un modelo específico sino que solo se describe que información se almacena en él. En este caso solo se aporta que se utilizara el algoritmo C4.5 que es un algoritmo de árboles de decisión.

También se explica que los tipos de datos de entrada son uno o más valores continuos y discretos ; y variable de salida es solo una variable del tipo discreta, de esta forma en procesos futuros no hará falta una investigación de dicho algoritmo.

5.1.2. APLICACIÓN DE LA ACTIVIDAD DE OBTENCIÓN Y EJECUCIÓN DE LOS CASOS TESTIGO

En esta sección se aplica al caso de validación actual la técnica utilizada para la ejecución de la actividad de obtención de los casos testigo: Para esto se deben efectuar los diferentes pasos de la técnica: Paso 1: Planteo de los casos testigo (sección 5.1.2.1), Paso 2: Generar Lista de Chequeo (sección 5.1.2.2), Paso 3: Test de los datos (sección 5.1.2.3) y Paso 4: Documentar conclusiones (sección 5.1.2.4).

5.1.2.1. PASO 1: PLANTEO DE LOS CASOS TESTIGO

Para esta tarea es necesario confeccionar la lista de chequeos para controlar si los datos son los correctos o si es necesaria la obtención de nuevos datos. Para esto es necesario analizar las técnicas de explotación de información a ejecutar para determinar qué tipos de datos serán requeridos.

Árboles de decisión (DT): Tiene como atributo de entrada valores continuos y discretos y atributo de salida una variable discreta.

El personal de Marketing nos informa que los datos necesarios para la predicción de potenciales clientes de depósitos a largo plazo son:

- Se selecciona la edad dado que es importante no contactar a menores de 18 años ya que la ley nacional no lo permite. También sirve como relación con la edad.
- La ocupación de la persona. Este datos es importante dado que nos dará información de posibles clientes si es que existe una relación entre las personas con el mismo trabajo o incluso separar a aquellos que no tiene o es desconocido.
- El estado civil. Hay que tener en cuenta, si está casado, que el cónyuge debe presentarse y estar de acuerdo. Por lo tanto es más difícil vender un producto a personas casadas que al resto.
- Nivel de educación. Se podrá ver el nivel educativo de las personas que se interesan en este producto.
- Si la persona posee un crédito en mora dado que estas personas son no confiables para el banco.
- Salario promedio anual en pesos. Para dar cuenta del poder adquisitivo de la persona.
- Si posee un crédito hipotecario para la vivienda: si es que este tipo de producto son los que más le interesa a estos clientes.
- Si posee préstamo personal: si es que este tipo de producto son los que más le interesa a estos clientes.
- Es importante tener en cuenta si hubo contacto con respecto a la campaña actual, por ende, estos datos serán importantes:
- El tipo de comunicación con el cual se contactó con la persona.
- El día del último contacto dentro del mes.
- El mes del último contacto dentro del año.
- La duración del último contacto, en segundos.
- Número de contactos realizados durante esta campaña y para este cliente.
- Número de días que han pasado desde que se puso en contacto con el cliente de una campaña anterior.
- Número de contactos realizados a este cliente antes de esta campaña. Es posible que algunas personas nunca se las haya contactado.
- Resultado de la campaña anterior.

En reuniones con el personal de Marketing y de bases de datos se llegó a la conclusión de que estos datos son los necesarios para continuar con el proceso.

Los datos entregados por el área de almacenamiento, están listos para continuar con el proceso de transformación de datos. Según el análisis efectuado sobre los datos dio como resultado la descripción que se encuentra en la tabla 5.1:

Campo	Tipo de dato
Fecha de nacimiento	Fecha.
Ocupación	Numero. Referenciado de tabla tipos de ocupación.
Estado civil	Numero. Referenciado de tabla Estados.
Educación	Numero. Referenciado de tabla Educaciones.
Tipo de préstamo	Numero. Referenciado de tabla tipos de préstamo.
Salario promedio	Numero flotante.
Saldo del préstamo	Numero flotante.
Contacto	Numero. Referenciado de tabla Tipos de contacto.
Fecha contacto	Fecha.
Duración	Hora.
Suscrito	Binario.

Tabla 5.1. *Descripción de los datos.*

5.1.2.2. PASO 2: GENERAR LISTA DE CHEQUEO

En este paso se genera el listado de chequeo con la información recolectada en el paso anterior. La tabla 5.2 muestra la lista de chequeo para este proceso.

Lista de Chequeo de evaluación de los datos

1=No Satisfactorio, 2=Parcialmente satisfactorio, 3=Completamente Satisfactorio, N/A=No aplica

		1	2	3	N/A
1. Controles Generales					
	A. Datos del tipo fecha	①	②	③	○
	B. Datos del tipo texto	①	②	③	○
	C. Datos numéricos	①	②	③	○
	D. Datos binarios	①	②	③	○
2. Controles sobre DT					
	A. Aplicabilidad del dominio	①	②	③	○
	B. Cuenta con las herramientas necesarias.	①	②	③	○
Comentarios:					

Tabla 5.2. *Lista de chequeo.*

5.1.2.3. PASO 3: TEST DE LOS DATOS

Una vez confeccionada la lista de chequeo se procede a la ejecución de las pruebas (test) para luego evaluar los resultados obtenidos. Este proceso se impactó en la tabla 5.3.

Lista de Chequeo de evaluación de los datos

1=No Satisfactorio, 2=Parcialmente satisfactorio, 3=Completamente Satisfactorio, N/A=No aplica

		1	2	3		N/A
1. Controles Generales						
	A. Datos del tipo fecha	✓	②	③		○
	B. Datos del tipo texto	①	②	✓		○
	C. Datos numéricos	①	②	✓		○
	D. Datos binarios	①	②	✓		○
2. Controles sobre DT						
	A. Aplicabilidad del dominio	①	✓	③		○
	B. Cuenta con las herramientas necesarias.	①	②	✓		○

Comentarios: Se reestructurarán los datos de la siguiente manera para obtener mejores y mas rápido los resultados.
Fecha de nacimiento -> Se agrega columna Edad
Fecha de contacto -> Se agrega columnas días y mes
Duración en horas -> Se agrega columna cantidad de minutos
Tipo de préstamo -> Se agregan columnas que identifica si tiene o no crédito personal o hipotecario
Saldo del préstamo -> Se agrega una columna que identifica si tiene o no mora (saldo + o -)
Según la cantidad de resultados de más de 6 meses se agrega una columna que suma estos.
Según campañas anteriores se suma los contactos y se agrega en una columna.
Se calcula la cantidad de días según llamados anteriores.

Tabla 5.3. *Lista de chequeo ejecutada.*

Se da como validos los datos. Esto no significa que los datos sean de calidad o que sean ya los correctos para ejecutar los algoritmos de modelado sino que los datos son los correctos con respecto a los tipos y rangos. Los datos quedan con el siguiente formato según la Tabla 5.4.

5.1.2.4. PASO 4: DOCUMENTAR CONCLUSIONES

En este paso se procede a documentar todo lo obtenido en el análisis de los pasos 1 y 2, la lista de chequeo ejecutada y la evaluación de los resultados obtenidos. Dado que los resultados son los mínimos necesarios para la siguiente actividad se procede a continuar con la gestión del conocimiento con el aporte del documento generado en este paso. Continuar con la actividad siguiente.

5.1.3. APLICACIÓN DE LA ACTIVIDAD DE DETERMINAR Y APLICAR LA ESTRUCTURA DE LOS DATOS

En esta sección se aplica al caso de validación actual la técnica utilizada para la ejecución de la actividad de determinar y aplicar la estructura de los datos: Para esto se deben efectuar los

diferentes pasos de la técnica: Paso 1: Determinar las fuentes de los datos (sección 5.1.3.1), Paso 2: Determinar las relaciones (sección 5.1.3.2), Paso 3: Unificar Tipos de datos (sección 5.1.3.3), Paso 4: Unificar Rangos de variables (sección 5.1.3.4) y Paso 5: Generar documento de integración (sección 5.1.3.5).

Campo	Tipo de dato y rango de valores
edad	Continuo
ocupación	Discreto: ("admin.", "Desconocido", "desempleado", "gestión", "criada", "empresario", "estudiante", "cuello azul", "por cuenta propia", "jubilado", "técnico", "servicios")
estado civil	Discreto: ("casado", "divorciado", "solo") (divorciado significa divorciado o viudo)
educación	Discreto: ("desconocido", "secundario", "primario", "terciario")
posee mora	Binario: ("sí", "no")
salario promedio	Continuo
crédito hipotecario	Binario: ("sí", "no")
préstamo personal	Binario: ("sí", "no")
contacto	Discreto: ("desconocido", "teléfono", "celular")
días	Continuo
mes	Discreto: ("enero" hasta "diciembre")
duración	Continuo
campaña	Continuo
pdays	Continuo
anteriores	Continuo
poutcome	Discreto: ("desconocido", "otro", "fracaso", "éxito")
suscrito	Binario: ("sí", "no")

Tabla 5.4. *Resultado del ajuste.*

5.1.3.1. PASO 1: DETERMINAR LAS FUENTES DE LOS DATOS

Los datos son de las mismas fuentes y se obtuvieron en tablas separadas y relacionadas por algún valor en especial.

5.1.3.2. PASO 2: DETERMINAR LAS RELACIONES

Las relaciones siguen la lógica de la tabla 5.5.

5.1.3.3. PASO 3: UNIFICAR TIPOS DE DATOS

El análisis de los datos arrojo que los mismos están unificados.

5.1.3.4. PASO 4: UNIFICAR RANGOS DE VARIABLES

El análisis de los datos arrojo que los rangos están unificados.

Campo	Tabla origen		Tabla	
	Nombre	Campo	Nombre	Campo
Ocupación	Personas	ocupacion	Ocupaciones	Descripción
Estado civil	Personas	Estado_civil	Estados	Descripción
Educación	Personas	educacion	Educaciones	Descripción
Tipo de préstamo	Personas	id_persona	Prestamos	id_persona
	Prestamos	Tipo_prestamo	Tipos de préstamo	Descripción
Salario promedio	Personas	id_persona	Salarios	id_persona
Saldo del préstamo	Personas	id_persona	Prestamos	id_persona
Contacto	Personas	id_persona	Historico contactos	id_persona
Fecha contacto	Personas	id_persona	Historico contactos	id_persona
Duración	Personas	id_persona	Historico campaña	id_persona
Suscrito	Personas	id_persona	Historico campaña	id_persona

Tabla 5.5. *Relaciones entre las tablas.*

5.1.3.5. PASO 5: GENERAR DOCUMENTO DE INTEGRACIÓN

Se documenta toda la información obtenida con el fin de aportar a la gestión de conocimiento de la organización.

5.1.4. APLICACIÓN DE LA ACTIVIDAD DE CONSTRUIR EL MODELO DE ENTRADA DE DATOS

En esta sección se aplica al caso de validación actual la técnica utilizada para la ejecución de la actividad de construir el modelo de entrada de los datos: Para esto se deben efectuar los diferentes pasos de la técnica: Paso 1: Efectuar análisis iniciales (sección 5.1.4.1), Paso 2: Ejecutar las distintas fases de transformación (sección 5.1.4.2) y Paso 3: Generar documento de Estructuración (sección 5.1.4.3).

5.1.4.1. PASO 1: EFECTUAR ANÁLISIS INICIALES

En esta paso se analiza los datos con el fin de encontrar posibles problemas como ser la detección de valores nulos, valores fuera de rango, valores duplicados, etc.

Para cada tipo de problema se toma valores derivados de los datos y de esta forma, una vez detectado cada problema en caso de que exista, se procede al tratamiento con el fin de lograr datos de una calidad suficiente para ejecutar los algoritmos de modelado.

En primer lugar por cada variable se analiza si existen datos perdidos utilizando las teorías referenciadas en el capítulo 2. Aplicando una de estas teorías se obtiene el porcentaje de datos faltantes el cual se denomina "%N". Luego en caso de encontrarse este valor mayor a cero se afirma que existen datos perdidos y se procede a categorizar el patrón de dato perdido y lo denomina "CP".

En el caso de validación actual el valor de %N de cada variable fue 0% dado que no se encuentran valores faltantes exceptuando a pdays y anteriores que arrojaron un %N igual con un calor de 82%.

Lo siguiente es analizar nuevamente la totalidad de las variables y de esta forma controlar la existencia de duplicados. Se aplican las teorías nombradas en el capítulo 2 y se determina, en cada caso si se encontró duplicados, la situación del problema "SP".

En el caso de validación actual no se encontraron duplicados.

Luego continúa con esta técnica mediante la detección de valores ruidosos con lo cual se determina para cada variable su gráfico de dispersión "GD". A partir de cada gráfico se analiza la dispersión de los valores y se determina si existe valores fuera de rango "FR". Este gráfico también determina el rango intercuartil "IQR".

Aplicando las herramientas de diagnostico de datos es posible visualizar las variables nombradas. Observando la figura 5.1 se visualizan algunos de estos.

				1		2	
				Perdidos		No de extremos	
	N	Media	Desviación típ.	Recuento	Porcentaje	Bajos	Altos
edad	4521	41,17	10,576	0	,0	0	44
Salario_promedio	4521	1,422.6578	3,009.63814	0	,0	2	504
días	4521	15,92	8,248	0	,0	0	0
duración	4521	263,9613	259,85663	0	,0	0	330
campaña	4521	2,79	3,110	0	,0	0	473
pdays	816	224,87	117,200	3705	82,0	0	7
anteriores	816	3,01	2,914	3705	82,0	0	34

Figura 5.1. *Datos estadísticos univariados.*

Para seguir este paso de la técnica es necesario determinar la distribución de las variables. En caso de no tener una distribución normal es conveniente en algunos casos que esto ocurra.

No es necesaria una distribución normal.

Luego se busca la detección de series en caso de existir se toma nota de las mismas y se denominarán "SE".

No se encontraron series.

En algunos casos se encuentran también variables con valores demasiados anchos o con demasiados caracteres, en estos casos es bueno tomar nota de los mismos para los siguientes pasos esto lo llamara "TD".

TD se encontró dentro de lo normal dado que no hay variables con anchos importantes.

Por ultimo se determina la cantidad de registros "TR".

TR es igual a 45212 por lo tanto es un valor adecuado de registros.

5.1.4.2. PASO 2: EJECUTAR LAS DISTINTAS FASES DE TRANSFORMACIÓN

En la figura 5.1 (1) es posible apreciar que las Variables pdays (cantidad de días desde el último contacto) y anteriores (cantidad de contactos anteriores) presenta una gran cantidad de valores perdidos pero en cantidades idénticas. En este caso se afirma que los valores faltantes en anteriores en realidad corresponden a no haber recibido llamadas anteriores y pdays es igualmente faltante dado que depende también de la anterior. Estos se completaran con "0" en caso de anteriores y "-1" en el caso de pdays. En el caso de los valores ruidosos de la figura 5.1 (2) se observa que existen gran cantidad de altos y solo un caso de bajos (salario_promedio), esto requiere un análisis por variable.

En el caso de edad toma como extremos valores de 73 a 87 lo que no corresponde a valores fuera de rango ya que pueden existir estos valores en edad.

En el caso de salario_promedio toma como extremos a los valores desde 3370 a 71188. Estos casos son posibles en un sistema económico como en el que está sumergida una entidad bancaria.

En el caso de días no existen extremos encontrados.

Duración se cuenta con un rango de extremos desde 667 a 3025. El valor mayor corresponde a 50 minutos aproximadamente. Estos extremos son posibles.

Campaña que es la cantidad de contactos realizados con anterioridad, contiene valores fuera de rango altos que rondan entre 6 y 50. Es posible que a personas específicas se los contacte con frecuencia.

La variable pdays contiene solo 7 casos extremos. Van desde 674 a 871. Son posibles valores dado que rondan entre un año y un año y medio aproximadamente.

Por último, la variable anterior cuenta con 34 extremos altos que se encuentran entre 9 y 25 contactos anteriores. Estos son normales dado que pueden haber sido contactados en varias oportunidades por razones especiales.

Para los casos de las variables discretas se reemplazó los valores faltantes por la palabra desconocido (del inglés unknown).

Dado que todos los valores mencionados se encontraron dentro de lo normal, no es necesario aplicar ninguna fase de transformación y por esta razón se da como finalizado este paso.

5.1.4.3. PASO 3: GENERAR DOCUMENTO DE ESTRUCTURACIÓN

Se encontraron los datos con una calidad óptima. Se documenta la herramienta Tanagra como herramienta de diagnostico de gráficos de dispersión. También es posible ingresar otras herramientas utilizadas para este proceso como la herramienta de IBM SPSS (Paquete estadístico para las ciencias sociales del ingles Statistical Package for the Social Sciences) y STATA (estadística y datos del ingles statistics & data) las cuales se utilizaron para analizar los datos en general.

5.1.5. APLICACIÓN DE LA ACTIVIDAD DE INSPECCIÓN DE LOS DATOS

En esta sección se aplica al caso de validación actual la técnica utilizada para la ejecución de la actividad de inspección de los datos: Para esto se deben efectuar los diferentes pasos de la técnica: Paso 1: Efectuar la inspección de los datos (sección 5.1.5.1), Paso 2: Actualizar el repositorio del conocimiento de la compañía (sección 5.1.5.2) y Paso 3: Preparar los datos para el siguiente paso del proyecto de EI (sección 5.1.5.3).

5.1.5.1. PASO 1: EFECTUAR LA INSPECCIÓN DE LOS DATOS

Se toma una serie de datos de pruebas para efectuar el test de los mismos utilizando la herramienta Tanagra, en la pestaña selección de instancia se selecciona el algoritmo muestreo (instance selection opción sampling), de esta forma se utiliza una reducida porción de los datos para efectuar el test. Se ejecuta el algoritmo C4.5 en búsqueda de errores.

Las pruebas efectuadas no arrojan errores en los datos.

5.1.5.2. PASO 2: ACTUALIZAR EL REPOSITORIO DEL CONOCIMIENTO DE LA COMPAÑÍA

En primer lugar, se actualiza la documentación con el fin de describir el proceso de testeo de los datos mediante la herramienta Tanagra, en la pestaña selección de instancia seleccionar el algoritmo

muestreo (instance selection opción sampling). Esta herramienta se utiliza para tomar muestras de los datos.

En este paso se almacenan todos los documentos realizados anteriormente dentro de un repositorio de conocimiento en la organización. Si no existiese, se debe crear un directorio (por ejemplo) distribuido mediante algún tipo de servicio (SVN por ejemplo) con el fin de que todo personal de la organización tenga acceso a él para utilizarlo en cuanto sea necesario extraer el conocimiento en el momento deseado.

5.1.5.3. PASO 3: PREPARAR LOS DATOS PARA EL SIGUIENTE PASO DEL PROYECTO DE EXPLOTACIÓN DE INFORMACIÓN

Los datos se encuentran preparados y listos para ser usados en etapas posteriores. Se entregan en formato .xls (Formato Excel) de tal forma de ser compatibles con distintos programas de cálculo. Este formato también es compatible con la herramienta Tanagra antes mencionada.

Se procede a entregar los datos y la documentación adecuada y de esta forma continuar con el proceso de explotación de información.

5.2. CASO DE VALIDACIÓN: DATOS DE PACIENTES INDIOS CON PROBLEMAS HEPÁTICOS

En esta sección se analiza el caso de validación correspondiente a un conjunto de datos de pacientes indios con problemas hepáticos.

El problema parte de la necesidad de obtener información de los datos sobre pacientes indios los cuales fueron diagnosticados con (o sin) problemas hepáticos. Este conjunto de datos se obtuvo a través de una recolección de pacientes desde el norte al este de Andhra, el mismo cuenta con datos de 441 personas de sexo masculino y 142 personas de sexo femenino [ICS, 2012]. Se agrega a estos datos un 10% de valores faltantes a una de las variables para analizar los resultados obtenidos en el proceso. Se ejecutará el proceso de transformación de datos propuesto ya que el problema no es más que un problema de explotación de información. Para comenzar se aplica la primer actividad, la actividad de enriquecimiento de los datos (sección 5.2.1), continúa aplicando la actividad de obtención y ejecución de los casos testigo (sección 5.2.2), luego se aplica la actividad para determinar y aplicar la estructura de los datos (sección 5.2.3), más adelante se aplica la actividad de construir el modelo de entrada de los datos (sección 5.2.4) y por último se aplica la actividad de inspección de los datos (sección 5.2.5).

5.2.1. APLICACIÓN DE LA ACTIVIDAD DE ENRIQUECIMIENTO DE LOS DATOS

En esta sección se aplica a este caso de validación la técnica utilizada para la ejecución de la actividad de enriquecimiento de los datos. Para esto se deben efectuar los diferentes pasos de la técnica: Paso 1: Conocer el Problema a Resolver (sección 5.2.1.1), Paso 2: Analizar Solución a Obtener (sección 5.2.1.2), Paso 3: Generación de Documento de Solución (sección 5.2.1.3), Paso 4: Analizar Técnicas de Modelado a Utilizar (sección 5.2.1.4) y Paso 5: Generación de Documento de Técnicas de Modelado (sección 5.2.1.5).

5.2.1.1. PASO 1: CONOCER EL PROBLEMA A RESOLVER

El hígado es el órgano más grande dentro del cuerpo. También es uno de los más importantes. El hígado tiene muchas funciones, incluyendo la transformación de los alimentos en energía y la eliminación del alcohol y las toxinas de la sangre. El hígado también produce bilis, un líquido amarillo verdoso que ayuda a la digestión.

Existen muchos tipos de enfermedades hepáticas. Algunas de ellas son causadas por virus, como la hepatitis A, la hepatitis B y la hepatitis C. Otras pueden ser a consecuencia de medicamentos, venenos o toxinas o por ingerir demasiado alcohol. Si el hígado forma tejido cicatricial por una enfermedad, se denomina cirrosis. La ictericia, o coloración amarilla de la piel, puede ser un signo de enfermedad hepática. Al igual que en otras partes del cuerpo, el cáncer puede afectar el hígado. Otras enfermedades hepáticas pueden ser hereditarias, como por ejemplo, la hemocromatosis.

Dada la gran cantidad de pacientes en el último tiempo con síntomas de enfermedades hepáticas, se necesita un sistema de explotación de información, con datos tomados de la India sobre pacientes con síntomas similares [ICS, 2012], para determinar con más precisión y certeza cuales son las características de los que realmente están enfermos. Los resultados de estos estudios son muy importantes para el desarrollo del sistema automático de diagnóstico médico.

5.2.1.2. PASO 2: ANALIZAR SOLUCIÓN A OBTENER

Dado que la solución a obtener es un árbol de decisión de las características de los pacientes con problemas hepáticos, es necesario obtener como salida del proyecto de explotación de información una determinada cantidad de reglas de pertenencia al grupo de personas con problemas hepáticos. Para esto es necesario aplicar a los datos las técnicas de modelado de árboles de decisión (mas precisamente el algoritmo C4.5). A través de esto se infiere que es necesario tener como datos de entrada valores continuos y discretos y como variable de salida discreta.

El proyecto plantea buscar reglas que determinen que tipo de valores debe tener cada paciente en sus variables con respecto a la variable de salida y con estos resultados crear un sistema automático de diagnóstico médico.

5.2.1.3. PASO 3: GENERACIÓN DE DOCUMENTO DE SOLUCIÓN

En este documento se describe todo lo planteado en el paso 1 y en el paso 2 con el fin de aportar a la gestión del conocimiento datos de este proceso de transformación de datos. En el mismo se ejecuta un algoritmo de árboles de decisión con el fin de detectar reglas de pertenencia a grupos que se dividen en dos: Los que tengan el valor "S" de la variable de selección son los que tuvieron enfermedades hepáticas y los que tengan el valor "N" no tuvieron ninguna enfermedad hepática.
La ejecución de este algoritmo devuelve un árbol con reglas de pertenencia a la "S" y a la "N" y con esto, es posible obtener reglas de pertenencia a los grupos con el fin de predecir posibles enfermedades en futuros pacientes.

5.2.1.4. PASO 4: ANALIZAR TÉCNICAS DE MODELADO A UTILIZAR

Este algoritmo fue analizado en proyectos anteriores por lo que se usa el mismo documento que en dichos proyectos.

5.2.1.5. PASO 5: GENERACIÓN DE DOCUMENTO DE TÉCNICAS DE MODELADO

Ya existe un documento en el repositorio de conocimiento de la organización por lo que se usa el mismo.

5.2.2. APLICACIÓN DE LA ACTIVIDAD DE OBTENCIÓN Y EJECUCIÓN DE LOS CASOS TESTIGO

En esta sección se aplica al caso de validación actual la técnica utilizada para la ejecución de la actividad de obtención de los casos testigo: Para esto se deben efectuar los diferentes pasos de la técnica: Paso 1: Planteo de los casos testigo (sección 5.2.2.1), Paso 2: Generar Lista de Chequeo (sección 5.2.2.2), Paso 3: Test de los datos (sección 5.2.2.3) y Paso 4: Documentar conclusiones (sección 5.2.2.4).

5.2.2.1. PASO 1: PLANTEO DE LOS CASOS TESTIGO

En este paso es necesario confeccionar la lista de chequeos para controlar si los datos son los correctos o si es necesaria la obtención de nuevos datos. Para esto es necesario analizar las técnicas de explotación de información a ejecutar para determinar qué tipos de datos serán requeridos.

Árboles de decisión (DT): Tiene como atributo de entrada valores continuos y discretos y atributo de salida una variable discreta.

Los datos obtenidos por los profesionales que estudiaron a los pacientes son 10 variables:

- Edad: Edad del paciente
- Género: El género del paciente
- TB: Bilirrubina Total
- DB: bilirrubina directa
- AlkPhos: fosfatasa alcalina
- SGPT: Alanina aminotransferasa
- SGOT: Aspartato aminotransferasa
- TP: totales proteínas
- ALB: albúmina
- Relación A / G: cociente albúmina y globulina
- Campo de selección: utilizado para dividir los datos en dos conjuntos

Los expertos en problemas hepáticos decidieron que estos datos son los mínimos e indispensables para continuar con el proceso.

Este paso se ejecutó dos veces ya que en el paso 3 se encontró algunos problemas y se solucionó.

Los tipos de los datos que fueron entregados son descriptos en la tabla 5.6:

Campo	Tipo de dato y rango de valores
edad	Continuo
sexo	Binario.
TB	Continuo
DB	Continuo
AlkPhos	Continuo
SGPT	Continuo
SGOT	Continuo
TP	Continuo
ALB	Continuo
Relación A/G	Continuo
selección	Binario.

Tabla 5.6. *Descripción de los datos.*

5.2.2.2. PASO 2: GENERAR LISTA DE CHEQUEO

En este paso se genera la lista de chequeo con la información recolectada en el paso anterior.

La tabla 5.7 muestra la lista de chequeo para este proceso.

5.2.2.3. PASO 3: TEST DE LOS DATOS

Una vez confeccionada la lista de chequeo se procede a la ejecución de las pruebas (test) para luego evaluar los resultados obtenidos. Este proceso se impactó en la tabla 5.6.

Dado que se va a analizar los datos respecto de la variable de selección se cambiaron todos los valores 2 por una "N" que significa que no posee enfermedad hepática y todo los valores 1 por una "S" que significa que tiene una enfermedad.

Lista de Chequeo de evaluación de los datos

1=No Satisfactorio, 2=Parcialmente satisfactorio, 3=Completamente Satisfactorio, N/A=No aplica

		1	2	3	N/A
1. Controles Generales					
	A. Datos del tipo fecha	①	②	③	○
	B. Datos del tipo texto	①	②	③	○
	C. Datos numéricos	①	②	③	○
	D. Datos binarios	①	②	③	○
2. Controles sobre DT					
	A. Aplicabilidad del dominio	①	②	③	○
	B. Cuenta con las herramientas necesarias.	①	②	③	○
Comentarios:					

Tabla 5.7. *Lista de chequeo.*

Lista de Chequeo de evaluación de los datos

1=No Satisfactorio, 2=Parcialmente satisfactorio, 3=Completamente Satisfactorio, N/A=No aplica

		1	2	3	N/A
1. Controles Generales					
	A. Datos del tipo fecha	①	②	③	✓
	B. Datos del tipo texto	①	②	✓	○
	C. Datos numéricos	①	②	③	○
	D. Datos binarios	①	②	③	✓
2. Controles sobre DT					
	A. Aplicabilidad del dominio	①	✓	③	○
	B. Cuenta con las herramientas necesarias.	①	②	③	○
Comentarios:					

Tabla 5.8. *Lista de chequeo ejecutada.*

También se modifica todas las "," del archivo CSV por ";" y todos los "." Por "," a fin de traducir los numero con punto flotante a coma flotante.

Por último se traduce la variable género a M para los valores Male y F para los valores Female.

Una vez encontrado los problemas se procede a solucionarlos.

Luego se comienza el paso 1 y se ejecuta nuevamente las listas de chequeo.

Se da como validos los datos dado que no se encentra nuevos problemas con respecto a la actividad actual. Esto no significa que los datos sean de calidad o que sean ya los correctos para ejecutar los algoritmos de modelado sino que los datos son los correctos con respecto a los tipos y rangos.

5.2.2.4. PASO 4: DOCUMENTAR CONCLUSIONES

En este paso se procede a documentar todo lo obtenido en el análisis de los pasos 1 y 2, la lista de chequeo ejecutada y la evaluación de los resultados obtenidos (paso 3). Dado que los resultados son los mínimos necesarios para la siguiente actividad se procede a continuar con la gestión del conocimiento con el aporte del documento generado en este paso.

5.2.3. APLICACIÓN DE LA ACTIVIDAD DE DETERMINAR Y APLICAR LA ESTRUCTURA DE LOS DATOS

En esta sección se aplica al caso de validación actual la técnica utilizada para la ejecución de la actividad de determinar y aplicar la estructura de los datos: Para esto se deben efectuar los diferentes pasos de la técnica: Paso 1: Determinar las fuentes de los datos (sección 5.2.3.1), Paso 2: Determinar las relaciones (sección 5.2.3.2), Paso 3: Unificar Tipos de datos (sección 5.2.3.3), Paso 4: Unificar Rangos de variables (sección 5.2.3.4) y Paso 5: Generar documento de integración (sección 5.2.3.5).

5.2.3.1. PASO 1: DETERMINAR LAS FUENTES DE LOS DATOS

Dado que los datos fueron obtenidos desde Internet en la fuente referenciada con anterioridad y además que los datos ya se encontraban integrados, este paso no aplica al proceso actual.

5.2.3.2. PASO 2: DETERMINAR LAS RELACIONES

Dado que los datos fueron obtenidos desde Internet en la fuente referenciada con anterioridad y además que los datos ya se encontraban integrados, este paso no aplica al proceso actual.

5.2.3.3. PASO 3: UNIFICAR TIPOS DE DATOS

Dado que los datos fueron obtenidos desde Internet en la fuente referenciada con anterioridad y además que los datos ya se encontraban integrados, este paso no aplica al proceso actual.

5.2.3.4. PASO 4: UNIFICAR RANGOS DE VARIABLES

Dado que los datos fueron obtenidos desde Internet en la fuente referenciada con anterioridad y además que los datos ya se encontraban integrados, este paso no aplica al proceso actual.

5.2.3.5. PASO 5: GENERAR DOCUMENTO DE INTEGRACIÓN

Dado que los datos fueron obtenidos desde Internet en la fuente referenciada con anterioridad y además que los datos ya se encontraban integrados, este paso no aplica al proceso actual.

5.2.4. APLICACIÓN DE LA ACTIVIDAD DE CONSTRUIR EL MODELO DE ENTRADA DE DATOS

En esta sección se aplica al caso de validación actual la técnica utilizada para la ejecución de la actividad de construir el modelo de entrada de los datos: Para esto se deben efectuar los diferentes pasos de la técnica: Paso 1: Efectuar análisis iniciales (sección 5.2.4.1), Paso 2: Ejecutar las distintas fases de transformación (sección 5.2.4.2) y Paso 3: Generar documento de Estructuración (sección 5.2.4.3).

5.2.4.1. PASO 1: EFECTUAR ANÁLISIS INICIALES

En esta paso se analiza los datos con el fin de encontrar posibles problemas como ser la detección de valores nulos, valores fuera de rango, valores duplicados, etc.

Para cada tipo de problema se toma valores derivados de los datos y de esta forma, una vez detectado cada problema en caso de que exista, se procede al tratamiento con el fin de lograr datos de una calidad suficiente para ejecutar los algoritmos de modelado.

En primer lugar por cada variable se analiza si existen datos perdidos utilizando las teorías referenciadas en el capítulo 2. Aplicando una de estas teorías se obtiene el porcentaje de datos faltantes el cual se denomina “%N”. Luego en caso de encontrarse este valor mayor a cero se afirma que existen datos perdidos y se procede a categorizar el patrón de dato perdido y lo denomina “CP”.

En el caso de validación actual el valor de %N de la figura 5.2 (1) para la variable TB arroja un valor de 11,8% por lo que en necesario analizar las razones, en cuanto al patrón de faltantes se observa un CP MCAR.

Los valores ruidosos extraídos de la figura 5.2 (2) se muestran según cada variable.

Lo siguiente es analizar nuevamente la totalidad de las variables y de esta forma controlar la existencia de duplicados. Se aplican las teorías nombradas en el capítulo 2 y se determina en cada caso si se encuentra duplicados, la situación del problema “SP”.

				1		2	
				Perdidos		No de extremos	
	N	Media	Desviación típ.	Recuento	Porcentaje	Bajos	Altos
Age	583	44,75	16,190	0	,0	0	0
TB	514	3,377	6,2193	69	11,8	0	76
DB	583	1,486	2,8085	0	,0	0	81
Alkphos	583	290,576	242,9380	0	,0	0	66
Sgpt	583	80,714	182,6204	0	,0	0	73
Sgot	583	109,911	288,9185	0	,0	0	68
TP	583	6,483	1,0855	0	,0	6	2
ALB	583	3,142	,7955	0	,0	0	0
AG	583	,941	,3280	0	,0	4	10

Figura 5.2. *Datos estadísticos univariados.*

En el caso de validación actual no se encontraron duplicados.

Se utiliza la misma herramienta que en procesos anteriores para generar los gráficos de dispersión y detectar así valores fuera de rango o ruidosos.

Prosigue este paso de la técnica determinado la distribución de las variables. En caso de no tener una distribución normal es conveniente en algunos casos que esto ocurra.

No es necesaria una distribución normal.

Luego se busca la detección de series en caso de existir se toma nota de las mismas y se denominarán “SE”.

No se encontraron series.

En algunos casos se encuentran también variables con valores demasiados anchos o con demasiados caracteres, en estos casos es bueno tomar nota de los mismos para los siguientes pasos se lo llama “TD”.

TD se encuentra dentro de lo normal dado que no hay variables con anchos importantes.

Por último se determina la cantidad de registros “TR”.

TR es igual a 584 por lo tanto es un valor adecuado de registros.

5.2.4.2. PASO 2: EJECUTAR LAS DISTINTAS FASES DE TRANSFORMACIÓN

Dado que se encontraron los valores eliminados adrede se procede a elegir una técnica con el fin de solucionarlos. Según las características de la variable (CP = MCAR y %N = 11,8%) se utiliza imputación múltiple para recuperar los valores.

Luego de ejecutar la imputación múltiple continua esta etapa.

Los valores ruidosos encontrados deben ser analizados para cada variable. Las variables Edad (Age) y TLB no presentan valores ruidosos. El resto presenta una cantidad menor de este tipo de registros.

Según las especificaciones medicas, todos los valores ruidosos se encuentran dentro de los posibles.

Se da como finalizado el paso dos correspondiente a la técnica de construir el modelo de entrada de los datos.

5.2.4.3. PASO 3: GENERAR DOCUMENTO DE ESTRUCTURACIÓN

Se almacena el procedimiento a seguir según los valores obtenidos en el paso 1. Las herramientas utilizadas para este propósito y para la solución. En este caso se utiliza: Tanagra y SPSS. Los datos estadísticos pueden servir para futuras ejecuciones.

5.2.5. APLICACIÓN DE LA ACTIVIDAD DE INSPECCIÓN DE LOS DATOS

En esta sección se aplica al caso de validación actual la técnica utilizada para la ejecución de la actividad de inspección de los datos: Para esto se deben efectuar los diferentes pasos de la técnica: Paso 1: Efectuar la inspección de los datos (sección 5.2.5.1), Paso 2: Actualizar el repositorio del conocimiento de la compañía (sección 5.2.5.2) y Paso 3: Preparar los datos para el siguiente paso del proyecto de EI (sección 5.2.5.3).

5.2.5.1. PASO 1: EFECTUAR LA INSPECCIÓN DE LOS DATOS

Se utiliza el mismo método que en procesos anteriores. Se toma una muestra de los datos con la herramienta Tanagra y se ejecuta el algoritmo C4.5 a esa porción reducida de los datos con el fin de buscar posibles diferencias.

La configuración usada para este algoritmo es:

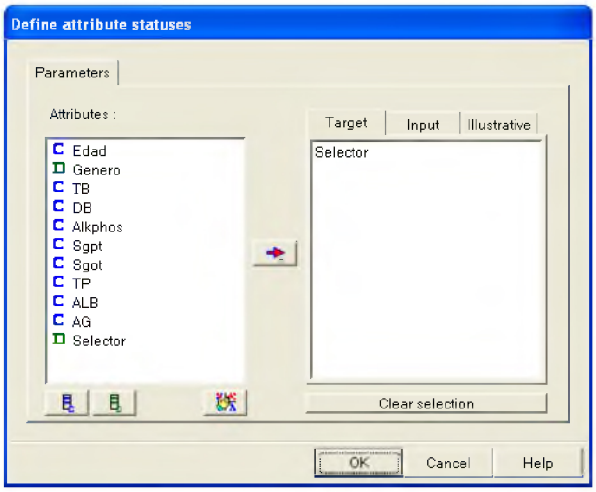

Figura 5.3. *Selección de atributo* ***Target****.*

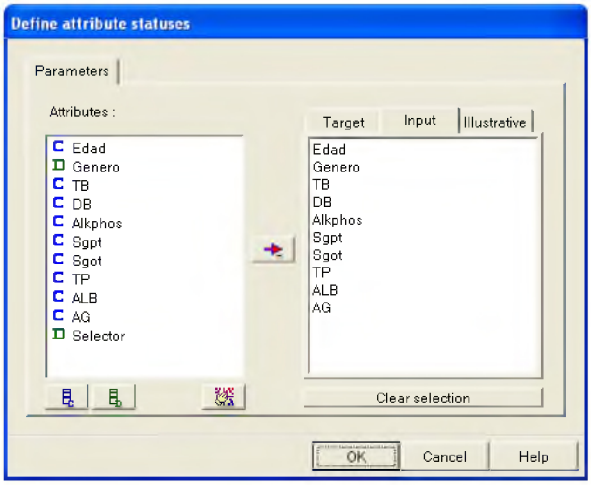

Figura 5.4. *Selección de los atributos* ***Input****.*

Para el atributo Target se utiliza la variable Selector como muestra la figura 5.3. Esta variable es del tipo Discreto. Luego se selecciona en los atributos Input el resto de las variables. Edad, TB, DB, Alkphos, Sgpt, Sgot, TP, ALB y AG que son del tipo continuas y la variable Genero que es del tipo Discreta (figura 5.4).

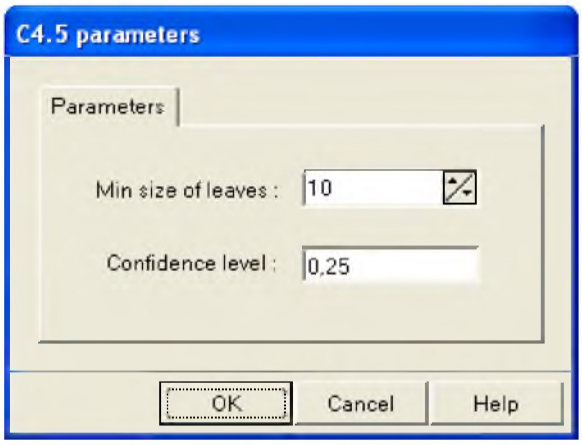

Figura 5.5. *Configuración de los parámetros del algoritmo* ***C4.5****.*

Para el algoritmo C4.5 elegir en Min size of Leaves el valor 10 (figura 5.5).

Ejecutando en las dos bases de datos (la original y la modificada) este algoritmo se obtiene los arboles de la figura 5.6 y 5.7.

Luego de efectuar un análisis sobre los árboles de ambas bases de datos se separan las reglas de los resultados “S” de la variable de selección, es posible observar que si bien cambian son muy similares. Cambian los órdenes y los porcentajes en cantidades menores pero de igual forma son muy similares las reglas para llegar al resultado positivo.

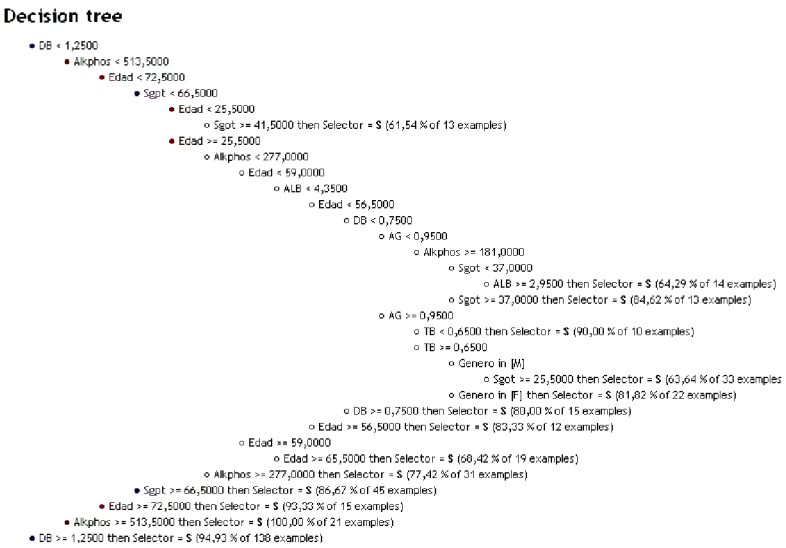

Figura 5.6. *Árbol de decisión para los datos Originales.*

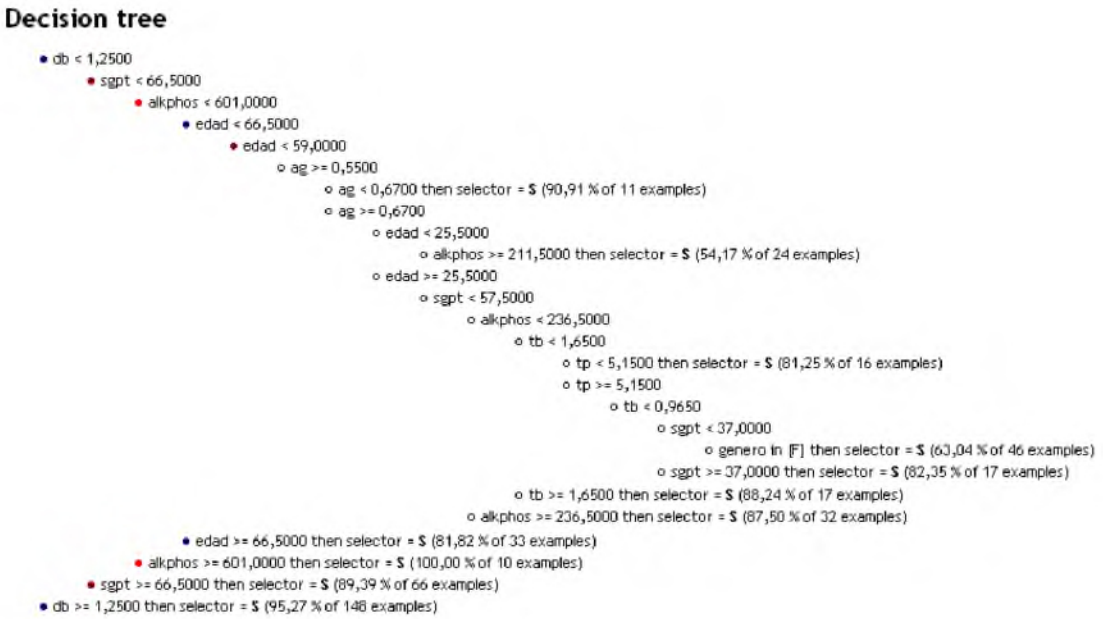

Figura 5.7. *Árbol de decisión para los datos reparados (solo los resultados selector = 'S').*

5.2.5.2. PASO 2: ACTUALIZAR EL REPOSITORIO DEL CONOCIMIENTO DE LA COMPAÑÍA

En este paso se almacenan todos los documentos realizados anteriormente dentro del repositorio de conocimiento de la organización. Éste ya debe de existir debido a que se ejecutaron procesos anteriores.

5.2.5.3. PASO 3: PREPARAR LOS DATOS PARA EL SIGUIENTE PASO DEL PROYECTO DE EXPLOTACIÓN DE INFORMACIÓN

Los datos se encuentran preparados y listos para ser usados en etapas posteriores. Se entregan en formato .xls de tal forma de ser compatibles con distintos programas de cálculo. Este formato también es compatible con la herramienta Tanagra antes mencionada.

5.3. CASO DE VALIDACIÓN: CONJUNTO DE DATOS PARA EL DIAGNOSTICO DE CÁNCER DE MAMA

En esta sección se analiza el caso de validación correspondiente a un conjunto de datos para el diagnostico de cáncer de mama.

Los valores obtenidos fueron calculados a partir de una imagen digitalizada de un aspirado con aguja fina (PAAF) de una masa en la mama. Se describen las características de los núcleos de células presentes en la imagen.

El problema parte de la necesidad de obtener información de los datos sobre pacientes femeninos con cáncer de mama y partiendo de esto intentar diagnosticar si el mismo es benigno o maligno [ICS, 2012]. Se agrega un 20% de error a cuatro variables distintas.

Se ejecuta el proceso de transformación de datos propuesto ya que el problema no es más que un problema de explotación de información. Para comenzar se aplica la primer actividad, la actividad de enriquecimiento de los datos (sección 5.3.1), continúa aplicando la actividad de obtención y ejecución de los casos testigo (sección 5.3.2), luego se aplica la actividad para determinar y aplicar la estructura de los datos (sección 5.3.3), más adelante se aplica la actividad de construir el modelo de entrada de los datos (sección 5.3.4) y por último se aplica la actividad de inspección de los datos (sección 5.3.5).

5.3.1. APLICACIÓN DE LA ACTIVIDAD DE ENRIQUECIMIENTO DE LOS DATOS

En esta sección se aplica a este caso de validación la técnica utilizada para la ejecución de la actividad de enriquecimiento de los datos. Para esto se deben efectuar los diferentes pasos de la técnica: Paso 1: Conocer el Problema a Resolver (sección 5.3.1.1), Paso 2: Analizar Solución a Obtener (sección 5.3.1.2), Paso 3: Generación de Documento de Solución (sección 5.3.1.3), Paso 4: Analizar Técnicas de Modelado a Utilizar (sección 5.3.1.4) y Paso 5: Generación de Documento de Técnicas de Modelado (sección 5.3.1.5).

5.3.1.1. PASO 1: CONOCER EL PROBLEMA A RESOLVER

Mediante una serie de estudios a pacientes con problemas de cáncer de mamas se han obtenido valores que fueron calculados a partir de imágenes digitalizadas de un aspirado con aguja fina (PAAF) de una masa que se encontró en la mama. Se describen las características de los núcleos de células presentes en la imagen.

Los resultados de estos estudios son muy importantes para el desarrollo del sistema automático de diagnóstico médico.

5.3.1.2. PASO 2: ANALIZAR SOLUCIÓN A OBTENER

Dado que la solución a obtener es un árbol de decisión de las características de los pacientes con problemas de cáncer de mama, es necesario obtener como salida del proyecto de explotación de información una determinada cantidad de reglas de pertenencia al grupo de personas con problemas de cáncer de mama benigno o maligno. Para esto es necesario aplicar a los datos las técnicas de modelado de árboles de decisión (mas precisamente el algoritmo C4.5). A partir de esto se descubre que es necesario tener como datos de entrada valores continuos y discretos y como variable de salida discreta. El proyecto plantea buscar reglas que determinen que tipo de valores debe tener cada paciente en sus variables con respecto a la variable de salida y con estos resultados crear un sistema automático de diagnóstico médico.

5.3.1.3. PASO 3: GENERACIÓN DE DOCUMENTO DE SOLUCIÓN

En este documento se describe todo lo planteado en el paso 1 y en el paso 2 con el fin de aportar a la gestión del conocimiento datos de este proceso de transformación de datos. En el mismo se ejecuta un algoritmo de árboles de decisión con el fin de detectar reglas de pertenencia a grupos que se dividan en dos: Los que tengan el valor “M” de la variable de selección son los que tienen un cáncer del tipo maligno y los que tengan el valor “B” son los que tienen un tipo de cáncer benigno. La ejecución de este algoritmo devuelve un árbol con reglas de pertenencia a la “M” y a la “B” y con esto es posible obtener reglas de pertenencia a estos grupos con el fin de predecir qué tipo de cáncer posee el paciente.

5.3.1.4. PASO 4: ANALIZAR TÉCNICAS DE MODELADO A UTILIZAR

Se usa un documento anterior.

5.3.1.5. PASO 5: GENERACIÓN DE DOCUMENTO DE TÉCNICAS DE MODELADO

Ya existe un documento en el repositorio de conocimiento de la organización. Se usa un documento anterior.

5.3.2. APLICACIÓN DE LA ACTIVIDAD DE OBTENCIÓN Y EJECUCIÓN DE LOS CASOS TESTIGO

En esta sección se aplica al caso de validación actual la técnica utilizada para la ejecución de la actividad de obtención de los casos testigo: Para esto se deben efectuar los diferentes pasos de la técnica: Paso 1: Planteo de los casos testigo (sección 5.3.2.1), Paso 2: Generar Lista de Chequeo (sección 5.3.2.2), Paso 3: Test de los datos (sección 5.3.2.3) y Paso 4: Documentar conclusiones (sección 5.3.2.4).

5.3.2.1. PASO 1: PLANTEO DE LOS CASOS TESTIGO

Para esta tarea es necesario confeccionar la lista de chequeos para controlar si los datos son los correctos o si es necesaria la obtención de nuevos datos. Para esto es necesario analizar las técnicas de explotación de información a ejecutar para determinar qué tipos de datos serán requeridos.
Árboles de decisión (DT): Tiene como atributo de entrada variables continuas y discretas y atributo de salida una variable discreta.
Los datos obtenidos por los profesionales que estudiaron a los pacientes son 32 variables:

- Número de identificación
- Diagnóstico (M = maligno, B = benigno)

Diez valores reales característicos se calculan para cada uno de los 3 núcleos celulares:

- Radio (media de las distancias desde el centro hasta puntos en el perímetro)
- La textura (desviación estándar de los valores de la escala de grises)
- Perímetro
- Área
- Suavidad (variación local en longitudes de radio)
- Compacto (perímetro ^ 2 / zona - 1,0)
- Concavidad (severidad de porciones cóncavas del contorno)
- Cóncavas puntos (número de porciones cóncavas del contorno)
- Simetría
- La dimensión fractal ("aproximación costa" - 1)

Los expertos en el análisis de las imagen digitalizadas de los aspirados con aguja fina deciden que estos datos son los mínimos e indispensables para continuar con el proceso.
Este paso se ejecuta dos veces ya que en el paso 3 se encuentra un problema ya solucionado.
Los tipos de los datos que fueron entregados son descriptos en la tabla 5.9:

Campo	Tipo de dato y rango de valores
Número de identificación	Continuo
Diagnóstico	Categórico: ("M", "B")
Radio	Continuo
La textura	Continuo
Perímetro	Continuo
Área	Continuo
Suavidad	Continuo
Compacto	Continuo
Concavidad	Continuo
Cóncavas puntos	Continuo
Simetría	Continuo
La dimensión fractal	Continuo

Tabla 5.9. *Descripción de los datos.*

5.3.2.2. PASO 2: GENERAR LISTA DE CHEQUEO

En este paso se genera el listado de chequeo con la información recolectada en el paso anterior. La tabla 5.10 muestra la lista de chequeo para este proceso.

Lista de Chequeo de evaluación de los datos

1=No Satisfactorio, 2=Parcialmente satisfactorio, 3=Completamente Satisfactorio, N/A=No aplica

		1	2	3	N/A
1. Controles Generales					
	A. Datos del tipo fecha	①	②	③	○
	B. Datos del tipo texto	①	②	③	○
	C. Datos numéricos	①	②	③	○
	D. Datos binarios	①	②	③	○
2. Controles sobre DT					
	A. Aplicabilidad del dominio	①	②	③	○
	B. Cuenta con las herramientas necesarias.	①	②	③	○
Comentarios:					

Tabla 5.10. *Lista de chequeo.*

5.3.2.3. PASO 3: TEST DE LOS DATOS

Una vez confeccionada la lista de chequeo se procede a la ejecución de las pruebas (test) para luego evaluar los resultados obtenidos.

Este proceso se impacta en la tabla 5.11.

Lista de Chequeo de evaluación de los datos

1=No Satisfactorio, 2=Parcialmente satisfactorio, 3=Completamente Satisfactorio, N/A=No aplica

		1	2	3		N/A
1. Controles Generales						
	A. Datos del tipo fecha	①	②	③		✓
	B. Datos del tipo texto	①	②	✓		○
	C. Datos numéricos	①	②	✓		○
	D. Datos binarios	①	②	✓		○
2. Controles sobre DT						
	A. Aplicabilidad del dominio	①	✓	③		○
	B. Cuenta con las herramientas necesarias.	①	②	✓		○
Comentarios:						

Tabla 5.11. *Lista de chequeo ejecutada.*

Se modifica todas las "," del archivo CSV por ";" y todos los "." Por "," a fin de traducir los numero con punto flotante a coma flotante.

Una vez encontrado el problema se procede a solucionarlo.

Luego se comienza el paso 1 y se ejecuta nuevamente las listas de chequeo.

Se da como validos los datos dado que no se encontró nuevos problemas con respecto a la actividad actual. Esto no significa que los datos sean de calidad o que sean ya los correctos para ejecutar los algoritmos de modelado sino que los datos son los correctos con respecto a los tipos y rangos.

5.3.2.4. PASO 4: DOCUMENTAR CONCLUSIONES

En este paso se procede a documentar todo lo obtenido en el análisis de los pasos 1 y 2, la lista de chequeo ejecutada y la evaluación de los resultados obtenidos. Dado que los resultados son los mínimos necesarios para la siguiente actividad se procede a continuar con la gestión del conocimiento con el aporte del documento generado en este paso.

5.3.3. APLICACIÓN DE LA ACTIVIDAD DE DETERMINAR Y APLICAR LA ESTRUCTURA DE LOS DATOS

En esta sección se aplica al caso de validación actual la técnica utilizada para la ejecución de la actividad de determinar y aplicar la estructura de los datos: Para esto se deben efectuar los diferentes pasos de la técnica: Paso 1: Determinar las fuentes de los datos (sección 5.3.3.1), Paso 2: Determinar las relaciones (sección 5.3.3.2), Paso 3: Unificar Tipos de datos (sección 5.3.3.3), Paso 4: Unificar Rangos de variables (sección 5.3.3.4) y Paso 5: Generar documento de integración (sección 5.3.3.5).

5.3.3.1. PASO 1: DETERMINAR LAS FUENTES DE LOS DATOS

Dado que los datos fueron obtenidos desde Internet en la fuente referenciada con anterioridad y además que los datos ya se encontraban integrados, este paso no aplica al proceso actual.

5.3.3.2. PASO 2: DETERMINAR LAS RELACIONES

Dado que los datos fueron obtenidos desde Internet en la fuente referenciada con anterioridad y además que los datos ya se encontraban integrados, este paso no aplica al proceso actual.

5.3.3.3. PASO 3: UNIFICAR TIPOS DE DATOS

Dado que los datos fueron obtenidos desde Internet en la fuente referenciada con anterioridad y además que los datos ya se encontraban integrados, este paso no aplica al proceso actual.

5.3.3.4. PASO 4: UNIFICAR RANGOS DE VARIABLES

Dado que los datos fueron obtenidos desde Internet en la fuente referenciada con anterioridad y además que los datos ya se encontraban integrados, este paso no aplica al proceso actual.

5.3.3.5. PASO 5: GENERAR DOCUMENTO DE INTEGRACIÓN

Dado que los datos fueron obtenidos desde Internet en la fuente referenciada con anterioridad y además que los datos ya se encontraban integrados, este paso no aplica al proceso actual.

5.3.4. APLICACIÓN DE LA ACTIVIDAD DE CONSTRUIR EL MODELO DE ENTRADA DE DATOS

En esta sección se aplica al caso de validación actual la técnica utilizada para la ejecución de la actividad de construir el modelo de entrada de los datos: Para esto se deben efectuar los diferentes pasos de la técnica: Paso 1: Efectuar análisis iniciales (sección 5.3.4.1), Paso 2: Ejecutar las

distintas fases de transformación (sección 5.3.4.2) y Paso 3: Generar documento de Estructuración (sección 5.3.4.3).

5.3.4.1. PASO 1: EFECTUAR ANÁLISIS INICIALES

En esta paso se analiza los datos con el fin de encontrar posibles problemas como ser la detección de valores nulos, valores fuera de rango, valores duplicados, etc.

Para cada tipo de problema se toma valores derivados de los datos y de esta forma, una vez detectado cada problema en caso de que exista, se procede al tratamiento con el fin de lograr datos de una calidad suficiente para ejecutar los algoritmos de modelado.

En primer lugar por cada variable se analiza si existen datos perdidos utilizando las teorías referenciadas en el capítulo 2. Aplicando una de estas teorías se obtiene el porcentaje de datos faltantes el cual se denomina "%N". Luego en caso de encontrarse este valor mayor a cero se afirma que existen datos perdidos y se procede a categorizar el patrón de dato perdido y lo denomina "CP".

En el caso de validación actual el valor de %N se divide en cuatro dado que son esa cantidad de variables las que cuentan con faltantes según la figura 5.9 (1). Para Concavidad_1 es 8,3, para Radio_2 es 4,7, para Area_2 también es 4,7 y por último la Dimensión_fractal_3 cuenta con un %N de 4. Para todas las variables el patrón de datos faltantes (CP) es MCAR.

Los valores estadísticos descriptivos de cada variable se pueden observar en la figura 5.8.

La mediana de Concavidad_1 es: 0,04736, para Radio_2: 0,286, para Area_2: es 20,525 y para Dimensión_fractal_3 es: 0,078485.

También es posible observar la cantidad de registros válidos del dominio que es 472. Este valor informa que algunas variables comparten datos perdidos según el registro.

	N	Mínimo	Máximo	Media	Mediana	Desv. típ.
Concavidad_1	522	.00000000	.21330000	.0718082580	.04736	5.5836E-2
Radio_2	542	.11150000	2.87300000	.3800837638	.0286	2.5852E-1
Area_2	542	6.80200000	5.4220000E2	3.78657158E1	20.525	4.5188E1
Dimensión_fractal_3	546	.05504000	.12330000	.0816719231	.078485	1.4050E-2
N válido (según lista)	472					

Figura 5.8. *Datos estadísticos descriptivos.*

Lo siguiente es analizar nuevamente la totalidad de las variables y de esta forma controlar la existencia de duplicados. Se aplican las teorías nombradas en el capítulo 2 y se determina en cada caso si se encontró duplicados la situación del problema "SP".

En el caso de validación actual no se encontraron duplicados.

Luego se continúa con esta técnica mediante la detección de valores ruidosos con lo cual se determina para cada variable su gráfico de dispersión "GD". A partir de cada gráfico se analiza la

dispersión de los valores y se determina si existen valores fuera de rango que es llamado “FR”. Este gráfico también determina el rango intercuartil “IQR”.

Se utiliza la misma herramienta que en procesos anteriores para generar los gráficos de dispersión y detectar así valores fuera de rango o ruidosos.

				1		2	
				Perdidos		No de extremos[a]	
	N	Media	Desviación típ.	Recuento	Porcentaje	Bajos	Altos
Radio_1	569	1.41272917E1	3.52404882E0	0	,0	0	14
Textura_1	569	1.92896485E1	4.30103576E0	0	,0	0	7
Perimetro_1	569	9.19690333E1	2.42989810E1	0	,0	0	13
Area_1	569	6.54889103E2	3.51914129E2	0	,0	0	25
Suavidad_1	569	.0963602812	1.4064128E-2	0	,0	1	5
Compacto_1	569	.1043409842	5.2812757E-2	0	,0	0	16
Concavidad_1	522	.0718082580	5.5836809E-2	47	8,3	0	0
Concavas_puntos_1	569	.0489191459	3.8802844E-2	0	,0	0	10
Simetría_1	569	.1811618629	2.7414281E-2	0	,0	1	14
Dimensión_fractal_1	569	.0627976098	7.0603627E-3	0	,0	0	15
Radio_2	542	.3800837638	2.5852632E-1	27	4,7	0	30
Textura_2	569	1.21685342E0	5.5164839E-1	0	,0	0	20
Perimetro_2	569	2.86605922E0	2.02185455E0	0	,0	0	38
Area_2	542	3.78657158E1	4.51887859E1	27	4,7	0	51
Suavidad_2	569	.0070409789	3.0025179E-3	0	,0	0	30
Compacto_2	569	.0254781388	1.7908179E-2	0	,0	0	28
Concavidad_2	569	.0318937163	3.0186060E-2	0	,0	0	22
Concavas_puntos_2	569	.0117961371	6.1702851E-3	0	,0	0	19
Simetría_2	569	.0205422988	8.2663715E-3	0	,0	0	27
Dimensión_fractal_2	569	.0037949039	2.6460709E-3	0	,0	0	28
Radio_3	569	1.62691898E1	4.83324158E0	0	,0	0	17
Textura_3	569	2.56772231E1	6.14625762E0	0	,0	0	5
Perimetro_3	569	1.07261212E2	3.36025422E1	0	,0	0	15
Area_3	569	8.80583128E2	5.69356992E2	0	,0	0	35
Suavidad_3	569	.1323685940	2.2832429E-2	0	,0	1	6
Compacto_3	569	.2542650439	1.5733648E-1	0	,0	0	16
Concavidad_3	569	.2721884833	2.0862428E-1	0	,0	0	12
Concavas_puntos_3	569	.1146062232	6.5732341E-2	0	,0	0	0
Simetría_3	569	.2900755712	6.1867467E-2	0	,0	0	23
Dimensión_fractal_3	546	.0816719231	1.4050617E-2	23	4,0	0	6

Figura 5.9. *Datos estadísticos univariados.*

Luego sigue este paso de la técnica, determinado la distribución de las variables.

En caso de no tener una distribución normal es conveniente en algunos casos que esto ocurra.

No es necesaria una distribución normal.

Luego se busca la detección de series en caso de existir se toma nota de las mismas y se denominarán “SE”. No se encontraron series.

En algunos casos se encuentran también variables con valores demasiados anchos o con demasiados caracteres, en estos casos es bueno tomar nota de los mismos para los siguientes pasos esto se llama “TD”.

TD se encuentra dentro de lo normal dado que no hay variables con anchos importantes.

Por último se determina la cantidad de registros "TR". TR es igual a 584 por lo tanto es un valor adecuado de registros.

5.3.4.2. PASO 2: EJECUTAR LAS DISTINTAS FASES DE TRANSFORMACIÓN

En este paso se procede a solucionar en primer lugar los cuatro problemas de valores perdidos. Primero con las variables Radio_2, Area_2 y Dimensión_fractal_3 que cuentan con un %N menor a 5% y un patrón o categoría CP MCAR. Teniendo en cuenta estos valores y utilizando las teorías del capítulo 2 es posible decir que es necesario asignarle a los valores perdidos la mediana de la variable.

En cambio para la variable Concavidad_1 con un %N es 8,3% y CP es MCAR también, se utiliza para completar, algoritmos para imputación múltiple.

De esta forma se soluciona el problema de valores perdidos.

Se encuentra que todas las variables menos Concavidad_1 y Concavas_puntos_3 poseen valores fuera de rango (Figura 5.9 (2)). Analizando cada variable en particular se observa que las medidas tomadas están dentro de lo norma y posible por lo que se descarta la necesidad de tratarlos.

La base de datos actualizada ya esta disponible para continuar con el proceso.

5.3.4.3. PASO 3: GENERAR DOCUMENTO DE ESTRUCTURACIÓN

Se genera un documento con las técnicas utilizadas según el caso a resolver. También es posible agregar al documento los valores estadísticos de los datos para futuras consultas.

5.3.5. APLICACIÓN DE LA ACTIVIDAD DE INSPECCIÓN DE LOS DATOS

En esta sección se aplica al caso de validación actual la técnica utilizada para la ejecución de la actividad de inspección de los datos: Para esto se deben efectuar los diferentes pasos de la técnica: Paso 1: Efectuar la inspección de los datos (sección 5.3.5.1), Paso 2: Actualizar el repositorio del conocimiento de la compañía (sección 5.3.5.2) y Paso 3: Preparar los datos para el siguiente paso del proyecto de EI (sección 5.3.5.3).

5.3.5.1. PASO 1: EFECTUAR LA INSPECCIÓN DE LOS DATOS

Se utiliza el mismo método que en procesos anteriores. Se toma una muestra de los datos con la herramienta Tanagra y se ejecuta el algoritmo C4.5 a esa porción reducida de los datos con el fin de buscar posibles errores.

Se utiliza la configuración por defecto del modelo de explotación de información.

Es posible observar en ambos árboles de decisión que los resultados son iguales por lo tanto el proceso fue exitoso.

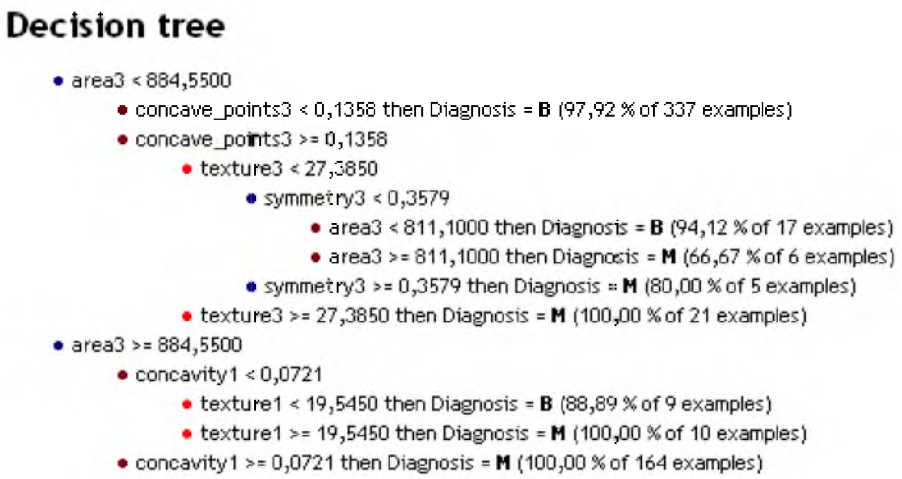

Figura 5.10. *Árbol de decisión para los datos Originales.*

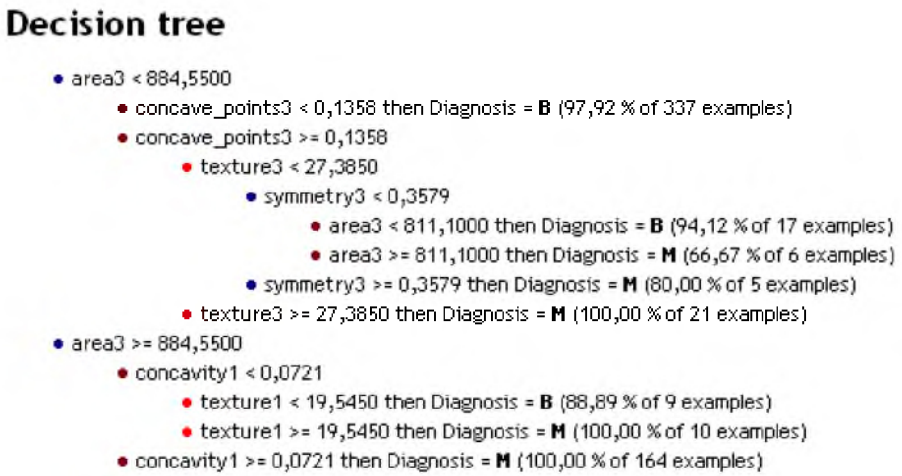

Figura 5.11. *Árbol de decisión para los datos reparados.*

En la figura 5.10 se muestra el árbol de decisión resultado de las pruebas con el modelo C4.5.

Por otra parte en la figura 5.11 se observa el árbol de decisión resultado de la aplicación de la técnica de modelado C4.5 con los datos reparados luego de adicionarle los datos faltantes.

Con estas pruebas satisfactorias se procede a continuar con el proceso.

5.3.5.2. PASO 2: ACTUALIZAR EL REPOSITORIO DEL CONOCIMIENTO DE LA COMPAÑÍA

En este paso se almacenan todos los documentos realizados anteriormente dentro del repositorio de conocimiento de la organización. Éste ya debe de existir debido a que se ejecutaron procesos anteriores.

5.3.5.3. PASO 3: PREPARAR LOS DATOS PARA EL SIGUIENTE PASO DEL PROYECTO DE EXPLOTACIÓN DE INFORMACIÓN

Los datos se encuentran preparados y listos para ser usados en etapas posteriores. Se entregan en formato .xls de tal forma de ser compatibles con distintos programas de cálculo. Este formato también es compatible con la herramienta Tanagra antes mencionada.

6. CONCLUSIONES

En este Capítulo se presentan los aportes de este trabajo (sección 6.1) y se destacan las futuras líneas de investigación que se consideran de interés en base al problema abierto que se presenta en este trabajo final de licenciatura (sección 6.2).

6.1. APORTES DEL TRABAJO FINAL DE LICENCIATURA

En este trabajo se ha corroborado que partiendo de una cantidad limitada de datos extraídos de alguna base de datos de interés en condiciones poco optimas para la ejecución de proyectos de explotación de información sobre los mismos, es posible tratarlos de alguna manera en especial con el objetivo de mejorar su calidad a fin de obtener resultados mas certeros posibles.
Inmerso en este contexto y mediante este trabajo se propuso:

- Un proceso para la transformación de datos para proyectos de explotación de información dividido en actividades: la primera es la actividad de *Enriquecimiento de los Datos*, sigue la de *Obtener y Ejecutar de los Casos Testigo*, lo siguiente es la de *Determinar y Aplicar la Estructura de los Datos*, luego *Construir el Modelo de Entrada de Datos* y por último la *Inspeccionar los Datos*.
- Dentro de la actividad de enriquecimiento de los datos se propuso, la *Técnica de Enriquecimiento de los datos* la cual espera como productos de entrada para su ejecución: los *Datos posiblemente sucios* y también a la *Información sobre el Proyecto de Explotación de Información,* aporta como productos de salida: los *Datos Sucios*, el *Documento de Solución* y el *Documento de Técnicas de Modelado*.
- Dentro de la actividad de obtener y ejecutar de los casos testigo se propuso la *Técnica de obtención y ejecución de los casos testigo* la cual espera como productos de entrada para su ejecución: Los *Datos Sucios*, el *Documento de Solución* y también a el *Documento de Técnicas de Modelado* y aporta, como producto de salida: Los *Datos Válidos* y el *Documento de Listas de Chequeo*.
- Dentro de la actividad de determinar y aplicar la estructura de los datos se propuso la *Técnica de Determinación y Aplicación de la Estructura de los Datos* la cual espera como productos de entrada para su ejecución: Los *Datos Válidos*, el *Documento de Solución* y también a el *Documento de Técnicas de Modelado* y aporta, como productos de salida: El *Documento de Integración* y los *Datos Integrados*.
- Dentro de la actividad de construir el modelo de entrada de datos se propuso la *Técnica de Construcción del Modelo de Entrada de Datos* la cual espera como producto de entrada para

su ejecución: Los *Datos Integrados* y aporta como productos de salida: El *Documento de Estructuración de los datos* y los *Datos Estructurados*.

- Dentro de la actividad de inspeccionar los datos se propuso la *Técnica de Inspección de los datos* la cual espera como productos de entrada para su ejecución: Los *Datos Estructurados*, el *Documento de Estructuración de los dato*, el *Documento de Integración*, el *Documento de Solución*, el *Documento de Técnicas de Modelado* y el *Documento de Listas de Chequeo* y aporta como producto de salida: Los *Datos de calidad*.
- La incorporación de la documentación nombrada proporciona información valiosa para el repositorio de conocimiento.

La propuesta de proceso de transformación de datos para proyectos de explotación de información, las actividades y las técnicas asociadas han sido validadas en tres dominios de conocimiento con características bien diferenciadas: El primero trata sobre la necesidad de predicción de clientes de depósitos a largo plazo, el segundo sobre datos de pacientes indios con problemas hepáticos y por último sobre un conjunto de datos para el diagnóstico de cáncer de mama.

6.2. FUTURAS LÍNEAS DE INVESTIGACIÓN

Durante el desarrollo de este trabajo final de licenciatura han surgido cuestiones que si bien no son centrales al tema abordado en la misma, constituyen temas concomitantes que (en consideración del alumno) dan lugar a las siguientes líneas de investigación futuras:

- En este trabajo se han utilizado teorías de descubrimiento de problemas en los datos que fueron propuestas por otros investigadores las cuales están diversificadas y además podrán surgir en instancias futuras nuevos problemas de calidad en los datos. Además según sea el caso de problema descubierto, se puede utilizar teorías diferentes para solucionar mismos casos de problema. Son de suma importancia la utilización de estas técnicas para proporcionar datos de calidad por lo que surge las siguientes preguntas:
  - ¿Cuáles son las teorías necesarias para el descubrimiento de problemas en los datos?
  - Según el caso descubierto, ¿Cuál es la teoría que se adapte mejor para la solución del problema en los datos?
- Este proceso propone mejor el rendimiento a la hora de necesitar datos de calidad en proyectos de explotación de información por lo que el aseguramiento de las entradas necesarias y la buena utilización de las salidas depende de los pasos previos y posteriores relacionados al proyecto en si por lo que se pregunta:

- ¿Cuál es la organización necesaria para el proyecto de explotación de información que garantice la buena utilización de este proceso?
- ¿Qué actividades se deben efectuar previamente para garantizar las entradas necesarias para este proceso?
- ¿Qué actividades se deben efectuar posteriormente para garantizar la buena utilización de las salidas de este proceso?

- Si bien el proceso propuesto en este trabajo aporta sistematicidad al proceso de transformación de datos y el mismo ha sido validado en dominios representativos, quedan como temas de trabajo abiertos:
  - La validación empírica más amplia del proceso de transformación de datos mediante la técnica de muestras apareadas basadas en grupos experimental y de control.
  - La validación empírica de las técnicas propuestas en un conjunto vasto y representativo de dominios de aplicación.

7. REFERENCIAS

[Allison, 2001] Allison, Paul D. Missing Data Techniques for Structural Equation Modeling, CA: Sage Publications. 2001.

[Botía, 2010] Juan A. Botía, Preprocesado de Datos. Departamento de Ingeniería de la Información y las Comunicaciones. Universidad de Murcia. Ingeniería Superior en Informática, UMU. 2010.

[Casal y Mateu, 2003] Jordi Casal, Enric Mateu. Tipos De Muestreo. CReSA. Centre de Recerca en Sanitat Animal / Dep. Sanitat i Anatomia Animals. Universitat Autónoma de Barcelona, 08193-Bellaterra, Barcelona. 2003.

[Cohen, 1998] Cohen, W.W. Integration of Heterogeneous Databases without Common Domains Using Queries Based on Textual Similarity. En: Proceedings of the SIGMOD International Conference Management of Data SIGMOD (Seattle, Washington, Junio 2-4), 1998.

[Davis y McCuen, 2005] Davis, A. y McCuen, R. Storm Water Management for Smart Growth. Springer. 2005.

[Elmagarmid et. al., 2007] A.K. Elmagarmid, P.G. Ipeirotis and V.S. Verykios, Duplicate Record Detection: A Survey, IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, 2007.

[Farhangfar et. al., 2007] Farhangfar, A., Kurgan, L.A. y Pedricks, W. A Novel Framework for Imputation of Missing Values in Databases. 2007.

[Farhangfar et. al., 2008] Farhangfar A., Kurganb L. and Dy J., Impact of imputation of missing values on classification error for discrete data. Pattern Recognition, vol. 41, 2008, 3692 - 3705. 2008.

[Gotoh, 1982] Gotoh, O. An Improved Algorithm for Matching Biological Sequences, Journal of Molecular Biology, vol. 162, no. 3, pp. 705-708, 1982.

[Grubbs, 1969] Grubbs, F. Procedures for Detecting Outlying Observations in Samples, Technometrics, Vol 11, No. 1, pp 1-21. 1969.

[16] [Iglewicz y Hoaglin, 1993] Iglewicz, B. y Hoaglin, D. How to detect and handle outliers. American Society for Quality. Statistics Division. 1993.

[Jöreskog, 2005] Jöreskog KG. Structural equation modeling with ordinal variables using LISREL. 2005.

[Jordi y Enric, 2003] Jordi Casal, Enric. Tipos de muestreo. CReSA. Centre de Recerca en Sanitat Animal / Dep. Sanitat i Anatomia Animals, Universitat Autònoma de Barcelona, 08193-Bellaterra, Barcelona. 2003.

[Levenshtein, 1966] Levenshtein, V.I. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals", Soviet Physics Doklady, vol. 10, no. 8, pp. 707-710, 1966.

[Li y Edwards, 2001] D. Li y E. Edwards. Automatic Estimation of Dixon's Test for Extreme Values Using a SAS Macro Driven Program. PharmaSug 2001.

[Little y Rubin, 1987] Little, R. y Rubin, D. Statistical Analysis with missing data. New York: John Wiley & Sons. 1987.

[Matsumoto et. al., 2007] Matsumoto, S., Kamei, Y., Monden, A., y Matsumoto, K. Comparison of Outlier Detection Methods in Faultproneness Models. En: Proceedings of the First international Symposium on Empirical Software Engineering and Measurement ESEM 2007 (Madrid, España, Septiembre 20 – 21), 2007.

[Merlino, 2004] Merlino, H, Un método de preprocesamiento de datos orientado al uso de explotación de información basado en sistemas inteligentes, Trabajo final especialidad en ingeniería de sistemas expertos, Instituto Tecnológico de Buenos Aires. 2004.

[Neftalí, 2006] Neftalí de Jesús Calderón Méndez. Minería De Datos Una Herramienta Para La Toma De Decisiones. Asesorado por el Ing. Edgar Mauricio Lone Ayala. Guatemala, abril de 2006.

[Neyman, 1934] Jerzy Neyman. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society Vol. 97, No. 4, pp. 558-625 Published by: Wiley. Disponible desde internet el día 20/02/2013 en http://www.jstor.org/stable/2342192. 1934.

[Nisselson et. al., 1983] Nisselson, H., Madow, W., Olkin, I. Incomplete Data in sample Surveys: Treatise. Wonder Book. New York. 1983.

[Rousseeuw y Leroy, 1996] Rousseeuw, P. y Leroy, A. Robust Regression and Outlier Detection. 3a Ed. New York, John Wiley & Sons, 1996.

[Scheffer, 2002] Scheffer Judi, Dealing with Missing Data. Res. Lett. Inf. Math. Sci. 3, 153-160. Disponible desde internet el día 20/02/2013 en http://www.massey.ac.nz/~wwiims/research/letters/. 2002.

[Smith y Waterman, 1981] Smith, T.F. y Waterman, M.S. "Identification of Common Molecular Subsequences", Journal of Molecular Biology, vol. 147, no. 1, pp. 195-197, 1981.

[Tukey, 1977] Tukey, John W . Exploratory Data Analysis. Addison-Wesley. Reading, Mass. : Addison-Wesley Pub. Co. 1977.

[Winkler, 1990] Winkler, W.E. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage", Proceedings of the Section on Survey Research Methods, pp. 354-359, 1990.

[Yancey, 2006] Yancey, W.E. "Evaluating String Comparator Performance for Record Linkage", Proceedings of the Fifth Australasian Conference on Data mining and Analytics, pp. 23-21, 2006.