Trabajo de Diploma

SISTEMA PARA LA SUSTITUCIÓN DE VALORES NULOS EN UNA BASE DE DATOS

AUTORES
Roxana Pérez Rubido
Yuriany Borges Roque
Dunia Taymí Machado La Paz

Tutores
MSc. Beatriz López Porrero
Dr. Ramiro Pérez Vázquez

2006

UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS

FACULTAD DE MATEMÁTICA FÍSICA Y COMPUTACIÓN

LICENCIATURA EN CIENCIAS DE LA COMPUTACIÓN

SISTEMA PARA LA SUSTITUCIÓN DE VALORES NULOS EN UNA BASE DE DATOS

2006

Santa Clara

Hago constar que el presente trabajo fue realizado en la Universidad Central Marta Abreu de Las Villas como parte de la culminación de los estudios de la especialidad de Ciencias de la Computación, autorizando a que el mismo sea utilizado por la institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos ni publicado sin la autorización de la Universidad.

_____

Firma de los autores

Los abajo firmantes, certificamos que el presente trabajo ha sido realizado según acuerdos de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

_____ _____

Firma de los tutores Firma del jefe del Seminario

PENSAMIENTO

El principio de la sabiduría es el temor de Jehová.

Pr.1:7

DEDICATORIA

A mis padres…

A José Antonio, mi esposo por no poder dedicarle más de mi tiempo…

Dunia Taymí Machado La Paz.

A mis padres por su esfuerzo y dedicación en todos estos años.

A Ernesto, mi esposo, por su comprensión y ayuda sin límite.

Yuriany Borges Roque

A mis padres por ser mi refugio y guía.

A mi hermano por su apoyo incondicional.

Roxana Pérez Rubido

AGRADECIMIENTOS

Les agradezco de todo corazón a todas las personas que colaboraron de forma científica, material y emocional a realizar este trabajo.

A Dios porque sin él nada hubiese sido posible.

A mis padres por su apoyo e incentivo.

A mi familia que siempre estuvo presente.

A mis profesores de siempre.

A Alejandro por las comidas y su ayuda incondicional.

A mis tutores Dr. Ramiro Pérez y Msc. Beatriz López por la asesoría prestada.

A todos mis compañeros de dominó que de una forma u otra colaboraron para aminorar el estrés.

RESUMEN

En los Sistemas de bases de datos frecuentemente se dejan de introducir datos, lo cual en principio puede traer graves consecuencias en las respuestas a solicitudes que se hagan sobre la información; pero esto puede tener aún mayor incidencia si las bases de datos de estos sistemas operacionales son fuentes de Almacenes de datos, pues la ausencia de valores puede influir negativamente en los procesos de toma de decisiones.

En este trabajo se expone el concepto de valor ausente y nulo; así como los tipos de ausencias. Además se destaca cómo el proceso de imputación de estos valores forma parte de la limpieza de datos que es necesario llevar a cabo cuando se produce la carga de los datos operacionales en un almacén o cuando se va hacer un proceso de minería de datos para la toma de decisiones.

Se explican algunas técnicas reportadas en la literatura para solucionar los problemas ocasionados por la existencia de datos ausentes y nulos dentro de las bases de datos y se presenta una herramienta en la que se han implementado algunas de las mismas.

ABSTRACT

In the data base systems some data are often not given, in principle it might cause bad consequences in the answers to questions about information, but it might be still worse if the data base of these operational systems is the source of date store, because the absent values could mislead in the process of taking decisions.

In this paper the concept of absent and missing values is introduced .Besides it is highlighted how the imputation process of this values is a part of the data cleaning that it is necessary to carry out when the operational data loading is done in a store or a data minning process is done in order to take decisions.

Some techniques used to obtain this replacement of missing values found in specialized papers are explained, and also we introduce a tool with some of these techniques included.

ÍNDICE

INTRODUCCIÓN

En las Bases de Datos Operacionales que nutren los Almacenes de datos, es usual que se dejen de introducir datos. Esto es un problema común y se debe a diferentes factores: desconocimiento de información, errores de diseño, no aplicabilidad, entre otros.

La ausencia o incompletitud de estos valores es siempre un problema cuando se realiza su análisis. Causan el deterioro de la calidad de los datos que alimentan al Almacén y como consecuencia la toma de decisiones que sobre esta información puede realizarse se ve también afectada. De ahí que tratar los valores ausentes dentro del proceso de limpieza sea de gran importancia y que la sustitución de los valores nulos sea uno de los aspectos tratados como fundamental en el proceso de Limpieza de Datos.

En la literatura se reportan algunas técnicas de sustitución de los valores ausentes. La más precisa es encontrar la información que falta pero evidentemente es la más difícil o engorrosa de utilizar, el uso de otras técnicas está determinado por la naturaleza del dato que se sustituye, si es numérico continuo, numérico discontinuo, cadena, fecha, u otro tipo.

Este proyecto tiene como objetivo general:

> Construir un sistema que ayude en la sustitución de valores ausentes y valores nulos en una Base de Datos utilizando diferentes técnicas.

Y como objetivos específicos:

- Estudiar y resumir la problemática de la aparición de los valores ausentes y valores nulos en las Bases de datos.
- Implementar métodos de sustitución de valores ausentes y valores nulos de dominio numérico por valores de tendencia central.
- Implementar métodos de sustitución para valores ausentes y valores nulos de dominio discreto.
- Realizar el Análisis, diseño e implementación del sistema.

Preguntas de Investigación.

¿Qué problemas trae la presencia de datos ausentes en un Almacén de datos?

¿Qué métodos se utilizan para la sustitución de valores ausentes?

¿Son aplicables todos los métodos en todas las situaciones?

¿Influye el tipo de dato, la naturaleza del dominio, en el método escogido?

¿Es necesaria la construcción de un software para la sustitución de nulos o bastará la utilización de paquetes profesionales de estadística?

En este trabajo se parte del hecho que la presencia de valores ausentes en las Bases de Datos de nuestro entorno es un problema común, por lo que cualquier intento de construcción de Almacenes de Datos a partir de Bases de Datos Operacionales en nuestras empresas requerirá en el proceso de limpieza de la sustitución de valores ausentes.

Suponemos que una herramienta de este tipo viabilizará el trabajo de reemplazo a pesar de que este pudiera ser realizado con otros paquetes, pero esto último requeriría conocimiento sobre los mismos y el tiempo de labor sería mayor.

El presente documento está estructurado de la siguiente forma:

Capítulo 1. Limpieza de datos. Tratamiento de valores ausentes.

Se presentan las características generales del problema, información sobre los diferentes tipos de datos ausentes, y la descripción de los métodos empleados en el tratamiento de los mismos.

Capítulo 2. Análisis y diseño de la herramienta "Sistema para la Sustitución de Valores Nulos en una Bases de Datos".

Se presenta el diseño y las características generales de la implementación del software.

Capítulo 3. Manual de usuario.

Se muestra información de cómo debe ser usado el software.

CAPÍTULO 1. LIMPIEZA DE DATOS. TRATAMIENTO DE LOS VALORES AUSENTES

En muchas aplicaciones relacionadas con los temas de investigación de Descubrimiento de conocimiento en los datos, Almacenes de Datos, y Toma de Decisiones, un aspecto crítico lo constituye el nivel de corrección de los datos con que se trabaja. Este tipo de aplicaciones generalmente se nutren de Bases de Datos Operacionales, y con frecuencia en las mismas se encuentran registros de datos con información incompleta o errónea, pues aunque se plantea que existen varios factores que pueden influir en la calidad, consistencia e integridad de los datos; muchas veces, el origen de los datos, constituye un factor crucial. Situaciones tales como "fecha de nacimiento desconocida", "conferencista por confirmar", "dirección actual desconocida" son comunes (Date, 2003 Tercera Parte). Aún cuando los desarrolladores de sistemas hagan ingentes esfuerzos por evitar los errores en los datos, la razón de error[1] es aproximadamente de un 5% (Redman, 1998: 78-79).

La existencia de "datos sucios", como también se le llama a estos errores en los datos, tiene un gran impacto en las instituciones, reflejándose esto en un alto costo operacional, toma de decisiones inadecuadas, incremento de la inseguridad y una desviación de la atención de las direcciones de las instituciones (Jonathan y Maletic, 2002).

La solución lógica a este problema es tratar de "limpiar" los datos de alguna forma; o sea, explorarlos para encontrar posibles errores y tratar de corregirlos. La realización de este proceso de forma manual es casi imposible por el número de horas-hombre requeridas para el mismo, además de ser por sí mismo un proceso muy laborioso, lento y susceptible de introducir nuevos errores en los datos; de ahí que la automatización de la limpieza de datos sea considerada una nueva e importante área de trabajo científico.

[1] Se define la razón de error como el número de errores por campos sobre el número total de campos

1.1. Limpieza de datos.

Un Almacén de Datos es una especie de punto focal que guarda en un único lugar toda la información útil, proveniente de sistemas de producción y fuentes externas. Una definición muy difundida es la dada por Inmon y Hakathorn: "El Almacén de Datos es una colección de datos orientados al tema, integrados, no volátiles e historiados, organizados para el apoyo de un proceso de ayuda a la decisión" (Inmon y Hakathorn, 2004).

El proceso de llevar los datos a un Almacén de Datos se conoce como carga del almacén. Este proceso de adquisición se desarrolla en tres fases: la extracción, la transformación y la carga (ETL, por sus siglas en inglés **E**xtract, **T**ransformation, **L**oad).

En el proceso de ETL la extracción de los datos consiste en la determinación de los datos relevantes que deben ser propagados al Almacén. Con vistas a disminuir la sobrecarga en el tiempo de procesamiento, este proceso incluye solamente los datos que han sido cambiados después del proceso previo de ETL, principalmente los nuevos artículos insertados y/o actualizados. La integridad de los datos extraídos es obligatoria y precisa la sincronización de los diferentes procesos de extracción.

La preparación de los datos corresponde a la transformación de las características de los datos del sistema operacional en la forma definida por el Almacén de Datos. Esta preparación incluye la correspondencia de los formatos de datos, la limpieza, la transformación y la agregación.

La Limpieza de datos es el proceso que se encarga de detectar y eliminar anomalías en los datos y la necesidad de llevarla a cabo aumenta cuando existen varias fuentes de datos que necesitan ser integradas (M. G. Ceruti and M. N. Kamel, 1999; An Extensible Framework for Data Cleaning, 2000). Las anomalías pueden ocurrir en Bases de Datos Operacionales, en Almacenes de Datos y en Conjuntos de Datos utilizados en sistemas de tomas de decisiones. La presencia de los errores o anomalías provoca que la información que se desea extraer de esos datos no tenga la calidad necesaria y entonces surge la necesidad de limpiar las fuentes de datos para que lleguen

coherentes a los Almacenes de Datos. Un error común es la ausencia de información y por eso en la Limpieza de Datos constituye una tarea fundamental el tratamiento de los valores ausentes.

Entre los tipos de errores más frecuentes en los datos al llenar una Base de Datos Operacional se encuentran:

- Valores contenidos dentro de atributos de formato libre o valores mal colocados.
- Valores atribuidos incorrectos.
- Diferentes representaciones para los atributos, y diferentes significados para los valores.
- Artículos duplicados.
- Valores ausentes y valores nulos.

1.1.1. Tipos de valores faltantes.

Los valores nulos y los valores ausentes pueden ser consecuencias de diferentes causas: ausencia de respuesta del cliente (por ejemplo en una encuesta), fallas en la trascripción de datos, fallas en el soporte físico de los datos, mal funcionamiento de los sistemas de adquisición de datos, no aplicabilidad del valor del campo al registro de información, entre otras.

La definición de estos involucra dos formas diferentes de ver la ausencia de información (Pyle, 1999):

- **Dato ausente**: es un valor que no está en nuestro conjunto de datos; pero que existe en el mundo real, sencillamente por algún error no aparece en nuestra Base de Datos Operacional.

- **Dato nulo(o vacío)**: es un valor que está fuera de la definición de cualquier dominio, el cual permite dejar el valor del atributo "latente". En otras palabras, un valor nulo no representa el valor cero, ni una cadena vacía, estos son valores que tienen significado; implica ausencia de información porque se desconoce el valor del atributo o simplemente para ese objeto no tiene sentido. Es un dato que falta, o sea, no existe en el mundo real.

Un ejemplo pudiera ser el siguiente: supongamos que controlamos en una empresa que vende sandwich la salsa que prefieren los usuarios (supongamos dos tipos de salsa: ketchup y mostaza y que son excluyentes) y además su sexo. Llega entonces un usuario que pide un sandwich sin ninguna salsa y supongamos que el operador del sistema olvida introducir al sistema el sexo del usuario. Habrá entonces dos campos sin valor: sexo y tipo de salsa, el primero existe (el usuario tiene un sexo determinado), sin embargo el segundo no existe (el usuario no quería ninguna salsa). Al cargar estos datos en el almacén se puede "sugerir" un dato para el sexo, no así para la salsa.

En los sistemas de bases de datos habitualmente no se hace diferencia entre un tipo de valor y otro, sencillamente se deja nulo el campo o en el mejor de los casos se utiliza el valor NULL. En el caso de los sistemas de Bases de Datos en nuestro país es común la no utilización de la marca NULL y en su lugar se escribe 0 (si es un dato numérico) o una cadena vacía en el caso de cadenas. Esto trae una complejidad adicional en el tratamiento de nulos pues el cero y la cadena vacía pueden tener significados concretos diferentes a la ausencia de valor.

1.1.2. Valores ausentes y su significación en los Almacenes de Datos.

Los valores ausentes son un problema común y la mejor implementación para minimizarlo es a través de una cuidadosa administración y asegurando la calidad de los datos. Cuando estos valores son menores que 1% son generalmente considerados triviales, de 1-5 % manejables, de 5-15% requiere de métodos sofisticados para manejarlos y más de un 15% afecta seriamente cualquier clase de interpretación.(Data clearing and replacement of missing values, 1999)

Estos valores en las Bases de Datos Operacionales deben ser reemplazados al cargarse en el Almacén de Datos, pues su presencia influye negativamente en los procesos de toma de decisiones, porque los modelos que se obtienen de estos datos pueden no expresar la realidad correctamente.

Un método empleado para el tratamiento de los valores ausentes consiste en eliminar los registros que contengan campos nulos, pero con esta solución podría ocurrir que la base de

datos resultante fuese demasiado pequeña o no refleje del todo la base original, además de que reduce drásticamente el rendimiento de algunas pruebas estadísticas. Por otra parte el hecho de que falte el valor de un campo en un registro podría tener algún significado en sí mismo.

1.2. Tipos de ausencia de datos.

Además de la diferenciación entre valor ausente y nulo expuesto anteriormente, en varios artículos (Jiménez, 2000; Garson, 2005;) se establece que es tan importante clasificar los tipos de valores faltantes como reconocer de qué manera se ha producido esa ausencia. De allí que se hable de tipos de ausencia de datos (no de tipos de datos ausentes). El tipo de ausencia será la que determine la estrategia analítica a emplear. De esta forma las ausencias se clasifican en (Little y Rubin, 1987):

- **Ausencia Completamente Aleatoria (MCAR[2]):** los datos perdidos son considerados MCAR cuando la probabilidad de ausencia del valor de una variable es completamente independiente del valor mismo, y del valor de cualquier otra variable. En otras palabras, la ausencia de los valores no está relacionada con las variables especificadas. Por ejemplo, suponiendo que peso y edad son variables de interés para un estudio en particular, si la probabilidad de que una persona proporcione su peso es igual para todos los individuos sin importar su peso o su edad, entonces el valor perdido es considerado MCAR. En el caso de una encuesta a un grupo familiar, la ausencia del valor "ingreso familiar" podría no ser considerada MCAR si se sabe que las familias con bajos ingresos tienden a omitir este dato en más oportunidades que las familias con altos ingresos(es decir, mientras más bajo el ingreso familiar, más alta la probabilidad de su ausencia). De esta manera, si la gente blanca es más proclive a omitir el dato de su "ingreso familiar" que la gente negra, entonces la ausencia del dato "ingreso familiar" no puede ser considerada MCAR, pues puede ser correlacionada con la variable "grupo étnico". Debe tenerse presente que lo que importa es el valor de la variable y no su ausencia, por ejemplo, si las personas que omiten reportar su "ingreso personal" son

[2] Siglas en inglés de Missing Completely At Random

también tendientes a omitir su "ingreso familiar", estas variables podrían ser consideradas MCAR, porque ninguna de esas dos circunstancias tiene nada que ver con el valor del ingreso en sí.

- **Ausencia al Azar (MAR[3]):** los datos perdidos son considerados MAR cuando la ausencia del valor de una variable no depende del valor mismo, pero sí del valor de alguna otra variable no recogida en la tabla. Usando nuevamente el ejemplo del peso y la edad, si la probabilidad de que una persona proporcione su peso varió acorde a un peso individual pero no a su edad, entonces estamos en presencia de un valor perdido tipo MAR. Por otra parte, si se sabe que los pacientes depresivos (la variable paciente depresivo no está en la tabla) tienden a no reportar su "ingreso personal", entonces esta ausencia no es MCAR (mientras más depresivos mayor la probabilidad de la ausencia), pero si dentro del grupo de los pacientes deprimidos la ausencia del dato no está relacionada al valor de ninguna variable, esa ausencia es al azar (MAR).

- **No ignorables**: los datos perdidos son considerados no ignorables cuando la ausencia está relacionada con alguna otra variable existente en la tabla. La ausencia no es aleatoria ni predecible desde ninguna variable en la base de datos. Ejemplo de esta ausencia es si la probabilidad de que un individuo proporcione la información de su peso varió acorde al peso de la persona en cada categoría de la edad, entonces estamos en presencia de una ausencia no ignorable. Este tipo de ausencia es la condición más dura a la hora de tratarla; y desafortunadamente es la más probable de que ocurra.

1.3. Tratamiento de la información faltante.

Una vez que la información sobre los modelos de valores perdidos es obtenida, estos pueden reemplazarse con valores apropiados. Numerosas técnicas han sido desarrolladas y reportadas en la literatura para solucionar los problemas ocasionados por la existencia de datos ausentes dentro de las Bases de Datos.

Por ejemplo un valor perdido puede ser ignorado, este camino puede ser el más fácil para el manejo de los mismos, pero no contribuye a la calidad de las Bases de Datos que contienen este valor (Redman, 1992; Wand y Wang, 1996). Otra forma típica de manejarlos es reemplazándolos por valores promedios, sustituyéndolos por valores mínimos, valores máximos. Estas técnicas pueden solamente ser aplicadas cuando los atributos son valores numéricos porque las medidas estadísticas usadas solo definen valores numéricos. La inferencia del valor más probable es también usada para el llenado de estos (Little y Rubin, 1987). Esta implementación es útil solamente cuando los valores son nominales pues trae problemas definir el valor más probable en un conjunto de valores continuos.

Si solamente un pequeño por ciento (<5%) de los datos es perdido, entonces estos se pueden sustituir usando la media (si es normal), mediana (si es sesgado) o moda (si es categórico), donde el objetivo es comparar varios grupos (las condiciones del género o tratamiento), y a menudo es deseable hacer este reemplazo dentro de cada grupo (Data clearing and replacement of missing values, 1999).

Cuando el porcentaje de los valores perdidos excede el 5%, un nuevo problema surge. Sustituir todos los registros por un solo valor disminuirá la varianza y aumentará la significación de cualquier prueba estadística basada en el mismo. Es, por tanto, recomendado sustituir los datos usando métodos más avanzados: imputación hot-deck, imputación múltiple (que modela la incertidumbre a los datos que faltan mientras usa los datos existentes (Rubin, 1987) o un modelo de regresión (que predice el valor que falta en los otros datos disponibles). Los modelos de regresión e imputación múltiple son más elegantes, pero mucho más difíciles porque cada variable requiere una ecuación diferente y en muchos casos múltiples ecuaciones por variable debido a que algunos predictores pueden estar perdidos (Data cleaning and replacement of missing values, 1999).

Algunas de estas técnicas prometen dar más información que otras, pero son muy complejas computacionalmente. Otras son muy poderosas bajo ciertas circunstancias pero pueden introducir sesgo en otras. La complejidad computacional es un problema. Además son matemáticamente

[3] Siglas en inglés de Missing At Random

complejas y varían según el tipo de datos a los cuales serán aplicados, consumiendo también mucho tiempo para un conjunto de datos muy grande (Pyle, 1999).

Analicemos ahora algunas de estas técnicas, las cuales se diferencian entre sí por su nivel de complejidad y por la calidad de los resultados:

- **Análisis de Casos Completos**[4] (Jiménez, 2000): Considerada segura y conservativa, esta técnica es una de las más empleadas y de hecho está fijada por defecto en muchas aplicaciones de Estadística y Minería de Datos. También es la más sencilla: simplemente se elimina todo registro que contenga algún campo con valor ausente (también puede ser flexibilizado para que elimine todo registro que tenga valores ausentes en determinados campos de interés para un análisis en particular). Cualquier otra técnica alternativa de manejo de datos ausentes debe ser estudiada y evaluada en comparación con esta técnica. En algunos casos es preferible regresar al Análisis de Casos Completos y asumir sus riesgos antes que aplicar alguna otra técnica más costosa o que genere datos de baja calidad.

  Objetivo:

  Crear una nueva tabla donde cada uno de los campos de cada uno de los registros almacena el dato correspondiente, siendo este dato extraído de la realidad (no aproximado o calculado, sino tomado directamente de la fuente de datos).

  Ventajas:

  Las ventajas están dadas por su simplicidad y por el poco tiempo computacional que requieren. Esta técnica no introduce ruido en el conjunto de datos y solo emplea observaciones "reales" (Stones Analystics, 2003).

  Desventajas:

  Esta técnica puede resultar una disminución significativa del volumen de datos a analizar, lo que resta mucha confiabilidad a los clásicos y tradicionales estudios estadísticos, en

[4] Traducción de Complete Case Analysis, también llamado Casewise o Listwise Deletion.

consecuencia, los resultados obtenidos pueden tener baja confiabilidad o sesgo (Little y Rubin, 1987; Rubin, 1987; Rubin, 1996).

Cuando la cantidad de casos de valores ausentes en grandes bases de datos es menor que un 5% es muy común eliminar casos de la base de datos (Garson, 2005).

- **Sustitución por la Media** (Jiménez, 2000): Esta técnica es una de las más simples y útiles; pues hace imputaciones (asignación de un valor calculado a un campo con dato ausente) y respeta completamente al subconjunto de los datos que sí están completos. La idea es bastante simple: se calcula el valor de la "media" para cada variable del conjunto de datos y el valor así obtenido es imputado a cada registro que carezca de un valor para esta variable. Es importante destacar que la "media" se calcula de manera diferente dependiendo de la naturaleza de la variable: para variables "categóricas" o "de clase" (por ejemplo, variables cuyos valores posibles sean A, B, C y D) se utiliza la moda (el valor con frecuencia más alta, es decir, el que esté presente en el conjunto de datos un número mayor de veces); para variables "enteras" u "ordinales", se emplea la mediana (el valor que se encuentra "en medio" de la lista de datos ordenada crecientemente, es decir, el valor que cumple que la mitad de los valores de la variable son menores a él y la otra mitad son mayores que él) y para variables "reales" o "continuas" se emplea el promedio (la sumatoria de los valores de la variable divididos por su frecuencia). Esta técnica está incluida en algunos sistemas de ayuda y soporte al proceso de Minería de Datos como un modelo de referencia; pero debe ser utilizado con cuidado pues puede introducir sesgo en el conjunto de datos.

  Objetivo:

  Pretende crear un nuevo conjunto de datos que contenga exactamente el mismo número de registros que el conjunto original, sin que haya en ellos algún dato ausente. Imputa en el conjunto de datos valores con propiedades estadísticas bien conocidas (moda, mediana y promedio) que se sabe de seguro pertenecen al espectro de valores posibles para esas variables y que tienen una buena probabilidad de aparecer, además de permitir hacer suposiciones sobre el resultado de técnicas de Minería de Datos aplicadas posteriormente,

pues aunque no genera variaciones en la media, sí genera cambios en la varianza y en la distribución de la variable imputada.

Ventajas:

Esta técnica es sencilla de implementar y requiere relativamente poco cálculo, imputa en el conjunto de datos valores con propiedades estadísticas bien conocidas y que permiten predecir de alguna manera el impacto que tendrán sobre futuros análisis, además que mantiene la validez de varias pruebas estadísticas de uso común (que trabajan sobre la esperanza, el coeficiente de correlación o la covarianza). Es una buena solución cuando la ausencia es aleatoria y está distribuida normalmente. También es ventajosa porque produce internamente un conjunto de datos de la matriz de correlación consistentes (Stones Analystics, 2003).

Desventajas:

La sustitución por la Media genera sesgo en los datos, pues imputa un mismo valor para cada ausencia de una misma variable (es decir, que si una variable está ausente en 20 registros diferentes, habrán 20 nuevas ocurrencias de la "media" en esa variable), lo cual puede afectar a otras técnicas de Minería de Datos hasta el punto de invalidarlas (por ejemplo, si se aplicara una "clasificación", la súbita aparición de más ocurrencias de los valores medios podría causar mucha "afluencia" de registros hacia una clase en particular).

- **Imputación por regresión** (Stones Analystics, 2003)**:** Predice los valores perdidos basados en la ecuación de regresión que usa las demás variables relevantes como predictoras. Para saber que tipo de regresión usar hay que tener en cuenta el tipo de la variable que tiene la información incompleta. Si el valor que ha de imputarse es número (p. ej. la edad, el salario o los valores de presión arterial), se puede emplear la regresión múltiple. En caso que sea una variable categórica, como sexo, el estatus socioeconómico o la práctica de ejercicio físico en el tiempo libre, podría emplearse la regresión logística y hacer la imputación según la probabilidad que el modelo de regresión estimado otorgue a cada categoría para el sujeto en cuestión (Cañizares, et al., 2003).

Ventajas:

Este método preserva la varianza y la covarianza de las variables con valores ausentes. Además permite trabajar con una Base de Datos completa, la que puede ser analizada empleando los procedimientos y paquetes estadísticos estándares (Cañizares, et al., 2003).

Desventajas:

Si los errores estándares son ignorados cuando los valores perdidos son predecidos, puede inflar el poder del modelo de predicción puesto que los valores perdidos de las variables dependientes fueron presentados perfectamente como predictores. Además esta técnica depende de un orden de acuerdo al cual las variables serán reemplazadas.

- **Método probabilístico basado en la distribución de los datos no perdidos** (Martina, 2005)**:** La idea de esta técnica es estimar la distribución de los datos disponibles en la variable donde queremos restablecer los valores perdidos y entonces, generar estos de acuerdo a dicha distribución. Puede ser implementado en 4 pasos:
    1. Calcular la frecuencia de los valores válidos. Para asegurar la menor acumulación de errores en el procedimiento los datos deben ser ordenados de forma ascendente de frecuencia de valores válidos.
    2. Calcular porciento válido (depende de la frecuencia) y porciento acumulado.
    3. Calcular el porciento de valores perdidos y el número de valores a ser reemplazados.
    4. Reemplazar los nulos por el valor válido correspondiente a la frecuencia analizada. Disminuir la cantidad de los valores perdidos.

  Este método se repite hasta que la cantidad de valores perdidos sea cero.

Ventajas:

Fácilmente automatizable desde cualquier lenguaje de programación o gestor de Bases de Datos.

- **Desviación Estándar** (Pyle, 1999): La forma de la variabilidad es un concepto importante en la decisión de qué valores usar para el reemplazo. La Desviación Estándar es una medida de variabilidad. Para usarla hay que calcular un nuevo valor para cada ausencia de la variable en la Base de Datos, de tal forma que mantenga la desviación estándar. El proceso del cálculo se realiza tantas veces como valores ausentes tenga la variable.

  Ventajas:

  La Desviación Estándar refleja más información sobre una variable que la media. Esta no solo refleja la tendencia central sino también información sobre la variabilidad con que se distribuyen las variables.

  Desventajas:

  Si la ausencia es múltiple hay que hacer el cálculo tantas veces como exista, requiriendo en Bases de Datos grandes de un tiempo de cómputo enorme.

- **Eliminación de Pares de Datos**[5](Jiménez, 2000): Esta técnica es mayormente empleada en los casos que requieren cálculos sobre dos variables. Usa la matriz de correlación donde la correlación entre cada par de variable es calculada desde todos los casos que han validado datos para estas variables (Stones Analystics, 2003).

  Objetivo:

  Esta técnica pretende ser una mejora a la técnica de Análisis de Datos Completos, ya que bajo la hipótesis MCAR, evitaría la drástica reducción del número de registros disponibles para su estudio mientras que mantendría la calidad de los datos. Por lo tanto, esta técnica no persigue la creación de un nuevo conjunto de datos que no tenga ausencias, ni aproxima o hace imputaciones en los datos, simplemente crea un nuevo conjunto de datos que será de utilidad solo para la realización de estudios sobre un par predeterminado de variables, ya que se eliminarán solo los registros donde falten los datos correspondientes a alguna de las dos variables seleccionadas, por lo que no habrá datos ausentes dentro de

ese contexto (aunque en términos generales, habría datos ausentes en el conjunto de datos, aunque no sean de notar para el estudio que se ha de realizar).

Ventajas:

No introduce ruido en el conjunto de datos y solo emplea observaciones "reales". Además es superior a la técnica Análisis de Casos Completos en cuanto a que genera un número mucho menor de eliminaciones, por lo que se dispone de un conjunto de datos de mayor significación.

Desventajas:

Las desventajas de esta técnica superan en mucho a sus ventajas, por lo que está muy desacreditada y ya casi no se emplea. Esta depende demasiado de la hipótesis MCAR, la cual es poco frecuente y difícil de demostrar. Además debido a que emplea diferentes conjuntos de datos para un mismo análisis dependiendo de qué variables se consideren, deben prepararse varios conjuntos para cada análisis (por ejemplo, se debe crear un conjunto de datos para estudiar la relación entre las variable X y Y y otro diferente para estudiar las variables X y Z) cuando lo que se desea es hacer análisis sobre un mismo conjunto de datos, los resultados obtenidos resultan sesgados y además costosos en tiempo, esfuerzo y espacio de almacenamiento.

- **Imputación múltiple** (Cañizares et al., 2003; Stones Analystics, 2003)**:** Este método genera una probabilidad máxima basada en la matriz de covarianza y el vector de las medias e introduce incertidumbres estadísticas en el modelo, usando la incertidumbre para reproducir la variabilidad natural encontrada en los datos de los casos completos. Se refiere a reemplazar cada valor ausente con más de un valor imputado. Es un enfoque basado en simulaciones donde a cada valor ausente se asignan m > 1 valores extraídos de una distribución predictiva, lo que produce m bases de datos. Después, en cada base de datos se realiza el análisis estadístico que responda al propósito del estudio, desde obtener estimaciones puntuales y sus intervalos de confianza hasta modelos de regresión. En este caso se obtienen tantos resultados del análisis realizado como imputaciones se hayan

[5] Traducción de Pairwise Data Deletion

hecho. La distribución predictiva se construye a partir de los valores observados; por ejemplo, usualmente se supone que el conjunto de variables sigue una distribución normal multivariada. Para construir la distribución se necesita estimar sus parámetros: vector de medias, matriz de varianzas y covarianzas. Estas estimaciones se obtienen a partir de las unidades que tienen todos los valores observados. Una vez estimados los parámetros de la distribución, se extraen muestras independientes de ella para asignar los valores en las observaciones que no están completas; según el número de muestras que se seleccione, se tendrá tantas bases de datos para analizar (Rubin, 1987)**.**

Ventajas:

Debido a que con este método no se predice el valor ausente, sino que se modela la incertidumbre que genera la ausencia de los datos, se preservan las relaciones entre las variables cuando se realizan las imputaciones de los valores ausentes (Shafter, 1997).

Desventajas:

Requiere la construcción de 5 a 10 Bases de Datos con valores imputados, las cuales son analizadas de forma individual, haciéndolo muy intenso en cuanto a tiempo.

- **Imputación <u>Hot Deck</u>** (Stones Analystics, 2003)**:** Los datos ausentes son reemplazados con valores seleccionados aleatoriamente presentados en un grupo de datos completos similares; o sea, identifica los casos más similares al caso del valor perdido y sustituye el valor perdido por ese más similar (Handling missing or incomplete data, 2004). El método introduce variaciones en el grupo de los datos de casos completos que producen menos tendencia hacia la media, porque la imputación se hace a partir de estos datos seleccionados aleatoriamente. Las dos áreas principales de importancia son: la selección de los conjuntos de características válidas para identificar el grupo potencial que contiene los valores con la variación razonable, y el aseguramiento de que ese conjunto de características permitirá para grandes grupos la variación razonable.

Ventaja:

Proporciona muestras exactas de población de estudio. Tiene una simplicidad conceptual. Mantiene el nivel apropiado de la medida de las variables (las variables categóricas siguen siendo categóricas y las variables continuas siguen siendo continuas) y la disponibilidad de una matriz completa de los datos en el final del proceso de la imputación que se puede analizar como cualquier matriz completa de los datos (Handling missing or incomplete data, 2004).

Desventajas:

El hecho de que los valores a reemplazar son seleccionados aleatoriamente hace que sea impropio para la predicción porque los valores dependerán de factores espurios tales como el orden de casos en los conjuntos de datos o los números seudo-aleatorios (Sarle, 1998). Además es difícil definir la "semejanza", pues puede haber cualquier número de maneras de definir qué "semejanza" está en ese contexto (Handling missing or incomplete data, 2004).

- **Esperanza máxima** (Stones Analystics, 2003)**:** Se implementa en dos pasos iterativos donde se estiman los parámetros de un modelo comenzando por una suposición inicial. Cada iteración consiste de dos pasos:
    1- **Paso de esperanza:** encuentra la distribución para los valores perdidos basados en los valores conocidos de las variables esperadas y la estimación actual de los parámetros.
    2- **Paso de maximización:** sustituye los valores perdidos con el valor esperado.

El método se reitera a través de estos datos hasta que sea obtenida su convergencia. La convergencia ocurre cuando el cambio de los parámetros estimados de una iteración a otra llega a ser insignificante (Handling missing or incomplete data, 2004).

Desventajas:

A pesar de ser una implementación poderosa y elegante requiere de una programación especializada lo que puede llevar a que sea costoso en tiempo. No agrega ningún componente de la incertidumbre a los datos estimados. Esto significa que mientras las estimaciones del parámetro basadas en la implementación del método son confiables, los errores estándar y los test estadísticos no lo son (Handling missing or incomplete data, 2004).

- **Información Completa de la Probabilidad Máxima (FIML[6])** (Stones Analystics, 2003)**:** Típicamente representado como una matriz de covarianza y el vector de medias, este método usa toda la información disponible sobre los datos observados, incluyendo las medias y las varianzas basadas en los puntos de los datos disponibles para cada variable. La probabilidad es calculada para la porción observada de cada caso de datos y entonces es acumulada y maximizada (SmallWaters Corporation, 2006).

  Ventajas:

  Tiene ventajas sobre la Esperanza Máxima y es que permite el cómputo directo de errores estándares apropiados y pruebas estadísticas. No requiere de la imputación (o paso esperanza) y típicamente converge rápidamente (SmallWaters Corporation, 2006). Además de la conveniencia del uso y la facilidad del conocimiento de las propiedades estadísticas (Handling missing or incomplete data, 2004).

  Desventajas:

  Es difícil incluir las nuevas variables para mejorar la exactitud de los parámetros estimados de los valores perdidos, pero puede no ser utilizado en el modelo estadístico final como predictores. Requiere también de una programación especializada lo que puede llevar a que sea costoso en tiempo.

[6] Siglas en inglés de Full Information Maximum Likelihood

1.4. Apreciaciones del capítulo.

En este capítulo se exponen los conceptos fundamentales de valor nulo y valor ausente en una Base de Datos y la necesidad de un sistema para el reemplazo de los mismos. La elección de las técnicas para el manejo de estos valores resulta una tarea compleja, pues un mismo método en determinadas situaciones produce estimaciones precisa y en otras no. En el capítulo siguiente se realizará el análisis y diseño de la herramienta implementada que permitirá conocer de una forma más detalla el desarrollo y uso de la aplicación.

CAPÍTULO 2 ANÁLISIS Y DISEÑO DE LA HERRAMIENTA "SISTEMA PARA LA SUSTITUCIÓN DE NULOS EN UNA BASE DE DATOS".

En este capítulo se desarrolla el análisis y diseño de la herramienta "Sistema para la sustitución de nulos en Bases de Datos" que nos permitirá conocer una descripción más detallada del problema, así como las herramientas computacionales y estadísticas utilizadas para el desarrollo del mismo.

2.1. Modelación del sistema.

En la modelación del sistema se utilizó la notación UML[7], Lenguaje de Modelamiento Unificado, que es un lenguaje gráfico para visualizar, especificar y documentar cada una de las partes que comprende el desarrollo del software, además este prescribe un conjunto de notaciones y diagramas estándar para modelar sistemas orientados a objetos y describe la semántica esencial de lo que estos diagramas y símbolos significan (*TLDP-ES/LuCAS,* 2006). UML ofrece varios diagramas que permiten modelar sistemas:

- Diagramas de Casos de Uso para modelar los procesos de negocio.
- Diagramas de Actividad para modelar el comportamiento de los objetos en el sistema.
- Diagramas de Clases para modelar la estructura estática de las clases en el sistema (*TLDP-ES/LuCAS,* 2006).

[7] Siglas en inglés de Unified Modeling Language

2.1.1. Diagrama de Casos de Uso.

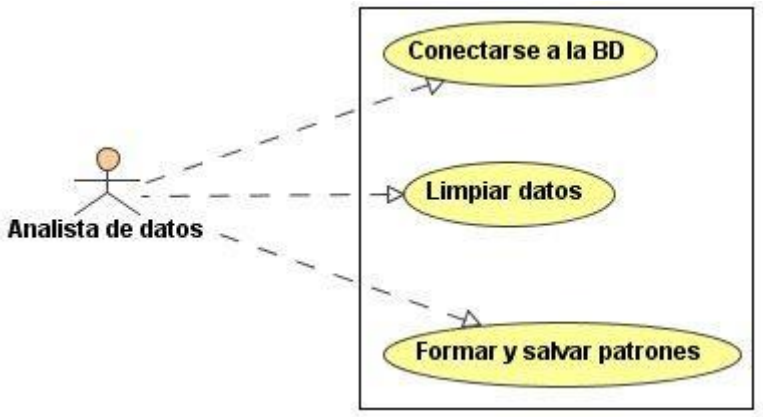

Figura 2.1 *Casos de uso para el actor Analista de Datos.*

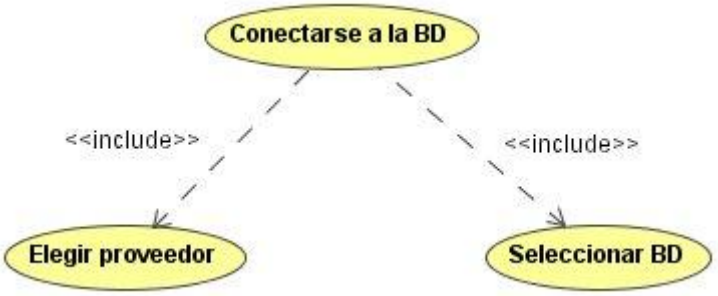

Figura 2.2 *Especificación del Caso de uso Conectarse a la BD.*

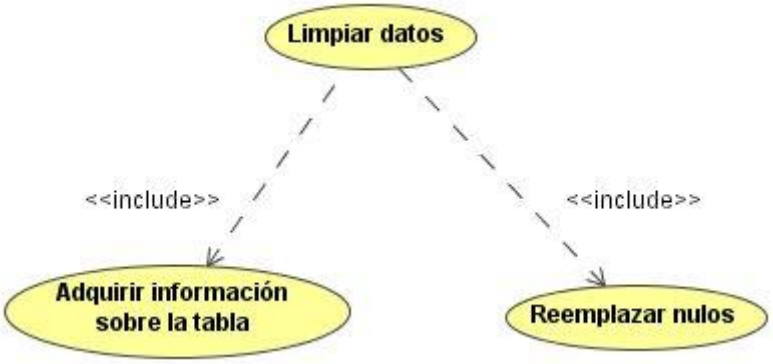

Figura 2.3 *Especificación del Caso de uso Limpiar Datos.*

Figura 2.4 *Especificación del Caso de uso Formar y salvar patrones.*

En el sistema que se propone hay tres casos de usos y un único actor que interactúa con él. En la figura 2.1 se presenta este diagrama para el actor Analista de Datos, a continuación describimos los casos de uso para este actor:

1. Conectarse a la BD (Base de Datos): este caso de uso es el encargado de realizar la conexión a la BD. El actor del sistema es el responsable de seleccionar el proveedor (Access, FoxPro y SQL Server) y la BD a la cual desea conectarse. (Ver figura 2.2).

2. Limpieza de datos: este caso de uso se responsabiliza, después que se logra la conexión a la BD, de realizar el reemplazo de nulos por uno de los métodos de sustitución, ya sea por el método de casos completos ,por regresión lineal simple o por otros de los métodos

3. estadísticos implementados en la herramienta (media, moda, mediana, desviación estándar), luego de haber seleccionado la tabla y el campo al que se le hará la limpieza.(Ver figura 2.3)

4. Formar y salvar patrones: este caso de uso se responsabiliza, después que se logra la conexión a la BD y la selección de la tabla, de formar los patrones asociados a dicha tabla y salvarlos para un fichero texto(extensión .txt).

2.1.2. Diagrama de Actividad.

Los diagramas de actividad proporcionan una forma de modelar el flujo de la información dentro de un proceso. Son típicamente usados para modelar la secuencia de actividades en un proceso. Un diagrama de actividad es considerado un caso especial de una máquina de estado en la cual la mayor parte de los estados son actividades y la mayor parte de las transiciones son implícitamente provocadas por la finalización de las acciones en las actividades (Mármol, 2005). Los diagramas de actividades de la figura 2.5, figura 2.6 y figura 2.7 muestran el conjunto de actividades que realiza un determinado objeto durante la ejecución de la aplicación.

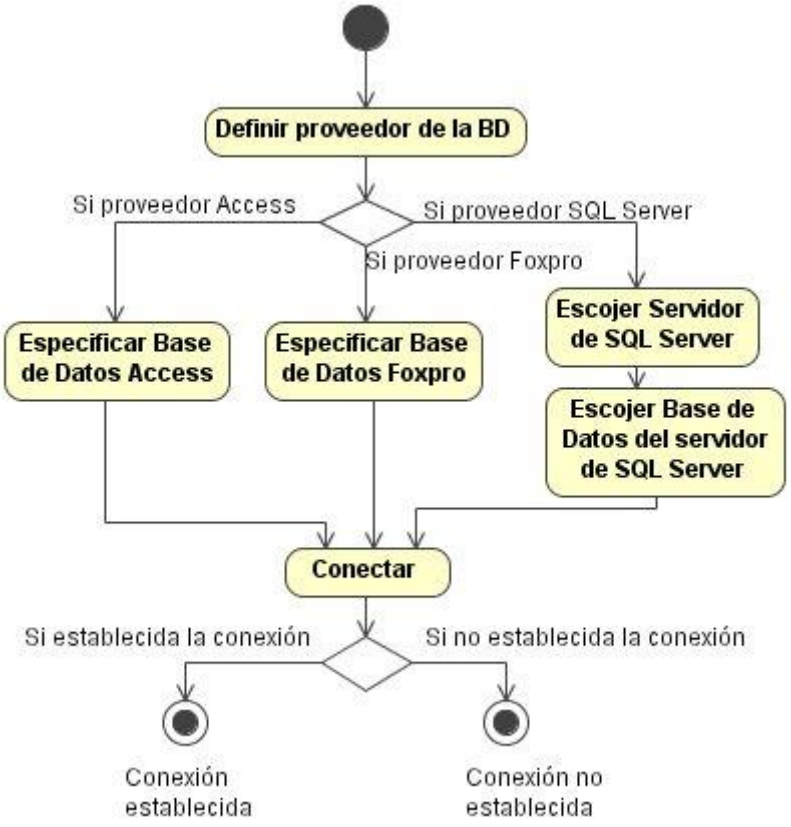

Figura 2.5 *Diagrama de Actividad para el caso de uso Conectarse a la BD.*

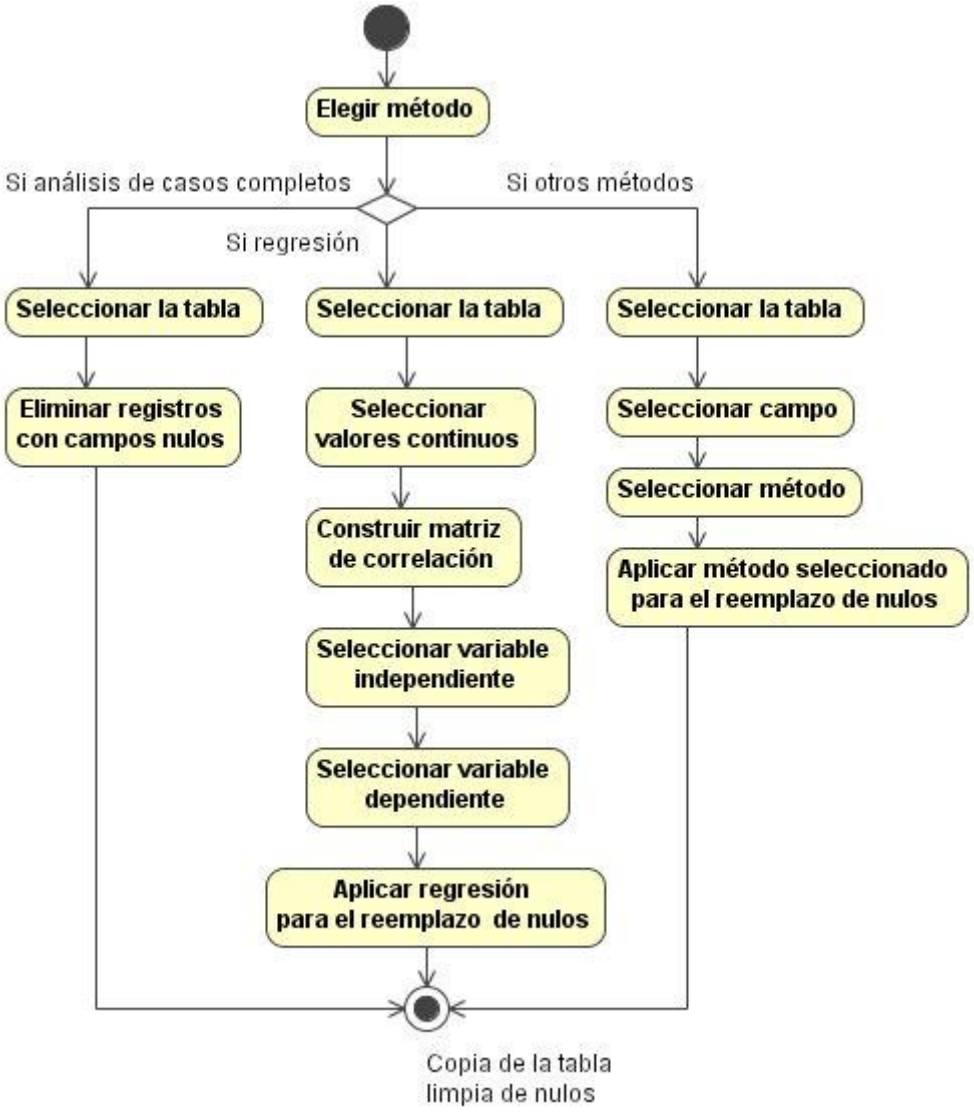

Figura 2.6 *Diagrama de Actividades para el caso de uso Limpieza de los Datos.*

Figura 2.7 *Diagrama de Actividades para el caso de uso Formar y salvar patrones.*

2.1.3. Diagrama de Clases.

El Diagrama de Clases es el diagrama principal de diseño y análisis para un sistema. En él, la estructura de clases del sistema se especifica con relaciones entre clases y estructuras de herencia. Durante el análisis del sistema, el diagrama se desarrolla buscando una solución ideal. Durante el diseño, se usa el mismo diagrama, y se modifica para satisfacer los detalles de las implementaciones (*TLDP-ES/LuCAS*, 2006).

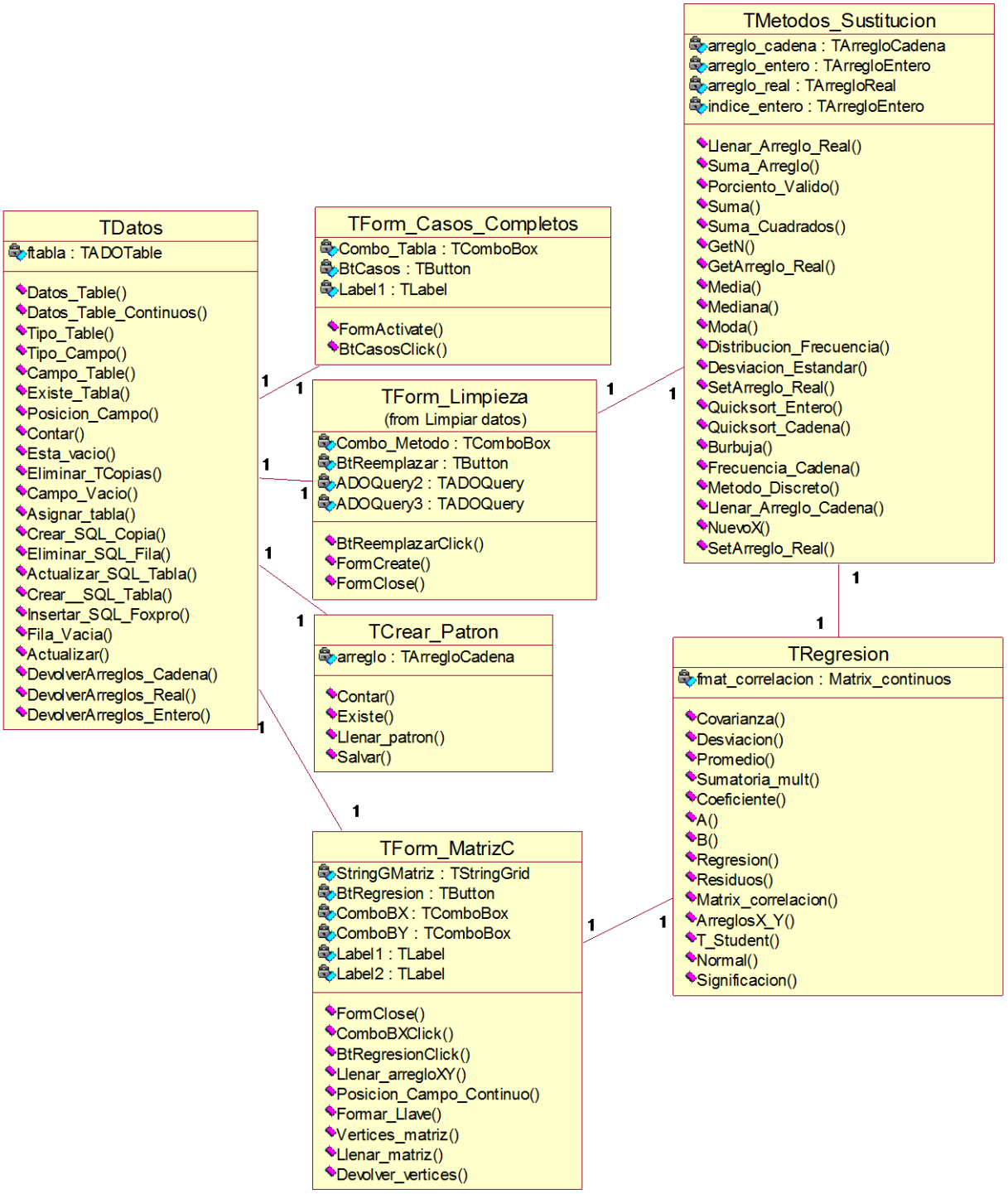

Figura 2.8 *Diagrama de Clases.*

Las clases de la aplicación desarrollada, juntamente con sus atributos y métodos son mostradas en la figura 2.8.

La herramienta de trabajo diseñada prepara la información, de forma tal que al leer de una determinada base de datos un campo, los contenidos de éste se guardan en arreglos en dependencia del tipo de dato, así, si el campo es de tipo entero, el arreglo de los elementos sería TArregloEntero, si es real, TArregloReal, si es cadena, TArregloCadena y si es cualquier otro tipo de dato como lo es la fecha, la hora, se guarda entonces en el arreglo TArregloField. También se usaron otros tipos de datos que pertenecen al conjunto de TFieldType.

TipoEntero = set of TFieldType; → Tipo de dato integer.
TipoReal = set of TFieldType; → Tipo de dato real.
TipoCadena =set of TFieldType ; → Tipo de dato string.
TipoFecha = set of TFieldType; → Tipo de dato ftDateTime, ftTime, ftDate.
TipoDinero = set of TFieldType; → Tipo de dato ftCurrency, ftBCD, ftFMTBcd.
TipoIndeterminado = set of TFieldType; → Tipo de dato ftUnknown, ftVariant.
Matrix=array of array of string; → Matriz de cadena.
Matrix_Regresion = array of array of array of real; → Arreglo tridimensional de reales
Matrix_Continuos = array of array of real; → Matriz de reales.

Las clases fundamentales que conforman el sistema junto con sus principales atributos y métodos, se relacionan a continuación:

1. TDatos

Clase que se encarga del manejo de las conexiones y todos los datos de la base de dato trabajando directamente desde el ADOTable.

Los atributos de esta clase son:

- ftabla: TADOTable; →A través de este atributo se acceden a los datos de las bases de datos.

Los métodos de esta clase:

- procedure Eliminar_TCopias (var Lista_Tablas: TStrings);→ Este método se utiliza

para eliminar de un ComboBox el nombre de las copias de las tablas.

- procedure Campo_Vacio (campo: String; var AVaciosIndex: TArregloEntero; var ANOVaciosValues: TArregloEntero); → Este método devuelve los índices de los campos que tienen al menos un valor nulo y los valores de los campos que no son nulos.
- function Datos_Table: Matrix;→ Este método devuelve todos los valores de la tabla seleccionada.
- function Datos_Table_Continuos (arreglo: TArregloCadena): Matrix_Continuos; → Este método devuelve dado un arreglo de campos continuos los valores que estos tienen en la tabla.
- function Datos_Table_ind (i, j: integer): Matrix_Continuos; →Este método devuelve los valores de los campos **i** y **j** en una matriz.
- function Tipo_Table: TArregloField; → Este método devuelve los tipos de todos los datos que posee una determinada tabla.
- function Tipo_Campo_Posicion (campo: string; i: integer): TFieldtype; → Este método devuelve dado el nombre de un campo y la posición que ocupa en la tabla de qué tipo es este campo.
- function Campo_Table: TArregloCadena; → Este método devuelve el nombre de los campos de una determinada tabla.
- function Posicion_Campo (nombre: string): integer; → Este método devuelve dado el nombre del campo, la posición que este ocupa en la tabla.
- procedure Asignar_tabla (nombre: string); → Este método asigna un nombre de Tabla al objeto ADOTable.
- procedure Crear_SQL_Copia (nombretabla: string; cadena: string); →Este método crea una copia de la tabla seleccionada.
- procedure Eliminar_SQL_Fila (cadena, nombretabla: string); → Este método elimina las filas donde al menos haya un valor nulo.
- procedure Actualizar_SQL_Tabla (nombretabla, campo, llave: string; valor: string); →Este método actualiza la fila del campo seleccionado donde se cumpla la llave.
- procedure Actualizar_SQL_Tabla_Completa (nombretabla, campo, valor: string);

→Este método actualiza la tabla completa en un campo determinado, imputa un valor siempre que este esté nulo.

- procedure Crear_SQL_Tabla (camino, campos: string); →Este método crea una tabla. Se usa específicamente para crear una copia de una tabla en FoxPro.
- procedure Insertar_SQL_FoxPro (camino, campos, valores: string); →Este método inserta un registro completo en una tabla creada. Se usa específicamente para FoxPro.
- function Existe_Tabla (nom_tabla: string): boolean; → Este método devuelve dado el nombre de una tabla si esta existe o no.
- function Esta_vacio (cadena:string): boolean; → Este método devuelve dado una cadena si está vacía completamente o no.
- function Contar: integer; → Este método cuenta la cantidad de filas de la tabla, es decir la cantidad de elementos que esta posee.
- function Fila_Vacia: matrix → Este método es el encargado de guardar en la matriz los elementos de cada fila que tenga al menos un nulo.
- procedure DevolverArreglos_Cadena (var VaciosIndex: TArregloEntero; var ANOVaciosValues:TArregloCadena);
- procedure DevolverArreglos_Real (var AVaciosIndex:TArregloEntero; var ANOVaciosValues:TArregloReal);
- procedure DevolverArreglos_Entero (var VaciosIndex: ArregloEntero; var ANOVaciosValues: TArregloEntero); → Los métodos anteriores llenan un arreglo con los valores de los campos que no tienen nulos y otro arreglo, de acuerdo a la naturaleza del dato, con los índices de los elementos que si lo están.
- function Actualizar: TArregloCadena; → Este método se utiliza para formar una llave compuesta para la consulta en SQL cuando hay que imputar un valor a la vez.

2. TMetodos_Sustitucion

Clase en la que aparecen los métodos de limpieza tales como la moda, media, mediana, Desviación estándar, método probabilístico basado en la distribución de los datos no perdidos.

Los atributos de esta clase son:

- arreglo_cadena: TArregloCadena;
- arreglo_entero: TArregloEntero;
- arreglo_real: TArregloReal; → Estos arreglos son los valores del campo seleccionado.
- indice_entero: TArregloEntero;→ Índices de los elementos que están nulos en el campo seleccionado.

Los métodos de esta clase:

- procedure Quicksort_Entero(inicio, fin: integer; var arreglo_original : TArregloEntero);
- procedure Quicksort_Cadena (inicio, fin: integer; var arreglo_original: TArregloCadena);
- procedure Burbuja (var arreglo_contador_entero:TArregloEntero; var arreglo_elementos_cadena: TArregloCadena); → Estos métodos son usados para ordenar los arreglos.
- procedure Frecuencia_Cadena (var arreglo_indice_entero: TArregloEntero; var arreglo_aux_cadena: TArregloCadena); → Es utilizado en el método probabilístico basado en la distribución de los datos no perdidos (Método Discreto) para calcular la frecuencia de los valores válidos.
- procedure Generar_Indices_Aleatorios (cant_aleat: integer; var arreglo_actual: TArregloEntero; var arr_indice:TArregloEntero); → Método que se usa para trabajar con el Método Discreto y que genera aleatoriamente los índices de los números que se van a reemplazar.
- function Llenar_Arreglo_Real (aux: real): TArregloReal;
- procedure Llenar_Arreglo_Cadena (var arreglo_aux: TArregloCadena; aux: string; arreglo: TArregloEntero); → Métodos que se encargan de llenar los arreglos según el tipo de dato.
- function Suma_Arreglo (arreglo: TArregloEntero): integer; → Calcula la suma de los elementos de un arreglo de enteros.

- function Porciento_Valido(total, parte: real): real; → Calcula el porciento válido que es usado en el Método Discreto.
- function Suma: real; → Calcula la suma de los elementos de un arreglo de números reales.
- function Suma_Cuadrados: real; → Calcula la suma de los cuadrados de los elementos de un arreglo de reales.
- function GetN: integer;
- procedure SetArreglo_Real (i: integer;x: real);
- function GetArreglo_Real (i: integer): real; → Métodos de cálculos auxiliares que se utilizan en otros métodos.

→ Métodos de sustitución

- function Media: real; → Calcula la media y devuelve un único elemento por el cual se sustituirán todos los valores nulos.
- function Mediana: integer; → Calcula la mediana y devuelve un único elemento por el cual se sustituirán todos los valores nulos.
- function Moda: string; → Calcula la moda y devuelve un único elemento por el cual se sustituirán todos los valores nulos.
- function Desviacion_Estandar: real; → Calcula la desviación estándar de un grupo de números reales.
- procedure Nuevos (var x1, x2: real); → Nuevo valor por el cual se sustituye un valor nulo para mantener la desviación estándar.
- function Metodo_Discreto: TArregloCadena; → Calcula varios elementos por los cuales se pueden sustituir los valores nulos para campos discretos.

3. TRegresion

Clase en la que es calculada la ecuación de regresión, la covarianza, el coeficiente de correlación, la significación a través del test de Students y los coeficientes de la ecuación de regresión.

Los atributos de esta clase son:

- fmat_correlacion: Matrix_continuos; → Este atributo es un arreglo bidimensional que se utiliza para almacenar los valores de la matriz de correlación.

Los métodos de esta clase:

- function Covarianza (arreglox, arregloy: TArregloReal): real; → Este método calcula la covarianza de los conjuntos *x* e *y*.
- function desviación (arreglo: TArregloReal): real; → Este método calcula la desviación estándar de un conjunto de valores.
- function Promedio (arreglo: TArregloReal): real; → Este método calcula la media de un conjunto de valores.
- function Sumatoria_mult(arreglox, arregloy: TArregloReal): real; → Este método se utiliza para hacer un cálculo auxiliar de hallar la sumatoria de las multiplicaciones de x_i e y_i.
- function Coeficiente (arreglox, arregloy: TArregloReal): real; → Este método calcula el coeficiente de correlación de Pearson *r*.
- function A(arreglox, arregloy: TArregloReal): real; → Este método calcula el término *a* de la recta de regresión lineal simple.
- function B(arreglox, arregloy: TArregloReal): real; → Este método calcula el término *b* de la recta de regresión lineal simple.
- procedure ArreglosX_Y(var arreglox, arregloy: TArregloReal; i, j: integer); → Este método se encarga de llenar los valores correspondientes de *x* e *y* según las variables que se escogieron para aplicarles el método de la regresión.
- function Regresion(X, Beta, Beta0: real): real; → Este método calcula el valor de la variable dependiente *y*, sustituyendo en la recta de regresión.
- function Residuos(arreglox, arregloy: TArregloReal): TArregloReal; → Este método calcula los residuos.
- function Matrix_correlacion: Matrix_Regresion; → Este método calcula la matriz de correlación formada por *r* y su significación.
- function Normal (x: real; var y: real): real; →Este método devuelve la significación para un valor dado mostrando si este es normal o no.

- function Significacion_T_Student(t, n: real): real; → Este método devuelve la significación a través de una prueba de t-Student para formar la matriz de correlación.
- function T_Student(r: real; n: integer): real; → Este método devuelve el valor de t.

4. TCrear_Patron

Clase encargada de crear el patrón y salvarlo en un fichero texto (.txt).

Los atributos de esta clase son:

- arreglo:TArregloCadena; → Este atributo se declara con el objetivo de poder utilizar un arreglo de cadenas en los métodos de la clase.

Los métodos de esta clase:

- function Contar(patron:string):integer; → Este método devuelve dado un determinado patrón la cantidad de veces que este se repite.
- function Existe(prueba:TArregloCadena;patron:string):boolean; → Dado un conjunto de patrones y un patrón en específico devuelve si este existe en el conjunto de patrones o no.
- procedure Llenar_patron; → Este método se encarga de formar el conjunto de patrones de una tabla.
- procedure Salvar(camino:string); → Este método se encarga de salvar los patrones para un fichero que se encuentra en el camino especificado.

Clases Visuales que consideramos necesarias para el diagrama de clases:

5. TForm_Limpieza

Esta clase visual es la encargada de utilizar los métodos de sustitución de valores nulos en una BD

Los atributos de esta clase son:

- Combo_Metodo: TComboBox ; → Este se llena con los métodos que se pueden usar para la sustitución de los valores del campo elegido.
- BtReemplazar: TButton; → En los eventos de este botón es donde se utilizan los métodos de sustitución de nulos según el método seleccionado en el ComboBox.

Los métodos de esta clase:

- procedure BtReemplazarClick(Sender: TObject); → En este método es donde se utilizan los métodos de sustitución y se reemplazan los valores nulos en la BD según la selección realizada en el ComboBox.
- procedure FormCreate(Sender: TObject);
- procedure FormClose(Sender: TObject; var Action: TCloseAction);

6. TForm_MatrizC

Esta clase es la encargada de utilizar y trabajar con todo lo relacionado con la matriz de correlación, usa además el método de regresión para la sustitución de valores nulos.

Los atributos de esta clase son:

- StringGMatriz: TStringGrid; → En el StringGrid se mostrará la matriz de correlación, es decir el valor de *r* y el valor de su significación.
- BtRegresion: TButton; → En este botón es donde se utiliza la recta de regresión para calcular el valor que sustituirá un valor nulo.
- ComboBX: TComboBox; → Este ComboBox se llena con el nombre de las variables o de los campos continuos para que se seleccione la variable que puede ser tomada como independiente.
- ComboBY: TComboBox; → Este ComboBox se llena con el nombre de las variables o de los campos continuos para que se seleccione la variable que puede ser tomada como dependiente, nunca se incluirá en esta lista la variable que se seleccionó como dependiente .

Los métodos de esta clase:

- procedure ComboBXClick(Sender: TObject); → Este método llena el ComboBY sin incluir en la lista de las variables aquella que fue tomada como independiente.
- procedure BtRegresionClick(Sender: TObject); → Este método calcula el valor a sustituir en los campos nulos usando para ello la recta de regresión calculada.
- procedure Llenar_arregloXY(var arreglo_real: TArregloReal; var arreglo_ind: TArregloEntero; cadena :string); → Este método devuelve el conjunto de valores y los índices de un campo determinado.

- function Posicion_Campo_Continuo(cadena: string; arreglo: TArregloCadena): integer; → Este método dado el nombre del campo y la lista de todos los campos continuos devuelve la posición de este en dicha lista.
- procedure Vertices_matriz(arreglo: TArregloCadena); → Este método llena en el StringGrid los nombres de los encabezados con el nombre de los campos continuos .
- procedure Llenar_matriz (arreglo: Matrix_Regresion); → Este método se encarga de llenar la matriz de correlación.

7. TForm_Casos_Completos

Los atributos de esta clase son:

- Combo_Tabla: TComboBox; → En este ComboBox se muestra la lista de las tablas de la BD seleccionada.
- BtCasos: TButton; → En este botón es donde se realiza la limpieza de los datos por el método de Casos Completos.

Los métodos de esta clase:

- procedure FormActivate(Sender: TObject); → Se llena el ComboBox con la lista de las tablas de la BD seleccionada .
- procedure BtCasosClick(Sender: TObject); → Se realiza el método de Análisis de Casos Completos donde se eliminan todos los registros de la tabla que contengan un valor nulo.

2.1.4. Tabla de eventos.

La Tabla de eventos muestra la secuencia de acciones que pueden causar la transición del objeto de un estado a otro en cada una de las ventanas de la aplicación y la respuesta que recibe del sistema dicho Analista de Datos al ejecutar cada una de estas acciones.

Ventana Principal:

¿Qué hace el actor?	¿Qué hace el sistema?
1. Seleccionar en el menú la opción Conexión a la BD.	1. Abrir la ventana de conexión a la BD.

2. Seleccionar en el menú la Ayuda.	2. Abrir la Ayuda de la aplicación.
3. Seleccionar en el menú la opción Acerca de …	3. Muestra la ventana de los Créditos.
4. Seleccionar en el menú la opción Salir.	4. Cerrar la aplicación

Menú Conexión

Ventana Conexión a la BD:

¿Qué hace el actor?	¿Qué hace el sistema?
1. Presionar el botón Construir.	1. Mostrar la ventana para elegir el proveedor para la construcción de la cadena de conexión.
2. Presionar el botón Aceptar.	2. Chequear si ha sido construida la cadena de conexión, en caso de haber sido construida muestra ventana principal, sino muestra un mensaje de error.
3. Presionar botón Cancelar.	3. Borra la cadena de conexión, si esta había sido construida previamente, y regresa a la ventana principal.

Ventana Propiedades de vínculo de datos:

¿Qué hace el actor?	¿Qué hace el sistema?
1. Seleccionar un proveedor de la lista de proveedores.	1. Cambiar el focus hacia su elección.
2. Presionar el botón Siguiente, dar doble click sobre el proveedor seleccionado o cambiarse manualmente para la paleta conexión.	2. Cambiarse hacia la paleta Conexión.
3. Presionar botón Aceptar estando activa la paleta Proveedores.	3. Verifica si todos los datos de la cadena de conexión fueron entrados

	completamente.
4. Presionar botón Abrir.	4. Llama al componente OpenDialog y muestra la ventana de Abrir.
5. Presionar botón Aceptar estando activa la paleta Conexión.	5. Verifica si todos los datos de la cadena de conexión fueron entrados completamente y si es así muestra la ventana Conexión a la BD.
6. Presionar botón Cancelar.	6. Muestra la ventana Conexión sin guardar los datos anteriores dando la posibilidad de formar otra cadena de conexión.

Menú Limpieza de Datos

Ventana Análisis de Casos Completos

¿Qué hace el actor?	¿Qué hace el sistema?
1. Elegir el nombre de una tabla y presionar el botón Eliminar Nulos.	1. Crea una copia de la tabla seleccionada sin los registros que contengan un valor nulo.
2. Cerrar la ventana.	2. Se muestra nuevamente la ventana Principal.

Ventana Métodos de Regresión:

¿Qué hace el actor?	¿Qué hace el sistema?
1. Seleccionar una tabla	1. Muestra el conjunto de todos los campos de esta tabla en un ListBox.
2. Seleccionar de la lista de todos los campos de la tabla aquellos campos que sean continuos y adicionarlos a un nuevo ListBox, luego presionar el botón Formar Matriz.	2. Muestra la ventana Matriz de Correlación

3. Cerrar la ventana.	3. Se muestra la ventana Principal.

Ventana Matriz de Correlación:

¿Qué hace el actor?	¿Qué hace el sistema?
1. Seleccionar la variable independiente y dependiente que formaran la recta de regresión lineal simple y presionar el botón Regresión para realizar el reemplazo usando este método.	1. Reemplazar en la tabla los valores nulos según el valor calculado por la recta de regresión.
2. Cerrar la ventana.	2. Vuelve a la ventana Principal.

Ventana Selección de Datos:

¿Qué hace el actor?	¿Qué hace el sistema?
1. Elegir una tabla de la lista.	1. Mostrar los datos de dicha tabla en un DBGrid y llena el otro ComboBox con los campos de esta tabla.
2. Elegir un campo de la tabla seleccionada.	2. Chequear si el campo seleccionado no tiene valores nulos o se encuentra completamente nulo (no tiene ningún dato), emite mensaje de error, de lo contrario, habilita el botón de Seleccionar Método.
3. Presionar botón Cancelar.	3. Vuelve a la ventana Principal, cancelando los cambios efectuados y haciendo no visible la opción de Limpieza de Datos.

Ventana Métodos de Reemplazo:

¿Qué hace el actor?	¿Qué hace el sistema?
1. Seleccionar método de reemplazo y presionar botón Reemplazar Nulos.	1. Verifica que se haya seleccionado un método de reemplazo y hace una llamada al procedimiento que da solución a este método.
2. Cerrar la ventana.	2. Vuelve a la ventana Selección de Datos.

Ventana Formar Patrones

¿Qué hace el actor?	¿Qué hace el sistema?
1. Elegir el nombre de una tabla y presionar el botón Formar Patrones	1. Salvar los patrones de la tabla en un fichero texto.
2. Cerrar la ventana.	2. Se muestra nuevamente la ventana Principal.

2.2. Herramientas computacionales utilizadas en el sistema.

El sistema fue desarrollado en el ambiente de programación Borland Delphi creándose una aplicación para Windows en lenguaje Pascal y haciendo uso de las facilidades que este brinda en la Programación Orientada a Objeto. Además se utilizó el lenguaje SQL para algunas consultas y actualizaciones en las tablas de las bases de datos en distintos gestores: SQL Server, Access y FoxPro.

2.3. Fórmulas y conceptos estadísticos utilizados.

Para el reemplazo de nulos en las Bases de Datos se utilizaron numerosos métodos estadísticos y sus respectivas fórmulas, manteniendo en cada caso el estadígrafo según el método escogido y el tipo de dato.

2.3.1. Medidas de tendencia central y de dispersión.

Otros de los conceptos estadísticos que se usan son los estadígrafos de tendencia central y medidas de dispersión.

Existen cuatro tipos de estadígrafos: los de posición, los de dispersión, los de apuntamiento y los de deformación. En la aplicación solo usamos los dos primeros ya que no fue necesaria la utilización de los demás.

Dentro del conjunto de estadígrafos de posición se distinguen dos tipos: los estadísticos de tendencia central y los de localización. Los de primer tipo brindan, de alguna forma, información sobre el centro de distribución, y los del segundo, señalan la localización de los valores extremos o valores más frecuentes (Guerra, 1987:33). En particular, se usó la media y la mediana como estadígrafos de posición del tipo de tendencia central y la moda, como estadígrafo de posición del tipo de localización. En relación con los estadígrafos de dispersión, son muy utilizados aquellos que indican la concentración de los valores del conjunto alrededor de su valor medio o promedio. El más importante de ellos es la varianza. Existen otros asociados con ella tal como la desviación típica, cuya fórmula y concepto fue de gran importancia para uno de los métodos de reemplazo implementado.

- **Media.**

  La media aritmética (o promedio aritmético) *(M),* es el punto de equilibrio del conjunto de valores. Con bastante generalidad se utiliza el símbolo $\bar{x}$ para denotarla. Esta se define como:

$$M = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + ... + x_n}{n}$$ (Guerra, 1987:37)

  Variables a las que se puede aplicar: Discretas y continuas.

- **Moda.**

  La moda *(M_o)*, es el valor más frecuente o el que más se repite en un conjunto de *n* valores, o sea, el valor que tiene la propiedad de poseer una frecuencia mayor (absoluta o relativa) (Guerra, 1987:36). La moda puede existir o no, incluso puede haber más de una moda en un mismo conjunto, por tanto no es única.

  Ejemplos:

Nota1:	3	3	3	4	5	Mo=3
Nota2:	3	2	4	5		No hay moda
Nota3:	3	3	4	4	5	Mo=3, Mo=4

  Variables a las que se puede aplicar: Discretas, tanto nominales como ordinales.

- **Mediana.**

  La mediana (M_e), se define como aquel valor que no es superado ni supera a más de la mitad de las observaciones del conjunto de valores. Es considerado el valor que divide a la muestra en dos partes iguales y representa el valor central del conjunto de observaciones. No tiene que pertenecer necesariamente al conjunto. En el cálculo de esta hay que distinguir dos situaciones atendiendo a que la cantidad de observaciones sea par o impar. Si *n* es impar basta con ordenar las observaciones $x_{(1)}, x_{(2)}\ldots\ldots x_{(n)}$ y tomar como valor de la mediana el valor central, es decir, $x_{(n+1)/2}$. Esto es válido tanto para variables continuas como para variables discretas. Si *n* es par, después de ordenada la muestra existen dos valores centrales, pudiéndose tomar como mediana cualquiera de ellos, o cualquiera entre ellos dos. Se acostumbra a tomar la semisuma de ambos valores centrales, o sea:

$$M_e = \frac{x_{(n/2)} + x_{(n+2)/2}}{2}$$ (Guerra, 1987:34)

Ejemplos:

Nota1: 3	3	4	5	5	Me=4
Nota2:	3	3	4	4	Me= (3+4)/2

Variables a las que se puede aplicar: Discretas y continuas.

- **Desviación típica.**

La desviación típica de un conjunto de valores es la medida de variabilidad que se define como la raíz cuadrada positiva de la varianza, por lo que siempre es un número no negativo (Guerra, 1987:50). Se calcula por la siguiente fórmula:

$$s = +\sqrt{V(x)} \quad s \geq 0 \quad \text{donde } V(x) = \frac{i}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \text{ es la varianza.}$$

Despejando y sustituyendo en la fórmula de la desviación se obtiene un nuevo valor de *x* que es utilizado en el reemplazo de los valores nulos en la base de datos.

La fórmula es la siguiente:

$$X_{n+1} = \frac{\frac{2*\sum_{i=1}^{n} x_i}{(n+1)^2} \pm \sqrt{\frac{4*\left(\sum_{i=1}^{n} x_i\right)^2}{(n+1)^4} - \frac{4*n}{(n+1)^2}} * \left(\frac{\sum_{i=1}^{n} x_i^{\,2}}{(n+1)} + \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{(n+1)^2} + S^2\right)}{\frac{2*n}{(n+1)^2}}$$

Variables a las que se puede aplicar: Continuas.

2.3.2. Regresión lineal simple.

En el estudio de la relación funcional entre dos variables poblacionales, una variable *x*, llamada independiente, explicativa o de predicción y una variable *y*, llamada dependiente o variable respuesta, presenta la siguiente notación:

$$y = a + b * x + e$$

Donde:

a es el valor de la ordenada donde la línea de regresión se intercepta con el eje Y

b es el coeficiente de regresión poblacional (pendiente de la línea recta)

e es el error.

Se denomina regresión lineal cuando la función es lineal, es decir, requiere la determinación de dos parámetros: la pendiente y la ordenada en el origen de la recta de regresión (Franco García, 2000), se dice entonces que *y* depende linealmente de *x*. La regresión nos permite además, determinar el grado de dependencia de las series de valores *x* e *y*, prediciendo el valor *y* estimado que se obtendría para un valor *x* que no esté en la distribución. Este método solo es aplicable a variables continuas.

Suposiciones de la regresión lineal.

1. Los valores de la variable independiente *x* son fijos, medidos sin error.
2. La variable *y* es aleatoria.
3. Para cada valor de *x*, existe una distribución normal de valores de *y* (subpoblaciones *y*).
4. Las varianzas de las subpoblaciones *y* son todas iguales.
5. Todas las medias de las subpoblaciones de *y* están sobre la recta.
6. Los valores de *y* están normalmente distribuidos y son estadísticamente independientes. (Monografía, 1997).

Coeficiente de correlación

El coeficiente de correlación es otra técnica de estudiar la distribución bidimensional, que nos indica la intensidad o grado de dependencia entre las variables *x* e *y*. El coeficiente de correlación *r* es un número que se obtiene mediante la fórmula.

$$r = \frac{S_{XY}}{S_X * S_Y}$$

Este puede estar comprendido entre -1 y +1.

- Cuando *r=1*, la correlación lineal es perfecta, directa.
- Cuando *r=-1*, la correlación lineal es perfecta, inversa

- Cuando *r=0*, no existe correlación alguna, independencia total de los valores *x* e *y*.

Puede demostrarse que el coeficiente de correlación es una magnitud que siempre estará en el intervalo (-1,1). Si r = ±1, todos los puntos de la muestra se encuentran exactamente sobre una recta. Si $r > 0$ la recta es creciente, si $r < 0$, la recta es decreciente y en la medida en que el valor absoluto de *r* se acerque a 1, la dispersión de los valores de *y* alrededor de las medias condicionales para cada *x* es menor. De esta forma el coeficiente de correlación nos da una medida de la fortaleza del enlace lineal entre *x* e *y*. El hecho de que *r* sea 0 ó cercano a este valor significa que no hay dependencia lineal entre las variables, pero no significa que no haya cualquier otro tipo de dependencia.

Este valor de *r* se usa en la matriz de correlación y se calcula su significación mediante un test de Student con *n-2* grados de libertad por la fórmula:

$$t = \frac{r * \sqrt{n-2}}{\sqrt{1-r^2}}$$

El problema fundamental radica en encontrar aquella recta que mejor ajuste a los datos. Tradicionalmente se ha recurrido para ello al método de mínimos cuadrados, que elige como recta de regresión a aquella que minimiza las distancias verticales de las observaciones a la recta (Pértiga y Pita, 2001). Más concretamente, se pretende encontrar *a* y *b* tal que se cumpla que:

$$Min \sum_{i=1}^{n} \left(y_i - a - bx_i \right)^2$$

Donde

$$a = \bar{x} - b * \bar{y} \qquad \text{y} \qquad b = \frac{S_{XY}}{S_Y^2}$$

Al calcular los valores respectivos de *a* y *b* se tiene la recta de regresión,

$$y = a + b * x$$

en la cual se puede sustituir el valor de la variable conocida *x* , y se obtiene el valor por el cual sustituir a la variable *y*.

2.4. Apreciaciones del capítulo.

En este capítulo se realizó el análisis y diseño de la herramienta implementada, y fueron descritos los conceptos y las fórmulas fundamentales utilizadas en el desarrollo de la misma. En el siguiente capítulo se muestra un manual de usuario que sirve de guía para proceder en la interfaz visual de la aplicación.

CAPÍTULO 3. MANUAL DE USUARIO.

Este capítulo está dedicado a la descripción del manual de usuario de la herramienta, mostrando la información de cómo proceder en la interfaz visual y hacer las acciones deseadas.

3.1. Requerimientos del software.

El sistema está implementado en Delphi y realiza la conexión con las bases de datos (SQL Server, Access, FoxPro) utilizando la tecnología ADO[8] y ODBC[9]; por lo que para examinar una base de datos determinada se requiere el driver ODBC para dicho gestor y construir una conexión con esos datos.

Un aspecto a tener en cuenta antes de manipular cualquier tabla de FoxPro (.dbf) es que se tiene que crear primeramente el ODBC correspondiente al driver de FoxPro, y para este ser creado correctamente, en el directorio 'c:\windows\system32' debe existir el archivo 'vfpodbc.dll' (versión 6.0 que tiene un tamaño de 912 KB).

3.2. Descripción de la herramienta y sus funcionalidades.

La herramienta se ha concebido para el reemplazo de nulos en diversos tipos de Bases de Datos: SQL Server, Access y FoxPro; y la creación del patrón de datos ausentes de una tabla. Para la implementación de la sustitución de los valores nulos en las **B**ases de **D**atos (BD) de tipo Access y SQL Server se usó la componente ADO que brinda Delphi, mientras que para analizar las BD de tipo FoxPro usamos ODBC por problemas de compatibilidad de ADO con algunos ficheros (.dbf).

Los elementos nulos son sustituidos por un(os) valor(es) en una copia que se realiza de la tabla seleccionada por el analista de datos y es(son) hallado(s) a través de uno de los métodos

[8] Siglas en inglés de ActiveX Data Objects

implementados: Casos Completos, Regresión, Desviación Estándar, Media, Mediana, Moda, Probabilístico Basado en la Distribución de los Datos no Perdidos (Método Discreto); que según el tipo de dato pueden escogerse.

3.3. Ambiente de trabajo

El ambiente de trabajo de la herramienta implementada es muy sencillo, posibilitando un buen desenvolvimiento en la ejecución y utilización del mismo, y permitiendo a su vez que el usuario realice la limpieza de sus datos en un corto período de tiempo.

3.3.1. Conexión con la Base de Datos.

La ejecución se inicia con una ventana como la mostrada en la figura 3.1, que contiene el menú inicial de la aplicación. Como se muestra la única acción que es posible realizar es la conexión a la base de datos:

Figura 3.1 Ventana Principal del sistema

[9] Siglas en inglés de Open DataBase Connectivity

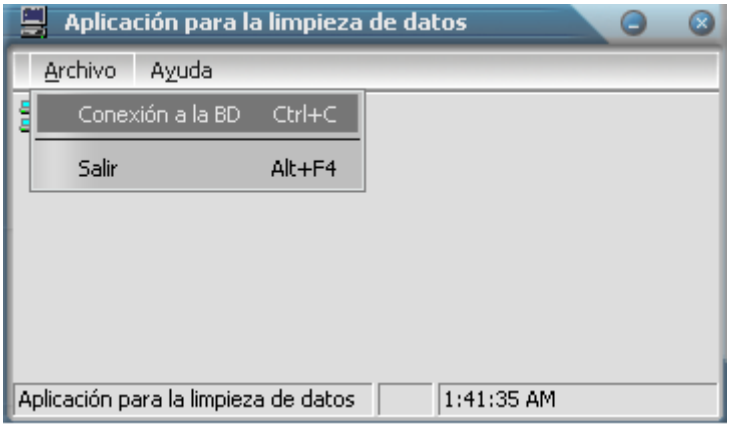

Figura 3.2 Ventana principal del sistema y menú conexión

Para conectarse a la BD (figura 3.2) es necesario construir la cadena de conexión (figura 3.3), para lo cual es requisito indispensable elegir un proveedor; que está en dependencia de si es Access, FoxPro o SQL Server (figura 3.4), y además la BD (figuras 3.5, 3.6, 3.7) en la que se hará posteriormente la limpieza de los datos (figura 3.8).

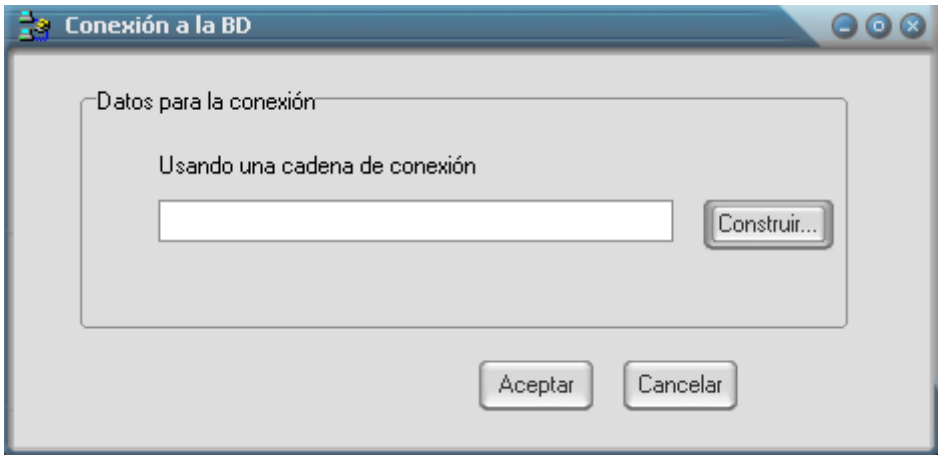

Figura 3.3 Construcción de la cadena de conexión

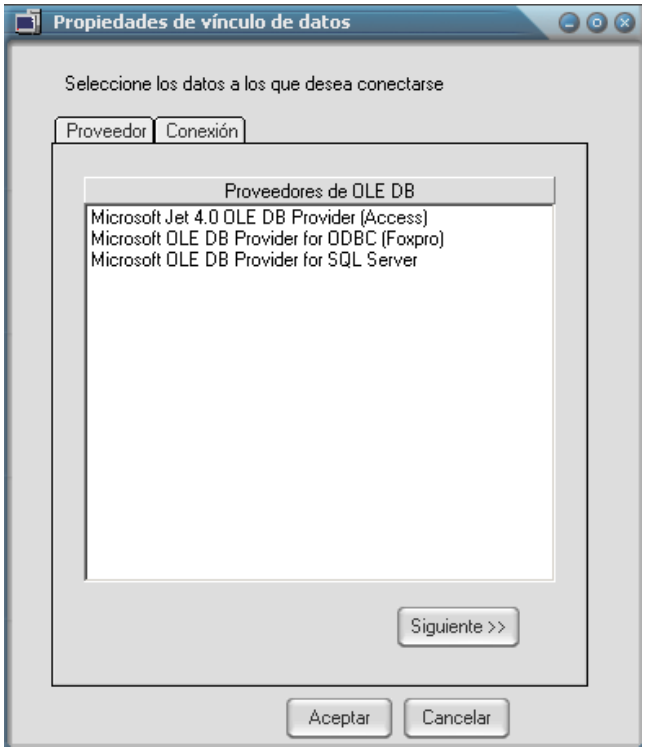

Figura 3.4 Selección del proveedor

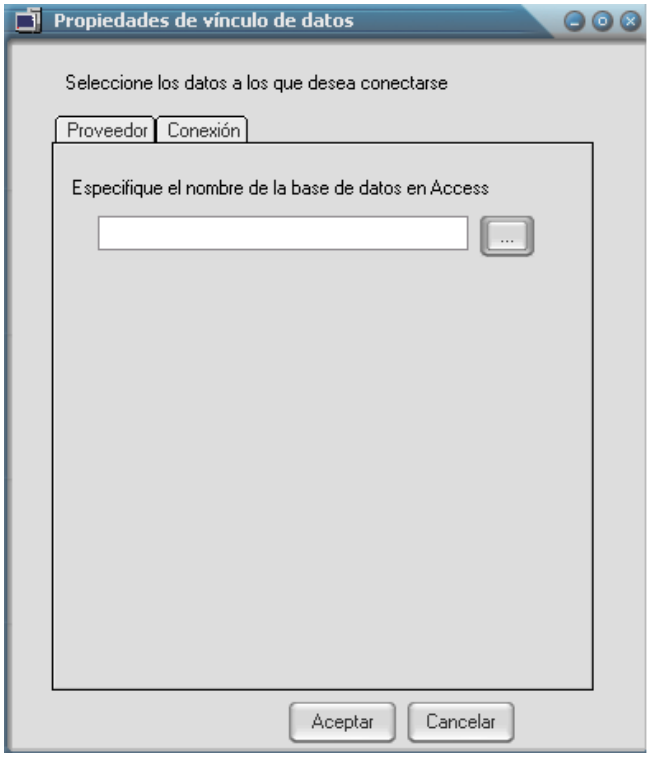

Figura 3.5 Selección de la BD en Access

Figura 3.6 Selección de la BD en FoxPro

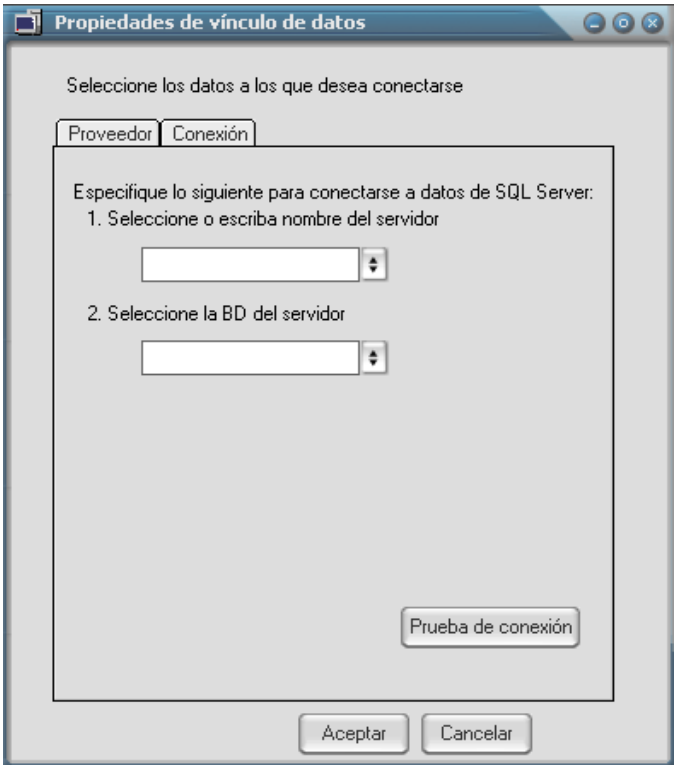

Figura 3.7 Selección de la BD en SQL Server

Es importante hacer notar que este diálogo es similar al empleado en **Delphi** para la creación de la cadena de conexión ADO.

3.3.2. Reemplazo de nulos.

Luego de realizar la conexión satisfactoriamente el menú inicial se amplía y aparece un submenú con los métodos (Casos Completos, Regresión, Métodos Estadísticos (media, moda, mediana, desviación estándar y método discreto)) que brinda la herramienta. El analista de datos selecciona uno a la vez y es entonces cuando ocurre el proceso de limpiar los datos.

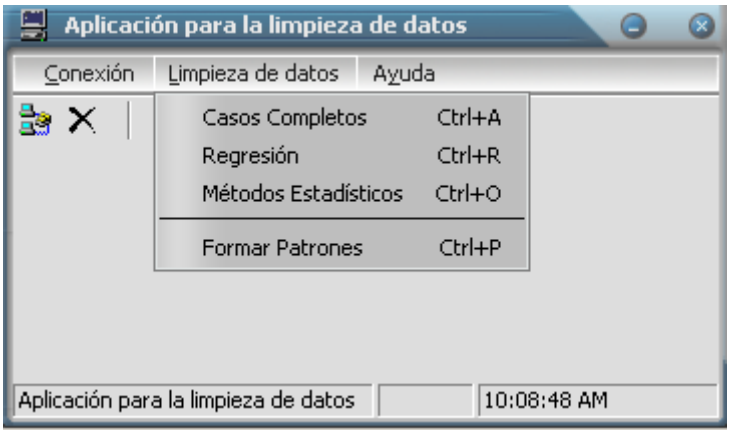

Figura 3.8 Ventana principal del sistema luego de la conexión a la base de datos

3.3.2.1. Casos Completos.

Para reemplazar los valores nulos mediante el método de Análisis de Casos Completos es necesario seleccionar la tabla (figura 3.9) y a partir de esta se crea la copia de la misma sin valores nulos.

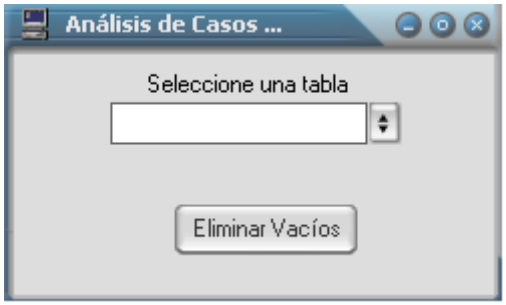

Figura 3.9 Análisis de Casos Completos

3.3.2.2. Regresión.

Para reemplazar los valores nulos a través del método de regresión es necesario seleccionar la tabla y los valores continuos (figura 3.10) para formar la matriz de correlación (figura 3.11) y a partir de esta el analista de datos selecciona la variable dependiente e independiente que forman parte de la recta de regresión y con esta ecuación hacer la imputación.

Figura 3.10 Método de regresión

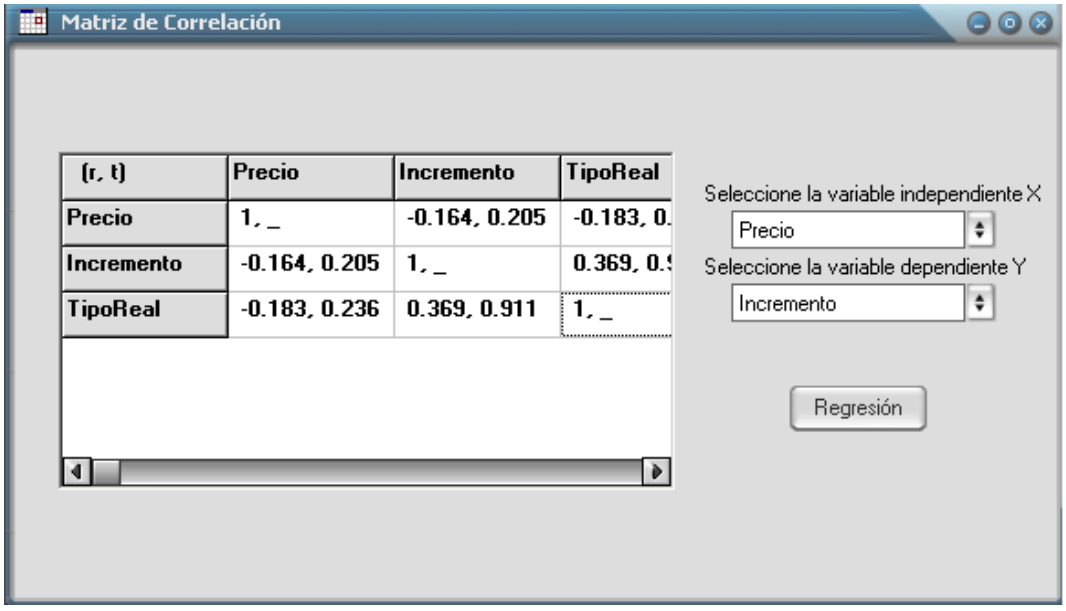

Figura 3.11 Matriz de correlación

3.3.2.3. Métodos Estadísticos.

Para reemplazar los valores nulos usando otros métodos es necesario seleccionar la tabla y el campo al que se le va a realizar la imputación (figura 3.12). Los métodos de limpieza: Moda, Mediana, Media, Desviación Estándar, Probabilístico Basado en la Distribución de los Datos no Perdidos (Método Discreto); dependen de la naturaleza del dato. Si el tipo de dato es:

- Entero: Mediana, Moda, Método Discreto (figura 3.13).
- Real: Media, Desviación Estándar (figura 3.14).
- Cadena: Moda, Método Discreto (figura 3.15).

Figura 3.12 Selección de datos

Figura 3.13Métodos para reemplazar valores enteros

Figura 3.14 Métodos para reemplazar valores reales

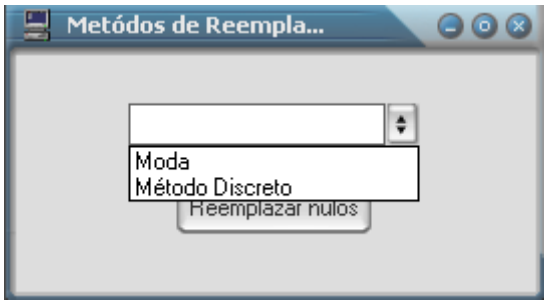

Figura 3.15 Métodos para reemplazar valores cadena

3.3.3. Formar Patrones.

Para formar los patrones y almacenarlos es necesario elegir la tabla (figura 3.16) desde la cual se va a formar el patrón, y el camino donde se ha de salvar el fichero **.txt** creado a partir de la información recopilada; luego puede ser utilizada esta información para un análisis de la tabla.

El patrón de nulos se considera un conjunto de vectores binarios definidos de la siguiente forma:

$\{(v_1, v_2,... v_n) \in Z_2 \mid v_i = 0$ si existe alguna tupla donde el atributo A_i sea nulo$\}$

n es el grado de la relación.

Para cada vector se almacena además su frecuencia de aparición.

Figura 3.16 Ventana Salvar Patrones

3.4. Áreas de aplicación

La herramienta puede ser usada en cualquier centro siempre y cuando tengan que trabajar con bases de datos en Access, FoxPro ó SQL que requieran sus datos de una limpieza, por ausencia de información en alguno de sus campos.

CONCLUSIONES

A partir del trabajo realizado se concluye que:

- Se realizó una revisión y recopilación de técnicas que en la literatura señalan la forma de predecir valores nulos.
- Se construyó la primera versión de una herramienta en que se utilizan algunas de las técnicas estudiadas.
- Se implementaron métodos de sustitución de nulos por valores de tendencia central como la moda, media, mediana, desviación estándar.
- Se implementó un método específicamente para los valores discretos llamado método probabilístico basado en la distribución de los datos no perdidos.
- Se comprobó la factibilidad de uso de la misma y aunque no se cuentan con datos estadísticos completos los primeros resultados aportan un balance positivo.

RECOMENDACIONES

- Cuando se manejen datos incompletos, hay que valorar previamente el uso de más de una alternativa para tratarlos y realizar un análisis de sensibilidad que permita una mejor elección del procedimiento a implementar.

REFERENCIAS BIBLIOGRÁFICAS

1. Cañizares, M.; Barroso, I. y K. Alfonso, (2003) *Datos incompletos: una mirada crítica para su manejo.* Instituto Nacional de Epidemiología y Microbiología (INHEM). La Habana. Cuba.

2. Ceruti, M. G. y M. N., Kamel, (1999) *Preprocessing and Integration of Data from Multiple Sources for Knowledge Discovery*. International Journal on Artificial Intelligence Tools

3. Date, C. J., (2003) *Introducción a los Sistemas de Bases de Datos*.

4. Franco García, Angel (2000) "Regresión". *Física con ordenador. Disponible en:* http://www.sc.ehu.es/sbweb/fisica/cursoJava/numerico/regresion/regresion.htm

5. Galhardas, H et al., (2000) *An Extensible Framework for Data Cleaning* In Proceedings of the International Conference on Data Engineering *(ICDE)*, San Diego, CA.

6. Garson, G. D., (2005) *Data Imputation for Missing Value.* Janvier 2005. Disponible en: http://www2.chass.ncsu.edu/garson/pa765/missing.htm

7. Guerra, Dra. C, (1987) Estadística.

8. Howell, David C., (1998) *Treatment of Missing Data.*

9. Information Technology Services (2004). "Handling Missing or Incomplete Data*"* at The University of Texas at Austin. Disponible en: http://www.utexas.edu/its/rc/answers/general/gen25.html

10. Inmon y Hakathorn, (2004) *Using the Data Warehouse.*

11. Jiménez, Dr. C., (2000) *Técnicas de Tratamientos de Datos Ausentes.*

12. Jonathan, I. y A.M. Maletic, (2002) *Automated Identification of Errors in Data Sets*. Memphis: The University of Memphis.

13. Little, R. J. y D. B. Rubin , (1987) *Statistical Analysis with Missing Data*. John Wiley and Sons, New York.

14. Mármol, D., (2005) *Determinación de una taxonomía de errores en los sistemas operacionales de nuestro entorno.* Tesis de Diploma. UCLV.

15. Martina, T., (2005) *Tratamiento de valores ausentes en las bases de datos.* Tesis de Diploma. Santa Clara, Facultad de Matemática-Física-Computación, Universidad Central Marta Abreu de Las Villas

16. McDermeit, M.; Funk, R. y M. Dennis, (1999) *Data cleaning and replacements of missing values.*

17. (2006) "Modelado de Sistemas con UML". *TLDP-ES/LuCAS.* Disponible en: http://es.tldp.org/Tutoriales/doc-modelado-sistemas-UML/multiple-html

18. Pértiga, S. y S. Pita, (2001) "Técnicas de regresión: Regresión Lineal Simple". *Fisterra.* Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Juan Canalejo. A. Coruña. Disponible en: http://www.fisterra.com/mbe/investiga/regre_lineal_simple/regre_lineal_simple2.pdf

19. Pyle, D., (1999) *Data Preparation for Data Mining.* Morgan Kaufmann Publishers, Inc. San Francisco, California.

20. Redman, T.C., (1992) *Data Quality: Management and Technology*, Bantam Books, New York.

21. Redman, T.C., (1998) *The Impact of Poor Data Quality on the Typical Enterprise*. Communications of the ACM. pp. 79-82.

22. (1997)"Regresión lineal simple".*Monografía.* Disponible en: http://www.monografias.com/trabajos27/regresion-simple/regresion-simple.shtml [2]

23. Rubin, D. B., (1987) *Multiple imputation for nonresponse in surveys*. New York: John Wiley.

24. Rubin, D. B., (1996) *Multiple Imputation after 18+year.* Journal of the American Statistical Association 91.

25. Sarle, W. S., (1998) *Prediction with Missing Inputs.* SAS Institute Inc., SAS Campus Drive, Cary, USA.

26. Shafter, J. L., (1997) *Multiple of incomplete multivariate data.* London: Chapman and Hall.

27. SmallWaters Corporation (2006) "Questions about Missing Data". *SmallWaters*. Disponible en http://www.smallwaters.com/amos/faq/faqa-missdat.html [Accesado el día 15 de marzo de 2006].

28. Stones Analystics, (2003) "Second Moment. The news and business resource for applied analystics". *Missing Data Method*. Disponible en: http://www.secondmoment.org/etal-column/index.php.

29. Universidad de Alberta (2006) "Tutorial de UML". *Department of computing science*. Disponible en: http://www.cs.ualberta.ca/~pfiguero/soo/uml/

30. Wand, Y. y R. Wang, (1996) *Anchoring Data Quality Dimensions Ontological Foundations.* Communications of ACM, 39.

BIBLIOGRAFÍA

1. Castilla, D. et la (2006) *Correlación*. Disponible en: http://www.uhu.es/44103/ficheros_datos/parcial2/spss2/correlacion.PDF

2. Hernández, M.A., (1995) The *merge/purge for large databases.* Proceedings of the ACM SIGMOD Conference.

3. (2006)"Lenguaje unificado de modelado". *Wikipedia.* Disponible en: http://es.wikipedia.org/wiki/Lenguaje_Unificado_de_Modelado

4. Mendoza, M. (2005) Estadística aplicada II. Análisis de regresión .Disponible en:http://cursos.itam.mx/mendoza/tem_ea_II.pdf

5. Pupo González, J. et al., (2004) *Análisis de regresión y series cronológicas.* Primera reimpresión. La Habana. Editorial Félix Varela.