Clase 2:
Consolidación de Bases de Datos

SEBASTIÁN MALDONADO

DIPLOMADO *BUSINESS INTELLIGENCE*

20 DE DICIEMBRE, 2011

DIAPOSITIVAS: CRISTIÁN BRAVO,
SEBASTIÁN MALDONADO

Nuestro Ejemplo Práctico

- Base de Datos 'Car Data'.
  - Entrega la percepción (de inaceptable a muy buena) para 1728 automóviles.
  - Posee información acerca de:
    - Precio: mantención y compra.
    - Comfort: Puertas, personas, espacio para equipaje, seguridad.
    - Percepción.
    - Accidentes: 3.000 accidentes registrados.
  - Objetivo: consolidar esta base de datos para uso apropiado en modelos de minería de datos.

Bases de Datos Relacionales: Lenguaje Consultas

- SINTAXIS SQL : 'Structured Query Language'

SELECT< Lista de atributos >

FROM< Lista de tablas >

WHERE< Condición >

Bases de Datos Relacionales: Lenguaje Consultas

- Recuperar todos los automóviles desde la tabla de percepciones.

```
SELECT ID
FROM PERCEPTION
```

- Recuperar los valores de todos los atributos de Comfort de los automóviles con seguridad baja ('low').

```
SELECT * FROM COMFORT
WHERE SAFETY = "low";
```

Bases de Datos Relacionales: Lenguaje Consultas

- Comando AND: Requiere ambas condiciones.

```
SELECT *
FROM COMFORT
WHERE SAFETY= "low" AND DOORS = "2";
```

- Comando OR: Requiere sólo una condición.

```
SELECT *
FROM PRICE
WHERE BUYING= "vhigh" OR MANT= "vhigh";
```

Información entre Tablas

- Si dos tablas están relacionadas a través de un campo, podemos extraer información de ambas.
- Operaciones:
  - Inner Join: Selecciona los campos que están en ambas tablas.
    - Ej: Id de los automóviles que han tenido un accidente y la última ocurrencia de cada uno.

```
SELECT ID, MAX(DATE_ACC) AS LAST_ACC
FROM PERCEPTION INNER JOIN ACCIDENT ON PERCEPTION.ID =
ACCIDENT.ID_CAR
GROUP BY PERCEPTION.ID
```

    - Consulta entrega sólo los automóviles que han tenido un accidente.

Información entre Tablas (2)

- Left/Right Join: Entrega todos los elementos de la tabla de la izquierda/derecha de la expresión y los elementos de la tabla de la derecha/izquierda sólo si estos están presentes.
  - Ej: Automóviles con su percepción y el total de accidentes que han tenido.

```
SELECT ID, PERCEPTION, COUNT(ID_CAR) AS N_ACC
FROM PERCEPTION LEFT JOIN ACCIDENT ON
PERCEPTION.ID = ACCIDENT.ID_CAR
GROUP BY PERCEPTION.ID, PERCEPTION
```

Bases para Modelos de Minería de Datos

- Modelos de minería de datos se componen de una tabla maestra.
  - Consolida TODA la información disponible.
  - Incorpora conocimiento del modelador.
  - Es una ÚNICA tabla.
  - Tabla DEBE cumplir con los requerimientos del modelo.
- Importante: Muestreo de la información disponible.
  - No se debe sesgar.
    - Casos únicos, a menos que se quiera sobredimensionar.

Fuentes de Datos

- 
  Bases de Datos Internas
  : Bases de datos internas de la entidad dónde se crea el modelo.

- 
  Fuentes Externas
  : Fuentes de información externa. Cualquier dato que se disponga, se crea útil y se pueda obtener en los períodos sucesivos.

- 
  Datos Generados
  : Datos generados. Toda variable construida partir de otras disponibles. Deben ser diseñadas teniendo en mente el fenómeno estudiado.

Consideraciones Archivo Maestro

- Tabla debe consistir en toda la información disponible.
  - Si tiene demasiados casos, eso se ve al momento de diseñar experimentos, NO al momento de diseñar la base de datos.
  - Se debe construir un archivo 'maestro' (en Excel por ej.) que permita considerar todas las variables que se utilizaron, sus fuentes y su manera de construirse.
- Variables a Considerar:
  - Provenientes de las fuentes de datos disponibles.
  - VARIABLE(S) OBJETIVO.
  - Variables construidas.

Ejemplo Archivo Maestro

- Nº: Nº de variable.
- Fuente: Desde qué fuente de datos (DW, fuente externa, etcétera) proviene.
- Tabla: Tabla de origen de la fuente.
- Nombre: Nombre variable.
- Procedencia: Interna/Externa/Generada.
- Tipo: Nominal/Ordinal/Continua.
- Descripción: Breve descripción de los datos que contiene.

Variables Internas

- Corresponden a las variables que están disponibles.
  - Suelen ser la gran mayoría.
  - Potencialmente, dieron origen al problema/proyecto.
- Consideraciones:
  - Disponibilidad.
    - ¿Existen cambios de bases de datos previsibles?
    - ¿Tendré esta variable los próximos dos años?
  - Credibilidad
    - ¿Es esta variable creíble?
    - ¿Está calculada correctamente?

Variables Externas

- Son variables que provienen de fuentes externas, independientes a la entidad que crea el modelo.
- Consideraciones.
  - Credibilidad Entidad.
    - ¿Son sus datos confiables?
  - Fechas.
    - ¿Cuándo fueron capturados estos datos?
  - Disponibilidad/Costo
    - ¿Cuánto cuesta obtener esta variable?
    - ¿Cuánto tiempo espero tenerla disponible?

Variables Generadas

- Variables construidas a partir de otras (Internas/Externas).
- Ejemplos:
  - Ratios: Deuda/Ingreso.
  - Transformaciones: Edad, Años Antigüedad, etc.
  - Transformaciones ad-hoc: 'Región' en 'Sector del País'.
- Consideraciones:
  - Crear variables que tengan sentido para el problema.
  - 'Ante la duda, constrúyanla'.

Construcción de Tablas Maestras

- La construcción debe hacerse considerando el tiempo dónde se obtienen los datos.
  - Ej: Si son solicitudes a la fecha X, entonces son variables disponibles EN X.
  - Error típico: Incorporar información del futuro. Ej: Edad actual en vez de edad al momento de generar el registro.
- Las variables deben ser:
  - Replicables.
  - Congruentes.
  - Asociadas al problema.

Variable Objetivo

- Corresponde al problema más importante en la definición del modelo.
- Consideraciones.
  - ¿Qué deseo modelar? (¿Cuál es mi problema?)
    - En la variable objetivo se debe definir esto con claridad.
    - ¿Puedo calcularla?
    - Horizonte de tiempo.
  - La variable objetivo debe ser estándar para todos los casos y estar adecuada al problema.
    - ¿Qué modelos se usarán?
    - ¿Cuáles son los requerimientos del modelo?

Ejemplos de variable objetivo

- Credit Scoring:
  - Si el cliente dejó de pagar su crédito (*defaulted*) en los primeros 12 meses desde la fecha de pago de su primera cuota.
- Fuga de Clientes:
  - Si el cliente dejó la compañía dentro del año siguiente.
- Marketing:
  - Si el cliente tomó el producto el mes siguiente.
- Importante: Esta es la información FUTURA (lo que quiero que prediga el modelo).

Consolidación Car Data

- Objetivo: Consolidar las variables asociadas a los autos, junto con una variable objetivo: si tuvo algún accidente a partir del 2001.

- Paso 1: Crear tabla “Con Accidente”

- Query:

```
SELECT DISTINCT ID_CAR INTO CON_ACCIDENTE
FROM ACCIDENT
WHERE DATE_ACC>#1/1/2001#;
```

Consolidación Car Data

- Paso 2: Consolidación

- Query:

```
SELECT Comfort.Id, Comfort.doors, Comfort.persons,
   Comfort.lug_boot, Comfort.safety,
   Perception.perception, Price.buying, Price.mant,
   IIf(IsNull([ID_CAR]),0,1) AS VAR_OBJ INTO
   TABLA_MAESTRA_CONSOLIDADA
FROM ((Comfort INNER JOIN Perception ON
   Comfort.Id=Perception.Id) INNER JOIN Price ON
   Perception.Id=Price.Id) LEFT JOIN CON_ACCIDENTE
   ON Price.Id=CON_ACCIDENTE.ID_CAR;
```

Diseño de Reportes de Proyectos de Business Intelligence

20

SEBASTIÁN MALDONADO

DIPLOMADO *BUSINESS INTELLIGENCE*

20 DE DICIEMBRE, 2011

Introducción

- Objetivos de un reporte de minería de datos.
  - RESUMIR el proceso realizado.
  - INFORMAR las conclusiones y los resultados relevantes.
  - RECOMENDAR políticas y usos de los resultados del modelo, según las necesidades del cliente.
- El reporte del proyecto debe ser AUTOCONTENIDO.
  - Debe contener información suficiente para que alguien entendido en el tema, pero no experto, sea capaz de comprenderlo cabalmente.
    - Incluir descripción breve de los modelos utilizados.
    - Incluir consideraciones técnicas importantes.
- El reporte debe permitir REPLICAR los resultados obtenidos por algún experto.

Secciones de un Reporte

- Portada
- Resumen ejecutivo.
- Índice(s).
- Introducción.
- Cuerpo del Informe.
- Conclusiones.
- Anexos.

Portada, Resumen e Índice

- Portada.
  - Debe incorporar información de los consultores, de a quienes va dirigido y la fecha.
    - Importante! Logos de las instituciones asociadas.
- Resumen Ejecutivo.
  - Sección importantísima del reporte. Resume el trabajo realizado, los resultados y conclusiones más importantes.
    - Largo máximo: UNA página.
- Índice.
  - Incluye todas las secciones del informe.

Introducción

- La introducción realiza una presentación del trabajo. Incluye:
  - La motivación del problema.
  - La descripción de quienes lo resuelven.
  - Una pequeña descripción de la estrategia de solución.
  - Alcance del proyecto. (Qué resolverá y qué no).
- El último punto de toda introducción es incluir una descripción del orden del trabajo.
  - 'En la siguiente sección veremos el modelo utilizado, para luego revisar los resultados en el capítulo 3...'

Cuerpo del Reporte

- Es dónde está la 'carne' del reporte.
- Se deben incluir todos los elementos relevantes para comprender la operación del proyecto.
- Incluye:
  - Descripción BREVE del proceso utilizado.
  - Incluye todas las partes del proceso, con su procedimiento y sus resultados.
    - Si es relevante, se pueden dividir los resultados en una sección aparte y final.
    - Resultados sólo si son RELEVANTES. Resúmenes de resultados son recomendados.
  - Se incluyen los elementos relevantes de esa fase de análisis.
  - Interpretaciones de los resultados.

Cuerpo del Reporte (II)

- NO incluye:
  - Tablas largas.
  - Descripción de modelos teóricos que no son relevantes.
  - Resultados de experimentos que se repiten varias veces.
  - COEFICIENTES DE MODELOS.
- TODOS estos elementos deben ir en la sección de Anexos.
  - Los anexos DEBEN estar referenciados en el cuerpo del informe.
    - 'El anexo X incorpora los resultados detallados de...'

Conclusiones

- Sector MÁS importante del informe.
  - Se resumen todos los resultados destacables de otras secciones.
    - Repetir no es malo si es importante.
    - Incorporar el conocimiento descubierto relevante.
  - Se incorporan recomendaciones relevantes con respecto a:
    - Usos del modelo o de los resultados.
    - Cambios relevantes a los procesos, datos, etcétera, que se hayan presentado en los experimentos.
  - Se incorporan comentarios generales con respecto al trabajo.
  - Si es relevante, las conclusiones son seguidas por una bibliografía.

Anexos

- Los anexos es dónde se incorporan todos los datos que quien solicita los datos debe saber, pero que no es indispensable para una correcta estructura del informe.
- Características:
  - Son referenciados en alguna parte del informe.
  - Tienen números de capítulo distintos. Utilizar letras. (Anexo A, Anexo B, etc…).
  - No existe problema en la cantidad de páginas que tenga, puede perfectamente ser más grande que el informe principal.