Práctica 5: Minería de datos sobre un Data Mart

Almacenes y Minería de Datos

Grado en Ingeniería en Informática
Dpto. de Informática e Ingeniería de Sistemas
Universidad de Zaragoza
Escuela de Ingeniería y Arquitectura

Sergio Ilarri y Jordi Bernad

16 de diciembre de 2020

1. Objetivos

En esta práctica se realizarán diversos procesos de minería de datos para intentar descubrir patrones de comportamiento en grandes volúmenes de conjuntos de datos.

Se utilizará el programa estadístico R, el cual puede conectarse directamente a la base de datos Oracle para obtener la fuente de datos, o hacer una consulta a la base datos y guardar en un fichero tsv los datos a analizar.

2. Contexto del problema

Se pretende analizar los datos almacenados en el data mart implementado en las prácticas anteriores que contiene información sobre vuelos comerciales en Estados Unidos.

2.1. Minería de datos en el data mart

Se requiere que se estudie con mayor profundidad la causa del retraso de los vuelos. Para ello, se realizará una selección y aplicación de técnicas de minería de datos para intentar resolver las siguientes cuestiones:

1. Comprobar qué factores son los que intervienen con relevancia en el retraso de los vuelos.
2. Clasificar los vuelos en tres grupos según el retraso que acumulan en función de:
   - *a*) aeropuerto origen
   - *b*) aeropuerto destino
   - *c*) aerolínea u operadora
   - *d*) modelo de avión

Se pueden hacer la 1 (reg paso a paso), 3 (anova) y el 2 (kmeans) es mañana

3. Analizar si los retrasos de los vuelos dependen según la franja horaria: mañana, tarde, noche.

4. Dado un vuelo concreto, determinado por su aerolínea, aeropuertos de salida y llegada, plantear la pregunta ¿voy a llegar con retraso? indicando con qué probabilidad (fiabilidad de la respuesta).

5. Estudiar si existen patrones frecuentes en los datos (reglas de asociación) e interpretar esos patrones. ¿Qué conclusiones se pueden sacar?

Será de especial interés reflejar en la memoria de la práctica una correcta interpretación y evaluación de los resultados obtenidos, explicando la extracción de conocimiento a partir de los datos.

3. Entrega de la práctica

La práctica se realizará en equipos de tres personas (salvo que existan problemas logísticos, que deberán comentarse y resolverse previamente con el profesor). Cuando se finalice la práctica se debe entregar un fichero .zip denominado,

pN_NIP1_NIP2_NIP3.zip

donde NIP1, NIP2 y NIP3 son los NIP de los autores de la práctica, con NIP1 $<$ NIP2 $<$ NIP3, y en pN la N representa el número de práctica, con el siguiente contenido:

1. Un fichero de texto denominado *autores.txt* que contendrá el NIP, los apellidos y el nombre de los autores de la práctica en las primeras líneas del fichero.

2. Un fichero de texto o PDF denominado informe.txt o informe.pdf, que contendrá la memoria de prácticas (respuestas a las cuestiones planteadas, esquemas y código desarollados, etc., según el caso). En dicho fichero se identificará claramente al comienzo los componentes del grupo de prácticas (nombre y apellidos y NIP de cada uno) y el número de práctica. El informe de prácticas debe contener al final un apartado de conclusiones personales que incluirá, entre otras cosas, información sobre el tiempo invertido por cada miembro del grupo de prácticas en la realización de la misma.

3. Todos los fuentes y programas de prueba desarrollados, si es el caso.

Al descomprimir el fichero .zip se deben extraer los ficheros y directorios necesarios en el directoriopN_NIP1_NIP2_NIP3. Es importante seguir las convenciones de nombrado y la estructura de ficheros y directorios descrita.

Para la entrega del fichero .zip, se utilizará Moodle 2 del Anillo Digital Docente de la Universidad de Zaragoza. **La fecha límite de entrega es el 15 de enero de 2021 a las 23:59.**

4. Recomendaciones

Para la evaluación de la práctica se considerarán los siguientes aspectos:

- La consecución de los objetivos planteados y la adecuada realización de las tareas correspondientes.
- La justificación de las afirmaciones incluidas en el informe.
- La inclusión de referencias a fuentes de información, si procede.
- La metodología para llevar a cabo la búsqueda de información necesaria para la realización de la práctica.
- La estructura y presentación del informe elaborado (completitud del trabajo, precisión, explicaciones adecuadas y completas, justificación de las decisiones tomadas, referencias adecuadas, coherencia de discurso, sin errores tipográficos y ortográficos, etc.).

Hay que asegurarse de que, en caso de que la práctica incluya algún tipo de desarrollo, funciona correctamente en los ordenadores del laboratorio o puede probarse fácilmente sobre una máquina virtual base proporcionada por el profesorado. También es importante someter código limpio (donde se ha evitado introducir mensajes de depuración y comentarios que no proporcionan información relevante).

Es importante extraer conclusiones: resultados obtenidos, indagaciones realizadas, diferencias observadas entre las distintas alternativas y/o herramientas, dificultades encontradas, valoración de los aprendizajes o mejoras de habilidades conseguidas, opinión personal, etc. Se recomienda destacar especialmente aquellos aspectos que muestren la capacidad de indagación, búsqueda de información, autonomía, y curiosidad personal que se hayan desarrollado.