Tarea 2 - Minería de datos descriptiva

EB. WG. ICD

Octubre 9, 2016

Objetivo:

El Ministerio de Educación y de Salud lo ha contactado a usted para analizar un conjunto de datos relacionado con universidades. Se requiere que usted aplique técnicas descriptivas de la minería de datos sobre los conjuntos de datos. Recuerde que usted es el experto, puede tomar las decisiones que usted quiera con tal que estén bien fundamentadas.

Las técnicas **Descriptivas** están orientadas a describir un conjunto de datos. **Clustering** es un proceso que consiste en la división de los datos en grupos de objetos similares. Las **reglas de asociación** se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos.

Requerimientos:

Se requiere que usted utilice **CRISP-DM** para resolver el problema, recuerde particionar los pasos de esta metodología en su informe. Debe realizar un análisis exploratorio (PCA), técnicas de agrupamiento (K-medias y clasificación jerárquica) y reglas de asociación.

Para realizar las tareas de clustering usted deberá utilizar los datasets, **Universidades.csv** y **Estudiantes.csv**, y para generar las reglas de asociación la matriz de transacciones, **Alimentacion.csv**. No es necesario que sepa el significado de las columnas de los datasets **Estudiantes** y **Universidades**, sin embargo puede obtenerla aquí.

Clustering:

Escoger y explicar, en base a los conocimientos adquiridos en clase, el mejor algoritmo de clustering según su criterio para los datasets **Estudiantes** y **Universidades**. Puede seleccionar las columnas que usted concluya que son de utilidad.

Para el dataset **Estudiantes** implemente y explique como el algoritmo de **k-medias** puede converger de manera más rápida (reducir el tiempo de ejecución), y en caso de que no pueda utilizar **clasificación jerárquica**, indique el motivo y la razón.

Ayuda: Recuerde que debe usar, al menos, **kmedias** y **clasificación jerárquica** (con **distintas distancias**). Usar métodos exploratorios que corroboren su decisión y de tal manera que la expliquen de mejor manera al usuario final. En caso de ser necesario indique manualmente los centroides iniciales.

Reglas de asociación:

Se desea analizar las compras realizadas por los estudiantes utilizando sus respectivas becas, para realizar esto usted deberá utilizar solamente la **matriz de transacciones**, **Alimentacion.csv**.

- Describa las reglas generadas e indique el valor de MinSupport, MinConfidence y MinLift, justique.
- Conocer las 10 transacciones con mayor número de apariciones en el dataset.
- Dado un estudiante nuevo que haya consumido N productos (N variables), poder recomendar un producto N+1. Para poder calcular las reglas es necesario definir un MinSupport, MinConfidence y MinLift, sin embargo, se desconoce cuáles son estos valores en consecuencia es tarea de usted determinar y justificar los mismos de acuerdo a su criterio.

- Si un estudiante consume chicles, ¿Cuál es la probabilidad que consuma refrescos?
- Ordene las reglas por confianza, soporte y lift e indique sus observaciones.
- Dado que el consecuente (RHS) de las reglas sea Naranja, indique el producto que más se repita en una transacción.
- Dado que un estudiante consumió Flips, Nutella y Vodka que producto usted le recomendaría.

Nota: Se tomará en cuenta otros análisis y conclusiones en base a las reglas generadas.

Consideraciones de forma:

Ingrese a la dirección ICDRepository-I-2016/mineria-de-datos-descriptiva y siga las instrucciones para crear un repositorio en GitHub perteneciente a la organización. Este repositorio será propiedad de la organización pero solo usted puede realizar cambios en el mismo. El repositorio debe poseer lo siguiente:

1. Scripts (.R) intradocumentados.
2. Crear un informe mediante Rmarkdown (.Rmd) y generar un PDF .
3. El informe debe almacenarlo en una carpeta **doc**.
4. Los scripts (.R y .Rmd) debe almacenarlo en una carpeta **src**.
5. README.md explicando la configuración del ambiente en el cual trabajó. Ejemplo: README.md de Bootstrap, GraphX u otro.

Consideraciones de contenido:

Se recomienda el uso de las funcionalidades de los paquetes **ArulesViz** y **ggplot2** para realizar análisis.

- La tarea es estrictamente individual. Se promueve la participación y discusión de la misma en un ambiente responsable. Sin embargo, cualquier evidencia de copia será severamente sancionada colocando una nota mínima de cero (0) puntos según lo establecido en la Ley de Universidades. Cualquier proyecto entregado debe ser fruto de su propio trabajo.
- Fecha de Entrega: Domingo 9 de Octubre de 2016.
- Hasta este día se aceptarán push's en los repositorios.
- No se recibirá ninguna tarea por correo electrónico.

Ayuda

- Libro Mining of Massive Datasets.
- Si lo desea puede utilizar otras técnicas para visualizar los conglomerados.

```
ggplot(NombreDataset,
       aes(x = columnaX,
           y = columnaY,
           color = factor(kmedias$cluster))) +
  geom_point(alpha = 0.50) +
  theme_minimal()
```