Programa Ciencia de los Datos

Curso Minería de datos e Inteligencia de Negocios

Marzo 2020

Prof. Lorena Zúñiga

Tarea #3 Minería de datos

Valor 15%

Objetivos

- Aplicar parcialmente la metodología CRISP-DM para resolver un caso de minería de datos
- Desarrollar un modelo descriptivo utilizando reglas de asociación
- Interpretar las reglas generadas por el modelo de asociación.

Descripción

Se tiene el siguiente dataset sobre población adulta residente en Estados Unidos, los datos se obtuvieron a partir de un censo en los años 90s. Se encuentran en el archivo AdultosUSA.csv

El dataset contiene los siguientes atributos:

- Edad : edad en años cumplidos
- TipoTrabajo: si el trabajo es en el sector privado, público, etc.
- NivelEducativo: se refiere al máximo nivel académico obtenido por la persona.
- AnnosEducacion: cantidad de años de escolaridad de la persona
- EstadoCivil
- Ocupación: ocupación actual de la persona
- Sexo: sexo de la persona
- HorasSemanales: cantidad de horas que labora por semana
- PaisOrigen: país de nacimiento
- Ingresos: indica si los ingresos de la persona son iguales o inferiores a $50mil o bien si son superiores a esa cantidad.

Por hacer:

Según lo visto en clase y el material disponible, debe encontrar reglas de asociación que permitan descubrir algún patrón interesante, por ejemplo si algunas características implican tener un ingreso superior a los $50mil. Documentar todo el proceso según las siguientes fases y actividades de la metodología CRISP-DM que se le solicitan.

Tome en cuenta que los datos NO se encuentran en un formato de transacción, por lo que deberá primero transformarlos antes de generar alguna regla.

Como parte de esa transformación se le recomienda que después de leer los datos originales, que transforme cada columna no numérica a factor. Para ello puede usar la función as.factor.

Por ejemplo: `dataframe$campo<- as.factor(dataframe$campo)`

Una vez que las columnas sean tipo factor, utilice la función `as()` del paquete `arules`, para convertir los datos del dataframe a un formato de transacción.
Por ejemplo: `trns<- as(dataframe,"transactions")`

Puede suponer que:

- el objetivo de negocio en este caso es descubrir características que podrían llevar a una persona a ganar tener ingresos superiores a $50mil.
- el objetivo de minería de datos es descubrir reglas de interés que permitan asociar características con el hecho de ganar más de $50mil.

Fases y actividades a desarrollar

- **Entendimiento del negocio**
  - **Determinar los objetivos de minería de datos**
    - Defina los criterios de éxito desde la perspectiva de minería de datos

- **Entendimiento de los datos**
  - Exploración de los datos
  - Verificación de la calidad de datos

- **Preparación de los datos**
  - Selección de los datos
  - Limpieza de los datos

- Construcción de nuevos datos (atributos)
- Transformaciones aplicadas a los datos

Si la preparación de los datos la realiza en un lenguaje de programación distinto a R, debe documentar los cambios o transformaciones aplicadas a los datos, aportando screenshots y agregándolos como figuras en el documento RMarkdown.

- **Fase de modelado**
  - Selección de técnicas
  - Construcción del modelo (reglas)
    - Selección de los parámetros
    - Ejecución (generación de reglas, eliminación de subconjuntos,etc)
    - Descripción de los modelos obtenidos (incluya al menos un gráfico)
  - Evaluación de los modelos
    - Muestre e interprete las mejores 3 reglas

Exporte el documento RMarkdown a html

Formato de entrega: archivo html únicamente

Forma de trabajo: individual o en parejas

Forma de entrega: enviar documento con la solución a través del TECDigital

Fecha de entrega: hasta el domingo 08 de marzo de 2020, a las 11:55 p.m.

Evaluación

Rubro	Valor
Entendimiento del negocio	**8**
Determinar los objetivos de minería de datos	4
Defina los criterios de éxito desde la perspectiva de minería de datos	4
Entendimiento de los datos	**15**
Exploración de los datos	7.5
Verificación de la calidad de datos	7.5
Preparación de los datos	**25**
Selección de los datos	7
Limpieza de los datos	6
Construcción de nuevos datos (atributos)	5
Transformaciones aplicadas a los datos	7
Fase de modelado	**40**
Selección de técnicas	3
Construcción del modelo (reglas)	7
Selección de los parámetros	4
Ejecución (generación de reglas, eliminación de subconjuntos,etc.)	15
Descripción de los modelos obtenidos (incluya al menos un gráfico)	11
Evaluación de los modelos (reglas)	**12**
Muestre e interprete las mejores 3 reglas	12
Total	**100**