Proyecto Kaggle Titanic

WG. EB. GDICD

I-2016

Objetivo:

Parte de su carrera como científico de datos consiste en mantenerse mejorando sus habilidades para afrontar nuevos y diferentes problemas, para ello Kaggle se ha vuelto un lugar común en el que se resuelven problemas con fines académicos o comerciales. En esta oportunidad, su trabajo consiste en usar sus conocimientos adquiridos para recorrer todo el ciclo de resolución de uno de los problemas más conocidos del área tiene como nombre Titanic.

Requerimientos:

Utilizando los datasets train y test, link:

- Realice las tareas de preprocesamiento que considere necesarias para mejorar la representación del dataset.
- Realice un análisis exploratorio.
- Aplique distintos algoritmos de clasificación o predicción (árbol de decisión, SVM, etc) y comparelos segun los criterios vistos en clase.
- La idea es que experimente con la plataforma y ponga a prueba sus habilidades, por lo que toda actividad que contribuya con dicho objetivo esta justificada.
- Aplique clustering (kmeans, hclust, etc) sobre las edades e indique cuantos grupos existen.
- Aplique reglas de asociación al siguiente DATASET e indique sus conclusiones de acuerdo a las reglas generadas (Ej: los varones que murieron fueron en su mayoría adultos (97%), toda la tripulación es adulta, las mujeres que murieron pertenecían a la tercera clase, etc).

```
load("titanic.raw.Rdata")
str(titanic.raw)
rules <- apriori(titanic.raw)
```

- Si considera útiles otras actividades no mencionadas en este documento o en la plataforma en cuestión, sientase libre de llevarlas a cabo, con la apropiada justificación y documentación.

Consideraciones de contenido:

Haciendo uso de los modelos de clasificación vistos en clase prediga que individuos sobrevivirán al hundimiento tras ser asignados a botes de salvación, para esto se recomienda que:

- Preprocese las características para adecuarlas al modelo que desee aplicar.
- Entrene los modelos requeridos.
- Ajuste los parámetros (justificando su decisión) para cada modelo con el fin de mejorar su desempeño (de ser posible, en caso contrario, justifique).
- Compare los modelos usando como referencia las respectivas matrices de confusión y curvas ROC (de ser posible, en caso contrario, justifique).
- Recomiende las características más influyentes en la decisión de su modelo (de ser posible, en caso contrario, justifique).
- Toda actividad que considere apropiada tomando en cuenta el contexto del problema, los requerimientos, sus conocimientos y/o curiosidad.

Estas actividades deben estar concisamente documentadas.

Consideraciones de forma:

Ingrese a la dirección Proyecto Kaggle Titanic 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 para realizacion de sus actividades.
2. README.md explicando la configuración del ambiente en el cual trabajó. Ejemplo: README.md de Bootstrap, GraphX u otro.
3. Explicación de su proceso **.Rmd**, puntos adicionales dependiendo de la presentación.
4. Recuerde mantener un orden en su repositorio (doc, src, data, etc).

Consideraciones de contenido:

- Se recomienda el uso de las funcionalidades de los paquetes usados en el laboratorio y
- Se tomará en cuenta el uso de otros paquetes (Puntos extras).
- 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 6 de Noviembre** de 2016.
- Hasta este día se aceptarán push's en los repositorios.
- No se recibirá ninguna tarea por correo electrónico.