Universidad del Valle

Data Mining

Fecha: 23/01/2021

**Proyecto Supervised Learning**

El objetivo del proyecto final consiste en permitirle a los estudiantes poner en práctica los conocimientos teóricos y prácticos adquiridos durante la clase. El proyecto se realizará de forma individual y se desarrollará con datos de ventas de seguros de vehículos.

Toda la información del caso la podrán encontrar en: <https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction>

Deberá responder los incisos de se enlistan a continuación, preparar una presentación con sus resultados, la cual deberá subir a youtube.com (el tiempo del video debe estar entre 5 y 10 minutos).

Incisos:

1. Explorar el dataset train
   1. ¿Qué tipos de datos identificamos?
   2. ¿Cuál es la distribución de las variables univariadas?
   3. Presentar por lo menos 3 tablas de contingencia que se relacionen con el target del caso.
   4. ¿Qué relaciones identifica entre las variables que podrían afectar en la decisión del cliente?
   5. Con los datos que ha visto plantear su hipótesis con relación al objetivo del caso (que los clientes adquieran el seguro)
   6. Presentar un listado de preguntas que ayuden a responder la hipótesis planteada
   7. Presentar gráficas para responder las preguntas planteadas en el punto anterior (puede utilizar Powerbi o ggplot):
2. Aplicar los modelos de supervised learning (mínimo 2 modelos de cada tema)
   1. Naive Bayes
   2. Árbol de decisión
   3. Bosque aleatorio
   4. Regresión logística
3. Explicar los resultados de cada modelo.
   1. Acurracy
   2. Precision
   3. Recall
4. ¿Con cuáles de los modelos que construyó se quedaría y por qué?
5. Conclusiones y recomendaciones

Rubrica:



Entregables:

* Rscript
* Powerbi (si aplica)
* Presentación PowerPoint con procedimiento
* Video en www.youtube.com