

Proyecto de clase

Contexto

El proyecto de clase tiene como propósito evaluar la calidad de los modelos que puedan desarrollar. En la primera entrega de proyecto se enfrentaron a un dataset (conductores de MetroCali) al que tuvieron que hacerle un EDA completo. En esta segunda entrega van a enfrentarse a un dataset del que no van a conocer nada. Lo único que se sabe es que se tiene un conjunto de variables independientes que describen el comportamiento y características de los clientes de una empresa, y que permiten predecir una variable objetivo (que se encuentra en archivo separado). La variable objetivo es binaria y los dos estados significan

- Estado 1: se pudo convencer al cliente de comprar unos servicios o productos determinados
- Estado -1: no se pudo convencer al cliente en cuestión

Disponen de un archivo llamado **training.tsv** que incluye la variable objetivo **CONVERSION** y de un archivo **test-STUD.tsv** que no la incluye. En cada uno de ellos hay una columna **ID** que corresponde al indicador del cliente.

Puntos a desarrollar

1. **Limpieza y EDA:** Solo tienen que realizar el EDA y la limpieza que sea necesaria para que sus modelos sean viables. Se tendrá en cuenta con un pequeño coeficiente en la nota final del proyecto. Se espera una sección en un jupyter notebook con SOLAMENTE lo MAS IMPORTANTE.
2. **Modelos predictivos:** Esto es lo más importante del proyecto. Se tendrá en cuenta no solo las métricas finales de evaluación de los modelos, sino el protocolo de evaluación y selección del modelo que se siguió (cross-validation, holdout, etc.). Se espera una sección en el jupyter notebook que describa el proceso y la estimación de las métricas de evaluación. Para la entrega de los resultados deben generar las predicciones de los registros del archivo **test-STUD.tsv** y generar un archivo **resultados-NOMBRE_APELLIDO.tsv** que contenga solamente las columnas **ID** y **CONVERSION**.
3. **Caracterización de los clientes:** Se deben encontrar perfiles para los clientes en estado 1 y -1 (máximo 5 para cada clase), y se debe poder caracterizarlos en términos de SOLO las variables más representativas. Se espera una sección en el jupyter notebook.

Rúbrica de puntuación

Preparación de los datos	Protocolo de entrenamiento y evaluación de modelos	Resultados enviados en archivo adjunto	Determinación del número de perfiles para cada clase	Caracterización de perfiles de clientes en los dos estados	TOTAL PROYECTO SEGUNDA ENTREGA
0.9	1.0	1.5	0.8	0.8	5.0