**Trabajo Práctico Módulo 2**

**Aprendizaje supervisado**

**Datos de un Banco**

Descripción:

Los datos utilizados corresponden a los clientes de una tarjeta de crédito de un banco. En el mismo podemos identificar las siguientes columnas.

1. CLIENTNUM: identificador único del cliente
2. Attrition\_Flag: Indica si el cliente sigue vigente o si ha cancelado la tarjeta
3. Customer\_Age: edad del cliente
4. Gender: género del cliente
5. Dependent\_count: cantidad de dependientes que posee el cliente
6. Education\_Level: nivel educativo del cliente
7. Marital\_Status: estado civil del cliente
8. Income\_Category: categorización de ingresos del cliente
9. Card\_Category: tipo de tarjeta que posee el cliente
10. Months\_on\_book: cantidad de meses desde que el cliente tiene la tarjeta de crédito
11. Total\_Relationship\_Count: Cantidad de productos que tiene el cliente
12. Months\_Inactive\_12\_mon: Cantidad de meses inactivos en los últimos 12 meses
13. Contacts\_Count\_12\_mon: Cantidad de veces que se contactó al cliente en los últimos 12 meses
14. Credit\_Limit: límite de crédito de la tarjeta
15. Total\_Revolving\_Bal: un saldo renovable es la parte del gasto de la tarjeta de crédito que no se paga al final de un ciclo de facturación.
16. Avg\_Open\_To\_Buy: Promedio del saldo disponible
17. Total\_Amt\_Chng\_Q4\_Q1: Cambios en los montos de las transacciones
18. Total\_Trans\_Amt: Monto total de transacciones
19. Total\_Trans\_Ct: Cantidad total de transacciones
20. Total\_Ct\_Chng\_Q4\_Q1: Cambios en la cantidad de transacciones
21. Avg\_Utilization\_Ratio: promedio de utilización de la tarjeta

Consigna:

El objetivo del proyecto es predecir aquellos clientes con mayor propensión a cancelar la tarjeta de crédito del banco, lo que se conoce como Churn. Es decir, predecir el valor de la columna Attrition\_Flag

El trabajo debe incluir:

1. Análisis Exploratorio de Datos
2. Responder algunas preguntas generales del dominio. Por ejemplo: Cantidad de personas que cancelan la tarjeta según sus ingresos, o dependientes. Edad promedio de los clientes, nivel educativo más frecuente, etc.
3. Análisis de Valores Nulos. Nota: los nulos se presentan con la palabra “unknown”
4. Análisis de Outliers
5. Transformación de variables categóricas. Evaluar cuándo usar one hot encoding o label encoding
6. Análisis de Distribución y Análisis de correlación de las variables
7. Dividir los datos en set de entrenamiento y en set de prueba. Trabajar con el set de entrenamiento. Evaluar desbalance de clases, probar hacer un balance un poco más equitativo, probar llevarlo a 50-50. (Oversampling, undersampling o SMOTE)
8. Entrenar 3 modelos de clasificación. Utilizar técnicas de k-fold y grid search encontrar los mejores hiper parámetros.
9. Reportar los resultados de cada modelo (Accuracy. Recall, Precision y F-score). Incluir matriz de confusión. Seleccionar el mejor modelo en base al análisis de las métricas.
10. Este punto es opcional: Incluir algún gráfico de Explainability.

Se debe entregar:

1. Jupyter notebook o un archivo .py con el Código organizado. Deben incluirse comentarios y breve interpretaciones de los resultados.
2. Una presentación de máximo 15 diapositivas, y elaborar un video de dicha presentación, explicando los hallazgos obtenidos (no más de 5 minutos).

El trabajo se puede hacer en grupos de hasta 4 personas.

**Fecha de Presentación del TP:**

**Cohorte 1: Lunes 21/06.**

**Cohorte 2: Lunes 5/07.**