

## Ejercicio 2

Diplomado Inteligencia de Negocios

Classificación de ordenes de compra

- Realizar Actividad en grupos de 4 personas.

Continuando con la misión de clasificar el potencial comportamiento de las ordenes de compra, en la presente tarea deberán desarrollar finalmente un modelo que estime el comportamiento de la base de datos RETAIL, tomando en consideración los puntos que se describen a continuación.

## 1 Descripción de la actividad

El problema de la presente tarea es estimar la probabilidad que una orden sea de alto riesgo ( $y = +1$ ) o bajo riesgo ( $y = -1$ ), sobre un listado de ordenes de compra que la empresa no tiene información y desea revisar. Para lograr este objetivo, ya se cuenta con una base de datos de cuyo comportamiento es conocido<sup>1</sup>.

- Utilizando la base de datos de las ordenes de compra cuyo comportamiento es conocido (i.e.  $(x, y)$ ,  $x \in \mathcal{X}$ ,  $y \in \{+1, -1\}$ ), limpia y transformada (obtenida en la tarea 1), deben crear un modelo que permita predecir el comportamiento de las 20.000 ordenes de compra sin información sobre su comportamiento  $y$  (**RETAIL\_eval.csv**)
- Pueden desarrollar un modelo que estime ya sea un puntaje de riesgo (*Score*) o una probabilidad de riesgo. Para ambos tipos de salida es sumamente importante que sean explícitos en el parámetro de corte (*threshold*) que define si en base a la predicción una orden será o no riesgosa. **Ejemplo:** Si la predicción de una orden de compra  $i$ -ésima es

$$P(\text{orden "Alto Riesgo"}|x_i) = 0,81$$

y el *threshold* para catalogarla como “alto riesgo” es definido por ustedes por

$$P(\text{orden "Alto riesgo"}|x) \geq 0,70, \forall x \in \mathbf{data}$$

entonces según el modelo, la orden  $i$ -ésima deberá ser revisada cautelosamente por la empresa.

- Dado que el modelo es generado con la base de datos **Retail.csv** limpia y transformada, para evaluar la nueva base de datos **RETAIL\_eval.csv** deberá ser procesada con el mismo procedimiento de limpieza y transformación.
- Los modelos desarrollados pueden ser de cualquier tipo (Árboles de decisión, Redes Neuronales, Support Vector Machines, naïve Bayes, regresión logística, combinaciones de los anteriores, etc.), pero **es necesario el desarrollo y evaluación de al menos 3 técnicas**.
- Para definir el *threshold* o entrenar algunos clasificadores, deben considerar el uso de la siguiente matriz de costo (tabla 1):

## 2 Entrega

La salida debe ser un archivo CSV construido de la siguiente forma:

<ID>;<CLASE>

Es decir, el parámetro de corte ya deberá ser aplicado a la predicción final que obtengan del modelo.

---

<sup>1</sup>archivo **Retail.csv** procesado en la actividad anterior.

	Orden Riesgosa	Orden Regular
Orden clasificada como “Alto Riesgo”	2	13
Orden clasificada como “Bajo Riesgo”	-25	15

Table 1: Matriz de costo.