

Ejercicio 1

Diplomado Inteligencia de Negocios

Transformación y Pre-procesamiento de Datos

- **Realizar Actividad en grupos de 4 personas.**

En comparación con el intercambio tradicional de bienes (donde un cliente paga por un producto y el vendedor se lo entrega), vía Web, estafadores de todo tipo han logrado operar en distintos niveles (desde estafadores ocasionales a bandas criminales organizadas). En este caso, la regla “bienes \times dinero” no puede ser aplicada directamente, a excepción de aquellas compras contra-entrega agendadas vía el sistema online. Claramente, esta modalidad puede ser un obstáculo para aquellos clientes interesados en un sistema más flexible. Según esto, varias tiendas están implementando distintos tipos de pago online, cuyo principal problema es la seguridad de que un cliente dado pague finalmente el producto que solicitó vía web.

Una de las grandes empresas de retail nacionales a detectado en los últimos años un aumento significativo sobre las ventas online en el período noviembre-diciembre de cada año (período navideño). Desafortunadamente, el nivel de perdidas en el negocio online es considerablemente mayor que el intercambio tradicional, a pesar de usar métodos simples de validación de tarjetas de créditos y direcciones de los clientes. Actualmente, el mecanismo de revisión de ordenes ha sido manual, causando graves errores y altos costos operacionales, donde además, el volumen de información es cada año mayor.

Dado esto, la empresa lo contrató para desplegar un modelo de minería de datos para automatizar una solución a este problema. En particular, se solicita el desarrollo de una aplicación que permita determinar el puntaje de riesgo individual con respecto al pago de un determinado cliente.

A continuación (tabla 1), se presenta una matriz de costo para el problema de clasificar una determinada orden cómo “Alto Riesgo” y “Bajo Riesgo”.

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

Table 1: Matriz de costo.

En este caso, la empresa considera el desarrollo de la herramienta que calcule la probabilidad de riesgo individual para cada orden, para luego, evaluando el costo esperado del problema, asignar una orden dada cómo de “Alto Riesgo” y “Bajo Riesgo”. La asignación a una de estas clases permitirá a la empresa tomar acciones con respecto a ciertas ordenes de compra, e.g. solicitar cambio de pago online a contra-entrega para aquellas ordenes más riesgosas, verificación vía telefónica, etc.

Una muestra de 30.000 ordenes fue seleccionada por la empresa para desarrollar el modelo solicitado, donde cada registro es una orden de compra con información específica almacenada por la empresa. Además se tiene la información a si efectivamente fue o no una orden de compra de “Alto Riesgo”¹.

¹Una descripción de los atributos se puede encontrar en el archivo *riesgo_atributos.pdf*

Descripción de la Actividad

El objetivo de la tarea es investigar e implementar distintas metodologías de pre-procesamiento de datos y selección de atributos, utilizando aquellas que estimen convenientes sobre la base de datos RETAIL.

Pregunta 1

Datos nulos e imputación de datos

- ¿Qué pasa si hay una cantidad considerable de valores nulos en una variable? Explique las técnicas que se pueden considerar para la imputación de datos. ¿En qué casos se podría considerar la alternativa de eliminar la variable?
- En la base de datos, ¿Cómo están relacionadas las variables que presentan valores nulos con la variable dependiente²?, ¿Como se comportan los datos nulos presentes en la base de datos de prueba?
- ¿Bajo qué supuestos se podría considerar aceptable la eliminación de un registro en una base de datos?, ¿Cómo se aplican estos supuestos en la base de datos en RETAIL?
- Considerando los puntos anteriores, proponga, desarrolle y documente una metodología de imputación de datos y eliminación de observaciones sobre la base de datos RETAIL.

Pregunta 2

Valores fuera de rango (*Outliers*)

- Mencione y explique al menos 3 técnicas que se pueden utilizar para identificar y tratar los valores fuera de rango o *Outliers*.
- Proponga, desarrolle y documente una metodología de tratamiento de Outliers sobre la base de datos RETAIL.

Pregunta 3

Preprocesamiento y re-codificación de variables

- Mencione y explique las alternativas y diferencias que se pueden considerar para el procesamiento y re-codificación de variables cualitativas nominales y variables cualitativas ordinales.
- En base a las variables cualitativas, explique y proponga estrategias que se pueden considerar para disminuir la dimensionalidad de la base de datos.
- Mencione y explique las alternativas que se pueden considerar para el procesamiento de variables cuantitativas continuas y variables cuantitativas discretas.
- Aplique y documente claramente las alternativas que consideradas sobre las distintas variables de la base de datos RETAIL.

² *Variable dependiente* es otro término utilizado para referirse a la variable objetivo.

Pregunta 4

Selección de atributos y Extracción de Atributos

- Proponga estrategias de selección de atributos utilizando cada una de las siguientes técnicas:
 - Análisis de correlación
 - Tablas de contingencia (*CrossTabs*)
 - Test χ^2 , test ANOVA.
 - *Information Gain*, *Gini Index*, Árboles de Decisión
 - *Forward*, *Backward Feature Selection*
- Explique detalladamente en qué consiste y cómo utilizaría la técnica Kernel-PCA para la extracción de atributos. ¿Cómo lo utilizaría para la selección de atributos?, ¿Cuales serían sus resultados en ambos casos con respecto a la base de datos? ¿Cuál es la diferencia con los resultados obtenidos con la técnica PCA?, ¿Cómo están relacionados estos resultados con la técnica de descomposición matricial SVD?
- En base a las técnicas mencionadas anteriormente, aplique y documente las estrategias que estime convenientes sobre la base de datos RETAIL.