LABORATORIO DE INTELIGENCIA COMPUTACIONAL
LABORATOIO 2: APRENDIZAJE NO SUPERVISADO PARA DESCUBRIMIENTO DE INSIGHTS
26 DE SEPTIEMBRE DE 2019

Siguiendo el uso de Python 3.7 de la distribución Anaconda, llevaremos a cabo esta actividad usando el algoritmo K-means para descubrir insights a partir de datos no etiquetados.

Ésta es una actividad muy necesaria en proyectos reales en que se requiere construir clasificadores predictivos pero que al comenzar con el proyecto nos encontramos con muestras no etiquetadas.

Usaremos como referencia las informaciones disponibles en este link: https://www.aprendemachinelearning.com/k-means-en-python-paso-a-paso/

Descargar de U-Cursos el notebook llamado Clustering.ipynb. Guardarlo en una carpeta del computador local. Desde la sesión web de Jupyter navegar hasta encontrar la carpeta donde se guardó el notebook. Dar click en el notebook para iniciar la experiencia.

Para esta sesión del laboratorio, debemos ejecutar las siguientes tareas:

1. Seguir paso a paso las informaciones del link arriba en el Notebook para comprender cada uno de los análisis que conlleva el aprendizaje no supervisado basado en K-means.

2. Usar el archivo de datos CREDITRISK_RAW_WithoutTarget.xlsx para replicar los mismos análisis del punto 1 con el fin de descubrir perfiles y grupos de similitud de clientes que solicitan créditos en una institución financiera. El objetivo de este punto será apoyar e ir dando solución a la problemática de un banco incorporando algunos conceptos de Machine Learning en un ambiente real y de negocio. Un objetivo secundario, pero muy importante, es el desarrollo de habilidades comunicacionales para explicar con mayor detalle la problemática de negocio, qué desafíos presenta hoy por ejemplo en las instituciones financieras chilenas este asunto, entender los datos disponibles a través de técnicas de aprendizaje no-supervisado. En aquellas variables donde considere que su descripción es limitada puede crearse su propio contexto en base al entendimiento del problema. El conjunto de variables inputs es el siguiente:

VARIABLES	DESCRIPCIÓN
ID	Identificador del cliente
GENERO	Genero del cliente
RENTA	Renta en pesos
EDAD	Edad en años
NIV_EDUC	Nivel Educacional
E_CIVIL	Estado Civil
COD_OFI	Código de la Oficina donde se realiza la solicitud
COD_COM	Código de la comuna donde está la Oficina
CIUDAD	Ciudad donde se realiza la solicitud
Crédito_1	Monto crédito 1
Crédito_2	Monto crédito 2
Crédito_3	Monto crédito 3
Crédito_4	Monto crédito 4
Monto solicitado	Monto actual solicitado
Días de Mora	Número de días que ha estado en mora (histórico)
Monto Deuda Promedio	Deuda promedio Anual
Número de meses inactivo	Número de meses en que no tiene el negocio activo
Número de cuotas	Número de cuotas que solicita el crédito actual
Aval	Con o sin aval
PAGA	Target

Tabla 1: Descripción de las variables

Entregables:

A. Una breve descripción con el entendimiento del problema de negocio de Riesgo Crédito (Credit Scoring) y por qué es importante para los bancos construir modelos predictivos de clasificación entre buenos y malos pagadores. Debe investigar en Internet sobre esta aplicación de negocio.
B. Basado en el punto 2 arriba, descubrir grupos de similitud y disimilitud entre clientes usando un subconjunto de variables del conjunto (a elección de cada alumno, >= 4 variables) usando K-means que permita distinguir entre comportamientos de buenos y malos pagadores. Documentar los perfiles que esos grupos resultantes tienen y generar algunas conclusiones acerca de datos (filas o columnas) útiles y no útiles del conjunto de datos que posteriormente permitan etiquetar a los clientes entre buenos y malos pagadores.

Entregar hasta el jueves 3 de octubre de 2019 por U-Cursos.