TAREA I de ESMA 4016 : Data Mining and Machine Learning

Fecha de entrega: Martes 18 de Febrero , 2020 por Piazza hasta las 7pm.
Enviar el notebook en formato html
Puntaje: 50 puntos

Datasets:

1-Scania: Air Presure System Failures in Scania Trucks (aps_failure_training_set.csv)
2- Estambul: ISTANBUL+STOCK+EXCHANGE

Los datos están disponibles en la UCI Machine learning repository
(https://archive.ics.uci.edu/ml/index.php) y en kaggle.com

Usar Python para responder a las siguientes preguntas:

1.(20) El conjunto de datos Air Presure System (APS) Failures in Scania Trucks contiene información acerca del sistema de Presion de Aire de los trucks Scania. Hay dos clases, que se debe a fallas relacionadas al Sistema APS la positiva y la negativa que se debe a falla por otras razones. Hay información faltante que aparece con "na". La clase es la primera columna de la tabla.

a) (4) Hacer un reporte de la información faltante, incluyendo visualización.
b) (5) Eliminar todas las columnas que tienen 60% o mas de "na" y filas con 10% o mas de "na".
c) En el dataset que queda en el paso b) Eliminar las columnas que tienen 50% o mas ceros y las filas que tienen valores mayores o iguales a 2 millones.
d) (6) Aplicar imputación usando la media o mediana y el metodo knn de imputación con k=3 vecinos para sustituir los datos faltantes.

2. (30 pts)

a) (6) Normalizar la data Estambul usando tres métodos visto en clase.
b) (4) Discretizar todas las columnas de la data en 10 intervalos de igual ancho
c) (6) Hacer un boxplot de los datos antes y después de la discretización. Comentar
d) (10) Insertar al azar al conjunto de datos Estambul n 5% , 10% y 15 % of missing values. No inserte missing values en la columna "date" que corresponde a fecha en que se tomaron los datos. Luego, imputar los missing values usando inputacion por la media, mediana y knn usando 5 vecinos mas cercanos.
e)(4) Calcular el valor

MSE=suma(valor verdadero-valor imputado)^2/numero de records de la tabla

y basado en este valor comparar los metodos de imputacion usados en d).