TAREA 2 de ESMA 4016: Data Mining y Machine Learning

Puntaje: 50 puntos

Fecha de Entrega: Marzo 10 hasta las 10 de la noche.

Usar Python para contestar las siguientes preguntas

1. (30 puntos) Usar el conjunto de datos asignados para responder las siguientes preguntas.

Pollution (Adel, Ambar, Cesar, Edwin): http://academic.uprm.edu/eacuna/pollution.txt La variable de respuesta is MORT
Crimen (Julio, Laura, Omar): http://academic.uprm.edu/eacuna/crimen.txt. La variable de respuesta es Crimen Rate.

a) (2) Hallar la variable que tiene correlación más alta con la variable de respuesta
b) (3) Hacer un plot correspondiente para ver si no hay outliers y determinar si el coeficiente de correlación es confiable.
c) (5) Hallar la regresión simple con la variable determinada en a) y graficarla. Interpretar el intercepto y la pendiente de la línea. Comentar su coeficiente de determinación.
d) (4) Hallar el modelo de regresión múltiple considerando todas las variable predictoras e interpretar dos coeficientes de regresión cualesquiera.
e) (2) Interpretar el coeficiente de Determinación R^2.
f) (5) Probar si cada uno de los coeficientes del modelo de regresión es cero. Comentar el resultado.
g) (4) Considerar valores adecuados de las variables predictoras y predecir la variable de respuesta.
h) (6) Aplicar dos métodos para seleccionar los mejores modelos y dar el R^2 de estos modelos.

II ((20)
Datasets:
Student Performance dataset (Labels es G3 pero binarizarla): (Adel, Laura, Omar) Disponible en la UCI y Kaggle
Default of credit cards clients : (Julio, Ambar, Cesar, Edwin) . Disponible en la UCI y en Kaggle.com

a) (8) Aplicar el metodo RELIEF a su conjunto de datos para seleccionar el mejor subconjunto de variables. Evaluar el subconjunto elegido usando la precisión del clasificación del clasificador LDA y Naive Bayes.

b) (6) Al conjunto que queda en la parte a) aplicar un metodo wrapper, junto los clasificadores LDA y Naive Bayes, para seleccionar variables en su conjunto de datos y comparar sus resultados de la parte a)
c) (6) Aplicar PCA al conjunto que queda de la parte a) y usando los clasificadores LDA y Naive Bayes comparar sus resultados con las parte a) y b).