TRABAJO TEÓRICO/PRÁCTICO INTEGRADOR PARA LA EVALUACIÓN DE LA PARTE TEÓRICA Y DE LA PARTE PRÁCTICA DE LA ASIGNATURA "INTRODUCCIÓN A LA CIENCIA DE DATOS"

(Representa el 95% de la nota de la asignatura)

Este trabajo consta de tres apartados: Análisis de datos, Regresión y Clasificación.

Cada estudiante contará con dos conjuntos de datos propios: datasetR – específico para realizar el apartado de regresión y datasetC para el apartado de clasificación. Además, todos los estudiantes contarán con el resultado de 16 conjuntos de datos para regresión y 20 conjuntos de datos para clasificación aplicados a los distintos algoritmos vistos en clase, tanto de regresión como de clasificación, para poder hacer los test estadísticos comparativos.

Se pide al estudiante que presente un informe único que contenga las respuestas a cada apartado. Este informe se puede entregar en .doc, .odt o .pdf. Para cada apartado se pide además incluir el código fuente que se ha utilizado para realizar todas las tareas requeridas. Es imprescindible que este código esté debidamente comentado de manera que ayude a entender los distintos pasos que se han realizado.

La fecha límite de entrega del trabajo será el día 22/12/2019 (Domingo) a las 23:55 a través de la plataforma PRADO en la sección "TRABAJO TEÓRICO/PRÁCTICO INTEGRADOR" (ítem "Entrega del Trabajo Teórico/Práctico Integrador").

APARTADO ANÁLISIS DE DATOS

En este apartado el estudiante debe realizar un estudio previo de sus dos conjuntos de datos asignados ($dataset_R$ y $dataset_C$). Este estudio debe incluir:

A.1 Descripción del tipo de datos de entrada (lista, data frame, ect., numero de filas, columnas, tipo de datos atómicos, etc.)

A-1. Cálculo de media, desviación estándar, etc.
A-2. Gráficos que permitan visualizar los datos adecuadamente.
A-3. Descripción del conjunto de datos a partir de los puntos anteriores.

APARTADO REGRESIÓN

En este apartado el estudiante debe utilizar el dataset$_R$ asignado para realizar lo siguiente:

R-1. Utilizar el algoritmo de regresión lineal simple sobre cada regresor (variable de entrada) para obtener los modelos correspondientes. Si el dataset$_R$ asignado incluye más de 5 regresores, seleccione de manera justificada los 5 que considere más relevantes. Una vez obtenidos los modelos, elegir el que considere más adecuado para su conjunto de datos según las medidas de calidad conocidas.
R-2. Utilizar el algoritmo para regresión lineal múltiple. Justificar adecuadamente si el modelo obtenido aporta mejoras respecto al modelo elegido en el paso anterior (en este apartado tenga también en cuenta la consideración de posibles interacciones y no linealidad).
R-3. Aplicar el algoritmo k-NN para regresión.
R-4. Comparar los resultados de los dos algoritmos de regresión múltiple entre sí, y adicionalmente mediante comparativas múltiples con un tercero (el modelo de regresión M5', cuyos resultados ya están incluidos en las tablas de resultados disponibles).

Nota: Al final de las transparencias para las clases de laboratorio de la parte de regresión, encontrará aclaraciones sobre lo que se pide para los apartados R-1 a R-4. Se incluyen ahí porque una vez realizado el trabajo de laboratorio es cuando mejor se puede entender lo que se pide.

APARTADO CLASIFICACIÓN

En este apartado el estudiante debe utilizar el dataset$_C$ asignado para realizar lo siguiente:

C-1. Utilizar el algoritmo k-NN probando con diferentes valores de *k*. Elegir el que considere más adecuado para su conjunto de datos. Analice qué ocurre en los valores de precisión en training y test con los diferentes valores de *k.*
C-2. Utilizar el algoritmo LDA para clasificar. No olvide comprobar las asunciones.
C-3. Utilizar el algoritmo QDA para clasificar. No olvide comprobar las asunciones.
C-4. Comparar los resultados de los tres algoritmos.