UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

Escuela de Post Grado
Facultad de Economía y Planificación
Departamento de Estadística e Informática

Técnicas de Minería de Datos
Practica calificada 1

Profesor: Enver Tarazona Vargas

Indicaciones Generales:

- Duración: 120 minutos
- Materiales a utilizar: Está permitido el uso de apuntes. Debe usar R para resolver los ejercicios planteados (incluir el código y programas usados para resolver las preguntas).
- El examen puede ser presentado usando cualquier procesador de textos (ver plantilla en Word). Grabe el archivo de la siguiente manera: PC01_EP7144_Apellido1 _ Apellido2 _ Apellido3.
- Al término de este examen el alumno deberá enviar la solución al correo enver.tarazona@pucp.edu.pe
- La presentación, la ortografía y la gramática influirán en la calificación.

Puntaje Total: 20 puntos

Pregunta 1 (14 puntos)

El conjunto de datos ***flas*** de la librería ***norm2*** contiene datos de un estudio para evaluar la confiabilidad de la Escala de actitud de lenguaje extranjero, un instrumento para predecir el éxito en el estudio de idiomas extranjeros (Raymond y Roberts, 1983). El cuestionario se entregó a 279 estudiantes matriculados en cuatro cursos de idiomas diferentes (francés, alemán, español, ruso) en la Universidad Penn State. Este conjunto de datos incluye el puntaje FLAS, la calificación final en el curso y otras variables para predecir el logro. Para mayores detalles consultar la ayuda en R. Realice un diagnóstico de datos perdidos y responda:

(a) ¿Cuantas filas con valores perdidos hay en el conjunto de datos? ¿Qué porcentaje de todos los valores es? (1.0 puntos)

(b) ¿Cuantos atributos (columnas) hay sin ningún dato perdido? Describa la cantidad y el porcentaje de datos perdidos que presenta cada atributo. (2.0 puntos)

(c) ¿Cuantos patrones de datos perdidos distintos presenta el conjunto de datos? ¿Cuál es el que ocurre con mayor frecuencia? Justifique su respuesta usando una gráfica apropiada. (3.0 punto)

(d) ¿Es posible identificar algún patrón de datos perdidos? (por ejemplo, la perdida de información en alguna variable parece estar asociada a otra.) Justifique su respuesta (4.0 puntos)

(e) Se desea realizar la imputación de la variable FLAS. Identifique a qué tipo de mecanismo se puede atribuir la presencia de datos perdidos. Justifique su respuesta usando técnicas de visualización y la prueba estadística t. Luego, sobre la base del mecanismo identificado, realice la imputación de los datos perdidos. (4.0 puntos)

Pregunta 2 (6 puntos)

La data ***humus*** se encuentra en el paquete ***mvoutlier***. Los datos fueron recogidos en el Proyecto Kola (1993-1998, Geological Surveys of Finland (GTK) and Norway (NGU) and Central Kola Expedition (CKE), Russia). Se analizaron más de 600 muestras en cinco capas diferentes, este conjunto de datos contiene la capa de humus. Para mayores detalles consultar la ayuda en R
Para el ejercicio considere las variables Ag, Al y Ba

(a) Sobre la base de un análisis univariado, ¿Existen evidencias de outliers en alguna de las variables? Justifique su análisis usando técnicas de visualización y la puntuación Z. (2.0 puntos)

(b) Evalué la presencia de outliers multivariados usando la distancia de Mahalanobis al cuadrado. Justifique su análisis usando técnicas de visualización (grafica de las distancias, grafica Q-Q y de frecuencias acumuladas usando la distribución chi-cuadrado) (2.0 puntos)

(c) Evalué la presencia de outliers multivariados usando el método basado en densidad local (LOF) considerando 10 vecinos. Compare los resultados obtenidos con el método anterior. (2.0 puntos)