Trabajo sobre preprocesamiento de datos

Máster Interuniversitario en Big Data: Tecnologías de Análisis de Datos Masivos

Minería de datos

Curso 2021/2022

El objetivo de este ejercicio es generar una serie de tablas sobre las que se puedan aplicar técnicas de minería de datos. Para ello se utilizarán los conocimientos adquiridos del tema de Preprocesamiento de datos.

Los datos de partida se encuentran en el fichero *echocardiogram.data* y la descripción de cada uno de los atributos se encuentra en el fichero *echocardiogram.names*. El proceso seguido debe ser debidamente documentado (hay que incluir comentarios y el código utilizado). En el proceso se debe realizar:

1. Importar la tabla correctamente. Esto implica que las columnas deben tener definidas correctamente su nombre y tipo de dato. Se debe mostrar en la memoria la estructura de la tabla.
    - Los nombres de las columnas mencionadas en el fichero echocardiogram.names deben tener los siguientes nombres: Survival, StillAlive, AgeAttack, PericardEffu, FracShort, EPSS, LVDD, WMS, WMI, Mult, Name, Group, AliveAt1.
2. Generar un fichero en formato CSV con la tabla importada y transformada correctamente. La tabla tendrá el mismo nombre y la extensión será CSV. El fichero creado se deberá entregar junto con la memoria.
3. Mostrar la distribución de los valores ausentes tanto por columnas como por filas, en el caso de que existan.
4. De acuerdo con la información suministrada, calcula los valores ausentes de la columna de clasificación (la última).
5. Antes del proceso de imputación, y en función de la distribución de NA's, indicar si sería conveniente eliminar alguna instancia o atributo. Razona tu respuesta.
6. Imputar los valores ausentes aplicando diferentes técnicas de imputación (como mínimo: imputación knn, media o mediana, y alguna del paquete mice). Muestra, mediante alguna gráfica las diferencias entre la tabla original y las diferentes técnicas de imputación utilizadas.
7. Generar diferentes ficheros CSV resultado de aplicar diferentes técnicas de imputación. Los nombres de los ficheros tendrán la siguiente estructura: *echocar.método de imputación.csv*. Por ejemplo, si utilizamos la imputación knn, el fichrero deberá llamarse *echocar.knnImpute.csv*.

Para que el trabajo sea evaluable se deben de entregar: la memoria en formato R Notebook/R Markdown con el código incrustado, la salida en formato HTML, y los archivos CSV generados.