

MIND: Minería de Datos Assignment 2

Preliminaries

- The assignment must be developed in R or Python using Jupyter Notebook
- You must submit only one document that contains the Jupyter Notebook
- Deadline to submit assignment: Thursday, 02-Sept at 23:55.

Data Set Information

The file *breast-cancer-wisconsin.data* provides some features about patients with breast cancer. The features are collected from images of a breast mass. The dataset is openly, you can find more information about it on Breast Cancer (Diagnostic).

Problem

Realizar un *Notebook* implementado en lenguaje *R* o *Python* (es de su libre elección) para solucionar cada uno de los siguientes numerales. El Notebook debe estar compuesto por dos formatos de celdas: i.fragmentos de código para dar solución a cada numeral y ii.texto para indicar el numeral que se esta desarrollando o adicionar información revelante al numeral.

Antes de proceder con el desarrollo del cuadernillo, le sugiero entender muy bien la información representada en el archivo, las características que allí se representan y que indica cada uno de los valores. El archivo *BreastCancerDS.names* tiene información detallada y además el link de la fuente: Breast Cancer.

1. Cargar el data set suministrado *breast-cancer-wisconsin.data*
2. Realice una exploración rápida de los datos cargados indicando el tipo de cada variable, una corta descripción de cada una y presente la dimensión de los datos.
3. Valide si algunos de los atributos tienen valores faltantes. Indique los atributos con valor faltante y el porcentaje por cada atributo.
4. Muestre la cantidad de pacientes por cada tipo de tumor (benign o malignant) junto con una representación gráfica.
5. Identifique relaciones/patrones encuentra entre los atributos y la posibilidad de tener cáncer, y visceer la posibilidad de no tenerlo.
6. Cuáles son los atributos más relevantes con la enfermedad del cancer de mama.
7. Realice gráfico de dispersión entre cada par de atributos con la intención de conocer la relación entre ellas y la posibilidad de tener cáncer o no.
8. Importante en cada gráfico hacer una interpretación sobre lo visualizado, además de la figura, es importante su respectiva interpretación a cada uno, es decir, qué revela el gráfico?. La interpretación es una redacción suya argumentada y coherente sobre los hallazgos observados en el gráfico.