CURSO: MINERÍA DE DATOS

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

LABORATORIO IV: Reducción de dimensionalidad (Preprocesamiento V. IV)

INTRODUCCIÓN

Esta práctica de laboratorio tiene como objetivo avanzar en la exploración de las técnicas de reducción de dimensionalidad de la etapa de Preprocesamiento, del Proceso de Descubrimiento de Conocimiento.

Para la exploración de estos temas, se utilizará el IDE R-Studio del lenguaje de programación R, a efectos de ejercitar los conceptos abordados en las clases teóricas.

CONSIGNAS

A partir del dataset *auto-mpg.data-original.txt*[1], se solicita trabajar sobre las siguientes consignas:

1. **SOBRE LOS DATOS**
    a. Cargue[2] y explore el dataset: explique en qué consiste el mismo y qué características posee.
    b. Con las técnicas abordadas en la práctica de laboratorio anterior, realice un breve análisis exploratorio para identificar cual es la distribución de sus variables y si existe relación entre las mismas.
2. **REDUCCIÓN DE DIMENSIONALIDAD**
    a. Indague sobre la varianza[3] de cada uno de los atributos que conforman el dataset. ¿Existen atributos que podrían ser eliminados de acuerdo a la técnica de *Low Variance Factor*? Actúe en consecuencia.
    b. Evalúe la relación entre atributos[4] a partir del coeficiente de correlación de Pearson y un análisis gráfico de *heatmap*[5] para estudiar la posibilidad de eliminar

[1] Disponibles en: https://archive.ics.uci.edu/ml/datasets/Auto+MPG
[2] Explore la instrucción *read.table().*
[3] Recuerde previamente normalizar el dataset, consulte la instrucción *scale*().
[4] Considere, además del cálculo del coeficiente, realizar un análisis gráfico en el caso de variables numéricas **-scatterplot-** o utilizar el **Test de Chi Cuadrado** para variables categóricas.
[5] Explore la instrucción *heatmap.2* de la librería *gplots.*

redundancia en el dataset. En caso de corresponder, aplique las técnicas de *Reducing Highly Correlated Columns* trabajadas en clase.

c. Por último, compare la importancia de cada uno de los atributos en función de la técnica de determinación de Random Forest[6] (suponiendo que intenta predecir la cantidad de cilindros de un auto). Analice la importancia de las variables de modo analítico y gráfico.

Referencias sugeridas:

García, S., Luengo, J., & Herrera, F. (2016). Data preprocessing in data mining. Springer.

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[6] Se sugiere utilizar las instrucciones *randomForest, importance* y *varImpPlot* de la librería randomForest.