CURSO: MINERÍA DE DATOS

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

Reducción de dimensionalidad

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

Utilizando las colecciones del primer TP *charts* y *artist_audio_features_solo_art*, se solicita trabajar sobre las siguientes consignas:

1. **SOBRE LOS DATOS**
    a. Utilizando como base las 3 queries compartidas como ejemplo en el *Slack*[1] el día 5 de Mayo, genere un *dataset* que contenga el promedio de medidas continuas por artista para tipos de álbumes *singles*, incluyendo el promedio de posición y de *streams* en los charts[2].
2. **REDUCCIÓN DE DIMENSIONALIDAD**
    a. Indague sobre la varianza[3] de cada uno de los atributos que conforman el dataset. ¿Cuáles son los dos atributos que podrían ser eliminados de acuerdo a la técnica de *Low Variance Factor*?
    b. Evalúe la relación entre atributos a partir del coeficiente de correlación de Pearson y un análisis gráfico de *heatmap*[4] para estudiar la posibilidad de eliminar redundancia en el dataset.

---

[1] Link al post en el canal primer-tp-2021

[2] *danceability, loudness, energy, speechiness, liveness, acousticness, instrumentalness, valence, avg_streams, avg_position*

[3] Recuerde previamente normalizar el dataset

[4] Explore la instrucción *heatmap.2* de la librería *gplots.*

c. Aplique la función *FindCorrelation* con un umbral de 0.75 e identifique las variables candidatas a ser eliminadas según esta técnica.

d. Suponiendo que quiere predecir si un artista ocupara un lugar entre las 100 mejores posiciones, compare la importancia de cada uno de los atributos utilizando la técnica de *Random Forest*[5] de forma gráfica y analítica.

   Trabaje la variable objetivo como tipo *factor* para generar un modelo de clasificación.

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.

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