UNIVERSIDAD AUTÓNOMA METROPOLITANA
Casa abierta al tiempo
Unidad Iztapalapa

# Temas selectos: BBDD
# Práctica 2: Selección de atributos
# Trimestre 25-I
# Dr. Orlando Muñoz Texzocotetla

1. Buscar y seleccionar un dataset en el sitio Kaggle de acuerdo a una problemática de aprendizaje supervisado elegida por el equipo. Para buscar el dataset usar las palabras claves: lda, pca, linear discriminant analysis, principal component analysis, pearson correlation, spearman correlation. El dataset debe tener más de 10 atributos números y una variable objetivo (la clase).

2. Limpieza de datos. Realizar el manejo de valores nulos y outliers. Sólo si es necesario.

3. Selección de atributos. Aplicar las siguientes técnicas para obtener un dataset con menos atributos (menor dimensión).

a. Correlación de Pearson (filtrado)
b. Correlación de Spearman (filtrado)
c. Análisis de componentes principales (reducción de dimensión)
d. Análisis discriminante lineal (reducción de dimensión)

Por cada técnica utilizada se deberá obtener un dataset.

4. Hacer un análisis comparando los datasets obtenidos con las 4 técnicas de selección de atributos, y contestar lo siguiente:

a. ¿Cuáles son las principales diferencias en los atributos seleccionados por los métodos utilizados? ¿Seleccionaron diferentes atributos o fueron los mismos?
b. ¿Qué limitaciones tienen los métodos de filtrado utilizados?
c. ¿El dataset generado por PCA es directamente interpretable? sí, no y por qué.
d. ¿Es adecuado usar más de una técnica de selección de atributos y por qué?

5. Compare los resultados al analizar los datasets (original sin selección de atributos, y los 4 datasets con selección de atributos) con un algoritmo de clasificación o regresión de acuerdo a la problemática elegida en el primer inciso. Analice y compare la exactitud (accuracy). En este caso usar el mismo porcentaje de cada dataset para entrenamiento y prueba. Conteste:

a. ¿Cómo cambió el desempeño del modelo al reducir el número de atributos?
b. ¿Cuál de las técnicas de selección produjo mejores resultados? ¿Fue ese resultado mejor que el análisis con el dataset original?
c. ¿Hubo algún método que eliminara información importante y afectara negativamente el desempeño del algoritmo de aprendizaje maquinal?

Entregables:

- Reporte en pdf: cuiden el formato (no letras de diferentes tipos y tamaños para el mismo texto, justificación del texto, etc.), agregar carátula, bien estructurado e incluir referencias.
- Archivo de programa en python. Incluyan el programa, no enlaces por favor.
- Dataset comprimido. Si el archivo es muy grande podría subirse a la nube e incluir el enlace.

NO COMPRIMIR LOS ARCHIVOS.
Nota: para los análisis se proporciona código de ejemplo en Python para los diferentes análisis ya sea de filtrado o de reducción de dimensionalidad, y también un pequeño ejemplo de un árbol de decisión para clasificar. Todos estos con el wine dataset que se describió en clase. Mejoren el código dado no sólo usen los ejemplos tal cual, sino adaptándolos al dataset y problemática elegida.