Laboratorio 1.
Análisis Exploratorio, Clustering y PCA

INSTRUCCIONES:

Utilice el data set House Prices: Advanced Regression Techniques que comparte Kaggle. Debe hacer un análisis exploratorio para entender mejor los datos, sabiendo que el objetivo final es predecir los precios de las casas. Recuerde explicar bien cada uno de los hallazgos que haga. La forma más organizada de hacer un análisis exploratorio es generando ciertas preguntas de las líneas que le parece interesante investigar. Genere un informe en pdf con las explicaciones de los pasos que llevó a cabo y los resultados obtenidos. Recuerde que la investigación debe ser reproducible por lo que debe guardar el código que ha utilizado para resolver los ejercicios y/o cada uno de los pasos llevados a cabo si utiliza una herramienta visual. Lleve a cabo un análisis de componentes principales y un agrupamiento. Este laboratorio debe realizarse de forma **INDIVIDUAL.**

DESCRIPCIÓN DEL DATASET

El dataset contiene datos de 1460 casas en Ames, Iowa y 79 variables que describen prácticamente todos los aspectos de estas. El archivo de descripción de los datos, que incluye nombre de variables y posibles valores lo puede encontrar en el link siguiente: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

EJERCICIOS

1. Haga una exploración rápida de sus datos para eso haga un resumen de su dataset
2. Diga el tipo de cada una de las variables del dataset (cualitativa o categórica, cuantitativa continua, cuantitativa discreta)
3. Aísle las variables numéricas de las categóricas, haga un análisis de correlación entre las mismas.
4. Utilice las variables categóricas, haga tablas de frecuencia, proporción, gráficas de barras o cualquier otra técnica que le permita explorar los datos
5. Haga un análisis de componentes principales, interprete los componentes
6. Haga un análisis de clustering, describa los grupos.
7. Haga un resumen de los hallazgos más importantes encontrados al explorar los datos y llegue a conclusiones sobre las posibles líneas de investigación.

EVALUACIÓN

(50 puntos) Análisis Exploratorio:

- Estudia las variables cuantitativas mediante técnicas de estadística descriptiva
- Hace gráficos exploratorios como histogramas, diagramas de cajas y bigotes, gráficos de dispersión que ayudan a explicar los datos
- Analiza las correlaciones entre las variables, trata de explicar los outliers (puntos atípicos) y toma decisiones acertadas ante la presencia de valores faltantes.
- Estudia las variables categóricas
- Elabora gráficos de barra, tablas de frecuencia y de proporciones

- Explica muy bien todos los procedimientos y los hallazgos que va haciendo.

(15 puntos) Análisis de componentes Principales

- Estudia la matriz de correlación, la agrega y explica lo que observa en ella
- Determina si es posible usar la técnica de análisis factorial para hallar las componentes principales
- Determina si vale la pena aplicar las componentes principales interpretando el test de esfericidad de Bartlett
- Obtiene los componentes principales y explica cuantos seleccionará para explicar la mayor variabilidad posible.
- Interpreta los coeficientes principales.

(15 puntos) Clustering

- Determina el mejor número de clusters a utilizar usando el método de Ward
- Hace el agrupamiento con cualquiera de los algoritmos estudiados
- Verifica la calidad del agrupamiento usando el método de la silueta.
- Interpreta los grupos, usando para eso las variables numéricas y categóricas dentro de cada grupo.

(20 puntos) Hallazgos y conclusiones.

- Hace un resumen de los hallazgos en el análisis exploratorio
- Le pone un nombre a los grupos que reflejen sus características principales
- Llega a conclusiones sobre los siguientes pasos a seguir.

MATERIAL A ENTREGAR

- Archivo .pdf con el informe de análisis exploratorio que debería tener:
- Script de R (.r o .rmd) o de Python que utilizó para responder las preguntas con el código utilizado o archivo de flujo de trabajo de KNime