Laboratorio 2.
Algoritmos de aprendizaje de máquinas

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 en **PAREJAS.** Para que se pueda calificar su laboratorio debe estar inscrito en algún grupo de canvas.

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. Divida el set de datos de entrenamiento que le provee kaggle en 2 conjuntos, entrenamiento (60%) y prueba (40%). Utilice el método de Muestreo Aleatorio Simple para seleccionar las filas que van a cada conjunto de datos.
2. Haga un modelo de regresión lineal para predecir el precio de las casas. Como ya hizo un análisis exploratorio del conjunto de datos, explique la selección de variables con los que hizo el modelo.
3. Haga un análisis del modelo generado, ¿Cuáles son las variables significativas? ¿Explica o no la variabilidad de los datos? Si considera necesario redefinir las variables del modelo, hágalo y explique las causas.
4. Compare el precio que predijo el algoritmo con el que ya se conoce, explique la efectividad del algoritmo ya sea definiendo una diferencia mínima o analizando el error en la predicción de los datos.
5. Utilice el conjunto de datos de prueba que provee kaggle para probar el algoritmo y explique los resultados obtenidos.
6. Cree una variable categórica que permita clasificar la fila de acuerdo al precio de venta en alto, medio y bajo. Explique la selección de los límites basado en la distribución del conjunto de datos.
7. Divida el conjunto de datos de entrenamiento que provee kaggle en dos conjuntos, entrenamiento y prueba. Utilice para esto el método de muestreo estratificado.

8. Haga un modelo de KNN (K nearest neighbors). Explique la elección del parámetro k que clasifique el precio de las casas en bajo, medio y alto.
9. Pruebe el algoritmo con sus datos de prueba, haga la matriz de confusión y explique los resultados.
10. Pruebe el algoritmo con el conjunto de prueba que provee Kaggle. Haga la matriz de confusión y explique los resultados.
11. Repita los pasos del 8 al 10 usando validación cruzada. Compare los resultados obtenidos. ¿Cuál de los modelos tiene un mejor desempeño?

EVALUACIÓN

(10 puntos) Conjunto de entrenamiento y prueba:
- Se dividió el conjunto de entrenamiento propuesto por kaggle en dos subconjuntos de entrenamiento (60%) y prueba (40%) de forma reproducible. Se utilizó el método de muestreo aleatorio simple en el paso 1 y muestreo estratificado en el paso 7.

(15 puntos) Modelo de regresión Lineal
- Se elaboró un modelo de regresión lineal multivariado.
- Explica la selección de variables que incluyó en el modelo, la explicación es lógica y relevante.

(12 puntos) Análisis del modelo de regresión lineal
- Se muestra un resumen del modelo, se explican las variables que son relevantes al modelo y si es un buen modelo o no dependiendo de los parámetros. Si vuelve a construir el modelo tiene en cuenta las variables que realmente son significativas en el modelo.

(13 puntos) Predicción con el modelo de regresión lineal.
- Predice el precio de las casas utilizando el modelo de regresión lineal que construyó
- Compara el precio que predijo el algoritmo con el que ya se conoce.
- Determina un criterio de éxito del modelo ya sea una diferencia mínima que es aceptable (explica en qué se basó para elegir ese número) o análisis del error.
- Explica la efectividad del algoritmo basado en el criterio de éxito que seleccionó.

(15 puntos) Modelo de los K vecinos más cercanos
- Se elaboró un modelo de knn.
- Explica la selección de variables que incluyó en el modelo así como la elección del número de vecinos a considerar, la explicación es lógica y relevante.
- Explica los límites de cada categoría de precio basado en la distribución de los datos.

(13 puntos) Predicción con el modelo de knn.
- Predice el precio de las casas utilizando el modelo de knn que construyó.
- Compara la categoría de precio que predijo el algoritmo con el que ya se conoce.
- Explica la efectividad del algoritmo usando la matriz de confusión.

(10 puntos) Validación cruzada.
- Ejecuta nuevamente los modelos usando validación cruzada. Compara los resultados obtenidos con la efectividad del algoritmo usando los conjuntos de entrenamiento y prueba generados de manera aleatoria.

(12 puntos) Comparación de los modelos

- Compara la efectividad de ambos modelos de KNN y determina cuál de los dos debería usarse para este problema y porqué.

MATERIAL A ENTREGAR

- Archivo .pdf con el informe que contenga, los resultados de los análisis y las explicaciones.
- Link de Google drive donde trabajó el grupo.
- 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
- Link del repositorio usado para versionar el código.

FECHAS DE ENTREGA

- **AVANCE:** Análisis Exploratorio y Modelo de Regresión Lineal Explicado: miércoles 21 de agosto 13:55.
- **DOCUMENTO FINAL COMPLETO:** sábado 24 de agosto de 2019 23:59
  NOTA: Solo se calificará el Documento Final si está entregado el avance con todo lo que se pide.